Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)



University of Tartu Library
NEALT Proceedings Series Number: 61
ISBN: 978-9985-4-1522-1
ISSN: 1736-8197
eISSN 1736-6305
2026

Proceedings of the

15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)

edited by
Arianna Masciolini, David Alfter, Herbert Lange,
Ricardo Muñoz Sánchez, Ildikó Pilán, Elena Volodina

Proceedings and all papers therein
published under a CC BY 4.0 license:
https://creativecommons.org/licenses/by/4.0

Front cover image by Pasi Mämmelä (mammela)
via Pixabay

University of Tartu Library

NEALT Proceedings Series Number: 61
ISBN: 978-9985-4-1522-1
ISSN: 1736-8197
eISSN 1736-6305

2026

Preface

The workshop series on Natural Language Processing (NLP) for Computer-Assisted Language Learning (NLP4CALL) is a meeting place for researchers working on integrating Natural Language Processing and Speech Technologies in CALL systems and exploring the theoretical and methodological issues arising in this connection. The latter includes, among others, the integration of insights from Second Language Acquisition (SLA) research and the promotion of "Computational SLA" through setting up Second Language research infrastructures.

The intersection of Natural Language Processing (or Language Technology / Computational Linguistics) and Speech Technology with Computer-Assisted Language Learning (CALL) brings "understanding" of language to CALL tools, thus making CALL intelligent. This fact has inspired the name for this area of research — Intelligent CALL, ICALL for short. As the definition suggests, apart from having excellent knowledge of Natural Language Processing and/or Speech Technology, ICALL researchers need good insights into second language acquisition theories and practices, as well as knowledge of second language pedagogy and didactics. Therefore, this workshop invites a wide range of ICALL-relevant research, including studies where NLP-enriched tools are used for testing SLA and pedagogical theories, and those where SLA theories (pedagogical practices or empirical data) are modeled using ICALL tools. The NLP4CALL workshop series is aimed at bringing together competences from these areas for sharing experiences and brainstorming around the future of the field.

Topics of Interest

We invited submissions:

- that describe research directly aimed at ICALL;
- that describe the ongoing development of resources and tools with potential use in ICALL, either directly in interactive applications, or indirectly in materials, application, or curriculum development, e.g. learning material generation, assessment of learner texts and responses, individualized learning solutions, provision of feedback;
- that discuss challenges and/or research agendas for ICALL;
- that describe empirical studies on language learner data.

This year, the workshop has a special focus on **process-oriented approaches to educational NLP**, including but not limited to work related to the collection and analysis of keystroke data from language learners.

We encouraged paper presentations and software demonstrations describing the above-mentioned themes primarily, but not exclusively, for the Nordic languages.

Invited speakers

This year, we had the pleasure to welcome two invited speakers: Andrea Horbach (Kiel University and the Leibniz Institute for Science and Mathematics Education (IPN)) and Rianne Conijn (Eindhoven University of Technology).

Andrea Horbach has been professor for Educational NLP at Kiel University and the Leibniz Institute for Science and Mathematics Education (IPN) since 2024. Her research interests include free-text scoring, argument mining, automated feedback, and the use of process data for assessment. She is particularly interested in developing explainable and fair methods that

support human agency in writing and assessment processes. The title of her talk was *A Focus on the Writing Process With and Without AI Support: Insights from the German PISA 2025 Foreign Language Assessment Data Collection.*

Rianne Conijn is an assistant professor in learning analytics in the Human-Technology Interaction group at Eindhoven University of Technology, the Netherlands. Her research focuses on developing methods to capture and model learning behavior (e.g., self-regulated learning) and writing behavior (e.g., revision processes), using fine-grained behavioral data, such as keystroke and clickstream data. In addition, she uses these methods to examine how educational technologies, such as generative AI, affect the learning process. Rianne recently received a NWO Veni grant on "Human-Centered AI in education" where she aims to improve human-AI collaboration in writing in higher education. The title of her talk was *Moving beyond the product: Keystroke Logging for CALL*

Previous workshops

This workshop follows a series of workshops on NLP4CALL organized by the NEALT Special Interest Group on Intelligent Computer-Assisted Language Learning (SIG-ICALL)[1]. The workshop series has previously been financed by:

- the Center for Language Technology at the University of Gothenburg;
- the SweLL project;[2]
- the Swedish Research Council's conference grant, Språkbanken Text;[3]
- the L2 profiling project;[4]
- itec;[5]
- the CENTAL;[6]
- the Analytics for Language Learning (A4LL) project[7] at LIDILE - Univ Rennes;
- the Mormor Karl project;[8] and
- the project "Expanding the scope of a multi-purpose lexicographic resource to grammar and L2 competence".[9]

Submissions to the fifteen workshop editions have targeted a wide range of languages, ranging from well-resourced languages (Chinese, German, English, French, Portuguese, Russian, Spanish) to lesser-resourced languages (Erzya, Arabic, Estonian, Irish, Komi-Zyrian, Meadow Mari, Saami, Udmurt, Võro). Among these, several Nordic languages have been targeted, namely Danish, Estonian, Finnish, Icelandic, Norwegian, Saami, Swedish and Võro. The wide scope of the workshop is also evident in the affiliations of the participating authors as illustrated in Table 1.

[1] https://spraakbanken.gu.se/en/research/themes/icall/sig-icall
[2] https://spraakbanken.gu.se/en/projects/swell
[3] https://spraakbanken.gu.se
[4] https://spraakbanken.gu.se/en/projects/l2profiles
[5] https://itec.kuleuven-kulak.be
[6] https://cental.uclouvain.be
[7] https://sites-recherche.univ-rennes2.fr/lidile/articles/a4all/
[8] https://mormor-karl.github.io/
[9] https://eki.ee/prg-1978/

The acceptance rate has varied between 44% and 83%, the average being 62% (see Table 2). Although the acceptance rate is rather high, the reviewing process has always been very rigorous with two to three double-blind reviews per submission. This indicates that submissions to the workshop have usually been of high quality.

Country	**Count**	**Country**	**Count**
Algeria	1	Italy	16
Australia	2	Japan	7
Belgium	21	Lithuania	1
Canada	4	Netherlands	4
China	5	Norway	16
Cyprus	3	Poland	1
Czech Republic	2	Portugal	8
Denmark	5	Romania	1
Egypt	1	Russia	10
Estonia	6	Slovakia	1
Finland	15	Spain	5
France	51	Sweden	101
Germany	145	Switzerland	16
Greece	6	Ukraine	2
Iceland	6	UK	35
Iran	2	Uruguay	5
Ireland	7	US	15
Israel	1	Vietnam	3

Table 1: NLP4CALL speakers' and co-authors' affiliations, 2012–2026

Workshop year	Submitted	Accepted	Acceptance rate
2012	12	8	67%
2013	8	4	50%
2014	13	10	77%
2015	9	6	67%
2016	14	10	71%
2017	13	7	54%
2018	16	11	69%
2019	16	10	63%
2020	7	4	57%
2021	11	6	55%
2022	23	13	56%
2023	18	12	67%
2024	23	19	83%
2025	16	7	44%
2026	51	25	49%

Table 2: Submissions and acceptance rates, 2012–2026

Program committee

We would like to thank our Program Committee for providing detailed feedback for the reviewed papers:

- Soroosh Akef, Leibniz-Institut für Wissensmedien, Germany
- David Alfter, University of Gothenburg, Sweden
- Stephen Bodnar, University of Tübingen, Germany
- Christopher Bryant, University of Cambridge, UK
- Andrew Caines, University of Cambridge, UK
- Xiaobin Chen, University of Tübingen, Germany
- Orphée De Clercq, Ghent University, Belgium
- Kordula De Kuthy, Leibniz-Institut für Wissensmedien, Germany
- Piet Desmet, KU Leuven, Belgium
- Irene Elmerot, University of Gothenburg, Sweden
- Thomas Francois, Université catholique de Louvain, Belgium
- Thomas Gaillat, Université Rennes 2, France
- Johannes Graën, University of Zurich, Switzerland
- Andrea Horbach, FernUniversität in Hagen, Germany
- Nikolai Ilinykh, University of Gothenburg, Sweden
- Jelena Kallas, Institute of the Estonian Language, Estonia
- Joni Kruijsbergen, Ghent University, Belgium
- Murathan Kurfalı, RISE Research Institutes of Sweden, Sweden
- Ronja Laarmann-Quante, Ruhr-Universität Bochum, Germany
- Herb Lange, University of Gothenburg, Sweden
- Peter Ljunglöf, Chalmers University of Technology, Sweden
- Arianna Masciolini, University of Gothenburg, Sweden
- Felix Morger, University of Gothenburg, Sweden
- Ricardo Muñoz Sánchez, University of Gothenburg, Sweden
- Lionel Nicolas, Eurac Research, Italy
- Robert Östling, Stockholm University, Sweden
- Ulrike Pado, HFT Stuttgart, Germany

- Magali Paquot, Université catholique de Louvain, Belgium
- Ildikó Pilán, University of Gothenburg, Sweden
- Egon W. Stemle, Eurac Research, Italy
- Maria Irena Szawerna, University of Gothenburg, Sweden
- Anaïs Tack, KU Leuven, Belgium
- Irina Temnikova, Sofia University "St. Kliment Ohridski", Bulgaria
- Ahmet Yavuz Uluslu, University of Zurich, Switzerland
- Sowmya Vajjala, National Research Council, Canada
- Elena Volodina, University of Gothenburg, Sweden
- Ekaterina Voloshina, University of Gothenburg and Chalmers University of Technology, Sweden
- Niklas Zechner, University of Gothenburg, Sweden
- Torsten Zesch, FernUniversität in Hagen, Germany

We intend to continue this workshop series, which so far has been the only ICALL-related recurring event based in the Nordic countries. Our intention is to co-locate the workshop series with the two major LT events in Scandinavia, the Swedish Language Technology Conference (SLTC) and the Nordic Conference on Computational Linguistics (NoDaLiDa), thus making this workshop an annual event. Through this workshop, we intend to profile ICALL research in Nordic countries as well as beyond, and we aim at providing a dissemination venue for researchers active in this area.

Workshop website

`https://nlp4call.github.io/`

Workshop organizers

- Arianna Masciolini, Språkbanken, University of Gothenburg, Sweden
- David Alfter, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg, Sweden
- Herbert Lange, Språkbanken, University of Gothenburg, Sweden
- Ricardo Muñoz Sánchez, Språkbanken, University of Gothenburg, Sweden
- Ildikó Pilán, Norwegian Computing Center, Norway
- Elena Volodina, Språkbanken, University of Gothenburg, Sweden

Acknowledgments

The 2026 edition of this workshop was supported jointly by:

- Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161);[10]
- Huminfra;[11]
- the Department of Swedish, Multilingualism, Language Technology;[12]
- the IncluEdS project[13]

[10] https://spraakbanken.gu.se/
[11] https://www.huminfra.se/
[12] https://www.gu.se/en/swedish
[13] https://inclueds.org/

Content

Preface i
Arianna Masciolini, David Alfter, Herbert Lange, Ricardo Muñoz Sánchez, Ildikó Pilán and Elena Volodina

How Much of Native Language Identification Is Corpus Artifact? A Cross-Corpus Audit of SVMs and LLMs on TOEFL11 and ICLEv3 1
Shima Rahimi

Lexical Profiling for Dutch: From Resource to Real-Life Implementation 16
Jasper Degraeuwe and Natalie Finlayson

Making Feedback Visible: Character-Level Error Localization in Handwritten CALL Input 34
Said Yasin and Torsten Zesch

Towards Semi-Automatic Benchmarking of Large Language Models for Self-Regulated Language Learning 40
Émile Alexandre, Yannick Parmentier and Guillaume Nassau

Adversarial Disentanglement of Speaker Identity and L2 Phonological Errors via Raw-Waveform Architectures 58
Amirhossein Ghanipour Amirhandeh

Investigating the role of revisions for learner level predictions 65
Ahood Al Sawar, Léo Nebel and Nicolas Ballier

Universal, but Grounded? Anchoring Cross-Lingual CEFR Assessment in the Framework's Own Descriptors 78
David Alfter

Assessing the Reliability of Automatic Speech-Processing Methods for L2 Speech: Evidence from an Automatic Pipeline, Posteriorgrams, and Whisper Probability Scores 99
Mehak Sohail and Nicolas Ballier

Multilingual language learning with Grammatical Framework 111
Adrian De Lon and Inari Listenmaa

Using LLMs for feature extraction: the case of language functions with deixis 119
Thomas Gaillat, Xiaobin Chen and Nelly Sagirov

Exploring the syntactic parsing of learner texts in Modern Greek: Variety trumps size, genre trumps grammaticality 130
Christina Klironomou, Arianna Masciolini and Aleksandrs Berdicevskis

Predicting Personalized Vocabulary Learning Burden from Intralexical and Interlexical Factors 140
Mykhailo Diachenko and John P. McCrae

A Multidimensional Evaluation of LLM-Generated Learning Materials for Greek Primary School Learners 160
Eleftherios Kalogeros, Sofia Stamou, Markella Kolocheri, Thaleia-Dionysia Kolovou, Manolis Gergatsoulis and Matthew Damigos

Etymology Can Inform Lexical Complexity Prediction 173
Noah-Manuel Michael, Anna Hülsing and Andrea Horbach

Data Augmentation for L2 English Speaking Assessment using TTS 197
Stefano Bannò, Penny Karanasou, Mengjie Qian, Kate M. Knill and Mark J. F. Gales

Automatic Analysis of Noun Phrase Complexity in Learner spoken Productions 214
Saint Joannes Monengue Essengo, Bi Yoman Ghislain Aymard Toa, Thomas Gaillat, Christine Evain and Fabienne Moreau

A corpus of written Swedish texts and keystroke data from deaf and hard-of-hearing children and adolescents 229
Moa Gärdenfors, Robert Östling, Lisa Rudebeck and Mats Wirén

CineLexi: A corpus-based bilingual resource of evaluative vocabulary in film reviews (German–Russian) 244
Mariia Kopylova

Translation Surprisal Predicts Learner Error in L2 French Production: A Language-Pair-Specific Difficulty Feature for ICALL 256
Matteo Lanza

Behind the Smokescreen: Uncovering Cross-LLM Bias and Reasoning Contradictions of LLM-as-a-judge for Generated Feedback 269
Céline Leuzinger

Reading Ahead: Pace-Based Prediction for Real-Time Tracking of Children's Oral Reading in the Browser 288
Bohdan Podziubanchuk, Einar Meister and Tanel Alumäe

ABC Diagnostics: A Web-Based Tool for Explainable CEFR Classification 295
Ricardo Muñoz Sánchez, Kostiantyn Kucher, Filip Kayar and Elena Volodina

Acoustic and Lexical Signals for Native Language Identification of Language Learners 306
Ahmet Yavuz Uluslu, Mengjie Qian, Kate Knill and Mark Gales

Poor Generalisation of Proficiency Classifiers Trained on Linguistic Complexity Features to New Writing Tasks 314
Emre Güçer, Soroosh Akef and Detmar Meurers

Automated L2 Speech Assessment with Audio-Based Self-Supervised Neural Networks 334
Eli Stafford and Sylvain Coulange

How Much of Native Language Identification Is Corpus Artifact? A Cross-Corpus Audit of SVMs and LLMs on TOEFL11 and ICLEv3

Shima Rahimi
Independent Researcher, Iran
shimarahimi1988@gmail.com

Abstract

We ask how much of the accuracy in native language identification (NLI) comes from real language signals and how much from problems in the corpus. We compare a supervised SVM, trained only on TOEFL11, with a zero-shot LLM (gpt-4o-mini) on eight native languages, testing both on TOEFL11 and on the separately collected ICLEv3. On TOEFL11 the LLM gains 35.6 to 48.5 percentage points from full text over content-masked text, depending on prompt format and filter tier, while the SVM gains 13.6 to 16.0. Normalising for chance, the SVM keeps about 77% of its above-chance accuracy under masking and the LLM about 4%, which matches the LLaMA-3.1-8B result of Uluslu and Schneider (2025) rather than their 70B. Letting the LLM reason first lowers its full-text accuracy but does not raise its masked accuracy. On ICLE, corpus-ID headers and essays that mention the writer's own country or language both inflate the LLM's accuracy, while topics tied to one language make essays harder for it. The SVM is far less affected, and its accuracy is unchanged by the header. After both filters, 71 essays still reach 91.5% (chance: 12.5%), an upper bound because the topic filter removes the essays the LLM finds harder.

1 Introduction

Native Language Identification (NLI) is the task of inferring a writer's first language (L1) from their writing in a second language (L2). The idea behind this task is cross-linguistic influence: a writer's first language affects how they write in the second language and leaves some indication of the L1 in their texts (Yu and Odlin, 2016). One main application of NLI is computer-assisted language learning software. If a system can identify the native language of a learner, it can predict the errors the learner might make in their writing, help in providing special materials for that specific learner so that the learner can work on those errors, and explain the differences between L1 and L2. Similar tools are already being developed using learner corpora, and L1-aware feedback is one of the practical promises of ICALL. Another application of NLI is forensic linguistics, in which an inference about a writer's background can help narrow a set of possible authors.

This application is the main reason why it matters whether an NLI system works for the right reason. Teachers can use the cross-linguistic patterns that a valid and accurate NLI system provides in identifying a learner's L1. A system that uses shortcuts, like a writer mentioning their own country or language, or a whole class being assigned to write about the same novel, looks the same as a system that uses cross-linguistic patterns if we consider only accuracy, but the first one does not explain anything about language, and its feedback does not convey useful information for a new learner or a new classroom. This paper is about the difference between those systems.

Two problems in current NLI research still need attention. First, most NLI studies test their models using cross-validation within a single corpus. This can cause the model to confuse the L1 signals with corpus-specific patterns that might correlate with L1 in that specific dataset. Brooke and Hirst (2013) showed that topic bias alone in the ICLE corpus can make NLI performance look much stronger than it really is. Second, large language models (LLMs) are now widely used, including for NLI (Zhang and Salle, 2023). The problem is that they have access to vast training data and general world knowledge, so they might guess a writer's L1 from hints in the text, such as a country name, that a model trained only on the target corpus could never use. Uluslu and Schneider (2025) studied the reliance of the model on lexical shortcuts directly, by masking the

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 1–15, November 19–20, 2026, University of Tartu Library

content words with part-of-speech tags and comparing LLaMA-3.1 on full text against masked text. The accuracy decreased greatly. Their study covered four closely related European languages, and we describe it in more detail in Section 2.

This paper asks two questions. RQ1: in a real cross-corpus setting, training only on TOEFL11 and testing on both TOEFL11 and the separately collected ICLEv3 corpus, across eight typologically different native languages, how much do a traditional SVM and a zero-shot LLM (gpt-4o-mini) rely on structure versus lexical content? Does letting the LLM reason freely, instead of forcing a fixed output format, change this reliance? RQ2: how much of the LLM's apparent accuracy on full-text ICLE comes from the real L1 signal, and how much comes from specific and measurable corpus problems, namely corpus-ID headers that state each essay's country of origin, essays mentioning the writer's own country or language, and essay topics that correlate with L1 because of how ICLE collects essays by institution? Do the SVM and the LLM differ in how much they are affected by each problem, and if so, why?

This paper makes three contributions. First, a cross-corpus comparison of an SVM and an LLM across eight languages from different language families, more than the four related languages Uluslu and Schneider (2025) used. Second, a comparison between a forced-format prompt and a free-reasoning prompt for the LLM, to test whether the format itself changes how much the model relies on lexical shortcuts. Third, three measured audits of corpus bias in ICLEv3: one on a corpus-ID header that states each essay's country of origin, one on essays mentioning the writer's own country or language, and one that tests whether Brooke and Hirst's (2013) topic-bias finding transfers from an SVM to a zero-shot LLM, and finds that in our data it does not. We show that the SVM and the LLM are affected differently by these problems, and based on how each system is built we suggest a reason for this difference.

2 Related Work

Classical NLI. Koppel et al. (2005) were among the first to treat NLI as an NLP task. They trained SVM models using part-of-speech bigrams, character n-grams, function words, and spelling-error types, and identified the L1 of writers from five language backgrounds with 80% accuracy, using the first version of ICLE. The field later built shared tasks around this problem. Tetreault et al. (2013) ran the first NLI shared task at the BEA-8 workshop and introduced the TOEFL11 corpus (Blanchard et al., 2013), which is the corpus we train on here and describe in Section 3. Malmasi et al. (2017) ran the second shared task, which combined written and spoken data, and drew submissions that used many feature types together, such as spelling errors, word and lemma n-grams, character n-grams, dependency trees, and morphological features.

Cross-corpus NLI and corpus bias. Malmasi and Dras (2015) showed that NLI systems trained and tested within one corpus through cross-validation do not always perform well on a second, separately collected corpus. This makes cross-corpus testing a harder and more informative test than cross-validation alone. Brooke and Hirst (2013) explained one reason why. Using four languages from ICLE (French, Spanish, Chinese, Japanese), they compared accuracy under a random train/test split against accuracy under a topic-controlled split, where no topic appears in both the training and test sets. Every feature type they tested dropped in accuracy under the topic-controlled split. This included part-of-speech n-grams (92.1% to 85.0%) and function words (85.2% to 77.1%), which earlier work had assumed were safe from topic bias. Character n-grams dropped the most (82.9% to 50.2%), and even the combined feature set dropped from 96.5% to 86.1% (all differences significant at $p < .001$). Brooke and Hirst argue this problem is deeper than shared topic words. In ICLE, each institution assigned its own essay topics, so topic is linked to both register and native language. Jarvis and Paquot (2012) suggested a way to fix this: remove four language groups (Chinese, Japanese, Turkish, and Tswana) that show the most variation in topic and skill level, and remove n-grams that also appear in the essay prompts. Brooke and Hirst point out that these four groups are exactly the non-European languages in ICLE, so removing them would limit any resulting system to European languages only. This is a problem for a task that should work across many languages. In this paper, we follow Brooke and Hirst's view rather than Jarvis and Paquot's filtering method. We keep eight typologically different languages, including Japanese, Korean, and Turkish, and instead offer an audit method (Section 6) as another way to address the same concern, without removing

non-European languages.

LLM-era NLI. Zhang and Salle (2023) showed that general-purpose LLMs can reach strong NLI accuracy in a zero-shot setting, without any task-specific training, while also giving natural-language explanations of their reasoning. Ng and Markov (2025) compared open-source and closed-source LLMs on NLI and found that open-source models do not match the accuracy of closed-source models without fine-tuning, but reach comparable accuracy once fine-tuned on labelled data. Most relevant to our work, Uluslu and Schneider (2025) tested whether an LLM's strong NLI performance reflects real structural analysis or reliance on content. They replaced content words with part-of-speech tags, keeping function words and punctuation, and compared LLaMA-3.1-8B and -70B on masked text against full text. On their TOEFL11 subset (TOEFL4: French, German, Italian, Spanish), the 70B model reached 96.5% accuracy zero-shot on full text, which dropped to 55.2% under masking. This is well above the 25% chance rate for their four-way task, but still a large drop. After fine-tuning an 8B model on the masked condition, performance recovered to 89.2% in-domain and 90.3% under cross-domain transfer to their ICLE subset, which shows the structural signal is considerable once the model is trained for it. The authors note that their own study is limited to four closely related European languages.

Gap. To our knowledge, no earlier work has applied this kind of masking-based audit to more than four typologically related languages in a true cross-corpus setting, or tested directly whether Brooke and Hirst's long-standing, SVM-based finding about ICLE's topic bias still holds for a zero-shot LLM that was never trained on ICLE. Goswami et al. (2024), in their survey of the NLI field, treat topic bias as a significant confounder for word-based features and note that NLI models may not generalise across datasets. This paper addresses cross-corpus generalisation and shortcut auditing together, using one experimental design. The two NLI shared tasks (Tetreault et al., 2013; Malmasi et al., 2017) are a useful point of comparison. Both were run as within-corpus tasks with cross-validation, which is the same setting that Malmasi and Dras (2015) later showed does not carry over to a second corpus collected separately. Neither shared task tested its systems with a masking-based check for lexical shortcuts, because this concern only became important once general-purpose LLMs, which have broad world knowledge, started being used for the task (Zhang and Salle, 2023). The gap we address is not only about using more languages than Uluslu and Schneider. It is also about running a shortcut check that was not a common concern when the main way to test NLI systems was cross-validation within TOEFL11 or ICLE.

3 Data

TOEFL11 (Blanchard et al., 2013) has 1,100 essays for each of eleven native languages. All essays respond to the same set of eight prompts, under standard exam-timing conditions. This paper uses eight of the eleven language groups: German, French, Italian, Japanese, Korean, Spanish, Turkish, and Chinese, split into 1,000 training and 100 test essays per language (8,000 training, 800 test).

ICLEv3 (Granger et al., 2020) has argumentative essays from more advanced learners across 25 native-language backgrounds. Unlike TOEFL11, ICLEv3 essays do not share one fixed set of prompts; topics vary by the contributing institution. This is an important design difference, and it motivates the audit in Section 6. This paper uses the same eight language groups from ICLEv3 as an external test set. The counts per language are: German 445, French 314, Italian 398, Japanese 366, Korean 400, Spanish 250, Turkish 286, Chinese 160 (2,619 essays in total).

Because TOEFL11 uses a shared, timed set of prompts and ICLEv3 does not, the two corpora are not equally exposed to topic-based shortcuts or to country and language names in the essay text. We state this design difference here, and we measure its effect directly in Section 6.

4 Methods

4.1 Part-of-speech masking

Following the general idea in Uluslu and Schneider (2025), though not their exact masking setup, we replace content words with part-of-speech tags using spaCy's `en_core_web_sm` model. NOUN, PROPN, VERB, ADJ, and ADV are mapped to single letters (N, P, V, A, R). All other tokens, including function words and punctuation, are left in place, so the masked version keeps the word order and the function words of the essay while removing

its content words. The full-text version keeps paragraph and sentence boundary markers (<P>, <S>). In the masked version spaCy splits each marker into three tokens, so <P> comes out as < P > or a similar form. The boundaries stay visible, but not always which kind of boundary it was. We found this after the runs were complete; a custom tokenizer rule would fix it. Both systems received the same masked text, but the masked LLM prompt describes the markers as intact and also uses P for proper nouns, so the damage may have affected the LLM more than the SVM. We also checked that masking keeps token counts aligned: for all 11,419 essays the masked text has the same number of tokens as the original.

4.2 SVM

We use TF-IDF features (word 1–3 grams, character 2–5 grams, with 5,000 and 10,000 maximum features) with a LinearSVC classifier. Both the masked and full-text vectorizers are fit only on the 8,000 TOEFL11 training essays, never on the TOEFL11 test essays and never on ICLE, in either condition. The classifier is likewise fit only on those 8,000 training essays. Every accuracy we report for the SVM is therefore measured on essays the model did not see during training. This means the comparison between masked and full-text isolates the effect of masking alone. It also matters for a finding in Section 6: since ICLE text is only ever transformed, not used for training, the SVM's vocabulary cannot pick up patterns that appear only in ICLE's raw text.

4.3 LLM

We use gpt-4o-mini in a zero-shot setting, under two prompt formats: a forced-JSON format with a small, 80-token output limit, and a free-text reasoning format, where the model reasons first and then gives a final answer on a separate line, with a 450-token limit. Both formats are tested on the same documents, which lets us test directly whether asking the model to reason changes how much it relies on structure versus content.

We called the model through the OpenAI API using the model name `gpt-4o-mini`. This is an alias rather than a dated snapshot, so the weights it serves can change over time. All four evaluation runs were made on 17 July 2026, and every API response in our logs records the model name `gpt-4o-mini`. Temperature was set to 0 in all runs.

4.4 Evaluation set

All systems and conditions are tested on the same 800-document set: 50 essays per language label per partition (`iclev3_external_test`, `toefl11_test`), reused across every SVM and LLM condition. Using the same set for every condition makes the comparisons in Section 5 fair.

The TOEFL11 half of this set is drawn only from the 800-essay TOEFL11 test split, never from the 8,000 essays used to train the SVM. We checked this directly: no document appears in both the training split and the evaluation set. The TOEFL11 numbers we report are therefore held-out in-domain results, and the ICLEv3 numbers are fully external, since no ICLE essay is used at any point in training.

The 50 essays per label were drawn at random from each test set with a fixed seed. We report results at three filter tiers. The *raw* tier uses all essays, with the ICLE corpus-ID header removed (Section 6.1). The *clean* tier also removes essays that mention their own L1's country or language (Section 6.2). The *doubly-clean* tier also removes ICLE essays on language-exclusive topics (Section 6.3). The topic filter uses only ICLE metadata, so for TOEFL11 the clean and doubly-clean tiers are identical.

4.5 Reproducibility

All code used in this paper is publicly available: the preprocessing and masking pipeline, the SVM and LLM evaluation, the three corpus audits in Section 6, and the scripts for the significance tests reported here. The code is available at `https://github.com/ShmRhm/nli-corpus-artifact-audit`.

We cannot release the corpora themselves. TOEFL11 and ICLEv3 are both licensed, so the released code assumes the reader has obtained them independently and documents the file layout it expects. There are two further limits on exact reproduction. The LLM results depend on the model version served by the OpenAI API at the time of our runs, and we cannot guarantee that the same model will return the same predictions later. The SVM results are deterministic given the same input corpus and a fixed random seed. The audits in Section 6 are deterministic and depend only on the corpus and the corpus metadata, so they can be re-run and checked directly by anyone with a licensed copy of ICLEv3.

5 Results

5.1 Main results

Table 1 shows the main results on the balanced 800-document evaluation set (50 essays per language per partition), which every system and condition shares. Two further SVM numbers are not in Table 1: on the complete, unbalanced sets, the SVM's full-text model reached 74.6% accuracy across all 2,619 ICLE essays and 83.0% across all 800 TOEFL11 test essays. We give these separately because they come from a larger set than the balanced 400-per-partition comparison the LLM is run on, so they are not directly comparable to the numbers in the table.

These results match the pattern discussed in Section 6. On the balanced set in Table 1, the LLM's full-text accuracy is higher on ICLE than on TOEFL11 (87.0% vs. 63.3%), while the SVM shows the opposite pattern (76.5% vs. 83.5%, lower on ICLE). This is an early sign that something drives the LLM's high ICLE accuracy that is not equally available to the SVM.

5.2 Reasoning format

Masked-text accuracy is 14.9% under forced-JSON and 14.6% under the reasoning format, a small decrease rather than an increase. The same small, flat, or slightly negative pattern appears on both partitions (ICLE: 15.0% to 14.8%; TOEFL11: 14.8% to 14.5%). Both formats stay close to the 12.5% chance rate for eight-way classification. Letting the LLM reason before answering does not recover masked-text performance under either format we tested. Uluslu et al. (2024) report a similar pattern on a different dataset: redacting named entities lowered an LLM's NLI F1 from 0.69 to 0.48 under a plain classification prompt, while a chain-of-thought prompt scored lower on unredacted text (0.54) but lost much less under redaction (0.47). Our reasoning format behaves in a similar way: it lowers full-text accuracy (Table 1) without raising masked accuracy. The reasoning prompt also adds a forensic-linguist role and a list of clue types (Appendix A), so this compares two prompts rather than two output formats alone. The near-chance masked result is not caused by the strict output format, because it remains when the model is free to reason. However, it does not tell us whether the model has no structural signal to work with, or is simply weak on this kind of input (Section 8).

5.3 Shortcut gaps on TOEFL11

Our main comparison is the gap between full-text and masked-text accuracy on the TOEFL11 partition, not on ICLE, and not on the unfiltered ICLE score. TOEFL11's shared-prompt, timed design controls for the topic and self-disclosure problems studied in Section 6, while ICLE's design does not.

Table 2 shows these gaps. We prefer the clean tier over the raw tier as our main number. The self-disclosure filter (Section 6.2) removes about 11 to 12 percent of the TOEFL11 test set as well, not only ICLE essays. This means raw TOEFL11 accuracy is already inflated by the same problem the audit is meant to catch. The clean and doubly-clean tiers are identical for TOEFL11, because the topic-exclusivity filter (Section 6.3) is based only on ICLE metadata and does not remove any TOEFL11 rows. The gap is stable from the clean tier onward, but not across every filter, so the filter matters even for TOEFL11, the corpus we use as a control.

All three raw-tier gaps are statistically significant. Because full-text and masked-text accuracy are measured on the same 400 TOEFL11 documents for each system and format, we test this with a paired exact McNemar test on the discordant pairs: LLM forced-JSON (197 documents correct only in the full-text condition versus 3 correct only in the masked condition, $p = 1.7 \times 10^{-54}$), LLM reasoning (177 versus 9, $p = 1.3 \times 10^{-41}$), and SVM (87 versus 23, $p = 6.0 \times 10^{-10}$). All three results are significant well beyond usual thresholds. The SVM's smaller effect size shows up as a less extreme p-value than the LLM's, though still highly significant.

The clean-tier gaps are also significant, using the same paired McNemar test restricted to the 354 TOEFL11 documents that survive the self-disclosure filter (Section 6.2): LLM forced-JSON (155 full-text-only versus 3 masked-only, $p = 3.6 \times 10^{-42}$), LLM reasoning (135 versus 9, $p = 5.5 \times 10^{-30}$), and SVM (70 versus 22, $p = 5.3 \times 10^{-7}$). Removing the self-disclosure confound shrinks the discordant-pair counts, as expected, but does not remove the effect: all three gaps remain significant well beyond conventional thresholds.

5.4 What the errors look like

Accuracy alone does not show whether a system fails in a linguistically sensible way. We therefore looked at which languages each system confuses on

System	Format	Condition	n (all)	Acc. (all)	n (ICLE)	Acc. (ICLE)	n (TOEFL11)	Acc. (TOEFL11)
SVM	—	full text	800	80.0%	400	76.5%	400	83.5%
SVM	—	masked	800	63.4%	400	59.3%	400	67.5%
LLM	forced-JSON	full text	800	75.1%	400	87.0%	400	63.3%
LLM	forced-JSON	masked	800	14.9%	400	15.0%	400	14.8%
LLM	reasoning	full text	798*	67.7%	398*	78.9%	400	56.5%
LLM	reasoning	masked	800	14.6%	400	14.8%	400	14.5%

Table 1: Main results across systems, prompt formats, and masking conditions on the 800-document evaluation set. *Two ICLE documents were dropped from the full-text reasoning condition because of a parse or API error, and were logged as such at evaluation time.

System	Format	Tier	Full	Masked	Gap
LLM	f-JSON	raw	63.3%	14.8%	**48.5**
LLM	f-JSON	clean	58.5%	15.5%	**42.9**
LLM	reasoning	raw	56.5%	14.5%	**42.0**
LLM	reasoning	clean	51.1%	15.5%	**35.6**
SVM	—	raw	83.5%	67.5%	**16.0**
SVM	—	clean	81.9%	68.4%	**13.6**

Table 2: Full-text vs. masked-text accuracy gap (in percentage points) on the TOEFL11 partition. "f-JSON" = forced-JSON. "clean" = "doubly-clean" tier (see Section 6).

the TOEFL11 partition, where prompts are shared and topic bias is controlled.

The SVM's errors follow typological lines. In the masked condition, where only structure is left, its most frequent confusions are Korean predicted as Japanese and Japanese predicted as Korean (9 essays each), together 18 of the 130 errors it makes on this partition, or about 14%. Next comes French predicted as German (6), then five pairs at 5 essays each: Japanese as Turkish, Japanese as Chinese, French as Spanish, Turkish as French and Turkish as Spanish. Japanese, Korean and Turkish are all verb-final and agglutinating, and French, German and Spanish are all Indo-European, so most of these errors stay within groups of related or structurally similar languages. Three do not. Japanese and Chinese differ in word order and morphology, although neither language has articles, and article use survives masking. Turkish predicted as French or Spanish does not fit the pattern at all. The same broad picture holds in the full-text condition, where the largest confusions on this partition are French as German (6), Korean as Japanese (5), Japanese as Korean (4) and Italian as German (4), out of 66 errors. An NLI system that mixes up Korean with Japanese, and French with Spanish, is making the kind of mistake that a linguist would expect from a system reading grammatical structure.

The LLM under masking fails in a different way. It does not spread its errors across typologically close neighbours; it stops using most of the label set. On masked TOEFL11 text under the forced-JSON format, it predicts Spanish for 298 of 400 essays (74.5%) and uses only four of the eight labels at all. Under the reasoning format it predicts Turkish for 190 of 400 essays (47.5%) and uses five labels. For comparison, the same model on full-text TOEFL11 uses all eight labels, with its most frequent prediction accounting for 41% of essays. The near-chance masked accuracy in Table 1 does not come from the model making fine-grained mistakes between similar languages. It comes from the model largely abandoning the task and falling back on a default answer.

Both systems lose accuracy when content words are removed, but only the SVM keeps making linguistically structured errors. The LLM's behaviour under masking looks less like weak structural analysis and more like the model abandoning the task. Section 8 discusses what this means for the model's structural ability.

6 Corpus-Bias Audit

6.1 Document-ID headers

The raw ICLEv3 text files begin with a corpus-ID header, for example `<ICLE-FR-UCL-0027.2>`. This header is part of the corpus file, not of the essay, but it encodes the country code and the institution code of each essay, so it states the answer to the classification task in the first few characters of the input. It is present in all 2,619 ICLEv3 essays, and TOEFL11 files never contain it.

In our first runs the header was passed to both systems as part of the essay text. This was a preprocessing oversight, but an easy one to make, since the header sits in the same file as the essay with no separator and nothing in the corpus documentation warns that it encodes the label. We report its effect because it shows how readily an LLM uses leaked metadata, and so that other users of ICLEv3 can

check for it.

We remove the header from the ICLE text and re-run both systems. This gives a clean comparison, because only one thing changes between the two runs. The effect on the two systems is very different. The LLM's full-text ICLE accuracy drops from 98.8% to 87.0% under the forced-JSON format, and from 98.0% to 78.9% under the reasoning format. The SVM's full-text ICLE accuracy does not change at all: it is 76.5% both before and after, on the balanced evaluation set.

We checked the SVM's fitted TF-IDF vocabularies directly to confirm why. Both vectorizers are fit only on TOEFL11 training text and only ever transform ICLE text (Section 4.2). To confirm this, we passed the header string alone through the fitted vectorizers: it produces no word features at all, and six character n-grams (`ic`, `cl`, `cle`, `le`, `fr`, `uc`). The header therefore carries almost no country evidence the SVM can act on: even `fr`, the country code in this example, is common in ordinary words such as *from* and *free*. This is why the SVM's accuracy is unchanged whether or not the header is present. The LLM reads the raw text at test time and does not make this distinction, so it can use the header's country code wherever it appears.

This is the clearest single case in our data of the wider pattern we report in this paper: the same corpus artifact costs one system between 11.8 and 19.1 accuracy points, depending on the prompt format, and the other nothing at all, for a reason that follows from how each system is built rather than from what each system knows about language.

6.2 Self-disclosure

We use a keyword search to flag essays that mention their own L1's country or language anywhere in the text. As the hand-check later in this section shows, some of these essays state the writer's nationality directly while others only discuss the country. The rate of flagging is uneven across languages, and much higher in ICLE than in TOEFL11 (Table 3).

Label	ICLE (external_test)	TOEFL11 (test)
JPN	71.3%	24.0%
KOR	55.5%	21.0%
SPA	25.2%	4.0%
ITA	25.1%	19.0%
TUR	24.1%	5.0%
ZHO	23.8%	5.0%
FRA	21.0%	11.0%
DEU	18.2%	8.0%

Table 3: Self-disclosure rate by native-language label.

On the 800-document evaluation set, the LLM's full-text forced-JSON accuracy is 67.9% on essays without a self-disclosure flag, and 99.5% on essays with one ($n = 616$ vs. 184). Across all 3,419 essays the SVM was run on (the 2,619 ICLEv3 external test essays plus the 800 TOEFL11 test essays), its full-text model shows a smaller but still clear gap: 73.4% without disclosure versus 84.5% with it ($n = 2{,}422$ vs. 997). Both splits are statistically significant: a chi-square test of self-disclosure flag against correctness gives $\chi^2(1) = 74.0$, $p = 7.7 \times 10^{-18}$ for the LLM and $\chi^2(1) = 47.8$, $p = 4.7 \times 10^{-12}$ for the SVM. The LLM's accuracy gap (31.6 points) is about three times the SVM's (11.1 points); the chi-square values themselves are not comparable, since the two sets differ in size. This is consistent with the pattern in Section 5: the LLM is more sensitive to this particular shortcut than the SVM is, though both are affected well beyond chance.

We checked this filter by hand on a random sample of 48 flagged essays drawn from both corpora. In 15 of them (31%) the writer links themselves to the country or language directly, as in *"My country, Korea"* or *"living in Italy, I know…"*. In 32 the essay discusses its own country or language without the writer saying they belong to it, as in *"In Japan, the death penalty is decided to be the ultimate punishment"*. The remaining one is a plain error: the phrase *"french fries"* in an essay by a French speaker, where the word has nothing to do with nationality. Uluslu et al. (2025) show the same effect in a controlled test: inserting a stereotype sentence about Italian coffee habits into an essay by a Spanish writer led an LLM to predict Italian, even though the prompt instructed it to disregard cultural references. This suggests that the effect comes from country- and culture-linked words in general, not from our keyword list. Our filter is better described as finding essays that mention their own L1's country or language than as finding direct self-disclosure. This does not change the accuracy differences above, which follow from the lexical cue being present in the text at all, but it does mean the filter is broader than the term "self-disclosure" suggests, and we use the term in this wider sense throughout.

The LLM is correct on 183 of the 184 flagged essays in the evaluation set, yet in only about a third of flagged essays does the writer say anything about

belonging to that country. The model is almost always right when an essay names its own L1's country or language, in any sentence and about anyone, whether or not the writer says they come from there. For a teaching tool this makes the problem harder: a learner does not have to write about their own background for the cue to appear.

The filter has two limits. It searches for the terms tied to each essay's own label, so it can only find cues that point the model toward the correct answer. A Spanish writer's essay about Japan is not flagged, even though the word *Japan* could push the model the wrong way. Our clean tier therefore still contains country and language names, just not the essay's own ones, and the 67.9% we report for unflagged essays may be lower than what the model would reach on text with no country names in it at all.

6.3 Topic exclusivity

ICLEv3's metadata includes a Title field, which records the essay prompt or title assigned to each text. We use each distinct title string directly as a topic identifier: we do no clustering or topic modelling, so two essays share a topic only if their Title fields match exactly. For every title we then compute the distribution of native-language labels across the essays carrying it, and flag an essay as topic-exclusive when a single label accounts for at least 80% of the essays under its title. Label shares are computed over all 2,619 ICLE essays, not over the filtered subset. Note that a title carried by a single essay satisfies the rule automatically. Of the 613 titles that meet it, 425 are used by only one essay, and those account for 425 of the 2,058 essays flagged, so "topic-exclusive" partly means "has a title no other essay shares". This is a deliberately simple rule, and Section 8 discusses what it misses. We compute this on the self-disclosure-clean ICLE subset (262 essays: 71 non-exclusive, 191 exclusive). The LLM's full-text forced-JSON accuracy is 91.5% on the 71 non-exclusive essays and 76.4% on the 191 exclusive essays. The exclusive-topic essays are actually harder for the LLM, not easier. This split is statistically significant but visibly weaker than the self-disclosure split above: $\chi^2(1) = 6.6$, $p = 0.010$ (Fisher's exact test, which is a better fit for the small cell counts here, gives $p = 0.0050$, so the effect is not just a side effect of the chi-square approximation). We treat this as an open question. Topic exclusivity is strongly linked to language label itself (German, Japanese, and Korean essays are 97 to 100 percent topic-exclusive, while French and Spanish are 32 to 44 percent), so any claim about "topic" driving this pattern needs to account for label distribution first.

We therefore checked whether label distribution alone explains the gap. It does not. In this subset, no German, Japanese, or Korean essay falls in the non-exclusive group, and those are among the labels the LLM handles worst (Korean 52.6%, Japanese 63.6%), so their absence should favour the non-exclusive group. But when we restrict the comparison to the five labels that appear in both groups (French, Italian, Spanish, Turkish, Chinese), the gap remains almost unchanged: 91.5% on non-exclusive essays ($n = 71$) against 76.4% on exclusive essays ($n = 123$). Whatever makes topic-exclusive essays harder for the LLM is therefore not only a matter of which languages they belong to. We do not have an explanation for it.

We ran the same split for the SVM. Here we use the full self-disclosure-clean ICLE set ($n = 1{,}719$), which is larger than the LLM's 262 essays because the SVM was scored on all of ICLE. Masked accuracy is 56.6% on non-exclusive essays ($n = 429$) against 58.9% on exclusive essays ($n = 1{,}290$). Full-text accuracy is 71.8% against 69.3%. Both gaps are small, and neither is significant (masked: $\chi^2(1) = 0.59$, $p = 0.44$; full text: $\chi^2(1) = 0.84$, $p = 0.36$). The SVM is not sensitive to this split. For the LLM, the topic effect is significant but much weaker than the self-disclosure effect.

6.4 Remaining signal

We removed the essays flagged for mentioning their own country or language, and also the essays on language-exclusive topics. The 71 ICLE essays left show 91.5% accuracy for the LLM's full-text forced-JSON condition, well above the 12.5% chance rate for eight-way classification, though $n = 71$ is a small sample (approximate 95% binomial confidence interval: 83 to 96 percent). This number needs care, because these 71 essays are not an average sample of the clean ICLE set. They are the same non-exclusive group reported in Section 6.3, where the LLM reached 91.5% against 76.4% on the 191 exclusive-topic essays. Because the topic filter takes out the essays the model found harder, 91.5% is the high end of the clean set, not an average of it, and should be read as an upper bound. A signal that stays well above chance af-

ter the self-disclosure and topic filters is what we would expect if some real cross-linguistic signal is present in the text, which is the basic assumption behind NLI. We do not read this residual as a sign that the audits failed, but as a limit on how much of the LLM's ICLE accuracy these audits can account for. Our audits explain part of the LLM's accuracy, but not all of it. We ran a further check. This time we did not look for essays that name their institution, but for essays whose words are much more common at one institution than in the rest of the corpus. For each institution with enough essays to test, we find short word tokens that appear far more often in that institution's essays than elsewhere, then flag any essay that contains at least one such token. In practice this mostly finds shared reading vocabulary, for example character and author names from a novel that a whole class wrote about, rather than any direct mention of the institution's name. We call this the institution-vocabulary check, not "institution self-mention," so as not to suggest the essays name their institution. This gives 74.7% accuracy without a flagged token ($n = 79$) and 83.1% with one ($n = 183$), computed on the same 262-essay self-disclosure-clean set as Section 6.3. Unlike the other splits we report, this difference is not statistically significant (chi-square test of institution-vocabulary flag against correctness: $\chi^2(1) = 2.0, p = 0.16$), so we treat it as suggestive only.

The two topic-based checks also point in opposite directions. Essays with institution-specific vocabulary are answered slightly better (83.1% against 74.7%), while essays on language-exclusive topics are answered worse (76.4% against 91.5%, Section 6.3). Only the second difference is significant, so we do not read this as two competing effects.

We also measured how closely this institution-vocabulary check is related to topic exclusivity, using the full 2,619-essay ICLE set. This test is between the two checks themselves, not against correctness, and it gives $\chi^2(1) = 159.0$, $p = 1.9 \times 10^{-36}$, $\phi = 0.246$. This shows a modest link between the two checks, so we do not treat the institution-vocabulary check as a fully separate third problem. Because its own accuracy split is not significant, we also do not treat it as support for the remaining signal. The two checks are related, since institution-specific words are themselves a kind of topic signal, just found from the data for each institution instead of read from the Title field used in Section 6.3.

6.5 Concrete examples

The audits above report numbers over many essays. To make them easier to picture, we give one real example of each, with any identifying detail removed.

Masking. In a TOEFL11 essay (label: Japanese), the first two sentences in full text are: *"I disagree. I have three opinion to support this reason."* After masking, the same text becomes: *"I V . I V three N to V this N ."* Each content word is replaced one for one (*disagree*→V, *have*→V, *opinion*→N, *support*→V, *reason*→N), while the function words, punctuation, and sentence structure stay the same. This shows why accuracy well above chance on masked text would point to real structural signal, not to the model using content words as a shortcut.

Self-disclosure. One ICLEv3 essay (label: Spanish) is flagged by our keyword search only because it contains the country name *"Spain"* in the text, which is not related to the essay's topic. This is the kind of essay the filter in Section 6.2 is meant to catch: one word matching the essay's own label is enough for the flag, and, as that section shows, usually enough for the model too, whether or not the writer says they come from there.

Institution vocabulary. A group of essays from one Belgian institution are flagged because they share words such as *briscoe*, *ramsay*, *lighthouse*, and *woolf*, which are character and author names from Virginia Woolf's *To the Lighthouse*. No essay in this group names its institution. The institution can still be guessed because a whole class was given the same novel to write about, so the shared book acts as a marker for the institution even though it was never meant to be one.

7 Discussion

The TOEFL11 gaps (Section 5) show that, even on a corpus built to control for topic and register, the zero-shot LLM depends on content words much more than the supervised SVM does: a 35.6 to 48.5 percentage-point gain from full text over masked text for the LLM, depending on prompt format and filter tier, against 13.6 to 16.0 points for the SVM. The smallest gap for the LLM is more than double the largest gap for the SVM. Uluslu and Schneider (2025) report both outcomes

in the same table, so the comparison depends on which of their models is used. Their LLaMA-3.1-70B keeps 55.2% zero-shot accuracy on masked TOEFL4 text, well above the 25% chance rate for their four-way task, while their LLaMA-3.1-8B falls to 26.0%, which is chance. Our zero-shot gpt-4o-mini reaches 14.8% on an eight-way task, close to its own 12.5% chance rate, so our result matches their smaller model rather than their larger one.

Because the two studies use a different number of classes, raw accuracy and raw gaps are not comparable. A measure that is comparable is the share of above-chance accuracy that survives masking, $(\text{masked} - \text{chance})/(\text{full} - \text{chance})$. On this measure, computed on the unfiltered TOEFL11 essays under the forced-JSON prompt so that it is comparable with their setup, our gpt-4o-mini keeps about 4% and our SVM keeps about 77%, against about 42% for their 70B model and about 3% for their 8B. Retention differs sharply between two sizes of the same model family and is similar for the two smaller models, which points to capacity rather than to a general property of large language models. Since gpt-4o-mini's size is not public we can say our result is consistent with that reading, not explained by it. The comparison is still not between identical situations (different models, a different number of classes, and a masking setup that is similar but not identical), so we do not claim that all LLMs rely on shortcuts in the same way. Their 70B model kept some structural signal under masking, while their 8B model and our gpt-4o-mini did not. Model capacity is the most likely reason, but our larger set of languages and our different masking setup may also play a part. Running open models of several sizes through the same pipeline would answer this question.

The masking example in Section 6.5 helps explain why the reasoning format does not close this gap. Under masking, the LLM sees a string like *"I V . I V three N to V this N ."*, which is grammatical form only, with no content words to give a clue. Asking the model to reason before it answers can only help if there is enough structural signal in that part-of-speech sequence, and Section 5 shows there is very little of it for gpt-4o-mini. The problem is not that the prompt forces the model to answer without reasoning; it is the input the model has to work from. This explains why masked-text accuracy stays close to chance in both prompt formats, and it also explains the error pattern in Section 5.4: the model does not make careful, meaningful mistakes between similar languages, it mostly stops using the label set and returns one default answer.

The restriction of the SVM's features to its TOEFL11 training vocabulary explains the difference between the SVM and the LLM on the corpus-ID header (Section 6.1). A corpus-ID string such as `<ICLE-FR-UCL-0027.2>` does not occur in TOEFL11 text, so it matches no word feature and only a handful of character n-grams that are common in ordinary English words (Section 6.1), while the LLM reads all the text it is given at test time. Since shared reading vocabulary is often ICLE-specific in the same way, the same argument explains part, but not all, of the institution-vocabulary effect.

It does not explain the self-disclosure result. Country and language names are ordinary English words that occur in TOEFL11 essays too, so they do have feature columns, and the SVM can use them. That is why Section 6.2 finds the SVM significantly affected by self-disclosure ($\chi^2(1) = 47.8$, $p = 4.7 \times 10^{-12}$), even though the effect is smaller than the LLM's. This reasoning therefore explains the header result, part of the institution-vocabulary result, and none of the self-disclosure result. This does not mean the SVM is a better NLI system in general: it was trained on in-domain data, while the LLM was used zero-shot. It is a narrower point about which kind of mistake each system might make, and it suggests that training only on corpus A and testing on corpus B is not only good practice in general, but also happens to keep this kind of corpus leakage out.

One practical suggestion follows from Section 6: full-text accuracy on ICLE, or on any untimed, variable-prompt learner corpus, should not be reported without a similar check for self-disclosure and topic correlation. Corpora with shared prompts, like TOEFL11, are structurally safer for this purpose, though Section 5 shows that even TOEFL11 is not fully free of self-disclosure leakage and benefits from the same check.

What this means for ICALL. These results matter for building L1-aware feedback into language-learning tools. A common way to build such a tool is to take a learner corpus, train or prompt a system to predict L1, and then use the prediction to select feedback. In a real tool the learner's first language can usually be asked directly, so what NLI adds is knowing which features carry the L1 signal,

since those are what feedback can point to. That makes the source of a system's accuracy more important than its level. Our results suggest two considerations for that pipeline. First, a high reported accuracy on a learner corpus is not by itself evidence that the system has learned anything about the learner's native language. On ICLE, a large part of the LLM's accuracy came from a corpus-ID header, from essays mentioning the writer's own country or language, and from class reading lists, none of which will be present when a real learner types an essay into a tool. A system adjusted and validated on those cues will work accurately in development but fail in real use. Second, the masking result shows that when the lexical cues are hidden, the accuracy of the LLM we tested falls close to chance, which is consistent with it not reading the grammatical patterns that cross-linguistic influence actually produces. Feedback based on such a prediction would not be reliable, and would not be explainable to a teacher or a learner in terms of any real property of the writing. We would therefore suggest that developers of ICALL systems that use L1 prediction audit their training and evaluation corpora in the way we describe here, and report accuracy on content-masked text alongside full-text accuracy, since the gap between the two is a simple, low-cost indicator of how much a system depends on clues that will not transfer to the classroom or to real educational use.

Uluslu and Schneider's measured amount of shortcut reliance came from a four-language, closely related subset of one corpus family, without a separate test for the kind of topic- or self-disclosure-based problems this paper measures. Part of the effect they report may therefore reflect corpus-specific problems in their evaluation data. We did not test their data, so this is only a hypothesis.

8 Limitations

This study uses one closed-source LLM, gpt-4o-mini. Results may not transfer to other model families or sizes. The same care is needed for the main comparison in this paper. Our LLM reaches 63.3% on full-text TOEFL11, below the 83.5% our own SVM reaches on the same essays (Table 1). A model that trails a linear classifier in-domain, and that under masking falls back on a single label for most essays (Section 5.4), may be showing general weakness on unfamiliar input rather than a specific reliance on lexical content. Our results are consistent with both readings, and separating them would need a stronger LLM run through the same pipeline.

Our design also does not separate a second confound. The two systems differ in supervision as well as in architecture and representation: the SVM is fit on 8,000 labelled TOEFL11 essays, while the LLM receives no task-specific adaptation at all. A zero-shot model may simply never have been given the chance to learn that part-of-speech sequences carry L1 information, which would produce the same masked-text collapse we report without any claim about what LLMs attend to in general. Three experiments would separate these factors, and we did not run them: few-shot prompting, to test whether a small number of examples recovers masked-text accuracy; an open-weights model such as LLaMA-3.1, which would also make the LLM side of this comparison reproducible in a way an API alias cannot; and a fine-tuned encoder such as XLM-RoBERTa, which would hold supervision constant and let the effect of architecture be read on its own. Until those are run, our results should be read as a measurement of what this zero-shot system does on this data, not as a general property of large language models. The LLaMA-3.1-70B model of Uluslu and Schneider (2025) kept much more of its accuracy under masking than gpt-4o-mini, while their 8B model kept about as little (Section 7). This points to model capacity, which we could not test with one closed model. Our LLM is also served through an API alias rather than fixed weights (Section 4.3), so running the same comparison on open-weights models with a fixed seed would make the LLM side of it reproducible in a way ours cannot be. We report statistical significance for each comparison: an exact McNemar test for the paired full-text against masked-text gaps in Section 5, and chi-square tests for the confound splits, which compare independent groups of essays (Sections 6.2, 6.3 and 6.4). Section 6.4 reports two of these tests: the accuracy split for the institution-vocabulary check is not significant, while the association between that check and topic exclusivity is. Section 6.3 also reports Fisher's exact test, because the cell counts there are small. The 80% threshold used for topic exclusivity in Section 6.3 is a simple rule, not a precise measure.

Our masking is coarse. Each content word be-

comes one of five letters, so its inflection is lost with it. In the example in Section 6.5, *"I have three opinion"* becomes *"I V three N"*, which hides the missing plural after a number, a common error for learners whose first language does not mark plural on nouns. Masking keeps some kinds of transfer visible, such as article and preposition use, pronouns, word order and punctuation, but removes lexical transfer (cognates, false friends, L1-influenced word choice), spelling and inflection. The masking gap is therefore an upper bound on shortcut use rather than a direct measure of it. Finer tags, or masking in steps starting with named entities, would show how much each kind of information contributes; Uluslu et al. (2025) find that removing named entities alone does not stop an LLM relying on surface cues.

Our topic audit also uses the Title field from the ICLEv3 metadata rather than topics inferred from the essay text. This is deliberate: the Title field is external to the essay body, so it does not carry the L1-correlated content that the audit is meant to detect. A topic model fit on the essays themselves would risk defining topics partly from the same signal we are trying to control for. The cost of this choice is coverage, since two essays on much the same subject can carry different titles and will not be grouped together. Embedding-based topic models such as Top2Vec (Angelov, 2020) could be used to test how much of the residual signal in Section 6.4 is explained by topical similarity that the Title field misses. We leave this to future work, and note that such an analysis would need to be read with the circularity above in mind.

The 71-essay ICLE subset that survives both audits still shows accuracy well above chance, which means our current audit does not fully explain the LLM's ICLE performance. That subset is small, and the confidence interval around its accuracy is wide (Section 6.4), so it can support the claim that some signal remains but not a precise estimate of how much. Answering this would need a larger cross-corpus set audited in the same way, which we hope to build in future work. This study covers eight of the languages shared between TOEFL11 and ICLEv3, chosen because both corpora include them, not through random sampling of all available languages.

9 Conclusion

For RQ1, the supervised SVM and the zero-shot LLM differ in how much they rely on lexical content, even on the TOEFL11 partition, which is built to control for corpus design problems. The LLM gains 35.6 to 48.5 percentage points from full text over masked text, while the SVM gains only 13.6 to 16.0 points. Each of these gaps is statistically significant on its own (Section 5); we did not test whether the LLM's gap is significantly larger than the SVM's, so we report the difference as a consistent ordering rather than as a tested contrast. Letting the LLM reason before answering does not meaningfully reduce this reliance. Our masked-text collapse matches the LLaMA-3.1-8B result reported by Uluslu and Schneider (2025), while their 70B model stayed well above chance, which points to model capacity rather than to a property of large language models in general (Section 7).

For RQ2, a real part of the LLM's apparent full-text accuracy on ICLE, though not all of it, comes from two measurable corpus problems: document-ID headers and self-disclosure (Sections 6.1 and 6.2). The third problem we audited, topic exclusivity (Section 6.3), does not inflate accuracy in our data; essays on language-exclusive topics are harder for the LLM, and we leave that unexplained. The SVM is much less affected by these problems, and in the case of the document-ID header not affected at all, because of how it is built rather than what it knows about language: its features and weights come only from TOEFL11. Some signal remains (91.5% accuracy on the 71 essays that survive both audits); it is well above chance and is not fully explained by our audits. Because the topic filter removes the essays the LLM found harder, this figure should be read as an upper bound on that unexplained signal rather than an average of it (Section 6.4). A further check for institution-specific vocabulary does not reach significance, so it neither supports nor weakens this residual; we report it because it is only moderately linked to topic exclusivity ($\phi = 0.246$) and so is not measuring the same thing twice. Together, these results suggest that NLI studies on untimed, variable-prompt learner corpora need this kind of audit before full-text accuracy can be read as evidence of real language transfer, and that this need applies differently depending on the system's design.

Ethics Statement

This study uses only licensed, de-identified learner corpora (TOEFL11, ICLEv3), obtained under their respective agreements. No new data involving human subjects was collected. All LLM evaluation used gpt-4o-mini through the OpenAI API: 3,200 API calls for the runs reported in Table 1 (800 documents times two text conditions times two prompt formats), which used 2,338,928 prompt tokens and 306,570 completion tokens, taken from the logs recorded with each API response. At gpt-4o-mini list pricing ($0.15 per million input tokens, $0.60 per million output tokens), this comes to about $0.53. SVM training and testing ran locally and needed no paid computing.

Acknowledgments

This paper builds on and substantially extends the author's 2023 thesis, *In-Corpus and Cross-Corpus Analysis of Native Language Identification Using Machine Learning*, supervised by Walter Daelemans and Ehsan Lotfi (Department of Linguistics, University of Antwerp). We thank them for their supervision of that earlier work.

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☒ Compression/expansion
☒ Generation from prompt

Details: An AI assistant (Claude, Anthropic) drafted and revised prose throughout the paper, including the abstract; suggested analyses and wrote the significance-test scripts; produced a first pass of the annotation in Section 6.2, which the author read in full and confirmed; and shortened and expanded text to meet the page limits. The research question, data preparation, pipeline and evaluation runs are the author's own work, and the author has verified every number reported here against the result files. Separately, gpt-4o-mini is an object of study in this paper, not a writing aid.

References

Dimo Angelov. 2020. Top2Vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Julian Brooke and Graeme Hirst. 2013. Native language detection with "cheap" learner corpora. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead*, pages 37–47. Presses universitaires de Louvain, Louvain-la-Neuve.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native language identification in texts: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3149–3160.

Sylviane Granger, Maïté Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *International Corpus of Learner English. Version 3*. Presses universitaires de Louvain, Louvain-la-Neuve.

Scott Jarvis and Magali Paquot. 2012. Exploring the role of n-grams in L1 identification. In Scott Jarvis and Scott A. Crossley, editors, *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*, pages 71–105. Multilingual Matters, Bristol, UK.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628.

Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1403–1409.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75.

Yee Man Ng and Ilia Markov. 2025. Leveraging open-source large language models for native language identification. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Ahmet Yavuz Uluslu, Tannon Kew, Tilia Ellendorff, Gerold Schneider, and Rico Sennrich. 2025. Robust native language identification through agentic decomposition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8398–8414.

Ahmet Yavuz Uluslu and Gerold Schneider. 2025. Investigating linguistic abilities of LLMs for native language identification. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 81–88.

Ahmet Yavuz Uluslu, Gerold Schneider, and Can Yildizli. 2024. Native language identification improves authorship attribution. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 297–303.

Liming Yu and Terence Odlin. 2016. *New perspectives on transfer in second language learning*, volume 92. Multilingual Matters.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

A Prompts

All four prompts are reproduced here exactly as used in the runs of 17 July 2026. Temperature was 0 in every run. The user message in all four conditions was `Essay:\n\n{text}`.

A.1 Forced-JSON format

Output was constrained with the OpenAI structured-output API using a strict `json_schema` with two required properties, `predicted_l1` (enumerated over the eight language codes) and `confidence_score` (a number between 0 and 1), and no additional properties. The completion limit was 80 tokens. Only `predicted_l1` is used in this paper.

Full-text condition.

```
You are an expert computational
linguist evaluating Native
Language Identification from
English essays written by
non-native speakers.

The text may contain <P> and
<S> tokens marking paragraph
and sentence boundaries; treat
them as structural formatting
markers, not content.

Choose exactly one
native-language code from:
DEU, FRA, ITA, JPN, KOR, SPA,
TUR, ZHO.

Return only the requested JSON
object.
```

Masked condition.

```
You are an expert computational
linguist evaluating Native
Language Identification from
content-word-masked English
essays.

The essay uses these
placeholders:
- N = noun
- P = proper noun
- V = verb
- A = adjective
- R = adverb

Function words, punctuation,
pronouns, determiners,
auxiliaries, conjunctions,
prepositions, and structural
markers such as <P> and <S>
remain visible.

Choose exactly one
native-language code from:
DEU, FRA, ITA, JPN, KOR, SPA,
TUR, ZHO.

Return only the requested JSON
object.
```

A.2 Free-reasoning format

The completion limit was 450 tokens. The final label was read off with the regular expression `FINAL_ANSWER\s*:\s*(DEU|...|ZHO)`; a response with no match was logged as a parse error and excluded. The prompt was assembled from an introduction, a boundary-marker legend, a part-of-speech legend used only in the masked condition, and a fixed instruction block.

```
You are an expert computational
linguist and forensic linguist
performing Native Language
Identification (NLI) on English
essays written by non-native
speakers.

The text may contain <P> and
<S> tokens marking paragraph
and sentence boundaries; treat
them as structural formatting
markers, not content.

[masked condition only:]
The essay's content words
have been replaced with these
placeholders:
- N = noun
- P = proper noun
- V = verb
- A = adjective
- R = adverb
Function words, punctuation,
pronouns, determiners,
auxiliaries, conjunctions, and
prepositions remain in their
original form.

Think step by step.  Identify
concrete linguistic clues -
```

word order, article/preposition use, function-word patterns, syntactic structure, collocations, or spelling/grammar patterns - that point to a specific native language. Write out your reasoning in a short paragraph.

After your reasoning, end your response with exactly one line in this exact format, with no extra text after it:
FINAL_ANSWER: <CODE>

<CODE> must be exactly one of:
DEU, FRA, ITA, JPN, KOR, SPA, TUR, ZHO.

Lexical Profiling for Dutch: From Resource to Real-Life Implementation

Jasper Degraeuwe[1], **Natalie Finlayson**[2,3]

[1]LT[3] and MULTIPLES, Ghent University, Belgium
[2]University of York, England
[3]Queen's University Belfast, Northern Ireland

jasper.degraeuwe@ugent.be natalie.finlayson@york.ac.uk

Abstract

Lexical profiling tools use natural language processing (NLP) software to measure the distribution of words in an input text across precompiled word lists. In Intelligent Computer-Assisted Language Learning (ICALL) contexts, foreign/second language (L2) teachers can use the output of lexical profilers to prepare their classes (e.g., by selecting or adapting texts for particular learner groups) or to analyse learner texts. The present study advances the state of the art in lexical profiling by introducing a first-of-its-kind lexical profiling resource for L2 Dutch (dubbed `NLLP-5000`) and demonstrating its use in two concrete ICALL applications. The resource, based on Tiberius and Schoonheim's *A Frequency Dictionary of Dutch* and made publicly available through an open repository, is characterised by the inclusion of inflected forms – obtained through a meticulous manual verification process – and the linguistic annotation of these forms in Universal Dependencies (UD) format. The UD-based annotations in `NLLP-5000` make the resource compatible with lexical profiling tools built on widely used NLP toolkits such as spaCy and Stanza.

`https://github.com/JasperDUGent/NLLP-5000`

1 Introduction

Since the advent of neural networks, deep learning, and (generative) large language models (LLMs), the "Intelligent" in Intelligent Computer-Assisted Language Learning (ICALL) has increasingly come to denote "artificially intelligent". Recent work in the field of ICALL has successfully used AI-driven methods to generate written corrective feedback (Wu et al., 2025), create conversational intelligent tutoring systems (Pérez-Ortiz et al., 2024), tailor texts to the needs of specific learners (Bodnar, 2025), build graded vocabulary lists (Alfter, 2024), and generate pedagogically relevant questions on existing educational materials (Gerald et al., 2026), among other applications and uses.

However, even though the use of AI for educational purposes offers considerable potential and added value, this does not imply that all existing ICALL tools should be replaced by AI or augmented with AI-based components. As argued by Stamper et al. (2024) and Soliar et al. (2026) in the context of intelligent tutoring systems, "one should not disregard previous knowledge and years of experience using expert-generated and traditional data-driven approaches" (Soliar et al., 2026). A similar trend can be observed in the field of corpus linguistics, where LLMs were found to be good at analysing and classifying frequency and collocation data, but to be doing a poor job of generating these data compared to traditional corpus methods (Davies, 2025).

In this paper, we follow this line of thought for the task of lexical profiling, defined by Finlayson et al. (2023) as "the practice of using software to measure the distribution of lexical items in texts across one or more word lists". In the field of ICALL, lexical profiling tools have two main applications: (1) helping teachers prepare foreign/second language (L2) classes and (2) supporting teachers (and researchers) in the analysis and assessment of learner texts.

The contribution of the paper is twofold. First, building on previous knowledge and ex-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 16–33, November 19–20, 2026, University of Tartu Library

isting resources, we introduce NLLP-5000[1], a first-of-its-kind resource for lexical profiling in Dutch. Second, we integrate NLLP-5000 into MultilingProfiler (Finlayson et al., 2023), a widely used lexical profiling tool, and provide a demonstration of the two above-mentioned applications (class preparation and learner text analysis/assessment).

2 Related Research

2.1 Frequency Dictionaries in SLA

When learning L2 vocabulary, frequency matters (Vilkaitė-Lozdienė and Schmitt, 2019). Although the exact numbers differ slightly across studies and languages, it has been shown that the top 5,000 most frequent word families in a language generally account for approximately 95% of the words in a written text (Green, 2020; Laufer and Ravenhorst-Kalovski, 2010). Consequently, drawing on methods from the domains of corpus linguistics (Sinclair and Renouf, 1988) and lexicography (Frankenberg-Garcia et al., 2021), the field of second language acquisition (SLA) has devoted considerable attention to developing frequency dictionaries that can inform course material developers, researchers, and teachers about which words to prioritise in vocabulary learning activities and test batteries (Nation, 2022).

A good illustration is provided by Brezina and Gablasova (2015), who introduced a new version of the General Service List (West, 1953) for English. Consisting of 2,494 lemmas obtained through a frequency-based selection method, the *new-GSL* covers about 80% of text in a collection of four large corpora (LOB, BNC, BE06, and EnTenTen12, with a combined size of more than 12 billion running words). As a result, L2 teachers and coursebook designers using the *new-GSL* can be reasonably confident that familiarising L2 English learners with its content will enable them to understand most of the words they encounter in running text.

2.2 Lexical Profiling

Apart from informing practitioners during the development of vocabulary learning activities, frequency dictionaries also constitute a valuable resource for lexical profiling tools. Lexical profiling can be used to characterise the vocabulary demands of a text, providing an indication of its lexical load, the potential for vocabulary learning through reading or listening, and the vocabulary knowledge required to understand different text types (Nurmukhamedov and Webb, 2019). Lexical profiling tools do this by comparing the lexical items in a text against one or more word lists representing the expected or targeted vocabulary of a particular learner group or proficiency level, including relevant inflected and derived forms. The proportion of the words in a text that is covered by a given word list is referred to as lexical coverage, and can be used to indicate the extent to which the vocabulary demands of a text are likely to be met by the assumed vocabulary knowledge of a learner group. In pedagogical contexts, such information can help teachers classify, select, and adapt texts for particular learner groups according to their vocabulary level (Dang, 2022).

3 Resource Creation

3.1 Input Data

The first contribution of the present paper is to describe the development of a first-of-its-kind lexical profiling resource for L2 Dutch. As input data to create the resource, *A Frequency Dictionary of Dutch* (Tiberius and Schoonheim, 2014) was used, one of the most extensive efforts to date to present core vocabulary to L2 learners of Dutch. As this resource is part of the *Routledge Frequency Dictionaries* series, it is henceforth referred to as the "Routledge dictionary". The dictionary consists of six separate word lists derived from four corpora[2], each with its own index (a sample from the "Core" list is presented in Figure 1). It is described as containing "the 5,000 top words of Dutch" (Tiberius and Schoonheim, 2014), but because the same headword is allowed to appear in multiple lists[3], the dictionary contains more than 5,000 entries in total. An entry consists of the following five elements:

- **Headword** (boldfaced): usually corresponds to the lemma as lexical unit (see Section 3.2.2 for more details)

[1]For **NL** (Dutch) **L**exical **P**rofiling.

[2]The corpora each represent a different genre: fiction (23 million words), newspaper (167 million), spoken (9 million), and web (91 million). For each corpus, an eponymous "genre list" was created based on a custom procedure (Tiberius and Schoonheim, 2014). The additional "Core" and "General" lists include words occurring frequently across all genres and across at least three of the genres, respectively.

[3]However, the frequency values attached to these duplicate entries are identical across lists. Only the example sentences provided differ depending on the list.

- **Part of speech** (italicised): adjective (`adj`), adverb (`adv`), article (`art`), conjunction (`conj`), interjection (`interj`), noun (`noun`), number (`num`), preposition (`prep`), pronoun (`pron`), or verb (`verb`)
- **English translation** (if the headword has multiple meanings, translation of the most common meaning)
- **Example sentence** (if the headword has multiple meanings, sentence illustrating the most common meaning)
- **Frequency value**: custom dispersion value based on the number of 2,000-word corpus samples in which the headword occurs, normalised across genres

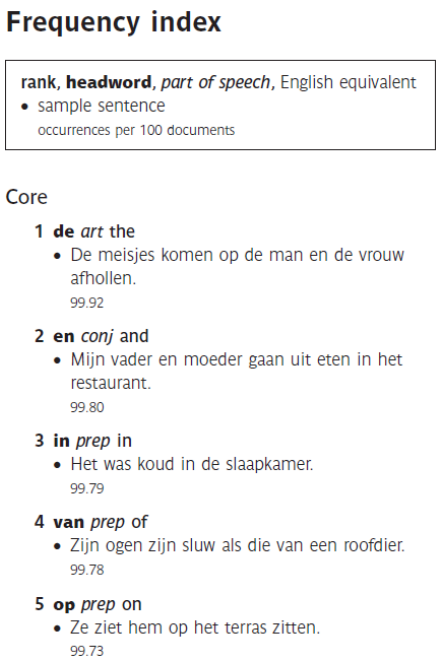


Figure 1: Sample from the "Core" list in *A Frequency Dictionary of Dutch* (Tiberius and Schoonheim, 2014).

3.2 Conversion into Lexical Profiling Resource

To improve consistency and enable seamless integration into existing lexical profiling tools, a series of modifications and additions were made to the input data[4] presented in Section 3.1.

[4]The official TXT file of the dictionary provided by Routledge contains some typographic errors and formatting issues; these were manually fixed. An overview of these fixes is presented in Appendix A. The resulting cleaned file served as input data to create `NLLP-5000`.

3.2.1 Removing Duplicate Entries

Process. Several headwords appeared in more than one of the six lists included in the Routledge dictionary, resulting in duplicate entries for one and the same headword (see Section 3.1). As a first step, these duplicate entries were removed.

Overview of modifications. Of the 5,838 entries included in the Routledge dictionary, 671 duplicate entries were removed. This yielded a provisional resource of 5,167 unique headwords.

3.2.2 Lemmatising Headwords

Process. To ensure consistency, word lists usually use a single type of lexical unit for their headwords. For our resource, we opted for the lemma[5] (see also Section 6 for a brief discussion of lexical units). All unique headwords in the input data which did not correspond to this lexical unit were modified. This task was performed manually by the first author of this paper, who is a native speaker of Dutch and holds a PhD in Linguistics. All decisions were taken based on the *Woordenlijst Nederlandse Taal* as reference work, accessed via Woordenlijst.org[6]. This resource prescribes the official Dutch spelling and provides extensive metadata on all words included in it (e.g., part of speech and inflected forms; see Figure 2 for an illustration). Table 1 illustrates the modifications made during this stage: 24 of the headwords retained after removing duplicates appeared to be inflected forms which needed to be changed into the base form, one entry had more than one part of speech (POS) and needed to be split into separate headwords, and for six entries the POS had to be changed.

After this lemmatisation step, 13 "lemma duplicates" were identified (i.e. word forms appearing as separate entries in the dictionary but belonging to the same lemma; see Table 2). In these cases, the base form was used as the headword and the highest frequency value among the individual word forms was assigned to the lemma. Since the Routledge dictionary includes dispersion-based frequency values and not raw frequencies (see also

[5]That is, a base form combined with a part of speech, including all of its inflected forms. We also consider diminutive forms of nouns (e.g., *dag* ['day'] → *dagje*) to be part of the lemma. Although traditional Dutch grammar treats diminutives as derivations (partly because they can change the noun's gender), they behave like inflections in many other respects (e.g., highly regular and no change in word class).

[6]`https://woordenlijst.org/`

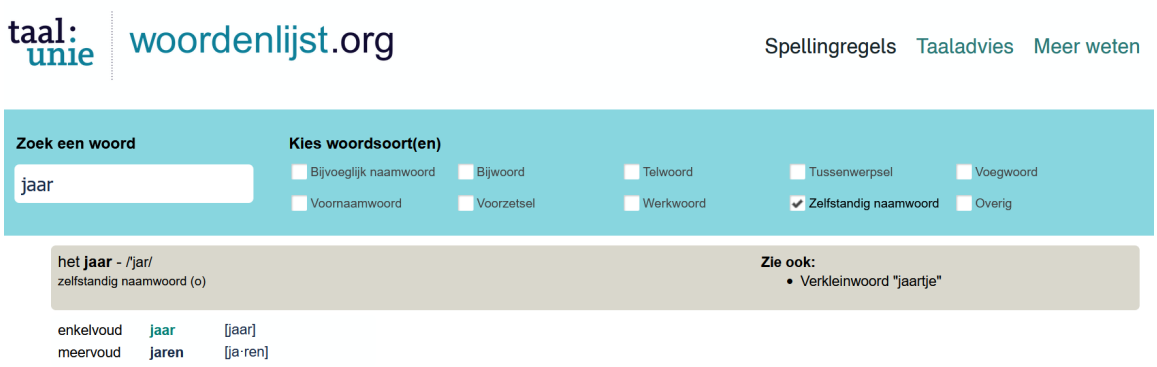


Figure 2: The *Woordenlijst Nederlandse Taal* entry for *jaar* ('year'), accessed via Woordenlijst.org (`https://woordenlijst.org/zoeken/?q=jaar`).

Type	Description	# cases	Applied modification	Instances
Inflected headword	Original headword is an inflected form, not the base form of word	24	Headword changed into base form	All inflected forms in Table 2; *beetje* (diminutive) → *beet*; *inkomsten* (plural) → *inkomst*; *jezelf* (accusative/dative) → *jijzelf*; *kleintje* (diminutive) → *kleine*; *kleren* (plural) → *kleed*; *meerder* as `num` → *meerdere*; *onszelf* (accusative/dative) → *wijzelf*; *praatje* (diminutive) → *praat*; *sommige* → *sommig*; *sprookje* (diminutive) → *sproke*; *werkzaamheden* (plural) → *werkzaamheid*
Multiple POS	Original headword has multiple POS tags assigned to it	1	Headword split into separate headwords	*het* → headword as `art` and as `pron`
Wrong POS	Original headword is tagged with an incorrect POS	6	POS changed	*ander* from `adj` to `num` and `pron`; *dergelijk* from `adj` to `pron`; *anno* from `prep` to `adv`; *alstublieft/alsjeblieft/a.u.b.* from ∅ to `interj`; *cc* from ∅ to `noun`; *o.a.* from ∅ to `adv`

Table 1: Modifications made to the original input data. The list of instances is exhaustive. The ∅ symbol means that no POS was assigned in the Routledge dictionary. Translations are provided in Appendix E.

Section 3.1), the combined frequency value of entries belonging to the same lemma could not be determined: because different word forms of the same lemma may occur in the same samples, summing their values would overestimate the lemma's dispersion. The highest value among the word forms was therefore the best available approximation, albeit a conservative one.

As an illustration, let us consider the forms *laat* ('late'), *later* ('later'), and *laatst* ('latest'), all included as separate entries in the Routledge dictionary. The three forms belong to the same lemma *laat*: *later* is the uninflected comparative form and *laatst* is the uninflected superlative form. *Laat* has a frequency of 30.43, *later* of 35.37, and *laatst* of 44.76. In our resource, the three separate entries are merged into one, namely the base form *laat* as the headword with a frequency value of 44.76 (i.e. the highest among the three individual values).

Overview of modifications. Of the 5,167 entries retained after the first step, 13 "lemma duplicates" were removed (Table 2), 17 entries were modified[7], and two entries were added (one for *het* and one for *ander*). After this step, our provisional resource contained 5,156 headwords in lemma form.

3.2.3 Adding Inflected Forms

Process. The original Routledge dictionary does not provide information on the inflected forms of its entries. However, information on inflected forms is essential for lexical profiling tools (see also Section 2.2). For example, once learners have acquired the basic morphological rules of the L2

[7]This number excludes modifications made for entries which appeared to be lemma duplicates.

frequency	headword_orig	M	O	V	M-O	M-O-V	M-V	O-V	eigennaam	standaardvorm_ENK	standaardvorm_MV	verkleinwoord_ENK	verkleinwoord_MV
1.18	goed		x							goed	goeden	goedje	goedjes
89.77	jaar		x							jaar	jaren	jaartje	jaartjes
73.56	mens	x								mens	mensen	mensje	mensjes
6.71	weer		x							weer	weren	weertje	weertjes
70.73	tijd	x								tijd	tijden	tijdje	tijdjes
7.14	haar									haar	haren	haartje	haartjes
66.05	dag	x								dag	dagen	dagje	dagjes
60.7	plaats						x			plaats	plaatsen	plaatsje	plaatsjes
2.34	enkel	x								enkel	enkelen	enkeltje	enkeltjes

Figure 3: Illustration of the forms automatically populated according to the predefined rules. Incorrect forms were manually corrected (e.g., **goeden* as plural of *goed* ['good'] should become *goederen*) and non-existing forms were replaced by "N/A" (e.g., **weren* as plural of *weer* ['weather'] does not exist).

Base form	**POS**	**Duplicate entries in Routledge dictionary**
laat	`adj`	*later* (comparative), *laatst* (superlative)
ver	`adj`	*verder* (comparative)
goed	`adj`	*best* (superlative)
een	`art`	*ene* (inflected form with *buigings-e*)
goed	`noun`	*goederen* (plural)
kost	`noun`	*kosten* (plural)
medium	`noun`	*media* (plural)
ik	`pron`	*mij/me* (accusative, dative)
hij	`pron`	*hem/'m* (accusative, dative)
zij/ze	`pron`	*hen* (accusative)
jij/je	`pron`	*jou* (accusative, dative)
ikzelf	`pron`	*mijzelf/mezelf* (accusative, dative)

Table 2: Lemma duplicates identified after the lemmatisation process. For translations see Appendix E.

they are studying, they are likely to know the plural form of a noun if they already know its singular form. Consequently, when an L2 teacher uses a lexical profiling tool to prepare instructional materials for such learners, the tool should be able to treat both forms as either known or unknown, even when their frequency values differ substantially.

Therefore, the resource was enriched with the inflected forms listed below. Note that not all headwords have all possible inflections: adjectives such as *huidig* ('current') and *vorig* ('previous') do not have comparative or superlative forms.

- **Adjectives**: positive degree (standard form, inflected form with *buigings-e*[8], and inflected form with *buigings-s*), comparative degree (standard form, inflected form with *buigings-e*, and inflected form with *buigings-s*), and superlative degree (standard form and inflected form with *buigings-e*)
- **Adverbs**: not applicable (in Dutch, adverbs have one invariable form)
- **Articles**: nominative case, accusative case, genitive case, and dative case (the article *een* also has an inflected form with *buigings-e* in nominative, accusative, and dative)
- **Conjunctions**: not applicable (one invariable form)
- **Interjections**: not applicable (one invariable form)
- **Nouns**: standard degree (singular form and plural form) and diminutive degree (singular form and plural form)
- **Numbers**: standard form, inflected form with *buigings-e*, and inflected form with *buigings-n*
- **Prepositions**: not applicable (one invariable form)
- **Pronouns**: nominative case, accusative case, genitive case, and dative case (some pronouns also have inflected forms with *buigings-e* and *buigings-n* in nominative, accusative, and dative)
- **Verbs**: infinitive, present tense forms, past tense forms, imperative, conjunctive, present participle, and past participle

The inflected forms were added using a semi-automatic process. First, a Python script (made available in the `NLLP-5000` repository[9]) was used to automatically populate all forms based on a set of predefined rules, as illustrated in Figure 3.

[8] More information on this and other *buigings* forms can be found in Appendix B.

[9] `https://github.com/JasperD-UGent/NLLP-5000` (`NLLP-5000_compilation`).

Original POS tag	UD tag(s)	Mapping type	Mapping details
adj (adjective)	ADJ	1-to-1	Mapping to fully equivalent UD ADJ tag.
adv (adverb)	ADV ADP PART	Manual	Regular adverbs were mapped to the ADV tag; verbal particles were mapped to the ADP tag; particles encoding negation (e.g., *niet* ['not']) were mapped to the PART tag.
art (article)	DET	1-to-1	In UD, articles are considered determiners (i.e. words that modify nouns or noun phrases and express the reference of the noun phrase in context). Therefore, the original art tag was mapped to the UD DET tag.
conj (conjunction)	CCONJ SCONJ	Manual	In UD, conjunctions are divided into coordinating conjunctions (CCONJ; e.g., *of* ['or']) and subordinating conjunctions (SCONJ; e.g., *hoewel* ['although']). All conjunctions in the dictionary were mapped to their corresponding type.
interj (interjection)	INTJ	1-to-1	Mapping to fully equivalent UD INTJ tag.
noun (noun)	NOUN PROPN	Manual	In UD, nouns are divided into common nouns (NOUN; e.g., *aanval* ['attack']) and proper nouns (PROPN; e.g., *Bijbel* ['Bible']). All nouns in the dictionary were mapped to their corresponding type.
num (number)	NUM	1-to-1	Mapping to fully equivalent UD NUM tag.
prep (preposition)	ADP	1-to-1	In UD, prepositions form part of the broader adposition category, which also encompasses circum- and postpositions. Therefore, the original prep tag was mapped to the UD ADP tag.
pron (pronoun)	PRON	1-to-1	Mapping to fully equivalent UD PRON tag.
verb (verb)	AUX VERB	Manual	In UD, verbs are divided into auxiliary verbs (AUX; e.g., *kunnen* ['can']) and main verbs (VERB; e.g., *creëren* ['to create']). Note that some verbs can be used as both types (e.g., *gaan* ['to go']). All verbs in the dictionary were mapped to their corresponding type(s).

Table 3: Overview of mapping from POS tags used in original dictionary to UD tags.

Second, all these forms were manually checked by the first author of the paper.

Overview of additions. In this step, 40,169 inflected forms were added to the 5,156 headwords obtained in the previous step. Manual verification ensured a high degree of data quality, while the exhaustive inventory of inflected forms gives the resource its unique, first-of-its-kind character.

3.2.4 Enriching Linguistic Information

Process. To allow for compatibility with lexical profiling tools offering more advanced features (e.g., removing certain morphological inflections), all inflected word forms added in the previous step were tagged with POS tags and morphological features in the Universal Dependencies (UD) format (De Marneffe et al., 2021). UD is a framework for cross-linguistically consistent annotation of human language at a grammatical level (i.e. labelling of parts of speech, morphological features, and syntactic dependencies). Covering more than 150 languages and used by popular natural language processing (NLP) toolkits such as spaCy[10] and Stanza (Qi et al., 2020), UD has become one of the most widely used frameworks for morphosyntactic analysis in the field of computational linguistics.

Table 3 shows how the POS tags used in the original dictionary were mapped to UD tags. One-to-one mappings were performed automatically; manual mappings were performed by the first author. More information on the definition of the morphological features is presented in Appendix C. The Python scripts used to add the UD POS tags and morphological features to the resource are publicly available in the above-mentioned GitHub repository.

Overview of additions. In this step, the 40,169 inflected forms (belonging to 5,156 unique headwords) from the previous step were enriched with UD POS tags and morphological features, allowing for the seamless integration of the resource into existing lexical profiling tools. This integration process will be described in Section 4 and demonstrated in Section 5.

[10] https://spacy.io/

3523-1	3,523	4	verlichten	verb	verlichten	VERB	VerbForm=Inf
3523-2	3,523	4	verlichten	verb	verlicht	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
3523-3	3,523	4	verlichten	verb	verlicht	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin
3523-4	3,523	4	verlichten	verb	verlicht	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin
3523-5	3,523	4	verlichten	verb	verlicht	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
3523-6	3,523	4	verlichten	verb	verlichten	VERB	Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin
3523-7	3,523	4	verlichten	verb	verlichten	VERB	Mood=Ind|Number=Plur|Person=2|Tense=Pres|VerbForm=Fin
3523-8	3,523	4	verlichten	verb	verlichten	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin
3523-9	3,523	4	verlichten	verb	verlichtte	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin
3523-10	3,523	4	verlichten	verb	verlichtte	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Past|VerbForm=Fin
3523-11	3,523	4	verlichten	verb	verlichtte	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin
3523-12	3,523	4	verlichten	verb	verlichtten	VERB	Mood=Ind|Number=Plur|Person=1|Tense=Past|VerbForm=Fin
3523-13	3,523	4	verlichten	verb	verlichtten	VERB	Mood=Ind|Number=Plur|Person=2|Tense=Past|VerbForm=Fin
3523-14	3,523	4	verlichten	verb	verlichtten	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin
3523-15	3,523	4	verlichten	verb	verlicht	VERB	Mood=Imp|VerbForm=Fin
3523-16	3,523	4	verlichten	verb	verlichte	VERB	Mood=Sub|VerbForm=Fin
3523-17	3,523	4	verlichten	verb	verlichtend|verlichtende	VERB	Tense=Pres|VerbForm=Part
3523-18	3,523	4	verlichten	verb	verlicht	VERB	Tense=Past|VerbForm=Part
3649-1	3,649	4	opinie	noun	opinie	NOUN	Gender=Fem|Number=Sing
3649-2	3,649	4	opinie	noun	opinies	NOUN	Gender=Fem|Number=Plur
3649-3	3,649	4	opinie	noun	opinietje	NOUN	Degree=Dim|Gender=Fem|Number=Sing
3649-4	3,649	4	opinie	noun	opinietjes	NOUN	Degree=Dim|Gender=Fem|Number=Plur
3683-1	3,683	4	badkamer	noun	badkamer	NOUN	Gender=Fem,Masc|Number=Sing
3683-2	3,683	4	badkamer	noun	badkamers	NOUN	Gender=Fem,Masc|Number=Plur
3683-3	3,683	4	badkamer	noun	badkamertje	NOUN	Degree=Dim|Gender=Fem,Masc|Number=Sing
3683-4	3,683	4	badkamer	noun	badkamertjes	NOUN	Degree=Dim|Gender=Fem,Masc|Number=Plur
4030-1	4,030	5	tevergeefs	adj	tevergeefs	ADJ	Degree=Pos
4030-2	4,030	5	tevergeefs	adj	tevergeefse	ADJ	Degree=Pos
4030-3	4,030	5	tevergeefs	adj	tevergeefs	ADJ	Degree=Pos
4036-1	4,036	5	erom	adv	erom	ADV	N/A

Created with Datawrapper

Figure 4: Sample of NLLP-5000. The visualisation was created with Datawrapper.

3.2.5 Converting Frequency Values

Process. Finally, the 5,156 unique headwords were ranked from highest frequency to lowest. Only the top 5,000 items were retained, meaning that the 156 items with the lowest frequency were removed from the resource. For the 5,000 selected items, we determined two new frequency-related values:

- **Frequency rank**: position of the headword in the ranked list (1 for headword with highest frequency; 5,000 for headword with lowest frequency)
- **Frequency band**: when dividing the list into five equal parts (1–1,000 | 1,001–2,000 | 2,001–3,000 | 3,001–4,000 | 4,001–5,000), part to which the headword belongs

Overview of modifications. In this step, the resource was trimmed down from 5,156 to 5,000 headwords and the original dispersion values were replaced by frequency rank and frequency band values. The set of 5,000 headwords in lemma form accounts for a total of 38,811 inflected forms.

3.3 Final Resource

Figure 4 shows a sample of the final version of our lexical profiling resource, dubbed NLLP-5000. The main resource statistics are presented in Table 4. The full resource is made available in the GitHub repository[11]. Note that the English translation and the example sentences (see Section 3.1) are not included in the resource, as the

[11] https://github.com/JasperD-UGent/NLLP-5000 (NLLP-5000_data).

Band	adj	adv	art	conj	interj	noun	num	prep	pron	verb	Inflections
1–1,000	156	123	3	18	4	359	30	38	50	219	7,397
1,001–2,000	177	67	0	7	6	480	16	5	5	237	7,625
2,001–3,000	167	54	0	4	4	507	4	6	3	251	7,759
3,001–4,000	172	36	0	4	3	524	6	2	3	250	7,780
4,001–5,000	161	61	0	5	7	461	7	1	4	293	8,250
Total	833	341	3	38	24	2,331	63	52	65	1,250	38,811

Table 4: Main statistics for `NLLP-5000`: number of headwords per POS and number of inflections, reported per frequency band and in total.

Inflected form	UD POS	UD features	FreeLing tag
komt ('comes')	VERB	Mood=Ind | Number=Sing | Person=3 | Tense=Pres | VerbForm=Fin	`VVIPS3`
kom ('come')	VERB	Mood=Imp | VerbForm=Fin	`VVM000`
gekomen ('come')	VERB	Tense=Past | VerbForm=Part	`VVPS00`

Table 5: Examples of the conversion from UD to FreeLing, three inflected forms of the headword *komen* (`verb`).

Routledge licence agreement prohibits redistributing these data.

4 Implementation in Existing Tools

Although lexical profiling tools serve the same purpose, they often employ different procedures for processing input texts and require their source lists to conform to tool-specific formats. As a result, integrating existing word lists into these tools can be cumbersome. When building the `NLLP-5000` resource, we anticipated this issue by enriching all word forms in the resource with detailed linguistic information (POS tag and morphological features) in the standardised UD format (see Section 3.2.4). With simple rule-based scripts, the UD-based entries from our resource can easily be converted into other formats with little or no loss of information.

In the present study, we illustrate this conversion process for the **MultilingProfiler** tool (Finlayson et al., 2023), which expects source lists in a FreeLing-based format (Padró and Stanilovsky, 2012). The FreeLing tags constitute essential information for the functionalities involving morphological inflections. MultilingProfiler's "Remove past tense forms" feature, for example, can prove to be very useful for teachers whose learners have yet to learn this verbal inflection and who therefore want the profiler to mark all past tense forms as unknown. To illustrate the format conversion, let us consider the POS category of verbs, for which the following FreeLing feature-value pairs[12] were required:

- **POS letter**: always `V`
- **Verb type**: `V` for lexical ("full") verbs, `A` for regular auxiliary verbs, `M` for modal verbs
- **Mood**: `C` for conjunctive/subjunctive, `I` for indicative, `M` for imperative, `N` for infinitive, `P` for participle, `0` if not applicable
- **Tense**: `P` for present, `S` for past, `0` if not applicable
- **Number**: `P` for plural, `S` for singular, `0` if not applicable
- **Person**: `1` for first person, `2` for second person, `3` for third person, `0` if not applicable

Using a rule-based script (Appendix D), the appropriate FreeLing tag could be assigned automatically to all verb entries (see Table 5 for examples).

[12]The final FreeLing tag is obtained by concatenating the values of all features into a single string (e.g., `VVIPS1` for a lexical verb in the indicative mood, present tense, and first-person singular).



Figure 5: Demonstration of preparing a text for a B2-level L2 Dutch class using MultilingProfiler and `NLLP-5000` (profiler output).

The same procedure was applied to the remaining POS categories.

5 ICALL Applications

5.1 Class Preparation

Description. Consider a teacher preparing a B2-level class on the topic of "nature". The school curriculum assumes that learners at this level are already familiar with the 5,000 most frequent lemmas in Dutch (i.e. the lemmas included in `NLLP-5000`). When selecting a text for classroom use, the teacher may therefore wish to identify words outside this assumed vocabulary. These words can then be addressed in a preparatory vocabulary activity or, in case of a reading-focused lesson, provided in a glossary.

To illustrate this process, let us suppose the teacher found an interesting news article online (see Appendix G for the full text). The teacher enters the article into MultilingProfiler, which then profiles the text against the `NLLP-5000` resource and highlights all words not covered by this list (Figure 5). The tool also provides summary statistics, including the percentage of running words covered by the list (Figure 6), and allows teachers to deselect certain inflections (see example of removing past tense forms discussed in Section 4). These outputs allow the teacher to quickly assess the lexical demands of the text and identify vocabulary that may require additional attention.

Discussion. In this application, it is assumed that the target learners are familiar with the 5,000 lemmas included in the `NLLP-5000` resource. Naturally, this assumption will only hold if learning these 5,000 lemmas forms part of the pre-B2 phase of the study curriculum. In language learning curricula taking learner-oriented corpus-based vocabulary lists (e.g., Kilgarriff et al., 2014) as their starting point, different assumptions will need to be made. To this end, MultilingProfiler offers the possibility to upload a custom list to be used as reference point for the profiling.

5.2 Learner Text Analysis and Assessment

Description. Alternatively, teachers can use `NLLP-5000` in a lexical profiling tool to analyse vocabulary use in learner-produced text (the results of which could then be used as a source of evidence in assessment). Consider the authentic learner text presented in Appendix F. By profil-

Global coverage

Total words covered by all lists (including multiword units)	446	82.1%

Word statistics (Tokens; each word counted separately. Figures show text coverage.)	count	% coverage
Total words from your chosen list	446	82.1%
Total words from your extended list	0	0.0%
Total cognates from outside your chosen list	0	0.0%
Total words from all lists (chosen list + extended list + cognates)	446	82.1%
Total words in the text	543	
Type-token ratio	0.5	

Figure 6: Demonstration of preparing a text for a B2-level L2 Dutch class using MultilingProfiler and NLLP-5000 (statistics).



Figure 7: Demonstration of analysing learner text using MultilingProfiler and NLLP-5000 (profiler output).

ing this text in MultilingProfiler, teachers can obtain an indication of how advanced the learner's vocabulary usage is. If learners are expected to know the 4,000 most frequent lemmas, for example, the teacher can select this threshold in the "Band" dropdown (Figure 7) and examine how many words in the text belong to frequency bands beyond this threshold. This information is also available in the statistics file, which can be downloaded in several different file formats and enables users to sort every word in the text into the selected frequency bands.

Discussion. This type of analysis should not be interpreted as an error analysis. While it is true that non-existing word forms will be classified as off-list, some learner errors involve incorrect combinations or uses of otherwise valid words and will therefore not be highlighted. This distinction is particularly important when lexical profiling is used to support assessment. A case in point is the phrase "met de lokale bevolking rijst" in Figure 7: *rijst* is an existing Dutch word meaning 'rice', but in this context the learner intended *reist* ('travels'). The lexical profiler, however, treats *rijst* as a valid lexical item even though its use in the sentence is erroneous.

Additionally, since NLLP-5000 does not distinguish word senses, the profiling results should not be interpreted as conclusive evidence of the learner's level of lexical proficiency. On the one hand, a learner showing productive knowledge of a polysemous lemma does not imply that this learner has acquired knowledge of all senses of this lemma. Especially for lemmas whose senses are substantially different in terms of frequency

and difficulty (e.g., *slot* can mean both 'lock' and 'castle'), this is an important observation to bear in mind. On the other hand, the use of relatively infrequent items alone does not necessarily mean that the learner has advanced lexical proficiency.

6 Discussion and Future Work

By introducing `NLLP-5000`, the present study advances the state of the art for lexical profiling in (L2) Dutch. Compared to the few existing Dutch word frequency lists designed for lexical profiling (e.g., Van Parys et al., 2025), our resource is distinguished by the inclusion of inflected forms (obtained through a meticulous manual verification process) and the linguistic annotation of these forms in Universal Dependencies (UD) format.

The implementation of `NLLP-5000` in an existing lexical profiling tool was demonstrated through an end-to-end workflow: from the conversion of UD-based labels into tool-specific tags (Section 4) to the illustration of concrete applications in ICALL contexts (Section 5). In both ICALL applications demonstrated, the focus was on L2 teachers as the primary stakeholders. This focus was deliberate: because recent research has shown that teachers are underrepresented in educational technology (EdTech) research (Gaudeau et al., 2026), we sought to contribute to a better representation of teachers in this field.

Future work should first of all address the limitations discussed in Section 8, such as the lack of word sense disambiguation and the limited size of the resource. In addition to this, extending the resource with the dimension of word families (Bauer and Nation, 1993) also constitutes an avenue worth exploring. Though not entirely undisputed as a measure for determining vocabulary size (Laufer, 2021), word family forms (e.g., *to use–usable*) have been argued to be the most transparent and efficient lexical unit from a pedagogical point of view, at least for more advanced learners (Webb, 2021). Conducting research on a word family resource for Dutch would constitute a valuable contribution to a field that is, apart from a few exceptions such as Volodina et al. (2024b) for Swedish, dominated by English both in terms of research and available resources.

7 Conclusion

This study built on previous knowledge and existing resources to develop a novel resource for lexical profiling in Dutch. Drawing on an extensive body of research emphasising the vital role of frequency in L2 vocabulary learning, we used *A Frequency Dictionary of Dutch* (Tiberius and Schoonheim, 2014) as input data. Several steps were undertaken to transform these input data into the novel `NLLP-5000` resource (**NL** **L**exical **P**rofiling, comprising 5,000 headwords), which is made publicly available through a GitHub repository. Duplicate entries were removed, all headwords were lemmatised, inflected forms were added and labelled according to the Universal Dependencies (UD) format, and the original dispersion values were converted into frequency ranks and bands.

The UD-based linguistic annotations in `NLLP-5000` make the resource compatible with lexical profiling tools built on widely used NLP toolkits such as spaCy and Stanza. The list is versatile in that it can be easily incorporated into tools that require a different format. We demonstrated this conversion process in Section 4 for MultilingProfiler, alongside two applications of the resource in ICALL contexts: class preparation and learner text analysis/assessment (Section 5).

Finally, although the present paper has mainly focused on the task of lexical profiling, the `NLLP-5000` resource can be integrated more broadly into ICALL environments. Whether in ICALL tools specifically dedicated to vocabulary learning (e.g., Goethals, 2018; for Spanish) or in full-fledged intelligent tutoring systems (e.g., Amaral and Meurers, 2011; for Portuguese), vocabulary lists constitute a valuable source of input data for generating well-informed vocabulary learning exercises tailored to the specific needs of both teachers and (groups of) learners.

8 Limitations

Word senses. In the current version of the resource, words with multiple senses appear under the same headword (e.g., *koper* as 'buyer' and as 'copper'). Extended work could introduce "word sense awareness" into the resource by using word sense disambiguation (WSD) methods for L2 learning purposes (Degraeuwe and Goethals, 2022).

Receptive versus productive knowledge. The class preparation application (Section 5.1) and the analysis of learner text (Section 5.2) address different aspects of vocabulary knowledge. Whereas

the former focuses on the vocabulary learners should recognise in educational materials (i.e. *receptive* knowledge), the latter aims to provide insights into their *productive* knowledge. These two aspects of vocabulary knowledge do not necessarily follow the same frequency thresholds. Following the work of Volodina et al. (2024a) on the Swedish Second Language Profile, future work should explore creating a suite of resources that clearly distinguish receptive from productive knowledge (e.g., with separate frequency indices).

Resource size. `NLLP-5000` is limited to the top 5,000 most frequent lemmas in Dutch. For beginner and intermediate learners, this range is large enough for the lexical profiler to return relevant output. For teachers working with advanced learners, however, the resource will be less useful. As advanced learners have likely mastered most or even all of the 5,000 lemmas included in the list, an additional set of frequency bands is required to profile texts in a meaningful way (e.g., up to the top 15,000 most frequent lemmas).

Validation. The applications presented in Section 5 constitute an *illustration* of how `NLLP-5000` can be used in an ICALL context; they should not be interpreted as *validated* pedagogical practices. To validate the resource, future work should study if the frequency ranks and bands correspond to learners' vocabulary knowledge and/or instructional needs.

9 Ethical Considerations

Role of paper author as annotator. The first author of the paper was responsible for (1) lemmatising the original Routledge input data, (2) verifying the automatically populated inflected forms, and (3) labelling the inflected forms according to the UD format. All decisions taken during these tasks (e.g., assigning base forms to POS categories) were taken based on the official prescriptive resource for Dutch, Woordenlijst.org (described in detail in Section 3.2.2). All other aspects of the tasks (e.g., checking if automatically populated forms were correct) did not entail any form of subjective judgement. We therefore conclude that the author's involvement in the methodology does not compromise the validity of the newly created resource.

Learner data. The writing assignment submission used to demonstrate the second ICALL application (Section 5.2) was fully anonymous. No personal data of the student were processed or stored in any way. The assignment itself also does not include any data that could potentially identify the student.

Acknowledgments

We wish to express our sincere gratitude to the CVO Groeipunt student who was willing to share the writing assignment submission presented in Appendix F and to the anonymous reviewers for their constructive and very valuable feedback.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: An AI tool was used for flagging language errors and spotting inconsistencies (e.g., using the same terms for the same concepts throughout the paper) as a final check prior to submitting the camera-ready version of the paper.

References

David Alfter. 2024. Out-of-the-Box Graded Vocabulary Lists with Generative Language Models: Fact or Fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

Luiz A. Amaral and Detmar Meurers. 2011. On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(1):4–24.

Laurie Bauer and I.S.P. Nation. 1993. Word Families. *International Journal of Lexicography*, 6(4):253–279.

Stephen Bodnar. 2025. A prototype authoring tool for editing authentic texts using LLMs to increase support for contextualised L2 grammar practice. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 63–71, Tallinn, Estonia. University of Tartu Library.

Vaclav Brezina and Dana Gablasova. 2015. Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1):1–22.

Thi Ngoc Yen Dang. 2022. Using VocabProfilers to select texts for extensive reading activities. In *Teaching English with Corpora*, pages 69–73. Routledge, London.

Mark Davies. 2025. CORPORA AND AI / LLMs: Overview. Online report, English-Corpora.org.

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jasper Degraeuwe and Patrick Goethals. 2022. Interactive Word Sense Disambiguation in Foreign Language Learning. In *Proceedings of the 11th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL 2022)*, pages 46–54.

Natalie Finlayson, Emma Marsden, and Laurence Anthony. 2023. Introducing MultilingProfiler: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts. *System*, 118:103122.

Ana Frankenberg-Garcia, Geraint Paul Rees, and Robert Lew. 2021. Slipping Through the Cracks in e-Lexicography. *International Journal of Lexicography*, 34(2):206–234.

Gabrielle Gaudeau, Aoife O'Driscoll, Jasper Degraeuwe, Andrew Caines, Donya Rooein, and Zeerak Talat. 2026. Incentives Of EdTech: A Systematic Review Of EduNLP Research. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 715–750, San Diego, California, USA. Association for Computational Linguistics.

Thomas Gerald, Sahar Ghannay, Julie Lascar, Paul Lerner, and Anne Vilnat. 2026. Can Multimodal LLMs Generate Pedagogical Questions? pages 5506–5515, Palma, Mallorca, Spain.

Patrick Goethals. 2018. Customizing vocabulary learning for advanced learners of Spanish. In *Technological innovation for specialized linguistic domains: languages for digital lives and cultures, proceedings of TISLID'18*, pages 229–240, Gent, Belgium. Éditions Universitaires Européennes.

Clarence Green. 2020. Extensive reading and viewing as input for academic vocabulary: A large-scale vocabulary profile coverage study of students' reading and writing across multiple secondary school subjects. *Lingua*, 239:102838.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Batia Laufer. 2021. Lemmas, flemmas, word families, and common sense. *Studies in Second Language Acquisition*, 43(5):965–968.

Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.

I.S.P. Nation. 2022. *Learning Vocabulary in Another Language*, 3 edition. Cambridge University Press.

Ulugbek Nurmukhamedov and Stuart Webb. 2019. Lexical coverage and profiling. *Language Teaching*, 52(02):188–200.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Roman Chernysh, Gabriel Mora-Rodríguez, and Lev Berezhnoy. 2024. A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions. pages 187–198.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

John McHardy Sinclair and Antoinette Renouf. 1988. A lexical syllabus for language learning. In Ronald Carter and Michael Mccarthy, editors, *Vocabulary and Language Teaching*. Routledge.

Mariia Soliar, Leona Colling, Stephen Bodnar, and Detmar Meurers. 2026. Using Interaction Log Data to Evaluate and Improve Feedback Accuracy in an Intelligent Language Tutoring System. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 190–207, San Diego, California, USA. Association for Computational Linguistics.

John Stamper, Ruiwei Xiao, and Xinying Hou. 2024. Enhancing LLM-Based Feedback: Insights from Intelligent Tutoring Systems and the Learning Sciences. Version Number: 2.

Carole Tiberius and Tanneke Schoonheim. 2014. *A frequency dictionary of Dutch: Core vocabulary for learners*. Routledge frequency dictionaries. Taylor & Francis Routledge, London.

Amaury Van Parys, Vanessa De Wilde, Lieve Macken, and Maribel Montero Perez. 2025. LexPro: A plurilingual lexical profiling tool to assist teachers and researchers in analysing vocabulary of L2 input. *Language Teaching Research*, pages 1–25.

Laura Vilkaitė-Lozdienė and Norbert Schmitt. 2019. Frequency as a Guide for Vocabulary Usefulness. In Stuart Webb, editor, *The Routledge Handbook of Vocabulary Studies*, 1 edition, pages 81–96. Routledge.

Elena Volodina, David Alfter, and Therese Lindström Tiedemann. 2024a. Profiles for Swedish as a Second Language: Lexis, Grammar, Morphology. pages 10–19.

Elena Volodina, Yousuf Ali Mohammed, and Therese Lindström Tiedemann. 2024b. Swedish word family resource: Construction, applicability, strengths and first experiments. *ITL - International Journal of Applied Linguistics*, 175(1):127–161.

Stuart Webb. 2021. Word families and lemmas, not a real dilemma: Investigating lexical units. *Studies in Second Language Acquisition*, 43(5):973–984.

Michael West. 1953. *A General Service List of English Words*. Longmans, Green & Co., London.

Jiahui Wu, Jianwei Li, Zigang Ge, Mingrui Xu, Li Lin, and Ru Zhang. 2025. Effectiveness of Generative AI in Automated Written Corrective Feedback With Prompting. *Journal of Educational Computing Research*, 63(6):1493–1527.

A Pre-processing of Original Routledge File

As explained in Sections 3.1 and 3.2, *A Frequency Dictionary of Dutch* (Tiberius and Schoonheim, 2014) was used as input data to create the `NLLP-5000` resource. The dictionary was accessed through the "Frequency-index.txt" file, provided by Routledge as an official supplement to the dictionary. However, upon processing the file, a series of small (formatting) issues were found. These issues, listed below, were manually fixed by the authors of this paper.

- **Line 7**: formatting error, no tab before second example sentence (missing tab was added and one of the trailing tabs was removed)
- **Line 1013**: typo in frequency of *kleden* (". 3.80" instead of "3.80")
- **Line 1344**: formatting error, only 10 elements in line instead of 11 (missing tab was added before "2.47")
- **Line 2471**: typo in frequency of *set* (". 3.98" instead of "3.98")
- **Line 3658**: typo in frequency of *ontwerp* (". 3.49" instead of "3.49")
- **Line 4157**: typo in frequency of *zichtbaar* (". 5.74" instead of "5.74")
- **Line 4865**: incomplete POS ("2) het" instead of "2) noun, het")
- **Lines 5202-5203**: formatting error (both lines belong to *vet*, so they were merged)
- **Lines 5405-5406**: formatting error (both lines belong to *rot*, so they were merged)

B Description of *buigings*-forms

- *Buigings-e* (adjectives): In some cases, an *e* is added to adjectives when they are used attributively (e.g., with plural nouns). For more detailed information, please refer to the official Taaladvies page.
- *Buigings-s* (adjectives): Inflected form needed when the adjective is combined with *iets* ('something').
- *Buigings-e* (articles): The article *een* ('a, an') has the inflected form *ene* in nominative, accusative, and dative.
- *Buigings-e* (numbers): In some cases, an *e* is added to numerals (e.g., *meest–meeste* ['most']), for example when the numeral is used attributively.
- *Buigings-n* (numbers): In some cases, *en* is added to numerals (e.g., *tweeën*), for example in the phrase *met z'n* ... (please refer to the Taaladvies page for more details).
- *Buigings-e* (pronouns): In some cases, an *e* is added to pronouns (e.g., *ieder–iedere* ['each']), for example when the pronoun is used attributively.
- *Buigings-n* (pronouns): In some cases, *(e)n* is added to pronouns (e.g., *al–allen* ['all']), for example when the pronoun is used to refer to a group of people.

C Definition of Morphological Features

The complete overview of the UD morphological features assigned to the inflected forms in `NLLP-5000` is presented in Table 6.

Original POS tag	UD feature	Applicable values for feature	Example(s)
adj (adjective)	Degree	Pos (positive), Cmp (comparative), Sup (superlative)	*kleiner* ('smaller'): Degree=Cmp
adv (adverb)	PartType (particle type)	Neg (negation particle), Vbp (verbal particle)	*niet* ('not'): PartType=Neg | Polarity=Neg *nooit* ('never'): PronType=Neg *weer* ('re-'): PartType=Vbp
	Polarity	Neg (negative)	
	PronType (pronominal type)	Neg	
art (article)	Case	Nom (nominative), Acc (accusative), Gen (genitive), Dat (dative)	*de* ('the'): Case=Nom | Definite=Def | PronType=Art *ener* ('of a, of an'): Case=Gen | Definite=Ind | PronType=Art
	Definite	Def (definite), Ind (indefinite)	
conj (conjunction)	N/A	N/A	N/A
interj (interjection)	N/A	N/A	N/A
noun (noun)	Degree	Dim (diminutive)	*aanbieding* ('offer'): Gender=Fem | Number=Sing *touwtjes* ('little ropes'): Degree=Dim | Gender=Neut | Number=Plur
	Gender	Fem (feminine), Masc (masculine), Neut (neuter)	
	Number	Sing (singular), Plur (plural)	
num (number)	Definite	Def (definite), Ind (indefinite)	*acht* ('eight'): Definite=Def | NumType=Card *ander* ('other'): Definite=Ind | NumType=Ord
	NumType (numeral type)	Card (cardinal), Ord (ordinal)	
prep (preposition)	AdpType (adposition type)	Circ (circumposition), Post (postposition), Prep (preposition)	*achter* ('behind'): AdpType=Prep
pron (pronoun)	Case	Nom, Acc, Gen, Dat	*aller* ('of all'): Case=Gen | Number=Sing | PronType=Ind *deze* ('this'): Case=Nom | Number=Sing | PronType=Dem
	Gender	Fem, Masc, Neut	
	Number	Sing, Plur	
	Person	1, 2, 3	
	Poss (possessive)	Yes	
	PronType	Dem (demonstrative), Ind (indefinite), Int (interrogative), Prs (personal), Rcp (reciprocal), Rel (relative)	
	Reflex	Yes	
verb (verb)	Mood	Imp (imperative), Ind (indicative), Sub (subjunctive)	*buigt* ('bends'): Mood=Ind | Number=Sing | Person=3 | Tense=Pres | VerbForm=Fin *geparkeerd* ('parked'): Tense=Past | VerbForm=Part
	Number	Sing, Plur	
	Person	1, 2, 3	
	Tense	Past, Pres (present)	
	VerbForm	Fin (finite), Inf (infinitive), Part (participle)	
	VerbType	Cop (copula), Mod (modal)	

Table 6: UD morphological features assigned to the inflected forms in NLLP-5000. "N/A" stands for not applicable.

D Script for Converting Verb Entries from UD to FreeLing Format

The script used to automatically convert verb entries in `NLLP-5000` from UD format to FreeLing format is included as supplementary data in the GitHub repository: https://github.com/JasperD-UGent/NLLP-5000/tree/main/supplementaryMaterials/NLP4CALL2026.

E Translation Glossaries

Translation glossary for Table 1

- *beetje* = 'bit'
- *beet* = 'bite'
- *inkomsten* = 'income, revenue'
- *inkomst* = 'entry'
- *jezelf* = 'yourself'
- *jijzelf* = 'yourself'
- *kleintje* = 'small/little one; small thing'
- *kleine* = 'little one'
- *kleren* = 'clothes'
- *kleed* = 'carpet; robe'
- *meerdere* = 'multiple'
- *onszelf* = 'ourselves'
- *wijzelf* = 'ourselves'
- *praatje* = 'chat'
- *praat* = 'talk'
- *sommig* = 'some'
- *sprookje* = 'fairy tale'
- *sproke* = 'tale'
- *werkzaamheden* = 'work, operations'
- *werkzaamheid* = 'activity'
- *het* = 'the; it'
- *ander* = 'other'
- *dergelijk* = 'similar'
- *anno* = 'in the year'
- *alstublieft/alsjeblieft/a.u.b.* = 'please'
- *cc* = 'carbon copy; cubic centimetre'
- *o.a.* = 'among other things, among others'

Translation glossary for Table 2

- *laat* = 'late'
- *later* = 'later'
- *laatst* = 'latest'
- *ver* = 'far'
- *verder* = 'further'
- *goed* = 'good'
- *best* = 'best'
- *een* = 'a, an'
- *ene* = 'a, an'
- *goed* = 'good, property'
- *goederen* = 'goods'
- *kost* = 'cost; living'
- *kosten* = 'cost(s), expense(s)'
- *medium* = 'medium'
- *media* = 'media'
- *ik* = 'I'
- *mij/me* = 'me'
- *hij* = 'he'
- *hem/'m* = 'him'
- *zij/ze* = 'she; they'
- *hen* = 'them'
- *jij/je* = 'you'
- *jou* = 'you'
- *ikzelf* = 'I myself'
- *mijzelf/mezelf* = 'myself'

F Input Text Learner Text Analysis and Assessment Demo

- Source: anonymous writing assignment submission from an L2 Dutch Level 4 course organised at CVO Groeipunt (Flemish adult education centre)
- Date: 28 February 2026

Voor mijn volgende reis naar Sri Lanka ben ik goed voorbereid omdat ik op een forum een artikel over de verschillende vervoermiddelen gelezen heb. Ik heb gelezen dat je met de bus, de trein, per tuktuk of taxi of met een privéchauffeur kunt reizen. Op het forum hebben toeristen verschillende manieren beschreven, met welke zij in Sri Lanka gereisd zijn. Er is een groot verschil tussen de vervoermiddelen, sommige zijn goedkoop maar misschien niet zo comfortabel en andere zijn een beetje duur maar ook heel flexibel. Blijkbaar is het het tofste om met de trein te reizen omdat het een avontuur is, je snel door het land komt en, het belangrijkste, dat je van een heel mooi uitzicht kunt genieten. Op de trein kan je tussen drie klassen kiezen maar de tweede is blijkbaar de beste manier om te reizen omdat je een stoel krijgt maar ook samen met de lokale bevolking rijst. Met de bus kan je ook reizen en tussen twee bussen, de regeringsbussen (SLTB) of de rode private bussen, kiezen. Je kunt met allebei bussen rijden, maar misschien moet je op de private bussen extra voor je bagage betalen en het kan ook echt bezig zijn. Alternatief kan je met een tuktuk of taxi rijden maar er is slim voor de reis met de rijder een prijs af te spreken. Wanneer je een Sri Lankaans rijbewijs hebt of ter plaatse wil maken, kan je ook zelf een tuktuk rijden. Vergelijkbaar, en ook flexibel, kan je ook een privéchauffeur voor de dag huren maar dat is het "duurste" vervoermiddel voor 30 euro per dag. Maar het is ook de flexibelste manier omdat je altijd kunt stoppen.

Ik vind de verschillende mogelijkheden heel interessant maar wanneer ik naar Sri Lanka zou gaan, zou ik met de trein of met een tuktuk rijden. Naar de bus wordt mij altijd misselijk en ik geloof dat de rijders heel snel, en misschien niet zo veilig, rijden om meer mensen, en meer geld, mee te nemen. Wanneer ik met het tuktuk rijd, heb ik ook verse lucht en misschien kan ik het plaatse een rijbewijs maken om zelf te rijden, dat vind ik een leuk idee! Om snel naar andere steden te gaan, zou ik met de trein rijden. Nu ben ik bereid om naar Sri Lanka te gaan!

G Input Text Class Preparation Demo

- Source: VRT NWS (official Belgian news provider)
- Date: 17 September 2026
- URL: https://www.vrt.be/vrtnws/nl/2026/09/17/nieuwe-kattensoort-uantwerpen/

Voor het eerst in 100 jaar nieuwe kattensoort ontdekt: zeldzaam gevlekt dier is kleiner dan een huiskat en werd ontdekt door biologen UAntwerpen

Biologen van de UAntwerpen hebben een nieuwe kattensoort ontdekt. Het gaat om een nieuwe soort tijgerkat, die in de bergbossen van Bolivia leeft. Het is meer dan 100 jaar geleden dat er een nieuwe kattensoort is ontdekt.

Maak kennis met de Leopardus tilcayo, het nieuwste familielid van de tijgerkatten. Het gevlekte katje is iets kleiner dan een huiskat en leeft in de bergbossen van de Andes in Bolivia. De soort blijkt zich 1,4 miljoen jaar geleden te hebben geïsoleerd van de andere tijgerkatten.

Dat blijkt uit genetisch onderzoek van biologen van de UAntwerpen, samen met een team van internationale wetenschappers. Ze ontdekten dat de tilcayo een volwaardige nieuwe kattensoort is.

Verborgen tijgerkat

De vorige keer dat een nieuwe kattensoort werd ontdekt, was in 1923, toen de Uruguayaanse graskat (Leopardus fasciatus) werd beschreven.

"Dat de tilcayo zo lang onopgemerkt is gebleven als nieuwe soort, heeft veel te maken met haar uiterlijk", zegt evolutiebioloog Jonas Lescroart (UAntwerpen), die het onderzoek mee heeft geleid.

"Tijgerkatten zijn kleine, gevlekte katachtigen uit Centraal- en Zuid-Amerika, die op basis van fysieke kenmerken amper onderling te onderscheiden zijn."

"Ook het zeldzame voorkomen en het schuwe karakter van het dier spelen mee in de late ontdekking. De tilcayo leeft in de Boliviaanse Yungas, een grillig, tropisch voorgebergte aan de oostelijke kant van de Andes."

"Dit ecosysteem staat bekend om een zeer grote soortenrijkdom, met fauna en flora die nergens anders voorkomen, maar het gebied is ook moeilijk toegankelijk voor onderzoekers."

Vreemde huiskat

Om te weten hoe de onderzoekers dan toch de nieuwe kattensoort op het spoor zijn gekomen, moeten we 10 jaar terug in de tijd.

"In 2016 werd een gevonden kitten binnengebracht in een Boliviaans lokaal opvangcentrum voor wilde dieren, omdat het toch wel erg vreemd gedrag vertoonde voor een huiskat en er heel anders uitzag", vertelt Lescroart.

Het dier werd toen geïdentificeerd als een noordelijke tijgerkat (Leopardus tigrinus) en kreeg spontaan de koosnaam Tigrino. Het katje trok al snel de aandacht van de Boliviaanse onderzoekers die betrokken waren bij de tijgerkatstudie.

Na een routinebloedafname door de dierenarts werd Tigrino's DNA geanalyseerd. Door die data te vergelijken met stalen van andere kattensoorten, konden de onderzoekers het bewijs leveren dat Tigrino tot een nieuwe kattensoort behoort.

Levend exemplaar

Volgens de onderzoekers is het ook zeer opmerkelijk dat een levend dier is gebruikt om een nieuwe diersoort te beschrijven. Normaal gebeuren de metingen, DNA-analyse en beschrijvingen op een dood dier. Nu gebeurde dat op een levend, verdoofd dier.

"Tigrino zal in het opvangcentrum in Bolivia verblijven tot zijn natuurlijke dood", zegt Lescroart. "Daarna zal het kadaver worden opgenomen in de collectie van het Nationaal Natuurhistorisch Museum van Bolivia."

"De ontdekking van de tilcayo maakt hopelijk meer middelen vrij voor het behoud van populaties tijgerkatten, die zich vaak in een kwetsbare situatie bevinden. Tegelijk vestigen we de aandacht op de ongelooflijke natuurwaarde die Bolivia rijk is."

Uiteindelijk leverden de onderzoekers overtuigend bewijs voor 5 verschillende soorten: de noordelijke en zuidelijke tijgerkat, die al langer als aparte soorten beschouwd werden, de oostelijke tijgerkat en de neveltijgerkat, die recent werden opgewaardeerd van ondersoort tot volwaardige soort, en de tilcayo, die het team vernoemde naar het lokale woord voor tijgerkat.

Making Feedback Visible: Character-Level Error Localization in Handwritten CALL Input

Said Yasin **Torsten Zesch**
CATALPA
Center of Advanced Technology for Assisted Learning and Predictive Analytics
FernUniversität in Hagen, Germany

Abstract

Feedback in CALL should be as precisely localized as possible. For example, instead of merely stating that "there is a spelling error in your statement", it is more helpful to indicate the exact location of the error in the text. In fact, an even higher degree of precision would be to identify the specific characters responsible for the error. This localization of erroneous characters is challenging for handwriting input, a modality still widely used in primary and secondary education especially in language learning contexts. Our demo presents an interactive end-to-end prototype that captures the handwriting input, transcribes it, generates feedback, and visually grounds it in the handwriting on character-level granularity.

1 Introduction

Feedback is a central component of computer-assisted language learning (CALL). In particular, it is most useful when learners can relate it directly to their own production (Ferris and Roberts, 2001). For written language learning, this means that feedback should not only state that an error exists, but also indicate precisely where it occurs. Precise localization reduces the effort required to interpret feedback, supports self-correction, and makes automated feedback more transparent to learners and teachers.

While localized feedback is comparatively straightforward for digitized text, it remains challenging for handwritten input. This is an important limitation, because handwriting continues to play a significant role in primary and secondary education, including both first-language (L1) and second-language (L2) learning contexts. If CALL systems only support keyboard input, they may fail to reflect authentic classroom practices and may exclude younger learners whose digital writing skills are still developing.

For written corrective feedback, Ellis (2009) distinguishes (i) metalinguistic feedback (such as "there is a spelling error in your sentence") from (ii) direct feedback (providing the correct form), and (iii) indirect feedback (only visualizing that an error occurred, but not providing the correct form). In Figure 1, we provide examples for each type. Metalinguistic feedback is not grounded in the output, so it basically works the same for digitized input or handwriting. To ground direct and indirect feedback in the handwriting it becomes at least necessary to localize word boundaries, so that feedback can be precisely placed. The most challenging to implement is indirect feedback at character-level granularity as this requires locating the exact position of characters within a handwritten word. However, it is also the potentially most rewarding as indirect feedback is often argued to promote more active processing and self-correction than direct feedback (Lalande, 1982), although empirical findings are mixed (Chandler, 2003; Bitchener and Knoch, 2010; Van Beuningen et al., 2012) depending for example on the learner's proficiency level (Kang and Han, 2015).

Grounding localized feedback in handwriting requires solving several connected problems. First, the system must capture the handwritten production. Second, it must transcribe the handwriting into machine-readable text. Third, it must generate pedagogically meaningful feedback on the transcribed content. Finally, and crucially, the feedback must be mapped back onto the original handwriting. This final step is non-trivial: even if the transcription is correct, the system must know which region of the handwritten text corresponds to the word and character that triggered the feedback.

In this demo paper, we present an interactive

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 34–39, November 19–20, 2026, University of Tartu Library

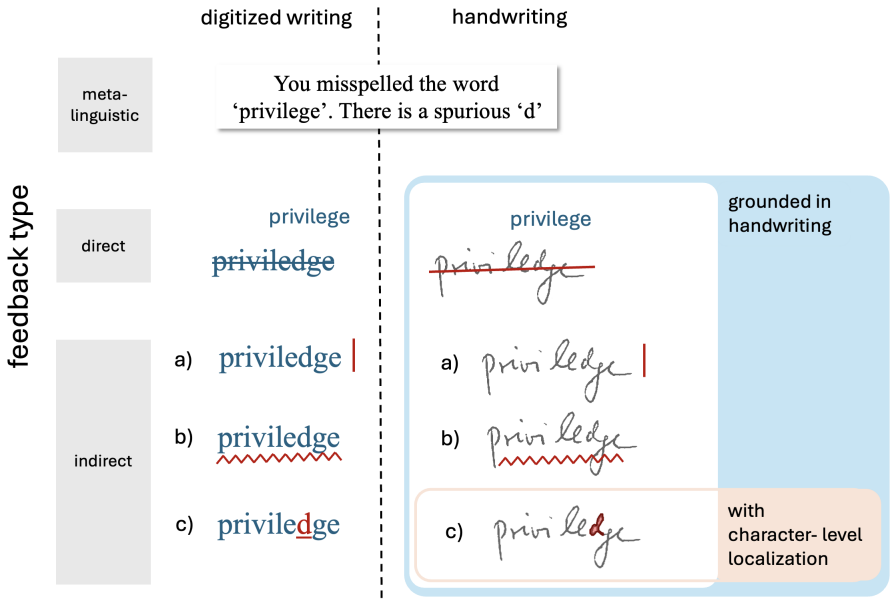


Figure 1: Examples of feedback types in digitized text and when grounded in handwriting.

end-to-end prototype for visually grounded feedback on handwritten learner input. The prototype is implemented as a mobile application for smartphones and tablets and runs fully offline on the device. Figure 2 provides a screenshot of the demo where corrective spelling feedback (as one example from a wider range of possible applications including morphology, syntax, lexical choice, etc.) is grounded in the handwriting and localized on character level.[1]

By keeping the learner's original handwritten production visible, the system preserves the modality in which the task was completed. At the same time, it connects this production to computational processing steps that are typically available only for typed input.

With our demo system, we aim to stimulate discussion about the role of handwriting in CALL, the design of localized feedback for young learners, and the technical requirements for grounding language feedback in multi-modal learner data. The groundink mobile application and its source code are available at https://github.com/said702/groundink.

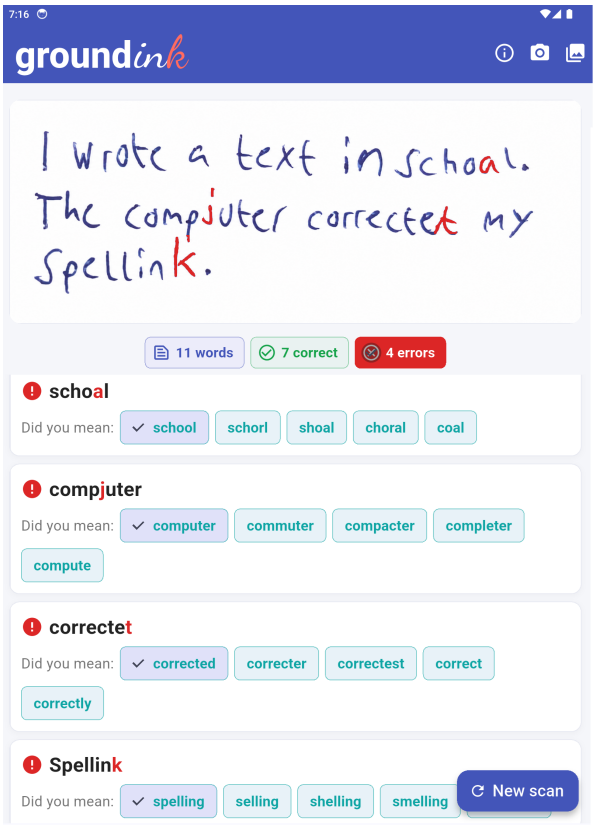


Figure 2: Character-level spelling feedback as implemented in the demo.

2 System Architecture

The prototype is implemented as an on-device mobile application. Its processing pipeline consists of

[1] We call our system *groundink*, a portmanteau of 'grounding' and 'ink'. At the same time, a misspelling of 'grounding' in need of localized feedback.

four main stages: word detection, character recognition and localization, feedback generation, and spatial grounding of the feedback in the original handwriting.

The application is implemented in Flutter, a cross-platform framework for mobile application development, using the Dart programming language. This allows us to use a single codebase for both Android and iOS.

Handwriting input is captured either through the device camera or from the local photo gallery using the Flutter image picker. All recognition and feedback components are executed locally on the device to support the privacy requirements of in-school applications. Therefore, the models are stored in the platform-independent Open Neural Network Exchange (ONNX) format and executed using ONNX Runtime, allowing the application to operate without an external server connection.

Word Detection. The first stage uses the nano-sized YOLO26n model[2] to detect handwritten words in the input image and predict a bounding box for each word. We fine-tune this model on the **IAM Handwriting Database** (Marti and Bunke, 2002), which contains 1,539 scanned handwritten English forms from 657 writers and 115,320 word instances. The predicted bounding boxes are used to extract the corresponding word regions from the original image. For display purposes, the detected words are ordered by line from top to bottom and from left to right within each line. Since each word is processed independently, the reading order does not affect the evaluation. While the model was trained on English data, it mainly learns to visually separate words without the textual content. This makes the resulting model largely language-independent. On an Android emulator (API 34), the word detection model took approximately one second to process a page containing around 200 words.

Character Recognition and Localization. Each word crop is processed by a YOLO26n-seg instance segmentation model for joint character recognition and localization. We fine-tune this model on **CharLoc** (Yasin and Zesch, 2026), which contains 37,010 handwritten English word images with more than 219,000 characters annotated with character labels, bounding boxes, and precise character contours. For each detected character, the model predicts its class and a contour representing its spatial extent. The predicted characters are ordered from left to right to reconstruct the recognized word, while the contours retain their locations in the handwritten image. On the Android emulator, the character model required approximately 175 ms per word.

Feedback Generation. The recognized transcription is passed to a feedback component, which generates feedback on the learner's production and identifies the corresponding target in the recognized text. The feedback target is represented by the character position or span to which the feedback refers.

Feedback Grounding. Once a character or character span has been identified as the target of the feedback, it must be mapped to its location in the handwritten image. This cannot be reliably inferred from its position in the transcription, since handwritten characters vary in width, spacing, and slant and may touch or overlap.

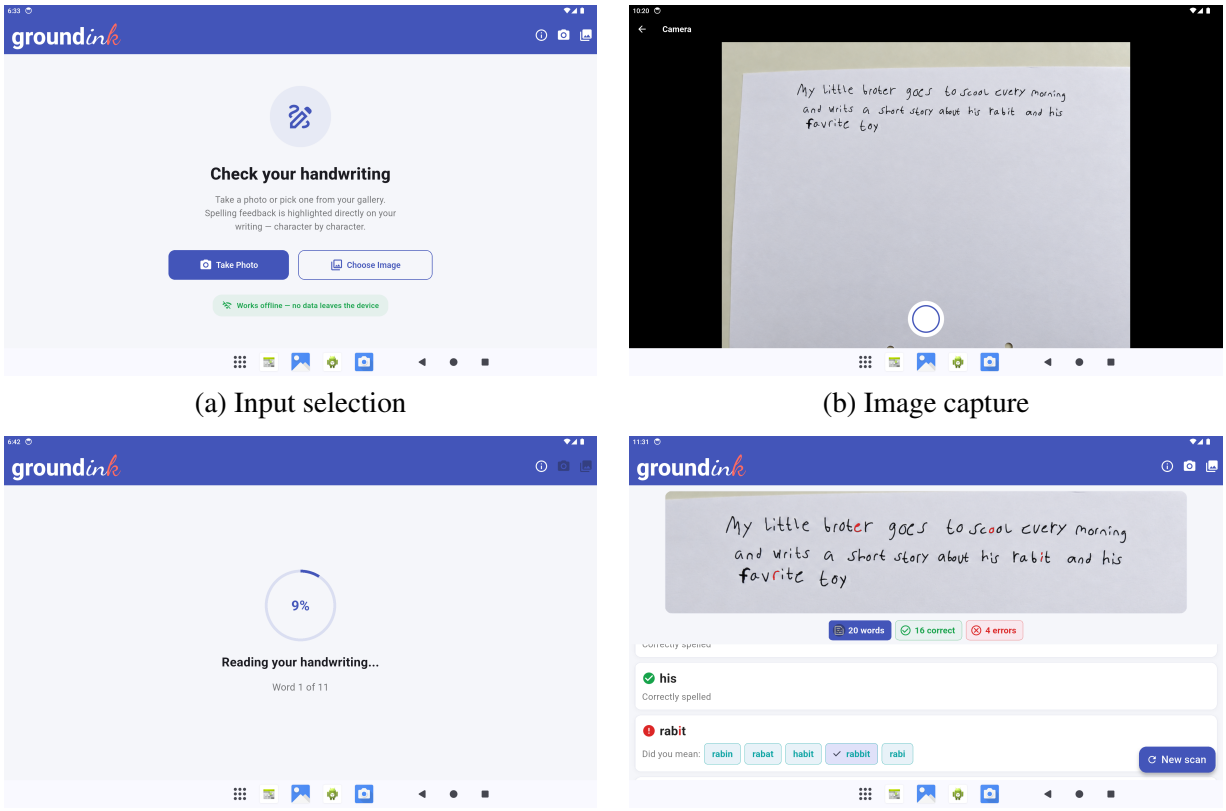
The identified character position is mapped to the corresponding predicted character contour. This contour is used as a mask to highlight the relevant character directly in the original handwritten image. Thus, the feedback is spatially grounded at character level rather than only at the level of the complete word.

3 Demo

To demonstrate character-level grounded feedback in a concrete application scenario, we use spelling feedback as an example. The feedback generation component computes the Levenshtein distance between each recognized word and the entries in an English word list bundled with the application and identifies several of the closest matches as correction candidates.[3] Among these candidates, the system automatically selects the closest match. If this does not correspond to the word intended by the student, they can choose from the remaining plausible alternatives. The selected word is then compared with the recognized word to identify and highlight the affected character in the orig-

[2] https://docs.ultralytics.com/models/yolo26

[3] Note that we are using English here for demonstration purposes. The application itself is not limited to English and has been successfully tested with various other languages that use the Latin alphabet. Highlighting is currently limited to the character set in CharLoc, i.e. the standard English alphabet without for example German umlauts (ä, ö, ü).



(a) Input selection (b) Image capture

(c) Local processing (d) Grounded feedback

Figure 3: Workflow of the mobile handwriting feedback system. The learner captures or selects an image of handwritten text, which is processed locally on the device before character-level grounded feedback and correction suggestions are presented.

inal handwriting. When a different candidate is selected, the highlighting is dynamically updated, allowing different error locations to be visualized in cases of ambiguous spelling errors.

Figure 3 shows how this character-level grounding is realized in the mobile demo. The learner first selects an image source, either by capturing a new image or choosing one from the gallery (a). The handwritten text is then captured through the camera interface (b), after which the application processes the image locally and displays the recognition progress (c). Finally, the relevant handwriting region is cropped and enlarged, and the interface presents it together with the recognized words, correction suggestions, and character-level feedback highlights (d). In this example, selecting "rabbit" as the correction for the recognized word "rabit" highlights the corresponding character in the original handwriting.

Although the current demo uses spelling errors as a concrete example, the grounding approach can also be widened to support other linguistic error types whenever the relevant feedback target

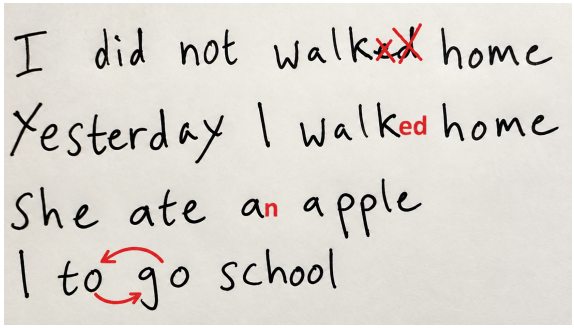


Figure 4: Illustrative visualizations of character- and word-level feedback beyond spelling.

can be mapped to the handwriting. Figure 4 illustrates such extensions: morphological or grammatical errors could be visualized by highlighting or crossing out superfluous characters, or by inserting missing characters or inflectional endings in a visually distinct typeset font. Word-order errors could be represented using arrows between word-level bounding boxes. Thus, the proposed approach can support a broader range of corrective feedback beyond spelling.

In future work, we plan to improve the robustness of the system and evaluate it in realistic classroom settings. Another promising direction is the integration of augmented reality.

4 Related Work

Previous systems for feedback on handwritten input either present the feedback separately from the handwriting or spatially ground it at the word level. We are not aware of any other research prototype that grounds feedback at character-level granularity in the original handwriting. Karlsson and Åkerman (2022) present Orthographer, which processes camera-captured handwriting and visually marks affected words together with correction suggestions. Bailey et al. (2020) developed a HoloLens-based application that analyzes handwritten words and provides immediate visual feedback in the user's field of view. Yasin and Zesch (2025) use a document camera to capture handwritten documents and spatially ground generated feedback using word-level bounding boxes.

Handwriting-based feedback has also been explored in language-learning applications. Pocket English Master (Imasha et al., 2023) combines real-time handwritten-text scanning with grammar and spelling correction for English learners. More recently, Paris et al. (2026) introduced FRANKI, a shared classroom device that scans handwritten student texts and generates AI-based feedback, which is presented either through a projection or as a printed response. Across these systems, feedback is either presented separately from the handwritten input or spatially associated with complete words. One factor contributing to this gap has been the limited availability of character-level annotations that support the localization of both block and cursive handwriting. Yasin and Zesch (2026) address this limitation with CharLoc, enabling the joint recognition and localization of individual handwritten characters.

Limitations

The demo is a proof of concept and has not yet been evaluated in user studies.

The feedback component has not been independently evaluated, as we focus in the demo on the localized grounding of feedback in handwriting. In other words, we consider for the purpose of this demo the generated feedback to be correct and focus on the visualization.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: We used OpenAI's GPT-5.6 Luna model through the ChatGPT app to improve grammar, sentence structure, clarity, and academic writing style (Editing), as well as to shorten selected author-written passages (Compression/expansion). All AI-assisted modifications were reviewed and verified by the authors.

References

Christine Bailey, Jacqueline Deprey, Abdulfatai Fakoya, John Nolan, Ivan Quiles-Rodriguez, Cameron Soderberg, Jeong-Yoon Wu, and Richard Yu. 2020. Exploration of augmented reality as an assistive device for students with dyslexia. Gemstone Team ART.

John Bitchener and Ute Knoch. 2010. Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, 19(4):207–217.

Jean Chandler. 2003. The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12(3):267–296.

Rod Ellis. 2009. A typology of written corrective feedback types. *ELT Journal*, 63(2):97–107.

Dana Ferris and Barrie Roberts. 2001. Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing*, 10(3):161–184.

Ashini Imasha, Kavindu Wimalaweera, Manohari Maddumage, Dilain Gunasekara, Kalpani Manathunga, and Devanshi Ganegoda. 2023. *Pocket English Master – Language Learning with Reinforcement Learning, Augmented Reality and Artificial Intelligence*, pages 74–85.

EunYoung Kang and Zhaohong Han. 2015. The Efficacy of Written Corrective Feedback in Improving L2 Written Accuracy: A Meta-Analysis. *The Modern Language Journal*, 99(1):1–18.

Viktor Karlsson and Aston Åkerman. 2022. Real-time detection of spelling mistakes in handwritten notes. Master's Thesis.

John F. Lalande. 1982. Reducing composition errors: An experiment. *The Modern Language Journal*, 66(2):140–149.

Urs-Viktor Marti and H. Bunke. 2002. The iam-database: An english sentence database for offline handwriting recognition. In *International Journal on Document Analysis and Recognition*, volume 5, pages 39–46.

Carla Paris, Valeria Herskovic, and Daniel Villaseñor. 2026. FRANKI: AI-Powered Feedback for Handwritten Assignments. In *Proceedings of the 25th Annual ACM Interaction Design and Children Conference*, IDC '26, page 1299–1302, New York, NY, USA. Association for Computing Machinery.

Catherine G. Van Beuningen, Nivja H. De Jong, and Folkert Kuiken. 2012. Evidence on the Effectiveness of Comprehensive Error Correction in Second Language Writing. *Language Learning*, 62(1):1–41.

Said Yasin and Torsten Zesch. 2025. Show, Don't Tell – Providing Visual Error Feedback for Handwritten Documents. In Zygmunt Vetulani, Patrick Paroubek, and Irakli Kardava, editors, *Human Language Technology as a Challenge for Computer Science and Linguistics – 2025*, pages 99–103. Adam Mickiewicz University Press, Poznań.

Said Yasin and Torsten Zesch. 2026. Accurate localization of characters in handwritten words. *International Journal on Document Analysis and Recognition (IJDAR)*.

Towards Semi-Automatic Benchmarking of Large Language Models for Self-Regulated Language Learning

Émile Alexandre
Université de Lorraine,
CNRS, LORIA,
F-54000 Nancy, France
emile.alexandre@loria.fr

Yannick Parmentier
Université de Lorraine,
CNRS, LORIA,
F-54000 Nancy, France
yannick.parmentier@loria.fr

Guillaume Nassau
Université de Lorraine,
CNRS, ATILF,
F-54000 Nancy, France
guillaume.nassau@univ-lorraine.fr

Abstract

Large Language Models (LLMs) have become increasingly popular personal assistants for language learning purposes. While exhibiting impressive capabilities on various Natural Language Processing tasks, including text summarisation, paraphrase generation, and question answering, their capacity to provide language learners with pedagogically adequate feedback is less clear. This is due, among other causes, to the lack of annotated data. In this paper, we describe a framework which aims at bridging this gap by (i) collecting manually annotated LLM-generated answers to learners' queries, and (ii) using this data to automatically characterise LLM skills along pedagogical dimensions with respect to a taxonomy of self-regulated LLM uses by language learners. This framework has been used to assess the ability of contemporary mainstream LLMs to generate pedagogically adequate answers to L2 French learners. Preliminary results on a small manually annotated dataset ($N = 49$) show that establishing a definitive assessment of these models remains elusive; rather, they reveal an absence of consensus among human experts when evaluating models. We release our code and dataset under open licences.

1 Introduction

Over the last few years, Large Language Models have gained remarkable popularity among citizens, especially students, due to their ever growing capability to generate fluent texts, together with their perceived ease of use and perceived usefulness (Shahzad et al., 2024). Education is no exception: LLM uses by language learners in particular increased, either in a self-regulated fashion (*e.g.* for conversational practice (Tyen et al., 2024)) or through dedicated learning applications (*e.g.* to support writing (Levine et al., 2025)). Even though LLM adoption by teachers is more limited (Galletti and Cesaroni, 2025; Sibug et al., 2026), LLM uses by teachers could likewise grow due to their ability to perform time consuming tasks requiring high human expertise, such as contextualising grammar practice (Bodnar, 2025), or building vocabulary exercises (Nikolova-Stoupak et al., 2024).

While potential benefits of LLMs – along with their corresponding limitations – were highlighted in the context of language learning (Vajjala et al., 2026), it remains unclear which models to use and what to *reliably* use them for. This is especially true in the context of Self-Regulated Language Learning (SRLL), where learners exhibit a high variability in their use cases of LLMs across different tasks. Learner interviews (Nassau and Molle, 2025) indeed revealed a wide range of practices adopted to enhance both language comprehension and production. These can roughly be split into three main categories: (i) using LLMs as conversational partners for *structured language practice* tailored to learners' level and interests, (ii) obtaining *meta-learning advice* on how to improve their learning strategies, and (iii) generating *personalised learning materials* (*e.g.* grammar exercises, vocabulary lists). These practices include roles traditionally associated with teachers and support learners' personal vision of a suitable learning process.

Unlike many Natural Language Processing tasks for which there exist benchmarks to evaluate LLMs' performances and compare models (see, *i.e.*, Natural Language Understanding (Liu et al., 2019)), the situation is less advanced for SRLL. Being able to assess LLMs' capabilities requires, among other things, representative datasets (Rudd et al., 2025) which may be challenging to collect. This work aims at bridging this gap by providing the community with a framework which can be used to collect representative data and use it to assess LLMs' capacity to deliver pedagogically adequate answers

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 40–57, November 19–20, 2026, University of Tartu Library

to highly diverse learner queries.

The paper is structured as follows. In Section 2, we present work related to the evaluation of LLMs in the context of language learning. In Section 3, we describe how we benchmarked LLM uses by L2 French learners. In particular, we describe the data that was collected and annotated, along with the language models of which the pedagogical feedback was evaluated, and the evaluation metrics we used to statistically assess the reliability of annotations. In Section 4, we present the implementation of our framework for semi-automatic benchmarking of LLM in self-regulated language learning. In Section 5, we comment on the results of our benchmarking study with L2 French learners. Finally, in Section 6, we conclude and present ideas for future work.

2 Related Work

There have been various attempts at measuring the effects of LLM-powered language tutors. These have led to the development of specific evaluation benchmarks based on curated taxonomies.

The closest approach to our work is that of Maurya et al. (2025), who proposed a comprehensive framework specifically tailored for mistake remediation in the context of mathematical problems. This framework builds on a taxonomy meant to unify the various dimensions used to assess the quality of pedagogical feedback in the field of education sciences. This taxonomy spans eight dimensions, ranging from diagnostic accuracy (mistake identification and location) to strategic feedback (revealing of the answer, providing guidance, actionability, coherence, tutor tone, and human-likeness). Maurya et al.'s framework has been implemented (Naeem et al., 2026) and used to build an evaluation benchmark of 1,596 responses from state-of-the-art LLM-based and human tutors. While sharing a common goal, namely assessing pedagogical ability of LLM-powered tutors, our work differs from that of Maurya et al. (2025) in the use case we consider (SRLL *vs.* math problems). In order to cover LLM uses for SRLL while keeping our taxonomy as simple as possible to facilitate data annotation, we restricted our work to a subset of the pedagogical dimensions used in their study (see § 3.3). Moreover, while Maurya et al. (2025) compiled a large benchmark of mistake remediation in mathematics by using existing datasets, we collected a new small dataset of openly available, real-world language-learner queries.

Another closely related work is the *L2-Bench* benchmark (Edgell et al., 2026), which has been designed to assess the broader picture of AI-mediated language tutoring. Unlike our approach, *L2-Bench* implements an evaluation taxonomy including interaction dynamics as well as planning, tutoring and evaluating competence. Another major difference with our approach lies in the fact that this benchmark builds on learning tasks designed by practitioners, while our interest lies in studying LLM capabilities in self-regulated contexts.

While these approaches provide comprehensive evaluation frameworks, scaling remains a challenging task, as highlighted by the findings of the *BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors* (Kochmar et al., 2025), which focused on selected dimensions (mistake identification, mistake location, guidance, and actionability). Indeed, the various approaches used – including sentence embeddings coupled with classification (Manlises et al., 2025) or fine-tuning of BERT-based architectures (Dadwal et al., 2025; Saha et al., 2025) – showed modest results while being subject to the bottleneck of computing resources, and subsequent costs.

In order to get around these limitations, recent evaluation approaches have been relying on LLM-as-a-judge techniques, which can reach high levels of agreement with human evaluators (Zheng et al., 2023). Without the need for compute-intensive fine-tuning, LLM-based evaluation can be significantly improved using In-Context Learning (ICL) (Min et al., 2022), few-shot learning (Brown et al., 2020), or other training techniques.

However, these approaches, often grounded in mathematical tutoring evaluation, rely on absolute scoring and classification – albeit with strong guidelines – which might be less suitable for more open-ended, subjective tasks, such as evaluating the pedagogical quality of AI tutors in a language learning context (Tack and Piech, 2022).

All in all, these observations highlight the inherent complexity and unresolved question of the right way to assess the pedagogical quality of LLM-based AI tutors, in particular in the context of L2 learning, and especially for subjective, nuanced tasks, such as determining fine-grained pedagogical dimensions. As we will see below in Section 3, our work is also subject to this complexity.

3 Evaluating LLM uses by L2 French learners

In this section, we describe an empirical study, which aimed at assessing LLMs' capabilities on queries from L2 French learners. This study is part of a wider project named *Littérat'IA*, of which the aim is to develop an ethical and methodological framework for LLM uses in language learning by examining current practices in university language centres, and evaluating the reliability of LLM-based tools on learning tasks. Ultimately, the project seeks to allow for informed, critical, and responsible use of LLM in language teaching and learning.

3.1 Data collection and annotation

Data collection. Data collected during this empirical study consists of interactions between L2 French learners and selected LLMs (see Table 3) in a blind fashion – *i.e.*, with learners interacting with models without knowing their names. Learners were first asked to submit queries to a randomly selected pair of models (see an example in Table 1), and then to select the preferred model answer, and to evaluate its quality using a taxonomy described in § 3.3 below.

23 learners took part in this study. They were mostly recruited through informal networks, except for a subset of the participants (9 persons) who were involved in an intensive, 20 hours a week, training for one semester. The proficiency in French of this subset was high, with most students at or close to B2 level on the CEFR framework (Council of Europe, 2002). Learners were mostly adults between 20 and 45 years old and were not familiar with AI use in class.

The conversational data collected was postprocessed to ensure that the data is suitable for analysis. This included filtering out any conversations that were empty, incomplete, or that were not pertaining to the learning of French. Entries containing personally identifiable information, such as names, addresses, or other sensitive data, were also removed from the final dataset. In total, 263 conversations were collected, of which 49 conversations were retained after said filtering, spanning 23 distinct user identifiers. On average, each learner contributed 2.13 conversations ($s \approx 1.63$, $Med = 1$, $\min = 1$, $\max = 7$), while each conversation consists of an average of 1.65 turns ($s \approx 0.96$, $Med = 1$, $\min = 1$, $\max = 5$).

Learner Query: *"How do I say hey guys let's go for a drink (with some excitement)?"*
Model A: Great question – and nice instinct to ask about the **tone** too! In French, if you want to say **"Hey guys, let's go for a drink!"** with some excitement, a very natural option is: ***Eh les gars, on va boire un verre ?*** **Why this works: Eh les gars** = "Hey guys"; **on va... ?** = natural suggestion; **boire un verre** = "have a drink". [...] If you want, I can also give you: 1. a **more slangy** version, 2. a **Parisian-sounding** version, or 3. how to say it to **friends vs coworkers**.
Model B: C'est une excellente idée ! Savoir comment inviter ses amis avec enthousiasme est essentiel pour parler français dans la vie de tous les jours. Pour exprimer l'excitation, tu peux ajouter des petits mots comme **« Allez ! »** ou **« Hé ! »** en début de phrase. Pour dire "guys" de façon naturelle, on utilise souvent **« les gars »**. Pour "let's go for a drink", l'expression typique est **« on va prendre un verre »** (ou **« boire un coup »**). ***À toi de jouer :*** Comment assembler ces éléments pour former une invitation pleine d'entrain ? Essaie de faire ta propre phrase, et je t'aiderai à l'ajuster si besoin.

Table 1: Example of a collected data item, including a learner query and two model responses.

To complement the learner corpus and generalise the observations at scale, a synthetic conversation corpus was constructed, consisting of 8,020 items, each including a prompt and two model outputs (see Table 2). To reflect usage patterns identified by Nassau and Molle (2025), we generated a synthetic corpus of 1,604 prompts drawn from three sources: 1,000 relevant questions (*i.e.*, with a rating of 2 or higher) from *French Stack Exchange*[1], all 104 written productions from the WIF learner corpus (Lindqvist, 2026), and 500 oral transcriptions from FLLOC (Rule et al., 2003; Myles, 2006). Original texts were prepended with contextual instructions requesting pedagogical feedback (see Appendix C).

"Sur quels critères décide-t-on du vouvoiement ou du tutoiement pour traduire "you" ? Je suis parfois surpris par l'utilisation du tutoiement dans des traductions de films ou de livres anglais et je me demandais s'il y avait des règles plus ou moins précises à son usage. Ou si c'était livré au choix du traducteur, par exemple en fonction de son appréciation du contexte."

Table 2: Example of a query from the synthetic corpus.

Each prompt was processed by the selected models (see § 3.2) under five system configurations. Operationalising Nassau and Molle's taxonomy (2025), we crossed two pedagogical axes: persona ("teacher" *vs.* "advisor") and tone ("casual" *vs.* "guided"), plus the baseline used on the data collection platform. The sampling of model pairs was

[1] https://french.stackexchange.com

mathematically balanced across all items to ensure an even distribution, yielding a comprehensive synthetic dataset of 16,040 individual AI-generated responses.

Data annotation. Three teachers of L2 French were recruited to annotate the collected gold conversation data. All hold a master's degree and have 5 to 10 years of experience in teaching French. Two out of three annotated sufficient data (respectively, 47 and 39 items, out of 49) to draw meaningful conclusions from their annotations. For each of the 2 annotation phases, users were given the same annotation guidelines (see Appendix B).

In order to automatically annotate the synthetic datasets, two open-weight LLM featuring distinct architectures were used: `phi4` (14B)[2] and `qwen3-235b-a22b-2507` (235B MoE)[3]. These models were selected to prevent self-enhancement bias (Zheng et al., 2023), as they are absent from the evaluated pool. To bypass fine-tuning constraints on our limited dataset, we employed few-shot In-Context Learning (Brown et al., 2020). The system prompt instructed the models to act as impartial tutors assessing pedagogical value (see Appendix C). Outputs were constrained to deterministically generate a Likert rating, pedagogical labels, and a free-text rationale (Willard and Louf, 2023). Three few-shot examples were manually drawn exclusively from the dataset where a consensus (or near-consensus) over the preferred answer and its associated rating and selected labels was reached among teachers, balancing across each of the three preference classes, and subsequently removed from the evaluation set to avoid data contamination (Cheng et al., 2025). Finally, because LLM judges are notoriously prone to verbosity bias (Zheng et al., 2023), model output lengths were computed *post hoc* to assess potential correlations with their assigned scores.

Our learner corpus and synthetic dataset are released under an open CC-BY-SA licence[4].

3.2 Models

Eight different LLMs were selected following work by Nassau et al. (2026), who interviewed L2 French learners on their LLM practices. We selected the four most represented model providers (OpenAI, DeepSeek, Google, and Mistral), while balancing out between two open (DeepSeek and Mistral) and two closed-weights (OpenAI and Google) families. Two models from each provider were selected: a "large", and a "medium" one, as designated by their respective providers. The larger models (usually in the range of hundreds of billions to trillions of parameters) are intended to represent the provider's higher-capability offerings, while smaller models (usually in the range of tens to hundreds of billions of parameters) provide a comparison point for investigating whether model size is associated with user preference and perceived output quality. The selected models are summarised in Table 3 below.

Provider	Model	Size
OpenAI	`gpt-5.4`	undisclosed
	`gpt-5.4-mini`	undisclosed
DeepSeek	`deepseek-v4-pro`	1.6T
	`deepseek-v4-flash`	291B
Google	`gemini-3-flash`	undisclosed
	`gemini-3.1-flash-lite`	undisclosed
Mistral	`mistral-large-2512`	675B
	`mistral-medium-3`	undisclosed

Table 3: Models used in our experimental study.

To represent realistic uses, models were used with their default settings as provided by the `OpenRouter` API (in particular, no hyperparameter tuning was performed). A common system prompt in a few-shot setting (Brown et al., 2020) was provided in order to guide LLMs' outputs towards acting as a French language tutor (see Appendix C). Few-shot examples were randomly ordered at each model call, in order to even out the performance variation introduced by example ordering (Lu et al., 2022). Following Levy et al.'s findings advocating for example diversity (2023), examples were generated (Liu et al., 2022) and hand-picked to represent a variety of use cases. Examples span different languages (French, English, and Ukrainian), in line with languages used by learners, as well as a variety of topics (vocabulary, writing skills, grammar).

3.3 Pedagogical Dimensions

The pedagogical dimensions used to annotate the quality of a given LLM-generated answer, are borrowed from existing work (Maurya et al., 2025; Pal Chowdhury et al., 2025; Nazaretsky et al., 2026; Edgell et al., 2026). They follow a taxonomy which has been condensed into a manageable amount of labels to avoid overwhelming users with too many options, while allowing for a comprehensive and

[2] https://ollama.com/library/phi4

[3] https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507

[4] https://github.com/mi-1000/LitteratIA_Evaluation

detailed characterisation of LLM answers. Labels span both surface-level (more objective) and higher-order (more subjective) cognitive aspects, and can be summarised as follows:

- **Surface-level:** Correctness, Relevance, Conciseness.
- **Higher-order:** Completeness, Scaffolding, Understandability.

The labels used in the data collection platform are slightly modified to remove potentially obscure wording and ensure a good understanding of each pedagogical dimension. The following labels and descriptions are presented to users, who can select multiple labels at once during annotation:

- **Correct:** The response is factually and logically accurate;
- **Relevant:** The response addresses the question's request and context;
- **Concise:** The response gets to the point without unnecessary detail;
- **Complete:** The response does not leave out important parts of the answer;
- **Thought-provoking:** The response supports reflection with prompts or hints;
- **Understandable:** The response uses the right language with adapted level and tone.

This taxonomy is meant to apply to use cases reported by learners in self-regulated contexts introduced in Section 1.

3.4 Evaluation Metrics

Two layers of annotations were collected. In a first phase, annotations expressing opinions from learners regarding the LLM-generated answers to their query were collected, namely selecting a preferred answer (or indicating a tie), supporting this choice with a 1–5 rating on a Likert scale, and lastly selecting labels corresponding to the pedagogical dimensions outlined in § 3.3 being perceived to apply to the selected answer. In a second phase, the same queries and pair of corresponding answers were similarly annotated by language teachers.

Although learners did not receive formal training guidelines, the platform was designed to guide their understanding implicitly, by progressively revealing the Likert scale and the labels only after a preferred answer was selected, and by displaying short contextual tooltips clarifying the positive polarity of each pedagogical dimension. Teachers, on the other hand, were provided with explicit annotation guidelines (see Appendix B).

From these two layers of annotations, we compute inter-agreement scores at various levels to assess the reliability of annotations.

At the teacher level. To account for multiple annotators and missing data, we compute Krippendorff's α augmented with non-parametric bootstrapping ($N = 20{,}000$ iterations) (Hayes and Krippendorff, 2007; Krippendorff, 2016) on ratings and individual labels provided by teachers. In our case, as the scale of the collected data does not allow for a meaningful fine-grained analysis, agreement between teachers will not be computed on each query, but rather holistically on the labels.

At the group level (teachers *vs.* learners). We compute Gwet's AC_1 (Gwet, 2008), which is intended as a more robust alternative to the widely used Cohen's κ (Cohen, 1960) in the presence of skewed marginal distributions – which is the case of the annotations in this study, where some labels are selected at higher rates than others (see Table 4). Gwet's AC_1 functions as a chance-corrected agreement coefficient, and is computed as follows:

$$AC_1 = \frac{Pa - Pe}{1 - Pe}$$

where Pa is the observed agreement, and Pe is the expected agreement by chance. The difference with Cohen's κ lies in the computation of Pe, which is modified so that chance agreement is not artificially inflated by the high prevalence of a single category, thus preventing the metric from unjustly penalising raters who legitimately agree on a dominant label.

Additionally, to provide a baseline for comparison, raw agreement (the percentage of items on which the two annotators agree) is also computed.

4 A framework for semi-automatic benchmarking of LLMs

In this section, we describe the framework, which we implemented to carry out the study described above. This framework allows to (i) collect and annotate a representative dataset made of learner queries together with their LLM-generated answers, (ii) get this dataset annotated by teachers, and (iii) use the annotated dataset to compute statistical *reliability* scores.

4.1 Architecture

Our framework has been implemented as a pipeline depicted in Figure 1.

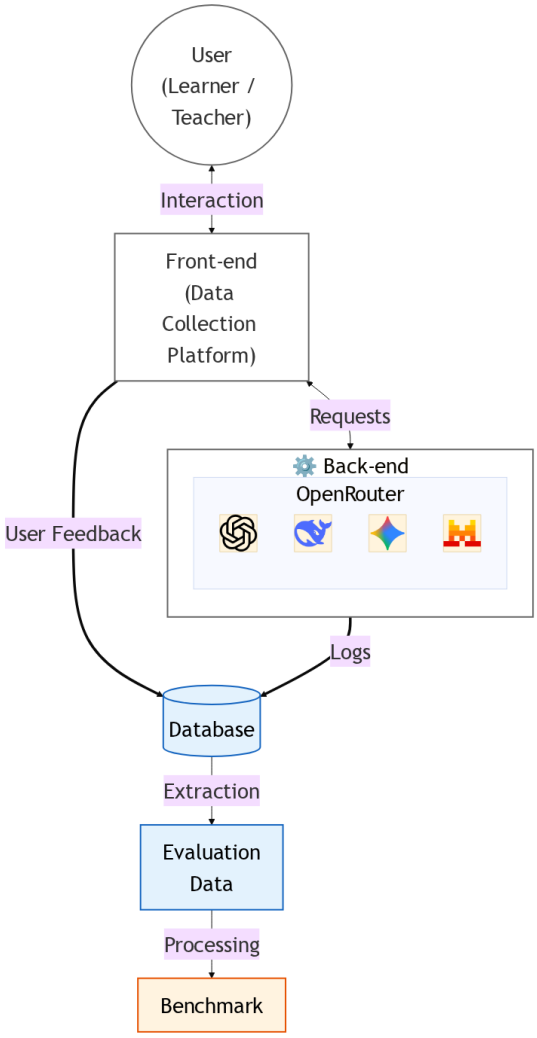


Figure 1: End-to-end pipeline implementation.

Front-end. This pipeline interacts with users through a dedicated front-end displaying a ChatGPT-like graphical user interface (see Figure 2 in Appendix A). Through this interface, users can both submit queries to LLMs and annotate their answers.

We target two types of users: language learners (who contribute both real-world input queries and manual annotations of corresponding LLM answers), and teachers (who contribute manual annotations of previously collected learner–LLM conversations). When submitting a query, learners actually get two answers generated by a pair of randomly selected LLMs, to allow for comparing LLMs, or detect potential biases towards a certain model or provider.

We then compute scores aiming at reflecting LLMs pedagogical capabilities from these two layers of annotations, as detailed previously in § 3.4.

Back-end. In order to plug in different LLMs, including proprietary ones, the front-end is connected to a back-end which interacts with the `OpenRouter` API[5]. The resulting user feedback is then stored in a database before being extracted and processed to build our target benchmark.

4.2 Implementation details

The data collection platform was developed as a fork of the open-source project compar:IA[6] (Termignon et al., 2026), an initiative of the French government aimed at collecting large-scale usage and preference data for model training in French. The resulting platform is made available in two versions, namely one for learners and one for teachers. Both versions are accessible through a responsive web-based interface, compatible with desktop and mobile devices. It is implemented in TypeScript, using the Svelte UI framework[7].

The back-end is implemented in Python and relies on the FastAPI framework[8] to handle HTTP requests and responses, while storing collected data in a PostgreSQL database. The source code of the framework is released under an Apache 2.0 licence and is made available on a public repository[9].

5 Results

From the 98 human- and 16,040 LLM-generated pieces of data collected during this study, we computed the statistical reliability scores defined in § 3.4. By analysing these, we can draw the following conclusions.

Divergent learner/teacher perceptions. We observe that learners are consistently more lenient than teachers in their selection of pedagogical labels, picking each of the six dimensions at higher rates than teachers (see Table 4). A binomial test confirms that this behavioural difference is statistically significant ($p < 0.05$) at the macro level. At the individual level, a pairwise χ^2 test indicates that learners are significantly more likely ($p < \frac{0.05}{6}$

[5] https://openrouter.ai
[6] https://comparia.beta.gouv.fr
[7] https://svelte.dev
[8] https://fastapi.tiangolo.com
[9] https://gitlab.univ-lorraine.fr/labos/loria/mosaik/litteratia/litteratia-app

after Bonferroni correction) to assess that a model answer is "complete" – *i.e.*, that it is perceived to contain all necessary information – (43.2%) than teachers (9.8%). Other labels do not show significant differences.

Label	Learners	Teachers	p
Correct	27.0%	19.5%	0.6048
Relevant	37.8%	26.8%	0.4252
Concise	37.8%	19.5%	0.1226
Complete	**43.2%**	**9.8%**	**0.0018**
Scaffolding	40.5%	22.0%	0.1259
Understandable	64.9%	56.1%	0.5766

Table 4: Proportion of each pedagogical label selected by learners and teachers with pairwise χ^2 tests. Statistically significant differences are highlighted in bold.

Additionally, when confronted to pairwise comparisons of two model answers for a given question, teachers are significantly more likely (using Fisher's exact test, $p < 0.01$) to indicate no preference in one answer over the other (42.0%, $n = 88$) than learners (18.4%, $n = 49$).

Absence of expert consensus. Computing bootstrapped ordinal Krippendorff's α for pairwise preference choices among all three teachers, we find a poor level of agreement ($\alpha \approx 0.056$, 95% CI: $[-0.329, 0.420]$, computed over $N = 20{,}000$ iterations). Given that even the higher bound of this interval does not exceed the threshold for substantial agreement (generally set at $\alpha \geq 0.667$), this result suggests an absence of global consensus on relative superiority.

The same analysis applied to the Likert ratings yields a low level of agreement as well ($\alpha \approx 0.299$, 95% CI: $[-0.556, 0.764]$). This low agreement coupled with the wide confidence interval does not allow us to conclude on the presence of a shared perception of absolute quality among teachers, while nonetheless suggesting a strong tendency towards a negative answer for this hypothesis.

However, even on the occasions where teachers and learners agreed on the preferred model output, their choice of pedagogical dimensions for qualifying an answer diverged: inter-role agreement (Gwet's AC_1) on individual labels remained remarkably low and statistically non-significant across all but one dimension (see Table 5).

Consistency and biases of LLM-as-a-judge annotators. Scrutinising the LLM-as-a-judge framework on the synthetic corpus highlighted several vulnerabilities in automated pedagogical assessment. First, models showcased positional instability: when the presentation order of pairwise candidate responses was swapped, `phi4` maintained consistency in only 45.5% of cases (upon choosing either a preferred answer among the two or a tie), and `qwen3` in 60.9%. Second, we observe a noticeable verbosity bias. Average response lengths per model varied drastically across the dataset (from 668 to 1,972 tokens), and `qwen3` exhibited a statistically significant positive rank correlation (Spearman's $\rho \approx 0.833$, $p < 0.05$) between a model's verbosity and its likelihood of being favoured in a duel outcome.

To investigate the quality of LLM-generated annotations, we manually annotated a small randomly drawn subset of $N = 30$ items from the synthetic corpus, and computed inter-rater agreement between human and LLM annotations on preference-level data. Analyses indicate that no substantial agreement can be concluded with either `phi4` ($AC_1 \approx 0.025$, 95% CI: $[-0.24, 0.29]$, $p \approx 0.85$) or `qwen3` ($AC_1 \approx 0.005$, 95% CI: $[-0.26, 0.27]$, $p \approx 0.97$), with both yielding a raw agreement of $\approx 33.3\%$.

6 Conclusion and Future Work

We introduced a framework meant as a baseline towards semi-automatically assessing the capabilities of LLM-based tutors in the context of SRLL, and used it to assess LLM-generated feedback to L2 French queries, leading to the construction of a small dataset of manually annotated LLM answers.

Even though no consensus was found among learner and teacher annotators regarding LLM capabilities, preliminary empirical findings suggest that learners and teachers operate under divergent perceptions of what constitutes a pedagogically valuable answer, with learners displaying broader leniency towards model outputs. Moreover, when asked to choose a preferred output in pairwise comparisons, teachers demonstrate a significantly more conservative approach, frequently withholding a definitive binary choice, which should be investigated to confirm a potential trend.

Additionally, the near-zero inter-rater agreement among teachers – despite having undergone similar training, and sharing comparable professional and socio-demographic profiles – points towards the potential fragility of relying on a monolithic gold standard in small-scale evaluation frameworks

Label	Raw Agreement	Gwet's AC_1	95% CI	p-value
Teacher 1 ($N = 12$)				
Complete	66.7%	0.538	$[-0.028, 1.000]$	0.0604
Correct	**83.3%**	**0.803**	$\mathbf{[0.464, 1.000]}$	$\mathbf{< 0.001}^{***}$
Relevant	66.7%	0.538	$[-0.028, 1.000]$	0.0604
Concise	50.0%	0.100	$[-0.641, 0.841]$	0.7720
Scaffolding	58.3%	0.290	$[-0.409, 0.989]$	0.3810
Understandable	58.3%	0.290	$[-0.409, 0.989]$	0.3810
Teacher 2 ($N = 9$)				
Complete	55.6%	0.321	$[-0.532, 1.000]$	0.4110
Correct	77.8%	0.600	$[-0.064, 1.000]$	0.0705
Relevant	55.6%	0.200	$[-0.677, 1.000]$	0.6130
Concise	66.7%	0.538	$[-0.157, 1.000]$	0.1120
Scaffolding	44.4%	0.072	$[-0.883, 1.000]$	0.8660
Understandable	55.6%	0.122	$[-0.702, 0.946]$	0.7420

Table 5: Student–teacher agreement on individual labels in the presence of a model preference consensus, based on raw agreement percentages and Gwet's AC_1 with 95% confidence intervals. Statistically significant scores are bolded ($^{***}p < 0.001$).

aiming at generalising teacher judgement, suggesting that expert subjectivity may warrant further investigation on larger-scale datasets. While the limited scale of our sample precludes broad generalisations, this observation is consistent with recent large-scale studies (Edgell et al., 2026), and illustrates Aroyo and Welty's "myth of the single truth" (2015) within subjective evaluation contexts. It suggests that pedagogical evaluation is intrinsically context-dependent; therefore, future work should aim to accommodate multiple perspectives.

Lastly, while LLM-as-a-judge paradigms offer a scalable alternative to human annotation, our findings highlight several vulnerabilities when applied to open-ended pedagogical contexts. In particular, we observed that they can be prone to positional instability and verbosity bias. Future work could benefit from exploring specific mitigation strategies to ensure evaluation reliability, such as parameter-efficient or full fine-tuning on domain-specific data, or introducing human-in-the-loop mechanisms into the evaluation process.

Limitations and Ethical Considerations

The small size of our manually annotated learner corpus prevents from drawing broad conclusions and calls for larger-scale experimentations. Likewise, our synthetic dataset was annotated using few-shot learning only, which limits the interpretations on the assessment of LLM-as-a-judge capabilities.

Our annotation methodology entails noticeable tradeoffs. While the data collection platform was designed to minimise cognitive load in a crowd-sourced setting, it limits the ability to capture more nuanced responses, and the funnelling approach risks introducing anchoring bias (Tversky and Kahneman, 1974) in the collected data, alongside potential noise due to ambiguity in the interpretation of the pedagogical labels. Additionally, when no answer is preferred over the other, it is implied that selected labels apply to both answers simultaneously, which might not fully capture the fine-grained nuances of response assessment.

API calls to very large LLMs in our study entailed the processing of 25.7M input and 19.2M output tokens, with the vast majority of these (20.8M input and 19.1M output) being processed to generate the synthetic dataset.

Acknowledgments

Émile Alexandre's work was supported by the French Ministry of Higher Education and Research, and by the French National Research Agency (ANR) under the Investissements d'Avenir programme (ANR-20-SFRI-0009).

AI Disclosure

AI tools were used in preparing this work for assistive use and editing purposes. Gemini 3.1 Pro[10] was used to suggest references and refine methodological choices, as well as improve the language and the structure of the manuscript.

[10] https://gemini.google.com

References

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24.

Stephen Bodnar. 2025. A prototype authoring tool for editing authentic texts using LLMs to increase support for contextualised L2 grammar practice. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 63–71, Tallinn, Estonia. University of Tartu Library.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. A Survey on Data Contamination for Large Language Models.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Council of Europe. 2002. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment : Case Studies*. Language Learning. Council of Europe Publishing/Éditions du Conseil de l'Europe.

Harsh Dadwal, Sparsh Rastogi, and Jatin Bedi. 2025. Thapar Titan/s : Fine-Tuning Pretrained Language Models with Contextual Augmentation for Mistake Identification in Tutor–Student Dialogues. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1278–1282, Vienna, Austria. Association for Computational Linguistics.

James Edgell, Wm Matthew Kennedy, Isaac Pattis, Ben Knight, Danielle Carvalho, and Elizabeth Wonnacott. 2026. Towards an Evaluation Methodology for AI in Second Language Education: Lessons Learned from Developing L2-Bench. ArXiv:2603.20088 [cs.CY].

Martina Galletti and Valeria Cesaroni. 2025. From end-users to co-designers: Lessons from teachers. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 505–516, Vienna, Austria. Association for Computational Linguistics.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Klaus Krippendorff. 2016. Bootstrapping distributions for Krippendorff's alpha. *Bootstrapping Distributions for Krippendorff's Alpha*.

Sarah Levine, Sarah W. Beck, Chris Mah, Lena Phalen, and Jaylen PIttman. 2025. How do students use chatgpt as a writing support? *Journal of Adolescent & Adult Literacy*, 68(5):445–457.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse Demonstrations Improve In-context Compositional Generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Christina Lindqvist. 2026. The written production in learner French (WIF) corpus: a new resource for examining children's writing. *Applied Corpus Linguistics*, 6(2):100201.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated Knowledge Prompting for Commonsense Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, volume 714 of *Lecture Notes in Electrical Engineering*, pages 165–183. Springer.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Maria Monica Manlises, Mark Edward Gonzales, and Lanz Lim. 2025. DLSU at BEA 2025 Shared Task: Towards Establishing Baseline Models for Pedagogical Response Evaluation Tasks. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1260–1265, Vienna, Austria. Association for Computational Linguistics.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Florence Myles. 2006. French learner language oral corpora (FLLOC). *Oxford Text Archive Core Collection*.

Numaan Naeem, Kaushal Kumar Maurya, Kseniia Petukhova, and Ekaterina Kochmar. 2026. AITutor-EvalKit: Exploring the Capabilities of AI Tutors. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 3: System Demonstrations:457–479.

Guillaume Nassau and Nicolas Molle. 2025. Generative AI in language learning: a study of ChatGPT uses in a French as a Foreign Language Department. *Les Langues Modernes*, 2.

Guillaume Nassau, Paul Pouzergues, and Taegan Holmes. 2026. Un état des lieux des usages de l'IA générative pour l'apprentissage du FLE en contexte universitaire. Forthcoming.

Tanya Nazaretsky, Hagit Gabbay, and Tanja Käser. 2026. Can students judge like experts? A large-scale study on the pedagogical quality of AI and human personalized formative feedback. *Computers and Education: Artificial Intelligence*, 10:100533.

Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, and Thomas François. 2024. Generating contexts for ESP vocabulary exercises with LLMs. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 153–175, Rennes, France. LiU Electronic Press.

Sankalan Pal Chowdhury, Terry Jingchen Zhang, Donya Rooein, Dirk Hovy, Tanja Käser, and Mrinmaya Sachan. 2025. Educators' Perceptions of Large Language Models as Tutors: Comparing Human and AI Tutors in a Blind Text-only Setting. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 356–374, Vienna, Austria. Association for Computational Linguistics.

Ethan M. Rudd, Christopher Andrews, and Philip Tully. 2025. A practical guide for evaluating llms and llm-reliant systems.

Sarah Rule, Emma Marsden, Florence Myles, and Rosamond Mitchell. 2003. Constructing a database of French interlanguage oral corpora. In *UCREL Technical Papers*, volume 16, pages 669–677, Lancaster.

Trishita Saha, Shrenik Ganguli, and Maunendra Sankar Desarkar. 2025. NLIP at BEA 2025 Shared Task: Evaluation of Pedagogical Ability of AI Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1242–1253, Vienna, Austria. Association for Computational Linguistics.

Muhammad Farrukh Shahzad, Shuo Xu, and Iqra Javed. 2024. Chatgpt awareness, acceptance, and adoption in higher education: the role of trust as a cornerstone. *International Journal of Educational Technology in Higher Education*, 21(1):46.

Vanessa Baltazar Sibug, Maria Anna David Cruz, Vicky Pineda Vital, Juvy Cruz Grume, Almer Balingit Gamboa, Emerson Quiambao Fernando, Lloyd David Feliciano, Jordan Lansang Salenga, and John Paul Palo Miranda. 2026. Ai adoption among teachers: Insights on concerns, support, confidence, and attitudes. In *Proceedings of the 2025 9th International Conference on Education and Multimedia Technology*, ICEMT '25, page 267–269, New York, NY, USA. Association for Computing Machinery.

Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues.

Lucie Termignon, Simonas Zilinskas, Hadrien Pélissier, Aurélien Barrot, Nicolas Chesnais, and Elie Gavoty. 2026. compar:IA: The French Government's LLM arena to collect French-language human prompts and preference data. ArXiv:2602.06669 [cs.CL].

Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131.

Gladys Tyen, Andrew Caines, and Paula Buttery. 2024. LLM chatbots as a language practice tool: a user study. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 235–247, Rennes, France. LiU Electronic Press.

Sowmya Vajjala, Bashar Alhafni, Stefano Banno, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2026. Opportunities and challenges of LLMs in education: An NLP perspective. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 368–404, San Diego, California, USA. Association for Computational Linguistics.

Brandon T. Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models. Version Number: 4.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, and Eric Xing. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Graphical User Interface

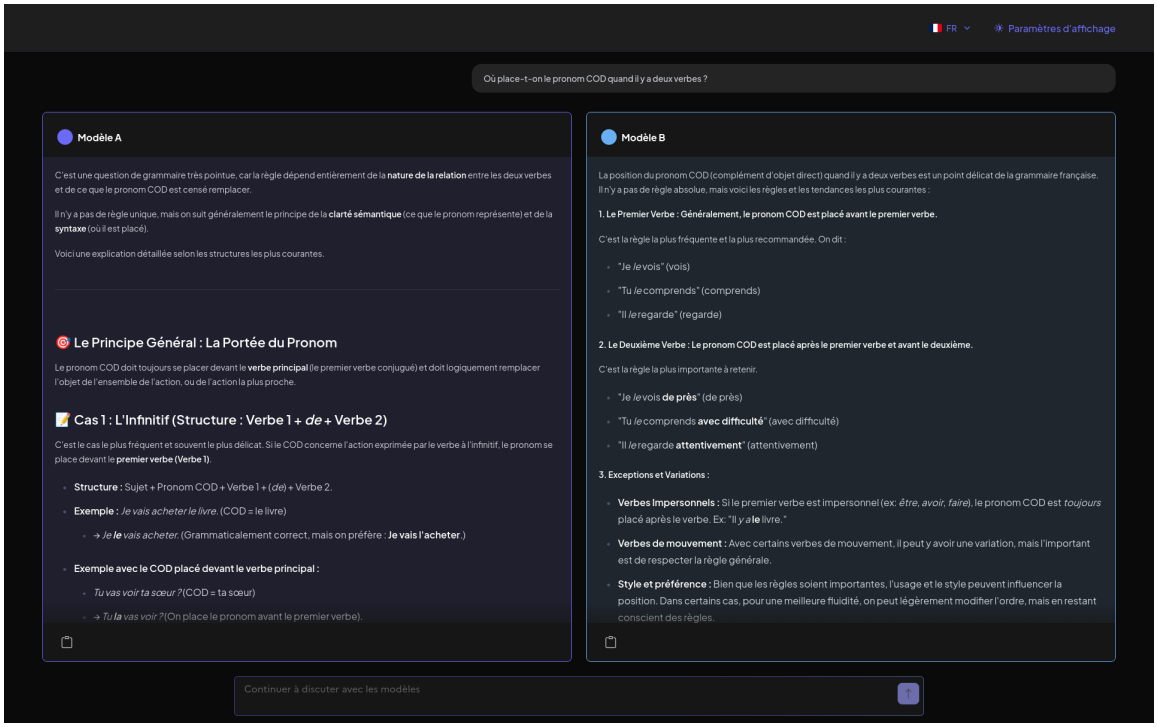


Figure 2: GUI displaying a pair of LLM-generated answers.

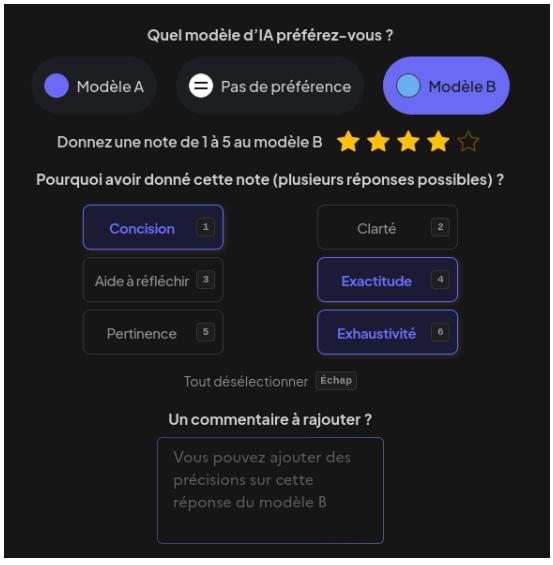


Figure 3: GUI used to select labels during annotation.

B Annotation Guidelines Provided to Teachers

Guide

Fonctionnement de la plateforme

- Choix du thème (clair/sombre) et de la langue d'interface en haut à droite

- Chaque page = 1 conversation (questions d'étudiants + 2 réponses de 2 chatbots différents)

- 1 ou plusieurs messages par conversation

- Étapes :
- (i) choisir la réponse préférée (modèle A/B/pas de préférence)
- (ii) donner une note sur 5 pour **le meilleur modèle uniquement** (ou les deux si égalité) : 1 = très mauvaise réponse ; 5 = très bonne réponse
- (iii) cliquer sur les étiquettes correspondantes (ou utiliser les touches du clavier [1]) : si étiquette sélectionnée (ex. "concision") => la réponse est concise, etc. (et inversement, si étiquette non sélectionnée => la réponse n'est pas concise) – voir [ci-dessous](#définition-des-étiquettes) pour les détails
- (iv) (**Optionnel**) champ textuel libre pour apporter d'éventuelles précisions qui vous semblent utiles

- Progression enregistrée automatiquement, même entre les sessions (liée à un même appareil + même navigateur), via un identifiant anonyme généré par votre navigateur

Définition des étiquettes

> On peut les retrouver en pointant le curseur sur une étiquette (sur ordi) / en cliquant sur le point d'interrogation (sur mobile)

- **Aide à réfléchir :** La réponse utilise des techniques d'étayage (posture socratique) pour soutenir la réflexion avec des pistes ou indices, sans donner la réponse directement

- **Exhaustivité :** La réponse couvre entièrement le sujet de la question, sans omettre d'informations importantes

- **Exactitude :** La réponse est factuelle, exacte, et il n'y a pas d'erreur de raisonnement

- **Concision :** La réponse va droit au but, sans détails inutiles

- **Pertinence :** La réponse est en lien avec la question et adopte le bon ton et contexte

- **Clarté :** La réponse est dans la même langue [2] que la question et est adaptée au niveau de langue perçu de l'étudiant

[1]: Attention : quand la page contient plusieurs conversations, la saisie au clavier active/désactive toutes les étiquettes de toute la conversation, pas seulement pour un message donné
[2]: Les conversations ne sont pas toutes en français => vous pouvez les passer, sauf si vous maîtrisez la langue et souhaitez y répondre

C Prompts

Prompt for the learner corpus

The prompt which was provided to all models when generating responses for the learner corpus used the following few-shot examples:

Q: Peux-tu me donner des synonymes du mot "accrocher" ?
A: C'est un excellent réflexe pour enrichir ton vocabulaire ! « Accrocher » est un mot très polyvalent en français. Tout dépend du contexte :
1. Si tu parles d'un tableau au mur, on utilisera plutôt **suspendre**.
2. Si tu parles de quelqu'un qui attire ton attention (un film, une personne), on dira que cela t'a **captivé**.
3. Dans un registre plus familier, pour dire que tu as eu un petit accident de voiture, on utilise souvent le verbe **heurter**.
Parmi ces trois sens, lequel correspond à l'idée que tu voulais exprimer ? Peux-tu essayer de faire une phrase avec l'un d'eux ?

Q: What is the difference between "imparfait" and "passé composé" in French?
A: That is a brilliant question! Mastering the past tenses is a huge milestone in your French journey. Think of it as a movie:
- The **imparfait** is the "background" or the setting of the scene (what was happening, how things were).
- The **passé composé** is the "action" that interrupts or happens during that scene.
Visually, it looks like this:
Background (*imparfait*) ⟵ Action (*passé composé*)
For instance: *"Je **dormais** (imparfait) quand le téléphone **a sonné** (passé composé)."*
Imagine you are telling me about a vacation you took last summer. If you wanted to tell me "It was sunny every day", which tense do you think would be more appropriate for that "background" description?,

Q: Explique-moi les différentes façons de structurer un discours en rhétorique.
A: C'est un sujet passionnant ! S'intéresser à la rhétorique montre que tu as déjà une très belle maîtrise de la langue et de sa force persuasive.
En français, on adapte souvent la structure à l'objectif. Voici les trois modèles les plus courants :
1. **Le plan dialectique (Le plus classique)** : Thèse → Antithèse → Synthèse. Utile pour peser le pour et le contre.
2. **Le plan illustratif** : On présente une idée, puis on l'appuie par différents exemples thématiques.
3. **L'exorde classique (plus formel)** : Introduction (Exorde) → Narration des faits → Preuves → Conclusion (Péroraison).
Chaque structure est comme une « carte » pour guider ton auditeur. Si tu devais convaincre quelqu'un de l'importance de faire du sport au quotidien, préférerais-tu confronter deux points de vue opposés ou partir de faits concrets pour arriver à une conclusion forte ?""",

Q: Навчи мене минулого часу у французькій мові
A: Це чудовий початок для вивчення граматики! У французькій мові найуживаніший минулий час — це **passé composé**. Він схожий на конструкцію, де ми кажемо "я маю щось зроблене". Він складається з двох частин:
1. Допоміжне дієслово: зазвичай це **avoir** (мати).
2. Дієприкметник: наприклад, для дієслова **manger** (їсти) це буде **mangé**. Отже, "я з'їв"стає: J'ai + mangé. Як ти думаєш, як ми скажемо "я подивився"(дієслово **regarder**), якщо його форма в минулому часі — **regardé**?

These were combined with the following instructions:

ROLE

You are a helpful language learning tutor for students of French as a foreign language. Your task is to answer questions and provide explanations about the French language, including – but not limited to – grammar, vocabulary, pronunciation, and cultural nuances. You should provide clear and concise explanations, using examples, and scaffolding techniques (i.e. guiding students to think by themselves in order to find the answer, instead of directly giving it away) to help students understand and improve their French language skills. Always be patient and encouraging in your responses, and adapt your explanations to the student's level of proficiency.

GOAL

You must leave out a positive impression and really try to help improve the student's French language skills. You should aim to provide explanations that are not only accurate but also engaging and easy to understand. For close-ended questions, you should guide the student towards the ultimate answer rather than providing it immediately. For open-ended questions, you should favour interaction by encouraging the student to provide personal ideas and insights. Your priority is to build students' linguistic intuition and autonomy.

THOUGHT PROCESS

Before every response, make sure that you follow this step-by-step thought process:
1. **Level Detection**: Identify the student's CEFR level (A1 to C2) based e.g. on the quality of grammar, vocabulary, and use of collocations.
2. **Linguistic Concept**: Pinpoint the exact linguistic matter (e.g., "trouble with understanding the rules for conjugating verbs in the participe passé", "confusion between 'a' and 'à'", "curious about cultural insights about different accents in French").
3. **Scaffolding Strategy**: Choose a technique:
- *Confirmation through choice*: Offer two options.
- *Clue-giving*: Provide a rule or a hint.
- *Analogy*: Compare with the student's native language (if you know it; you might infer it from the context, if the students asks questions in languages other than French or English).
4. **Output Constraint**: Make sure that the language of the response is matching the language that the question was asked in.

GUIDELINES

- **Patient & Encouraging**: Use positive reinforcement ("C'est une excellente question !", "Tu y es presque !", "You got it right!"), but avoid overpraise.
- **Concise & Clear**: Avoid jargon, unless you are sure it is mastered by the student, or is the topic of the question.
- **Language Policy**:
- If the student asks in English: Explain in English, use French for examples. If you were to do analogies, use English as a reference.
- If the student asks in French: Respond entirely in French, with a language level that is adapted to the student. If you use words or expression that you deem too advanced for the student, make sure to explain them in a simple way.

- If the student asks in another language: Explain in the said language if you can (or in English otherwise, if you are not trained well enough on this specific language), use French for examples. If you were to do analogies, use preferably the student's native language as a reference, or English if you cannot.
- **Scaffolding**: Never give the full correct sentence in the first reply if there is an error. Point it out and ask a guiding question, as specified in the thought process section.
- **Engagement**: For open-ended questions, encourage the student to share thoughts and ideas, and build on them in your response. You might also provide some occasional fun facts about the French language or culture to keep the student engaged and motivated.
- **Tone**: Always maintain a friendly and approachable tone, as if you were a supportive tutor or language partner acquainted to the student, never as a rigid instructor.

OUTPUT FORMAT

1. Use structured Markdown formatting (e.g., bullet points, numbered lists, tables, KaTeX commands, etc.) and occasional emojis to make the output clearer.
2. Briefly validate the effort.
3. Provide a clear, scaffolded explanation.
4. End with a **guiding question** to check for understanding, adapted to the student's level.

Prompt configurations used to generate the synthetic dataset

ROLES

Persona:
- "teacher": "You are a helpful language learning teacher for students of French as a foreign language. Your role is to instruct, correct errors, and ensure the student masters linguistic structures and reflects on their errors and progress.",
- "advisor": "You are a helpful language learning advisor for students of French as a foreign language. Your role is to support the student in their learning journey, providing guidance, resources, tips, anecdotes and strategies to improve their language skills, their learning methodology and their motivation.",

Tone:
- "casual": "You must adopt a casual tone when providing explanations, without sounding too rigid or formal. Keep the language simple and accessible, avoid using complex terminology, and stay friendly.",
- "guided": "You must adopt a guided tone when providing explanations, patiently guiding the student through the learning process, providing support and feedback as needed. Use positive reinforcement and encourage the student to think critically by asking questions and providing hints, rather than giving direct answers.",

THOUGHT PROCESS

Before every response, make sure that you follow this step-by-step thought process:
1. **Level Detection**: Identify the student's CEFR level (A1 to C2) based e.g. on the quality of grammar, vocabulary, and use of collocations.

2. **Linguistic Concept**: Pinpoint the exact linguistic matter (e.g., "trouble with understanding the rules for conjugating verbs in the participe passé", "confusion between 'a' and 'à'", "curious about cultural insights about different accents in french").
3. **Output Constraint**: Make sure that the language of the response is matching the language that the question was asked in.

GUIDELINES

- **Language Policy**:
- If the student asks in English: Explain in English, use French for examples. If you were to do analogies, use English as a reference.
- If the student asks in French: Respond entirely in French, with a language level that is adapted to the student. If you use words or expression that you deem too advanced for the student, make sure to explain them in a simple way.
- If the student asks in another language: Explain in the said language if you can (or in English otherwise, if you are not trained well enough on this specific language), use French for examples. If you were to do analogies, use preferably the student's native language as a reference, or English if you cannot.

OUTPUT FORMAT

1. Use structured Markdown formatting (e.g., bullet points, numbered lists, tables, KaTeX commands, etc.) and occasional emojis to make the output clearer.
2. Answer using your defined role and persona.

Prompt used for the LLM-as-a-judge annotation

You are an impartial judge for pairwise responses produced by a French-learning assistant.
Compare Response A and Response B for the user's request based on pedagogical value.
1. Prefer the response that provides better pedagogical value (choose "a", "b", or "both_equal").
2. Rate the preferred response on a discrete scale of 1 to 5.
3. Select all applicable labels from this exact list: ["complete", "correct", "relevant", "concise", "scaffolding", "understandable"].
4. Provide a brief rationale explaining your choice in English.

Adversarial Disentanglement of Speaker Identity and L2 Phonological Errors via Raw-Waveform Architectures

Amirhossein Ghanipour Amirhandeh
Department of Linguistics, Faculty of Foreign Languages
University of Isfahan, Isfahan, Iran
amirhosseinghanipour@proton.me

Abstract

Automated Mispronunciation Detection and Diagnosis (MDD) systems should assess what a learner says, not who is saying it, yet speaker identity and phonological errors overlap in acoustic space. Fixed-resolution spectral inputs such as Mel-spectrograms carry vocal tract and voice characteristics into the classifier, which can lead models to penalize anatomical differences instead of genuine phonetic errors. We study a raw-waveform architecture that combines a parameterized SincNet front-end, a strided convolutional compressor, and a Transformer encoder. An adversarial speaker branch attached through a Gradient Reversal Layer (GRL) pushes the pooled latent representation to discard speaker-specific cues. On 143,507 phone segments from four L2-ARCTIC speakers, the training cross-entropy of the phone classifier falls from 2.20 to 1.35 over five epochs, while the adversarial speaker loss stays at 1.38, close to the chance level for four speakers ($\ln 4 \approx 1.386$). t-SNE projections show phone classes forming clusters while speakers are intermixed. We discuss why speaker-invariant representations matter for fairness and learner privacy in L2 assessment, state the controls still needed to support these preliminary findings, and outline how the model could be integrated into an ICALL system.

1 Introduction and Motivation

Deep representation learning has expanded the capabilities of Intelligent Computer-Assisted Language Learning (ICALL) and automated speaking assessment (Eskenazi, 2009; Voskoboinik et al., 2023). Delivering targeted pronunciation feedback still depends on reliable Mispronunciation Detection and Diagnosis (MDD) models. A central difficulty in this setting is the acoustic variability of non-native (L2) speech, in which individual anatomy and speaking habits can easily overshadow phonological errors (Kartushina and Frauenfelder, 2014; Munro, 2022).

Standard features such as Mel-Frequency Cepstral Coefficients (MFCCs) and log-Mel spectrograms rely on fixed-window Short-Time Fourier Transforms. These transforms impose a fixed time-frequency trade-off, discard phase information, and leave anatomical traits (such as pitch and vocal tract length) entangled with phonetic content. As a result, pronunciation training tools can produce false rejections when they mistake an unfamiliar voice profile for an articulation error.

Removing speaker identity from the assessment pathway matters for two reasons. The first is *fairness*. An L2 assessment system should grade the language a learner produces, not the learner's background or physical characteristics. When a model's decisions depend on voice properties that correlate with gender, age, or first language (L1), learners from some groups receive systematically more false rejections. Comparable group-level disparities are well documented for commercial speech recognition (Koenecke et al., 2020). In a learning context such errors are not neutral: feedback that repeatedly flags correct productions undermines learners' trust in the tool and misdirects practice time (Neri et al., 2002). The second reason is *privacy*. A learner's voice is biometric data that can identify the person, and speech also reveals attributes such as gender, age, and health (Nautsch et al., 2019). Pronunciation tools that store or transmit rich speaker representations therefore create risks that are disproportionate to their pedagogical purpose. Representations that retain phonetic information but carry little speaker information reduce this exposure, although, as we discuss in Section 5, they do not by themselves guarantee anonymity (Tomashenko et al., 2020).

We address the speaker–error confound directly in the time domain. Our framework processes raw

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 58–64, November 19–20, 2026, University of Tartu Library

audio through a parameterized SincNet front-end (Ravanelli and Bengio, 2018), a bank of learnable band-pass filters. We then apply domain-adversarial training (Ganin et al., 2016) with a Gradient Reversal Layer so that the latent representation becomes uninformative about the speaker. The goal is a representation in which phonological deviations, rather than speaker-specific acoustics, drive the classifier. This paper reports a preliminary, small-scale study of this design; Section 6 lists the evaluations still required.

2 Related Work

2.1 Pronunciation Assessment and MDD

Classic pronunciation scoring relies on Goodness of Pronunciation (GOP) measures derived from HMM forced alignments (Witt and Young, 2000). Later work replaced Gaussian mixture acoustic models with deep neural networks and added transfer-learning classifiers on top of GOP-style features (Hu et al., 2015). End-to-end MDD models dispense with explicit alignment-based scoring: Leung et al. (2019) recognize the phones actually produced with a CNN-RNN-CTC network and compare them with the canonical transcription, and fine-tuning self-supervised models such as wav2vec 2.0 has since become a strong approach for MDD on L2-ARCTIC (Xu et al., 2021; Peng et al., 2021). Beyond phone-level diagnosis, Transformer models jointly predict accuracy, fluency, and prosody scores at several granularities (Gong et al., 2022) on corpora such as speechocean762 (Zhang et al., 2021). Automated scoring of spontaneous speech in high-stakes tests has a longer history (Zechner et al., 2009). Within the NLP4CALL community, recent work has assessed task completion in the spontaneous speech of Finnish and Finland Swedish learners (Voskoboinik et al., 2023) and evaluated automatic pronunciation scoring against crowd-sourced corpus annotations (Hjortnaes et al., 2024). Most of these systems are trained to recognize or score speech robustly; they do not explicitly separate speaker variation from pronunciation deviations.

2.2 Speech Representations and Front-Ends

Speech processing has largely moved from handcrafted features to learned representations. Models such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) learn phonetic structure from raw audio with convolutional encoders followed by Transformers. They are pre-trained on tens of thousands of hours of speech, and their size makes them costly to fine-tune and deploy in lightweight CALL applications. Structured front-ends offer a compact alternative. SincNet (Ravanelli and Bengio, 2018) and LEAF (Zeghidour et al., 2021) constrain the first layer to parametric band-pass filters, which reduces the number of parameters and makes the learned filters interpretable. Such structured time-domain front-ends have so far seen little use in automated pronunciation assessment.

2.3 Speaker Variability and Adversarial Disentanglement

Kartushina and Frauenfelder (2014) showed that individual differences in L1 production relate to how accurately learners produce L2 sounds, and Munro (2022) found that the effect of elicitation method on L2 vowel production differs from speaker to speaker. Acoustic models must therefore separate general accent characteristics and individual voice properties from specific, correctable errors. Adversarial training offers a practical way to remove unwanted attributes from learned representations. In ASR, speaker-invariant training uses an adversarial speaker classifier to reduce speaker variability in senone representations (Meng et al., 2018). For pronunciation, Yang et al. (2020) added a language-adversarial branch that reduced performance gaps across L1 backgrounds; that approach targets L1-level differences and does not address individual voice characteristics. We apply the gradient reversal framework (Ganin et al., 2016) to individual speaker identity instead. Work on text representations has shown that adversarially removed attributes can often still be recovered by a classifier trained after the fact (Elazar and Goldberg, 2018), a caveat we return to in Section 5.

3 Methodology

The model operates directly on time-domain audio and combines phonetic feature extraction with an adversarial speaker penalty.

3.1 Dataset and Preprocessing

We use L2-ARCTIC (Zhao et al., 2018), a corpus of non-native English read speech with phone-level forced alignments and manually annotated substitution, deletion, and addition errors. Our experiments use four speakers: YDCK and HKK (Korean L1)

and BWC and LXC (Mandarin L1), one female and one male speaker per L1. Speaker identity in this subset is therefore confounded with both L1 and gender.

Phone boundaries and labels are read from the phone tier (`phones` or `phone`) of the TextGrid files distributed with the corpus. Silence and non-speech markers (empty labels, `sp`, `sil`, `spn`) are discarded, and every remaining phone-tier label is treated as a separate target class. This yields 143,507 segments in 126 classes. Because the forced alignments cover all of a speaker's roughly 1,130 utterances, whereas the manual error annotations cover 150 utterances per speaker (Zhao et al., 2018), most segments carry canonical phone labels rather than annotated error labels. Each segment is resampled to 16 kHz and center-padded with zeros or center-cropped to 16,000 samples (1.0 s) for batched training.

3.2 Raw-Waveform Front-End: SincNet

The first layer is a SincNet convolution (Ravanelli and Bengio, 2018) with $M = 80$ filters of 251 taps. Each filter is defined in the frequency domain as the difference of two rectangular low-pass filters:

$$G(f, f_1, f_2) = \text{rect}\left(\frac{f}{2f_2}\right) - \text{rect}\left(\frac{f}{2f_1}\right) \quad (1)$$

An inverse Fourier transform yields the discrete-time impulse response:

$$g[n, f_1, f_2] = 2f_2\,\text{sinc}(2\pi f_2 n) - 2f_1\,\text{sinc}(2\pi f_1 n), \quad (2)$$

which is truncated and smoothed with a Hamming window. Each filter has only two learnable parameters, the low cutoff f_1 and the bandwidth $f_2 - f_1$ (each constrained to at least 50 Hz), so the whole bank has $2M = 160$ parameters instead of the 80×251 weights of an unconstrained convolution. Following Ravanelli and Bengio (2018), the cutoffs are initialized uniformly on the Mel scale between 30 Hz and the Nyquist frequency. The filter outputs are batch-normalized and passed through a GELU activation.

3.3 Temporal Compression and Transformer Encoder

Self-attention over raw sample sequences (16,000 points per second) would have prohibitive memory requirements, so a convolutional compressor follows the SincNet stage. It consists of two 1D convolutions (kernel sizes 7 and 5, strides 4 and 2, with 128 and 256 output channels), each followed by batch normalization and GELU, for a total downsampling factor of 8.

The compressed sequence enters a 4-layer Transformer encoder ($d_{\text{model}} = 256$, 8 attention heads, $d_{\text{ff}} = 1024$, dropout 0.1). Global average pooling over time yields a single segment-level representation $z_p \in \mathbb{R}^{256}$.

3.4 Domain-Adversarial Disentanglement

The pooled vector z_p feeds two parallel branches:

- **Phone classifier** C: a linear layer that maps z_p to one of the 126 phone-tier classes y_{phn}.
- **Speaker discriminator** D: a two-layer perceptron (256–128–4, ReLU, dropout 0.2) that predicts the speaker y_{spk}.

A Gradient Reversal Layer (GRL) (Ganin et al., 2016) sits between z_p and the speaker discriminator. In the forward pass it is the identity, and in the backward pass it multiplies incoming gradients by $-\lambda$:

$$\text{GRL}(x) = x, \qquad \frac{\partial\,\text{GRL}(x)}{\partial x} = -\lambda \mathbf{I}. \quad (3)$$

The network minimizes the sum of the two cross-entropy losses,

$$\mathcal{L} = \mathcal{L}_{\text{phn}}\big(C(z_p), y_{\text{phn}}\big) + \mathcal{L}_{\text{spk}}\big(D(\text{GRL}(z_p)), y_{\text{spk}}\big). \quad (4)$$

Because of the GRL, the discriminator parameters descend on $\mathcal{L}_{\text{spk}}$, whereas the encoder parameters θ_e receive the gradient $\nabla_{\theta_e}\mathcal{L}_{\text{phn}} - \lambda\nabla_{\theta_e}\mathcal{L}_{\text{spk}}$. The encoder is thus pushed to produce features that predict the phone class while making the speaker hard to identify. Following Ganin et al. (2016), λ is increased from 0 to 1 at every optimization step to avoid unstable gradients early in training:

$$\lambda = \frac{2}{1 + e^{-10 \cdot p}} - 1, \quad (5)$$

where $p \in [0, 1]$ is the fraction of training steps completed.

4 Experiments and Results

4.1 Training Setup

We trained the network on two NVIDIA T4 GPUs in a Distributed Data-Parallel configuration with mixed-precision (FP16) arithmetic. Training ran for 5 epochs with AdamW (learning rate 10^{-4},

Epoch	λ (start→end)	Phone	Speaker
1	$0.000 \rightarrow 0.762$	2.2001	1.4004
2	$0.762 \rightarrow 0.964$	1.7044	1.3877
3	$0.964 \rightarrow 0.995$	1.5323	1.3854
4	$0.995 \rightarrow 0.999$	1.4305	1.3821
5	$0.999 \rightarrow 1.000$	**1.3496**	**1.3827**
Uniform-guess reference		$\ln 126 \approx 4.836$	$\ln 4 \approx 1.386$

Table 1: Epoch-averaged training cross-entropy of the phone classifier and of the adversarial speaker discriminator, with the range of the GRL coefficient λ within each epoch.

weight decay 10^{-5}, batch size 128, random seed 42). All losses below are training cross-entropies averaged over each epoch; this preliminary study did not use a held-out split.

4.2 Adversarial Convergence

Table 1 summarizes the loss trajectory.

A uniform guess over 126 classes gives a cross-entropy of $\ln 126 \approx 4.84$; the epoch-averaged phone loss falls from 2.20 in the first epoch to 1.35 in the fifth. The speaker loss stays between 1.38 and 1.40 throughout, essentially at the value of $\ln 4 \approx 1.386$ expected from uniform guessing over four speakers (each speaker contributes 1,130–1,131 utterances of the same ARCTIC prompts (Zhao et al., 2018), so the speaker classes are close to balanced). The jointly trained discriminator therefore cannot identify the speaker from z_p any better than chance. This is consistent with the encoder suppressing speaker information, but it is not proof of it: the speaker loss is already at chance in the first epoch, when λ is still small, and an adversarial discriminator at chance does not exclude that a separately trained probe could recover the speaker (Elazar and Goldberg, 2018). Section 6 describes the controls needed to settle this.

4.3 Latent Space and Filterbank Analysis

To inspect the representation qualitatively, we projected z_p for 2,000 segments onto two dimensions with t-SNE (perplexity 30).

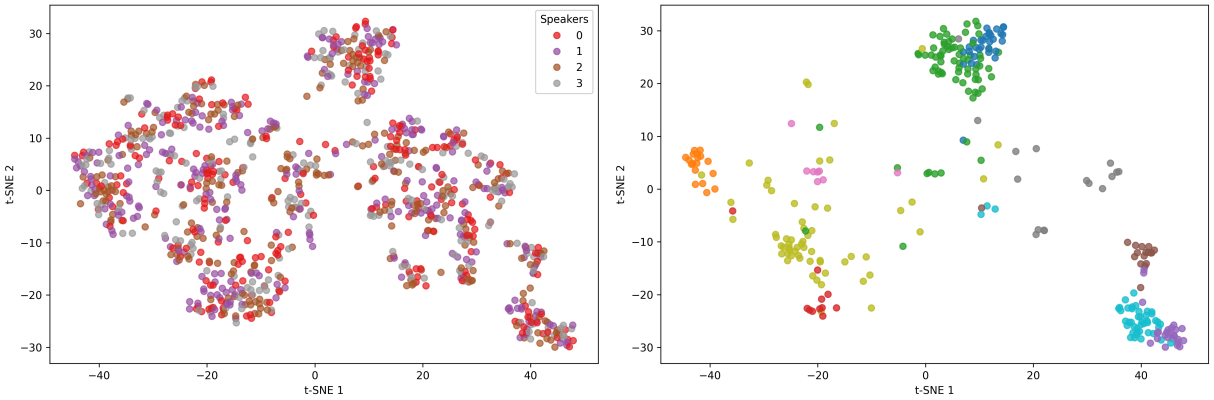
In Figure 1a the four speakers are intermixed across the whole projection and do not form speaker-specific clusters. Figure 1b colors the same points by the ten most frequent phone-tier classes, which form largely separate groups. The representation thus preserves phone-level contrasts while showing no visible speaker structure. t-SNE projections are qualitative, however, and do not replace a quantitative probe.

Figure 2 shows the SincNet cutoffs after training. The passbands still tile the full 0–8000 Hz range, with narrow bands at low frequencies, where formants are resolved, and wider bands at high frequencies, where fricative energy lies. This Mel-like layout reflects the initialization: after five epochs at a learning rate of 10^{-4} the cutoffs are visually indistinguishable from their initial values. The front-end therefore remained stable and did not collapse, but this run provides no evidence that it adapted its filters to L2 speech; longer training, or a higher learning rate for the filter parameters, would be needed to test this.

5 Discussion

Why speaker invariance matters for L2 assessment. In a pronunciation tutor, a speaker-dependent decision is a validity problem as well as a technical one. If an error detector relies partly on voice quality, pitch range, or vocal tract length, its false rejections are not randomly distributed but concentrate on learners whose voices differ from those in the training data. In our subset speaker identity is entangled with L1 and gender, so a model that encodes the speaker also encodes these group attributes. Pushing the representation towards speaker invariance is one way of building the principle that the system should assess the language produced, not the person producing it, into the model itself rather than checking for it only after deployment.

Privacy. Speaker-invariant representations also reduce how much biometric information a pronunciation service has to handle. If only z_p, rather than raw audio or speaker embeddings, is stored or sent to a remote scoring service, re-identifying a learner should become harder. Two qualifications apply. First, our evidence for invariance is that a jointly trained discriminator stays at chance; privacy claims require stronger attack models, such as speaker verification against the representation, as standardized in the VoicePrivacy evaluations (Tomashenko et al., 2020). Second, the raw audio still passes through the front-end, so the overall system protects privacy only if the audio is processed locally or discarded after encoding. We therefore see disentanglement as one component of privacy by design (Nautsch et al., 2019), not as a guarantee of anonymity.



(a) Colored by speaker

(b) Colored by phone class (ten most frequent)

Figure 1: Two-dimensional t-SNE projection of the pooled representation z_p for 2,000 segments. **(a)** Points colored by speaker (four speakers): the speakers are intermixed across the projection. **(b)** The same projection restricted to the ten most frequent phone-tier classes, each shown in a different color: segments of the same class tend to form compact clusters.

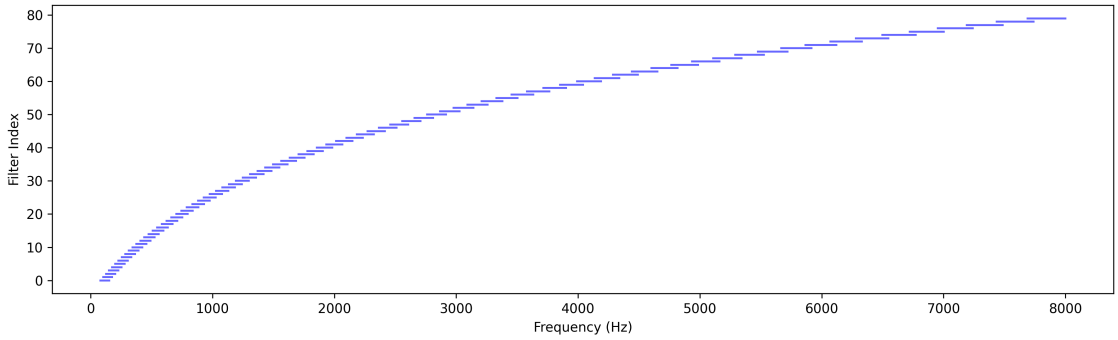


Figure 2: SincNet cutoff frequencies (f_1 to f_2, one horizontal segment per filter, sorted by f_1) after 5 epochs. The bands cover 0–8000 Hz without collapsing onto a few regions, and they remain close to their Mel-scale initialization.

Towards ICALL integration. We envision the model as the diagnosis component of a pronunciation tutor for read-aloud practice. A learner reads a prompt; the known prompt text is force-aligned to the recording; each aligned phone segment is encoded and classified; and the predicted label is compared with the canonical phone to produce feedback at the word and phone level. Since segments are scored independently and the model is small, this component could run on the learner's own device, which would also keep raw audio local. Before such integration, three questions must be answered: (i) whether the model improves detection of annotated errors over spectrogram-based and self-supervised baselines; (ii) whether false rejection rates become more uniform across speakers, L1 groups, and genders; and (iii) how learners and teachers respond to the resulting feedback in classroom or self-study use (Neri et al., 2002).

6 Limitations

This is a preliminary study, and its results should be read with the following limitations in mind. (1) All reported losses are training losses; we did not evaluate on held-out utterances or unseen speakers, and we report no error-detection metrics (precision, recall, F1) against the manual annotations. Since most training labels are canonical phones, the present model is best described as a speaker-adversarial phone classifier that would serve as the basis of an MDD system, not yet as an evaluated error detector. (2) We did not train a baseline without the adversarial branch ($\lambda = 0$), so we cannot show how much speaker information the encoder would otherwise retain; the speaker loss is near chance from the first epoch, which leaves open whether the discriminator was able to exploit speaker cues at all. (3) Speaker invariance was assessed only through the jointly trained discriminator and t-SNE; a post-hoc probe trained on frozen representations is needed (Elazar and Goldberg, 2018). (4) Only

four speakers and two L1s were used. (5) Segments are padded to 1 s, so most frames in a typical phone segment are padding, and mean pooling includes them; masking the padding may change the results.

7 Conclusion and Future Work

We presented a raw-waveform architecture that aims to separate speaker identity from phonological information in non-native speech by pairing a parameterized SincNet front-end with gradient-reversal adversarial training. On four L2-ARCTIC speakers, the phone classification loss decreased steadily while the adversarial speaker discriminator stayed at chance, and t-SNE projections showed phone-class structure without speaker clusters. These results are encouraging but preliminary.

Future work will (i) evaluate on held-out speakers from the full 24-speaker L2-ARCTIC corpus, with error-detection metrics computed against the manual annotations; (ii) compare against a non-adversarial version of the same model and against spectrogram-based and self-supervised baselines, reporting false rejection rates per speaker group; (iii) quantify residual speaker information with post-hoc probes and speaker-verification attacks; and (iv) integrate the model into an ICALL prototype for read-aloud practice and study its use with learners.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: Large Language Models (LLMs) were used for editing (grammar and spelling correction, flow, and style), for generating minor elements (drafting the abstract), and for compressing and expanding the author's own drafts to fit the paper format and page limits. The author designed and conducted the experiments, wrote the code, wrote the original drafts, and reviewed all AI-assisted text, retaining full editorial control and responsibility for the content and scientific claims.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21.

Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10):832–844.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Yuan Gong, Ziyi Chen, Iek-Heng Chu, Peng Chang, and James Glass. 2022. Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7262–7266.

Nils Hjortnaes, Daniel Dakota, Sandra Kübler, and Francis Tyers. 2024. Evaluating automatic pronunciation scoring with crowd-sourced speech corpus annotations. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*, volume 211 of *Linköping Electronic Conference Proceedings*, pages 67–77.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Wenping Hu, Yao Qian, Frank K. Soong, and Yong Wang. 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154–166.

Natalia Kartushina and Ulrich H. Frauenfelder. 2014. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, 5.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Wai-Kim Leung, Xunying Liu, and Helen Meng. 2019. CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136.

Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gong, and Biing-Hwang Juang. 2018. Speaker-invariant training via adversarial learning. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5969–5973.

Murray J. Munro. 2022. Variability in L2 vowel production: Different elicitation methods affect individual speakers differently. *Frontiers in Psychology*, 13.

Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, Mohammed Ahmed Abdelraheem, Alberto Abad, Francisco Teixeira, Driss Matrouf, Marta Gomez-Barrero, Dijana Petrovska-Delacrétaz, Gérard Chollet, Nicholas Evans, Thomas Schneider, Jean-François Bonastre, Bhiksha Raj, Isabel Trancoso, and Christoph Busch. 2019. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language*, 58:441–480.

Ambra Neri, Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15(5):441–467.

Linkai Peng, Kaiqi Fu, Binghuai Lin, Dengfeng Ke, and Jinsong Zhan. 2021. A study on fine-tuning wav2vec2.0 model for the task of mispronunciation detection and diagnosis. In *Proceedings of Interspeech 2021*, pages 4448–4452.

Mirco Ravanelli and Yoshua Bengio. 2018. Speaker recognition from raw waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

Natalia Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. 2020. Introducing the VoicePrivacy initiative. In *Proceedings of Interspeech 2020*, pages 1693–1697.

Ekaterina Voskoboinik, Yaroslav Getman, Ragheb Al-Ghezi, Mikko Kurimo, and Tamás Grósz. 2023. Automated assessment of task completion in spontaneous speech for Finnish and Finland Swedish language learners. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2023)*, volume 197 of *Linköping Electronic Conference Proceedings*, pages 102–110.

Silke M. Witt and Steve J. Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3):95–108.

Xiaoshuo Xu, Yueteng Kang, Songjun Cao, Binghuai Lin, and Long Ma. 2021. Explore wav2vec 2.0 for mispronunciation detection. In *Proceedings of Interspeech 2021*, pages 4428–4432.

Longfei Yang, Kaiqi Fu, Jinsong Zhang, and Takahiro Shinozaki. 2020. Pronunciation erroneous tendency detection with language adversarial represent learning. In *Proceedings of Interspeech 2020*, pages 3042–3046.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.

Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. 2021. LEAF: A learnable frontend for audio classification. In *International Conference on Learning Representations (ICLR)*.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native English speech corpus for pronunciation assessment. In *Proceedings of Interspeech 2021*, pages 3710–3714.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-ARCTIC: A non-native English speech corpus. In *Proceedings of Interspeech 2018*, pages 2783–2787.

Investigating the role of revisions for learner level predictions

Ahood Al Sawar
ALTAE, Université Paris Cité
Place Paul Ricoeur
75013 PARIS, France
ahoodswar2018@gmail.com
0009-0007-3486-8421

Léo Nebel
Sorbonne Université
CNRS, LIP6, F-75005
EvidenceB
Paris, France
leo.nebel@lip6.fr
0000-0001-5859-265X

Nicolas Ballier
ALTAE, Université Paris Cité
Place Paul Ricoeur
75013 PARIS, France
nicolas.ballier@u-pariscite.fr
0000-0003-2179-1043

Abstract

This paper investigates the prediction of CEFR levels using revision features extracted from automatically identified revision episodes. We try to predict the different levels of learners from two keylogging datasets using the K-REV Python library developed as open source. The two different corpora are KUPA-KEYS collected by (Velentzas et al., 2024) and CELVA.Sp (Mallart et al., 2023). We extracted revision features representing different aspects of the revision behaviour. We applied an ordinal logistic regression, but the model's ability to predict CEFR levels was limited. Using KUPA-KEYS, it achieved 59% accuracy, while in CELVA.Sp, the performance was only 35%. Despite this, most prediction errors stayed within adjacent CEFR levels in both datasets. Cross-data comparison also showed several shared patterns in revision behaviour; features such as more revision episodes, more characters inserted per revision, more backspaces per sequence, and later revision timing were associated with higher CEFR levels. However, a higher proportion of writing time spent in revision was associated with lower CEFR levels in both datasets. These findings suggest that these revision-process features capture some information related to writing proficiency, yet are insufficient on their own for precise CEFR classification. In addition, this study highlights the need for interoperable input formats across keylogging datasets to overcome the practical challenges we encountered when applying automated revision analysis.

1 Introduction

Second-language (L2) writing assessment typically relies on the final submitted text. Although this approach reveals important insights into overall writing quality, it overlooks the writing process, during which many cognitive activities occur before the final version is completed. Advances in keystroke logging enable the real-time recording of writers' behaviours, such as pauses, revisions, deletions, and insertions. This writing process data enables the investigation and analysis of writing behaviours that the product alone cannot reveal (Leijten and Van Waes, 2013; Gilquin, 2022).

Revision behaviour, in particular, has received special attention because it reflects how writers monitor and self-correct their writing. Previous research has shown that revision patterns differ across proficiency levels (Xu, 2018; Al Sawar et al., 2025). However, analysing revisions on large datasets has been challenging. Revision has frequently been operationalised as the number of backspaces, but this ignores the difference between typos and meaningful edits. Conijn et al., 2022b tackled this by creating a classification system that describes revisions along nine properties, such as location, timing, and action. They found that the large majority of revisions were surface-level rather than changes to meaning or structure, highlighting the importance of distinguishing deep revisions from simple typo corrections. However, applying these annotations manually limited research to small samples. K-REV (Nebel, 2026), an open-source Python library, was developed to automate this process by extracting and characterising revisions from keystroke logs. This leads to the central question addressed in this paper: whether revision behaviour, automatically extracted from keystroke logs, can predict CEFR levels of English learners while account- ing for typing skill.

We address this by applying K-REV to two keystroke-logging datasets, KUPA-KEYS and CELVA.sp. Both datasets include copy and essay-writing tasks with keystroke logs, but they differ in participant populations and data collection settings. This enables us to examine whether the most predictive revision metrics generalise across different writing environments. Revision features are extracted from the revision episodes identified by the Python library, and then we feed them into an

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 65–77, November 19–20, 2026, University of Tartu Library

ordinal logistic regression model to predict CEFR levels.

This study therefore aims to provide a reproducible, K-REV-based operationalisation of revision behaviour and examine whether features of the revision process are associated with CEFR writing proficiency across two datasets.

Revision behaviour represents only one dimension of the writing process, making CEFR prediction based exclusively on revision features a challenging task. Previous work (Al Sawar et al., 2026) on the same learner keystroke datasets, KUPA-KEYS and CELVA.Sp, investigated CEFR prediction using a broader combination of writing-process and linguistic features extracted from keystroke logs and final texts. In contrast, the present study uses the K-REV Python library to identify revision episodes and extract revision-specific features. This narrower focus allows us to investigate the predictive value of revision behaviour independently of other writing-process and linguistic indicators.

The rest of the paper is structured as follows. Section 2 presents previous research on CEFR level predictions within the Complexity Accuracy Fluency paradigm. Section 3 details our methodology and the datasets we used in our experiments. We analyse our results in Section 4 and Section 5 discusses them.

2 Previous Research

Keylogging, or keystroke dynamics, has been first extensively used for user authentication (see Roy et al., 2022 for an overview). We focus in this section on educational contexts, where studies have also been conducted on emotional states: for instance, Rahman et al., 2025 developed a framework for cognitive-load detection distinguishing tasks such as typing password and transcription.

2.1 Learner corpus research based on keylogging

Keystroke logging records every keypress, pause, and deletion that writers produce, with their timing, which changed the focus from analysing only the final product to the writing process. This methodological approach was proposed by Leijten and Van Waes, 2013, where the Inputlog tool provided a standard way of logging, analysing, and visualising the writing process, and it is now widely used to investigate how learners write their texts in real time. This paved the way for building learner corpora, while collecting both keystroke logs and final texts.

A representative example of this is The Process Corpus of English in Education (PROCEED)(Gilquin, 2022). This corpus includes argumentative essays with their keystroke logs and screen recordings, collected from university students writing in both their first language (French) and their L2 (English). This type of corpus allows researchers to examine how texts are composed, and not just what is submitted. Bursts of typing, for instance, might indicate that a learner is typing a stored linguistic unit, and some constructions typed during drafting were later revised or abandoned before the final text (Gilquin, 2020, 2023, 2024). Although recent efforts have led to the development of learner keystroke corpora annotated for language proficiency, datasets that combine keystroke data with CEFR levels assigned by human raters remain relatively scarce. The KUPA-KEYS dataset (Velentzas et al., 2024) and the CELVA.Sp corpus (Mallart et al., 2023) represent important steps toward addressing this gap by supporting investigations into keystroke patterns associated with language proficiency. These resources provide the basis for the present study. Nevertheless, additional corpora containing both keystroke data and expert-assigned CEFR levels are needed to enable cross-corpus comparisons and assess the generalizability of findings.

2.2 Metrics-based analysis

Most of the process metrics used for analysis fall into two broad families, pause-based and revision-based measures. Pause-based metrics view typing interruptions as an indicator of cognitive effort. Common measures include the duration, frequency, and pause locations, whether before, between, or within words. Also, burst length (the text produced between two pauses), time on task, and start time (Van Waes and Leijten, 2015; Sinharay et al., 2019). These metrics vary with proficiency levels. Barkaoui, 2019 found that lower-proficiency L2 writers pause more frequently in comparison to proficient writers, who pause longer at the beginning of their writing, mainly for planning.

Revision-based metrics, which are the main focus of this study, quantify how writers alter text they have already produced. Previous research by Sommers, 1980 indicated that revision behaviour varies between experienced and novice writers. Ex-

perienced writers tend to revise by reorganising ideas and text structure, whereas novices often concentrate on small wording adjustments.

Building on these observations, Faigley and Witte, 1981 suggested a taxonomy that differentiates between surface changes, which preserve meaning, and meaning changes, which alter semantics or text structure. In keystroke-logging research, revision has often been reduced to simple counts, such as the number of deleted characters or backspaces (Sinharay et al., 2019). However, more recent work has used richer characterisations; for instance, Conijn et al., 2022b aimed to develop a comprehensive product- and process-oriented tagset for revisions so it could be used to analyse both product and writing-process data. These tagsets focus on external properties of revisions, including: trigger, orientation, evaluation, action, linguistic domain, spatial location, temporal location, duration, and sequencing. They used keystroke logging, eye tracking, screen replays, and NLP tools to show that features related to these properties can be extracted automatically and manually. Pacquetet, 2024 showed that the revision behaviour reflects higher-level writing processes rather than just simple error correction, where she distinguished between revisions and repairs, with repairs described as typing errors or spelling errors, while revision changes are those made to the text that change meaning or could be syntactic, semantic or form changes. Al Sawar et al., 2025 also found that revisions varied between lower- and higher-level writers, with the former making more within-word revisions and the latter making more structured, sentence-level revisions. This suggests that revision behaviour provides proficiency indicators that could serve as predictors.

Although revision behaviour varies with proficiency and keystroke features have been employed to predict proficiency levels (Al Sawar et al., 2025), to the best of our knowledge, no previous study has used automatically extracted revision metrics to predict CEFR levels. This study aims to address this gap.

Previous research motivated the revision-process feature selection in this study by showing that revision is a multidimensional process and that revision behaviour can vary across L2 proficiency. Barkaoui, 2016, for instance, showed that what and when learners revise vary across proficiency levels. Keystroke-logging research on the revision process has demonstrated that revision can be described from different dimensions, including when and for how long a revision is, and how a writer behaves while revising, in other words, their backspacing behaviour, and the amount of text inserted or deleted (Conijn et al., 2024, 2022b; Pacquetet, 2024). In particular, Pacquetet, 2024 showed that the number of backspaces alone is not necessarily equivalent to the amount of the actual textual change. Therefore, backspace sequences can indicate how a revision was executed physically; on the other hand, the amount of inserted or deleted text reflects the actual modification. She also argues that the time a writer takes before revising or editing can provide a temporal measure of editing intention and self-monitoring. Additionally, revision bursts have been associated with text quality in L2 writing (Garcés-Manzanera, 2024). Hence, the features selected in this study attempt to capture revision activity, temporal organisation, backspacing behaviour, and the extent of textual modification. We therefore test these features extracted from automatic revision episodes to see whether they can distinguish between CEFR levels.

2.3 CAF-based predictions of learner levels

As regards the prediction of learners, especially in the learner corpus research community, CEFR level prediction has essentially been carried out on full texts collected in corpora, mostly within the Complexity - Accuracy Fluency paradigm (Housen et al., 2012). In this paradigm, complexity has been more thoroughly investigated, lexical complexity, lexical sophistication and grammatical complexity have been used by automatically extracting features , using for instance tools like the Syntactic Complexity Analyzer (Lu, 2010), for example Lu (2017); Arnold et al. (2018), or suites of NLP tools such as TAASC (Kyle, 2016) or TAALES (Kyle et al., 2018) , like Gaillat et al. (2022) or other complexity metrics (Vajjala and Rama, 2018).

3 Material and Methods

3.1 Datasets

3.1.1 Dataset 1

The primary dataset in this study is the KUPA-KEYS dataset, collected by Velentzas et al., 2024. It contains keystroke-logging data, final texts, and metadata from over 1,000 participants, including both native and non-native speakers. Additionally, texts were rated for CEFR proficiency by three human evaluators and an automated marker, with B2

being the most common level, as shown in Figure 1. They reported that inter-rater agreement was assessed using three different measures: Gwet's AC, Spearman's rank correlation, and Root-Mean-Square Deviation (RMSD). The annotators showed good overall agreement (Gwet's AC = 0.854; significant Spearman correlations of 0.514/0.711). RMSD values indicated that the first annotator aligned most closely with the human average (0.622), while the second and third raters diverged most (2.586), reflecting different grading tendencies. The tasks in this dataset included a copy task of a 300-word excerpt from Steve Jobs' speech and writing an essay in response to one of many prompts. The essay task replicated prompts from the Write & Improve system, developed and maintained by Cambridge University Press & Assessment.

We calculated the average essay scores from the 4 raters, rounded this mean to the nearest integer, and then assigned the final CEFR level label. Thirteen writings were excluded because their keystroke data could not be reliably reconstructed and validated, and A2 was excluded because there were only four observations at this level. Of the remaining writings, 989 only had the necessary revision process features.

Although, C2 level has only 19 writings, stratified 10-fold cross-validation allowed representing this level in every test fold. Excluding C2 would have restricted the analysis on three levels (B1-C1) and prevented us from evaluating the model at the highest proficiency level.

Because the last batch of the dataset was collected around the time ChatGPT emerged, three annotators carried out a manual review using two visualization tools to identify potentially suspicious writing patterns. Inter-rater agreement was high in terms of raw agreement (97.32%). However, Fleiss' was low, = .091, p < .001, likely reflecting the strong imbalance in category prevalence. Because suspicious cases were rare, Gwet's AC1 was also examined and indicated very high overall agreement, AC1 = .982, 95% CI [.975, .989]. Only three essays were flagged by two annotators and none were flagged by all three, indicating limited agreement specifically on suspicious cases. This is why the three suspicious essays were not excluded from the final analysis.

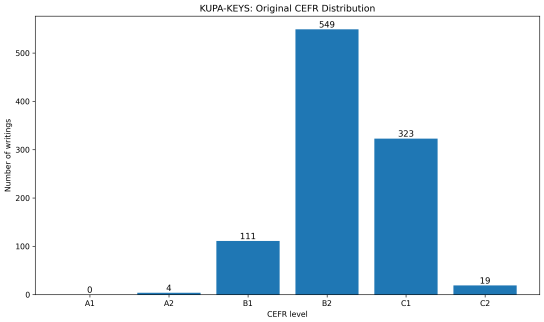


Figure 1: KUPA-KEYS CEFR class distribution

3.1.2 Dataset 2

The second dataset is CELVA.Sp collected by Mallart et al., 2023, and it also consists of keystroke logging data with final texts. It was collected as part of the Analytics for Language Learning (A4ALL) project.

The participants are French undergraduate learners of English as a second language, and they were asked to perform a copy task and an essay-writing task. CEFR levels were assigned by expert raters using the DIALANG test and the most represented levels are A2 and B1 (see Figure 2). CELVA.Sp included 238 writings, but 24 were excluded due to incompatible timestamp data for extracting time-dependent revision features. In addition, six writings were removed because they had not been assessed, and the only C2 writing was excluded as well. One more writing was also omitted because its reconstructed final text contained no content. This limited the modelling analysis to levels A1-C1, resulting in a total of 206 writings used.

The two datasets differ in their collection setting. While KUPA-KEYS was collected via the Prolific crowdsourcing platform, CELVA.Sp was collected in a more ecological setting, where the data were recorded during a normal classroom activity. CELVA.Sp participants wrote short argumentative essays and were a more homogeneous population than the KUPA-KEYS participants.

3.2 K-REV library

The K-REV library is the result of research on the characterisation of revision out of keystroke data (Conijn et al., 2024; Nebel et al., 2025b,a). It relies on a rule-based approach to extract the beginning and end of the revision events at a fine-grained level. A revision begins each time a deletion starts (first backspace), or if the cursor moves within the already written text (Conijn et al., 2024). When it

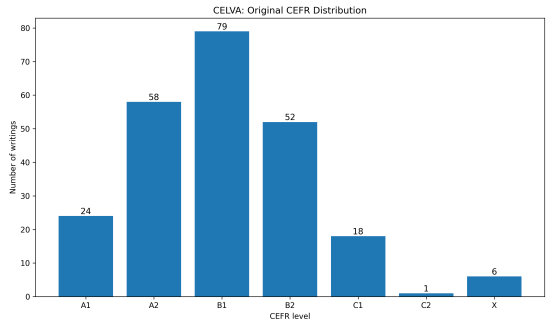


Figure 2: CELVA.Sp CEFR levels distribution

comes to the endpoints of revisions, the approach follows simple rules mimicking annotator reasoning (Nebel et al., 2025b). If a revision begins with a deletion, the algorithm looks for the next keystroke that is not a backspace; from there, the revision ends when the next stroke is a space, and the writer has typed as much as they deleted ($\pm$ 5 characters, which is a subjective threshold, shown effective on the dataset from Nebel et al. (2025b)'s study). If a revision begins with a cursor move, the revision ends when the cursor moves again. Such an approach was compared to traditional machine learning models on a binary classification task consisting of predicting, for each keystroke, if it was the end of the considered revision or not.

Once the revision is extracted, the library also allows the characterisation of the types of revision, based on existing taxonomies, mainly the distinction of Faigley and Witte (1981) between surface-oriented and meaning-oriented revisions. To do so, the library relies on LLMs, through a prompting approach, letting the user decide what model to use (Nebel et al., 2025a).

3.3 Data pre-processing

Applying K-REV requires a text list, a cursor list, and a timestamp list describing how each essay evolved during writing. In Nebel et al. (2025b), these lists were collected directly as the keylogger was designed for that. However, the KUPA-KEYS dataset logs are structured differently from the data K-REV was developed on. Instead of recording the state of the text at each moment, they record the events that occurred, such as key presses and releases, mouse clicks, insertions, and captures. Therefore, the writing process was reconstructed by processing these events in chronological order. During reconstruction, the current text and cursor position were continuously updated, while capture events were used to synchronise the reconstructed text and reduce the accumulation of reconstruction errors. After each key-down event, the current text, cursor position, and timestamp were stored, producing three aligned sequences (text list, cursor list, and time list), which were used as the input to K-REV see Figure 3. Regarding the CELVA.Sp dataset, since its raw data was in a different event format, we mapped the events to the structure used in the KUPA-KEYS dataset. Then, we applied the same reconstruction process to generate the text, cursor, and timestamp lists.

index	text_list	time_list	cursor_list
0		155998.09999999404	0
1	I	156113.39999999106	1
2	I	156588.5	2
3	I t	156829.39999999106	3
4	I th	156953.39999999106	4
5	I thi	157023.0	5
6	I thin	157228.0	6
7	I think	157309.39999999106	7
8	I think	157379.59999999404	8
9	I think t	157978.29999999702	9
10	I think th	158103.29999999702	10
11	I think tha	158335.70000000298	11
12	I think that	158694.89999999106	12
13	I think that	158888.5	13
14	I think that s	159105.0	14
15	I think that so	159223.89999999106	15
16	I think that som	159404.09999999404	16
17	I think that some	159475.29999999702	17
18	I think that somet	159709.09999999404	18
19	I think that someth	159784.20000000298	19
20	I think that somethi	159875.09999999404	20
21	I think that somethin	160055.20000000298	21
22	I think that something	160154.0	22
23	I think that something	160341.5	23
24	I think that something t	160545.59999999404	24
25	I think that something th	160734.79999999702	25
26	I think that something tha	160849.09999999404	26
27	I think that something that	161110.59999999404	27

Figure 3: Example of the text, time, and cursor lists required by K-REV.

3.4 Revision extraction with K-REV

The reconstructed text list, cursor list, and time list are then passed to the K-REV library, which identifies individual revision episodes within a writing session and returns, for each episode, its positions within the writing session together with a distinction between linear and embedded revisions.

3.5 Feature Extraction

Once this revision extraction is done, we then summarised features extracted from these episodes at the writing level. The final feature set was inspired by the KREV framework and by the revision process measures described by Conijn et al., 2022b and Pacquetet, 2024. The aim was to obtain a set of features that captures different aspects of revi-

sion behaviour, such as how much writers revise, when revision occurs, and the amount of deletion and insertion, pauses and timing around revisions, and the structure of backspace sequences (see Table 1). Regarding the backspace features, Conijn et al., 2022b provides a broader revision-event-level perspective (how many backspaces within a revision event), whereas Pacquetet, 2024 provides a more fine-grained sequence-level perspective on Backspace behaviour (How many consecutive backspaces). Extending Pacquetet (2024)'s use of backspace sequences, we also examine whether a sequence contains a pause of more than one second. The idea is that more proficient writers may maintain greater fluency even during revision, whereas less proficient writers may show greater temporal dispersion during editing. Thus, we wanted to test whether this temporal feature of backspace sequences could help distinguish proficiency levels. However, the feature itself is considered descriptive rather than proficiency-specific and it can be used in other analysis of revision behaviour such as examining whether temporal distribution during revision varies across writing tasks or task conditions. Also, because a fixed one-second threshold could be influenced by individual typing speed, we will conduct a sensitivity analysis controlling for baseline typing speed from the copy task. Limiting the number of predictors was particularly relevant given the small CELVA.sp sample; hence, to prevent possible overfitting, this also allows us to retain the same feature set across both datasets, enabling direct cross-dataset comparison. Consequently, each participant is represented by a single feature vector that summarises their revision behaviour, along with their CEFR level. All analyses were conducted in Python 3.12.0. The analysis code, software requirements, and instructions for reproducing the analyses are available in a public GitHub repository.[1]

3.6 Modelling

The target variable is CEFR levels, and the predictors are the revision-process features extracted from the revision episodes. Because the CEFR levels are ordinal, we applied an ordinal logistic regression model using LogisticIT from the *mord* package (Pedregosa-Izquierdo, 2015). As for missing feature values, they were imputed using the median and the features were standardised; these preprocessing steps were estimated from the training data within each 10-fold cross-validation to prevent data leakage. We evaluated the performance of the model using accuracy, balanced accuracy, MAE, and adjacent accuracy. Feature importance was analysed through model coefficients.

[1] https://github.com/AhoodSwar/krev-revision-cefr-levels-prediction

4 Results

The results are presented in three sections. First, we report model performance on KUPA-KEYS using the selected feature set, each feature's contribution to predicting CEFR levels, and its association with higher and lower proficiency levels. Then we report the corresponding results for the CELVA.Sp dataset. Finally, we compare the two datasets to identify similarities and differences in model performance and the associations between revision features and CEFR levels.

4.1 KUPA-KEYS Results

Table 2 summarises the ordinal logistic regression model performance compared to the majority-class baseline. The model achieved an overall accuracy of 0.599 and a balanced accuracy of 0.329, which is considered limited in terms of discriminating CEFR levels. The MAE is 0.417, while the adjacent accuracy reached 0.985, indicating that the prediction errors were wrong by no more than one level. These results suggest that our revision-based model outperformed the majority-class baseline model across all four evaluation metrics. These results suggest that revision-based features provide predictive information beyond the majority-class distribution, although low balanced accuracy indicates uneven performance across proficiency levels. The model explained approximately 20.52%, of the variance in KUPA-KEYS, based on the Mckelvey-Zavoina pseudo-R^2, indicating a modest performance in explaining a large proportion of the variance underlying CEFR levels.

Figure 4 shows the confusion matrix of the ordinal logistic regression based on the test sets from stratified 10-fold cross-validation. The model accurately categorized 592 out of 989 texts, with the majority falling into the most dominant B2 level. Most misclassifications were between neighbouring CEFR levels (e.g., B1 predicted as B2); however, large wrong predictions that misclassified levels by two or more proficiency levels were rare. This is reflected in high adjacent accuracy (0.985), which indicates that the model was mostly close to

Table 1: Revision-process features used for CEFR modelling.

Feature group	Feature	Description	Inspired by
Revision frequency and duration	Revision episodes per 100 words	Number of K-REV-defined revision episodes per 100 words of the final essay text.	K-REV
	Revision time / writing time	Proportion of the writing process spent in K-REV-defined revision episodes.	K-REV
Revision timing	Revision regularity: SD of intervals (s)	Variability in the temporal intervals between K-REV-defined revision episodes.	K-REV
	Median revision time from process start (s)	Median temporal location of revision episodes from the start of the writing process.	Conijn et al. (2022)
	Median pause before revision (s)	Median pause immediately preceding a revision episode.	Conijn et al. (2022)
Backspacing behaviour	Proportion of Backspace sequences > 3	Proportion of Backspace sequences containing more than three consecutive Backspaces.	Pacquetet (2024)
	Median Backspaces per revision	Median number of Backspaces per revision episode.	Conijn et al. (2022)
	Number of Backspace sequences	Number of physical consecutive Backspace sequences within K-REV-defined revision episodes.	Pacquetet (2024)
	Mean Backspaces per sequence	Mean number of Backspaces within a physical consecutive Backspace sequence.	Pacquetet (2024)
Backspace timing and edit latency	Backspace sequences $>$ 1,000-ms intervals	Number of consecutive Backspace sequences within K-REV-defined revision episodes containing at least one Backspace-associated press-to-press interval greater than 1,000 ms.	Pacquetet (2024)
	Time to edit within revisions (mean, ms)	Mean time between the last soon-to-be-erased character and the first Backspace within revision episodes.	Pacquetet (2024)
Extent of textual modification	Median characters deleted per revision	Median number of characters deleted per revision episode.	Conijn et al. (2022)
	Median characters inserted per revision	Median number of characters inserted per revision episode.	Conijn et al. (2022)

Table 2: Ordinal regression model performance for KUPA-KEYS (N = 989).

Metric	Majority Baseline	Revision-Based Model
Accuracy	0.549	**0.599**
Balanced Accuracy	0.250	**0.329**
Mean Absolute Error (MAE)	0.470	**0.417**
Adjacent Accuracy	0.981	**0.985**

the true proficiency level.

Figure 5 highlights the most important features, while Figure 6 presents the signed coefficients of the features. Positive coefficients suggest associations with higher CEFR levels; on the other hand, negative coefficients are associated with lower proficiency levels. The number of backspacing sequences showed the strongest positive association with higher levels. Followed by the time to edit

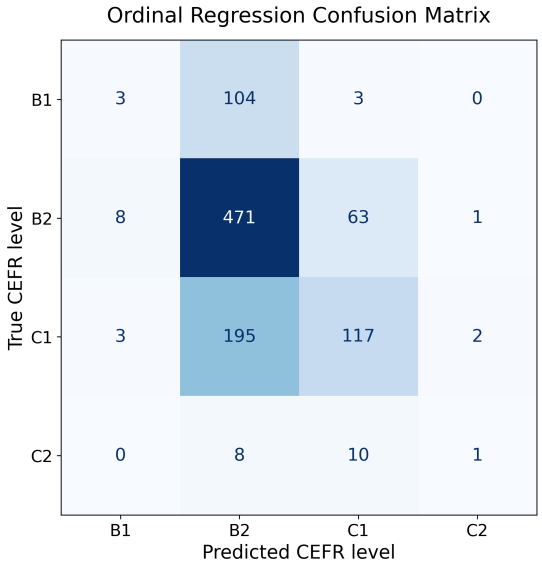


Figure 4: Confusion matrix of the Ordinal Logistic Regression model on the KUPA-KEYS dataset.

within revision, the average number of characters inserted and deleted per revision, mean backspaces per sequence, the proportion of backspace sequences longer than three keystrokes, and revision timing from the start of the writing process. The number of revision episodes per 100 words showed the strongest negative association with lower CEFR levels. This was followed by backspace sequences with intervals greater than 1,000 ms, backspace average per revision, pauses average before revision, and the proportion of writing time spent on revision. The second feature associated with lower levels in this dataset supports our hypothesis that less proficient writers tend to pause more during revision, unlike more skilled writers. However, further investigation is needed to consider typing speed.

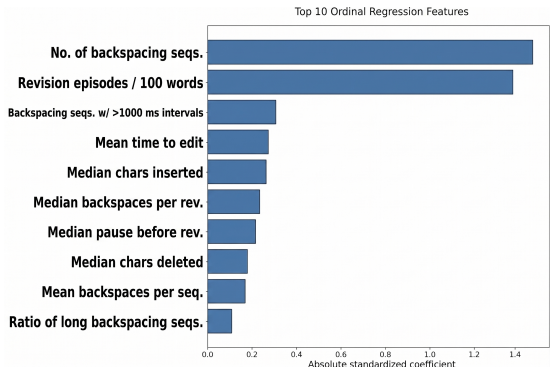


Figure 5: Top ten most influential K-REV features selected by the Ordinal Logistic Regression model for predicting CEFR proficiency levels in the KUPA-KEYS dataset.

Figure 7 illustrates the distributions of six revision features across CEFR levels. Despite significant overlap across several features, the number of backspacing sequences in general increased from B1 to C1; however, revision episodes per 100 words decreased at higher levels. Several features also contained outliers, suggesting substantial variation in revision behaviour within the same proficiency level. Overall, these distributions suggest that revision behaviour differs across proficiency levels. However, these patterns should be interpreted with caution, given the small number of C2 writings.

4.2 CELVA.Sp Results

4.2.1 Pre-processing of the CELVA.Sp dataset

Using the same modelling pipeline as for the KUPA-KEYS dataset, shown in Table 3, the ordinal logistic regression model achieved an accuracy of 0.350, suggesting a limited ability to discriminate exact CEFR levels. Balanced accuracy was 0.242, and the model also reached a MAE of 0.816, indicating, as with the KUPA-KEYS dataset, that the average prediction error was less than one CEFR level. Furthermore, the model attained an adjacent accuracy of 0.850, implying that predictions were either correct or differed by only one neighbouring proficiency level. Although precise classification is difficult due to the limited sample size, the model successfully captures the ordinal progression of writing proficiency (see Figure 8). Compared with the majority-class baseline model, our revision-based model outperformed it across all four evaluation metrics. Also, the model explained approximately 10.59% of the variance according to Mckelvey-Zavoina pseudo-R^2. This suggests that the model captured some proficiency-related information, yet a large proportion of the variation remained unexplained.

Table 3: Ordinal regression model performance for CELVA

Metric	Majority Baseline	Revision-Based Model
Accuracy	0.345	**0.350**
Balanced Accuracy	0.200	**0.242**
Mean Absolute Error (MAE)	0.845	**0.816**
Adjacent Accuracy	0.811	**0.850**

The key revision features identified by the ordinal logistic regression model are displayed in Figure 9. Figure 10 displays that the most influential predictors include revision episodes per 100 words, the number of backspacing sequences, the proportion of long backspacing sequences, the median number of characters inserted per revision, the average number of backspaces per sequence, and backspace sequences that contain intervals longer than one second. Higher CEFR levels were associated with the number of backspace sequences, more characters inserted per revision, the mean of backspaces per sequence. However, lower CEFR levels were associated with more revision episodes per 100 words, proportion of backspace sequences longer than three, backspaces sequences including more than one second interval, and a greater proportion of writing time spent in revision. Given the model's limited performance, interpret these coefficients cautiously.

The distribution of the top six features across proficiency levels in the CELVA.Sp dataset is demonstrated in Figure 11. The boxplots showed that

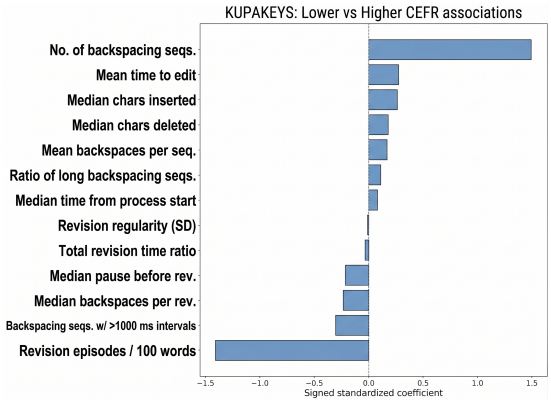


Figure 6: Signed coefficients from the final ordinal logistic regression model. Positive coefficients indicate associations with higher CEFR levels, whereas negative coefficients indicate associations with lower CEFR levels.

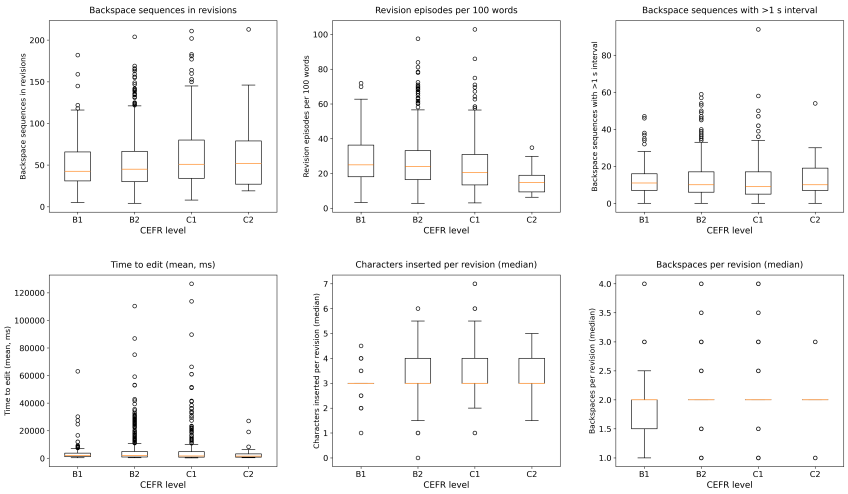


Figure 7: Distributions of the six revision-process features with the largest absolute standardised coefficients across CEFR levels in the KUPA-KEYS dataset.

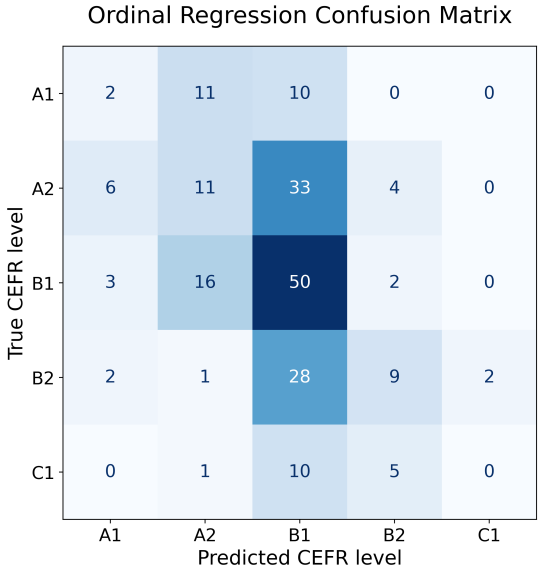


Figure 8: Confusion matrix of the Ordinal Logistic Regression model on the CELVA.Sp dataset.

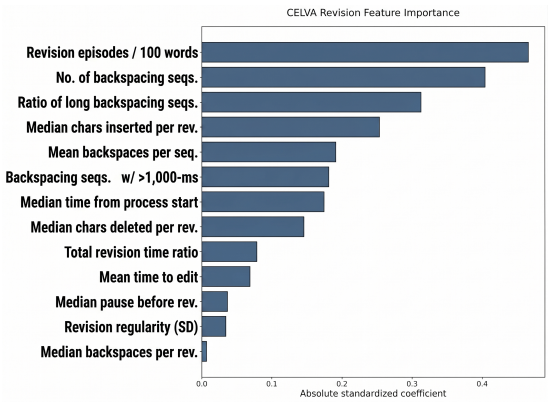


Figure 9: Top ten most influential K-REV features selected by the Ordinal Logistic Regression model for predicting CEFR proficiency levels in the CELVA.Sp dataset.

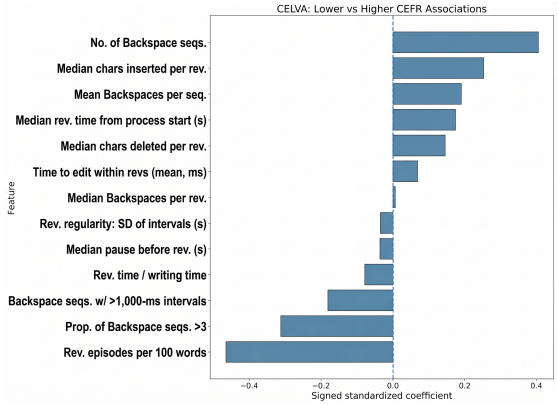


Figure 10: Signed regression coefficients of the top K-REV features in the Ordinal Logistic Regression model. Positive coefficients indicate an increase in the likelihood of higher CEFR levels, whereas negative coefficients indicate an association with lower proficiency levels.

revision episodes per 100 words decreased as the proficiency levels increased, while the number of backspacing sequences tended to increase. The other features showed smaller differences and substantial overlap across CEFR levels. Regarding the number of backspace sequences with intervals exceeding 1,000 ms, there was some overlap, but it was higher at higher CEFR levels which is inconsistent with our hypothesis. These results suggest that some revision behaviours vary with proficiency levels. However, this should also be interpreted with caution, due to the small size of this dataset and its imbalance.

4.3 Cross-corpus comparison

The ordinal logistic regression model performs better on the KUPA-KEYS dataset than on the CELVA.Sp dataset, with 59% accuracy versus 35%. This difference may be attributed to sample size variations. Hence, the better performance in KUPA-KEYS should be interpreted with caution because the two datasets differ in their sample size and CEFR distribution. The balanced accuracies in both datasets indicate that both models struggle with balanced predictions across all levels due to dataset imbalance. The adjacent accuracy using KUPA-KEYS is notably high at 98%, compared to 85% for CELVA.Sp. However, this implies that in both datasets, predictions were either correct or mainly off by only one level. Across both datasets, more backspacing sequences, more characters inserted and deleted per revision, more backspaces per sequence, later revision timing, and longer editing time were associated with higher CEFR levels. Also, more revision episodes per 100 words, a higher proportion of writing time spent revising, longer pauses before revision, and more backspacing sequences with intervals longer than one second were associated with lower CEFR levels. However, some features showed different patterns, such as the proportion of long backspacing sequences was associated with higher levels in KUPA-KEYS but it was associated with lower levels in CELVA.Sp. Median backspaces per revision showed the opposite pattern. Overall, several revision features showed similar patterns across the two datasets and captured some differences across CEFR levels, but they were limited in predicting the exact CEFR level.

4.4 Documentation and Interoperability

Applying K-REV to different keylogging datasets also highlighted the need for clear documentation and interoperable input formats. KUPA-KEYS and CELVA.sp were stored in different structures, requiring dataset-specific preprocessing before revision episodes could be extracted, as explained in the text reconstruction section. This could increase the effort required to apply the library to new keylogging datasets and make replication more difficult.

5 Discussion

Our findings suggest that higher-proficiency writers tend to revise less frequently relative to text length, but their edits are more extensive, with greater text insertion and deletion and more backspacing sequences. This may suggest that more proficient writers have stronger language skills and a wider vocabulary, allowing them to make larger changes while revising less often. In contrast, lower-proficiency writers tend to revise more frequently and spend more time revising. They also tend to pause and hesitate while making revisions. This may suggest that they could be having difficulty finding suitable replacements, leading to more frequent revisions and longer pauses during editing. Also, we found that keystroke-derived revision metrics are moderate indicators of CEFR levels, despite the limited performance of the models (Conijn et al., 2022a). However, our findings align with previous research showing that writing process features are associated with proficiency and writing quality. Barkaoui, 2016, for example,

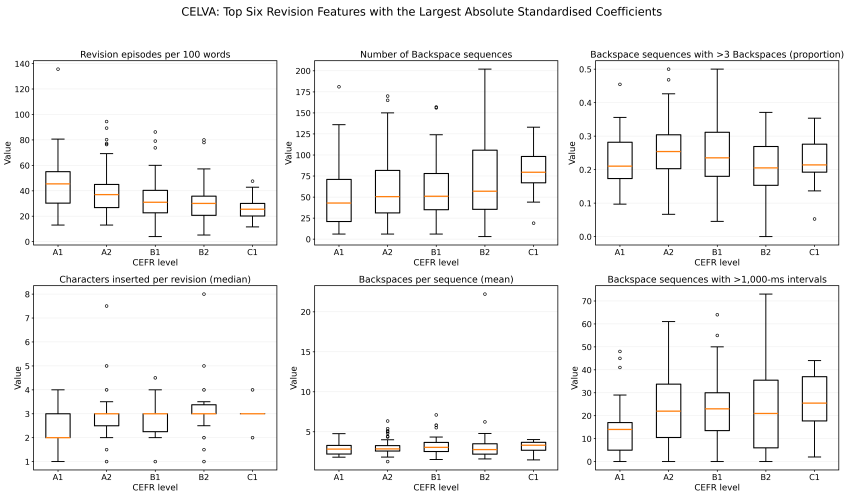


Figure 11: Distributions of the six revision-process features with the largest absolute standardised coefficients across CEFR levels in the CELVA.Sp dataset.

found that proficiency was related to learners' revision behaviour. More generally, previous studies have shown that process features extracted using keystroke data can provide information related to writing proficiency (Sinharay et al., 2019; Al Sawar et al., 2025). This study extends this by using automatically extracted revision episodes to predict CEFR levels across two datasets. However, the limited performance in predicting the exact CEFR levels and the differences between the two datasets we used suggest that revision behaviour should not be treated as a direct indicator of CEFR levels.

The cross-corpus evaluation provides valuable insights into the generalisability of our approach. By testing on two distinct datasets (KUPA-KEYS, a diverse crowdsourced corpus, and CELVA.sp, collected in an ecological and homogeneous setting), we identified several consistent trends. However, the difference in the resulting exact accuracies underlines a limitation in model transferability. This discrepancy, likely due to the variations in sample size and class distribution, suggests that, while the correlations between revision behaviours and proficiency seem to generalise well, the predictive thresholds remain highly context-dependent and require further validation on larger learner corpora.

From an educational perspective, these prediction models could significantly complement traditional product-oriented evaluations by introducing process-oriented insights. However, deploying such predictions for high-stakes summative assessment is premature. Because our models' performance is still moderate, and the causal mechanisms linking specific predictive features to distinct CEFR levels require more rigorous exploration, using them for grading could result in inaccurate evaluations. Instead, their true utility lies in formative assessment, where they can render some of the often-invisible cognitive mechanics of writing visible. By providing granular feedback on the writing process, these models can help teachers identify learners who struggle with fluency or over-rely on surface-level, local revisions rather than global planning. For students, visualising their own writing rhythms can foster metacognitive awareness and encourage better self-regulation (Vandermeulen et al., 2020). Furthermore, it would be highly valuable for future work to explore the early prediction of CEFR levels using real-time keystroke logs. Post-hoc predictions delivered after a task is completed are fundamentally less actionable for learners. Developing early detection models would allow Intelligent Tutoring Systems to provide timely feedback (Shute, 2008) while students are still actively engaged in the writing process.

6 Conclusion

In this paper, we have attempted to predict CEFR level of learners on the basis of revision types as systematically assigned by the K-REV library. Although our results are promising, another obvious candidate for CEFR level prediction on the basis of keylogging data is large language models; exploratory analysis is currently carried out (Gan and Adams, 2026).

Limitations

An obvious limitation of this study is the number of corpora we used to evaluate the benefits of the K-REV library. While we are currently testing a sub-component of the PROCEED corpus, we can only re-assert the (current) scarcity of CEFR-graded keylogging data for learners. We also acknowledge that C2 has a small sample size (n = 19) in the KUPA-KEYS dataset, which limits the reliability of predictions for this level. However, we retained C2 in the main analysis to examine revision-based prediction across the full B1–C2 proficiency range. To address this limitation, we also conducted a sensitivity analysis excluding C2 (N = 970). Although accuracy was slightly higher without C2 (0.612 vs. 0.599), the results are not directly comparable because the classification task changes from four levels to three. Future studies with larger C2 samples are needed to determine whether revision features can reliably distinguish advanced proficiency levels.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: ChatGPT was used for language editing, including grammar and syntax correction, spell-checking, and improving the clarity and readability of the manuscript. Additionally, ChatGPT provided partial assistance with code development and refinement. All AI-assisted code was reviewed and validated by the authors, who retain full responsibility for the implementation, scientific content, and results.

References

Ahood Al Sawar, Thomas Gaillat, and Nicolas Ballier. 2026. Predicting CEFR writing levels from writing process and linguistic features: a cross-corpus comparison. In *Actes de l'atelier Intelligence Artificielle et ÉDUcation (IA EDU)*, pages 107–125, Nantes, France. ATALA.

Ahood Al Sawar, Erin Pacquetet, Cyriel Mallart, Andrew Simpkin, and Nicolas Ballier. 2025. Predicting CEFR levels for learners of English with keylogging metrics, an exploratory study. In *Actes de l'atelier Traitement de données langagières dynamiques par les outils et méthodes du TAL 2025 (DYN-TAL)*, pages 7–11, Marseille, France. ATALA & ARIA.

Taylor Arnold, Nicolas Ballier, Thomas Gaillat, and Paula Lissón. 2018. Predicting CEFR levels in learner English on the basis of metrics and full texts. *arXiv preprint arXiv:1806.11099*.

Khaled Barkaoui. 2016. What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, 100(1):320–340.

Khaled Barkaoui. 2019. What can L2 writers' pausing behavior tell us about their L2 writing processes? *Studies in Second Language Acquisition*, 41(3):529–554.

Rianne Conijn, Christine Cook, Menno Van Zaanen, and Luuk Van Waes. 2022a. Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 32(4):835–866.

Rianne Conijn, Emily Dux Speltz, and Evgeny Chukharev-Hudilainen. 2024. Automated extraction of revision events from keystroke data. *Reading and Writing*, 37(2):483–508.

Rianne Conijn, Emily Dux Speltz, Menno van Zaanen, Luuk Van Waes, and Evgeny Chukharev-Hudilainen. 2022b. A product- and process-oriented tagset for revisions in writing. *Written Communication*, 39(1):97–128.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College Composition & Communication*, 32(4):400–414.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146.

Qiao Gan and Benjamin Adams. 2026. From keystrokes to scores: Toward a multidimensional predictive model of writing evaluation by humans and large language models across linguistic, cognitive, and social dimensions. *Assessing Writing*, 69:101090.

Aitor Garcés-Manzanera. 2024. Language bursts and text quality in digital writing by young EFL learners. *Journal of New Approaches in Educational Research*, 13(1):12.

Gaëtanelle Gilquin. 2020. In search of constructions in writing process data. *Belgian Journal of Linguistics*, 34:99–109.

Gaëtanelle Gilquin. 2022. The process corpus of English in education: Going beyond the written text. *Research in Corpus Linguistics*, 10(1).

Gaëtanelle Gilquin. 2023. Causative constructions in process: How do they come into existence in learner writing? *Yearbook of the German Cognitive Linguistics Association*, 11(1):105–120.

Gaëtanelle Gilquin. 2024. Keylogging and screencasting to help investigate L2 writing processes. pages 285–296. Routledge.

Alex Housen, Folkert Kuiken, and Ineke Vedder. 2012. Complexity, accuracy and fluency. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 32:1–20.

Kristopher Kyle. 2016. *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Ph.D. thesis, Georgia State University.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior research methods*, 50(3):1030–1046.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30(3):358–392.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2017. Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, 34(4):493–511.

Cyrielle Mallart, Nicolas Ballier, Jia-Yu Li, Andrew Simpkin, Bernardo Stearns, Rémi Venant, and Thomas Gaillat. 2023. A New Learner Language Data Set for the Study of English for Specific Purposes at University. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 281–287, Portugal. NOVA CLUNL.

Léo Nebel. 2026. *Automation of Feedback on the Writing Process: Construction and Evaluation of Process-Oriented Feedback Based on the Granular Characterization of Revision Episodes*. Ph.D. thesis, Sorbonne Université.

Léo Nebel, François Bouchet, and Vanda Luengo. 2025a. Characterizing revision events in students' writing processes using llms. *Workshop - Writing And Literacy Instruction for Educational Data Mining*. Hal-05186500.

Léo Nebel, François Bouchet, Vanda Luengo, and Mathilde Couraud. 2025b. Towards automated characterization of revision events in student writing. In *Proceedings of the European Conference on Technology Enhanced Learning (ECTEL)*, pages 397–411, Cham. Springer Nature Switzerland.

Erin Pacquetet. 2024. *The Effect of Linguistic Properties on Typing Behaviors and Production Processes*. Ph.d. dissertation, University at Buffalo, State University of New York.

Fabian Pedregosa-Izquierdo. 2015. *Feature extraction and supervised learning on fMRI : from practice to theory*. Theses, Université Pierre et Marie Curie - Paris VI.

Mahmudur Rahman, Pranto Goswamee, Maria Sultana, and Kamruddin Nur. 2025. Typing under pressure: Estimating cognitive load from keystroke dynamics and textual patterns. In *2025 28th International Conference on Computer and Information Technology (ICCIT)*, pages 2521–2526. IEEE.

Soumen Roy, Jitesh Pradhan, Abhinav Kumar, Dibya Ranjan Das Adhikary, Utpal Roy, Devadatta Sinha, and Rajat Kumar Pal. 2022. A systematic literature review on latest keystroke dynamics based models. *IEEE Access*, 10:92192–92236.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Sandip Sinharay, Mo Zhang, and Paul Deane. 2019. Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education*, 32(2):116–137.

Nancy Sommers. 1980. Revision strategies of student writers and experienced adult writers. *College Composition and Communication*, 31(4):378–388.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 147–153.

Luuk Van Waes and Mariëlle Leijten. 2015. Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition*, 38:79–95.

Nina Vandermeulen, Mariëlle Leijten, and Luuk Van Waes. 2020. Reporting writing process feedback in the classroom: Using keystroke logging data to reflect on writing processes. *Journal of Writing Research*, 12(1):109–140.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

Chunyan Xu. 2018. Understanding online revisions in L2 writing: A computer keystroke-log perspective. *System*, 78:104–114.

Universal, but Grounded? Anchoring Cross-Lingual CEFR Assessment in the Framework's Own Descriptors

David Alfter
Gothenburg Research Infrastructure in Digital Humanities
Department of Literature, History of Ideas, and Religion
University of Gothenburg, Sweden
david.alfter@gu.se

Abstract

We explore whether automatic CEFR assessment of learner writing can be grounded in the CEFR itself rather than in its labels. Recent work towards a universal, cross-lingual CEFR classifier treats the levels as a shared label set and leaves out the descriptors that make it language-independent. We assess 11,194 learner texts in Swedish, French and English against 56 CEFR writing descriptors, with a fulfillment score and evidence statement per descriptor from a locally deployed open language model, and derive levels from the resulting profiles. The profiles track the human reference when it rates the text but barely when it tests the writer; the descriptors are heterogeneous: 14 are retained in all three languages whenever descriptors are selected to fit the labels, whereas task-bound descriptors are rarely assessable and fit the scale poorly. A shared 30-descriptor subset doubles exact agreement in French, and rules on selected subsets transfer across languages at least as well as models trained on the descriptor scores, though neither beats fine-tuned classifiers. The aim is grounding rather than prediction: every level decomposes into descriptor-level judgments, and corpora not labeled on the CEFR can be placed, relative to one strict rater, on the same scale as CEFR-labeled corpora.

1 Introduction

The Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001, 2020) describes what learners can do in a language at six levels, A1 to C2, by means of illustrative descriptors, i.e., "can do" statements calibrated on a common scale from teacher judgments with the help of the Rasch model (North and Schneider, 1998; North, 2000). The framework is explicitly language-independent, which is arguably the source of its appeal: a B1 writer of Swedish and a B1 writer of French are meant to be able to do the same things. Consequently, the CEFR levels have become the label set of choice for automatic proficiency assessment in many languages (e.g., Vajjala and Lõo, 2014; Volodina et al., 2016a; Tack et al., 2017; Wilkens et al., 2023; Yancey et al., 2023).

Vajjala and Rama (2018) took the language independence of the framework seriously and asked whether one classifier could serve several languages, an idea that has since been reproduced and extended (Caines and Buttery, 2020; He and Li, 2024) and has recently culminated in Universal-CEFR (Imperial et al., 2025), a standardized collection of 505,807 CEFR-labeled texts in 13 languages (Section 2.1). The field is converging on cross-lingual CEFR assessment with one model and one label set.

However, in all of this work the CEFR enters the modeling only as a set of labels. The universal element is either the feature representation or the multilingual encoder; the descriptors, i.e., the part of the CEFR that is language-independent, are not part of the model. Even the "descriptor-based prompting" of Imperial et al. (2025) supplies one overall descriptor per level as context and asks for a single level label "without explanation or justification", and studies that rate essays with generative language models (Yancey et al., 2023; Mizumoto and Eguchi, 2023; Ahlers et al., 2025) likewise ask for the level directly. A level predicted in this way is universal in name only: nothing in the prediction can be traced back to a statement in the framework, and a B1 in one language and a B1 in another are comparable only insofar as the training labels were. This is not only a methodological concern: the assessment of learning outcomes is a high-risk use of AI under the EU AI Act (European Parliament and Council, 2024), yet the best current essay scorers are fine-tuned transformers that deliver a level without a ratio-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 78–98, November 19–20, 2026, University of Tartu Library

nale (e.g., Wilkens et al., 2023).

In this paper, we turn the descriptors into the unit of assessment. We assess every text in three learner corpora with a text-level reference – 10,813 texts in Swedish, French and English – against 56 writing descriptors of the CEFR, and a fourth corpus with a writer-level reference as a control. An open instruction-tuned language model acts as the rater and returns, for each text and descriptor, a fulfillment score on a five-point scale and a short evidence statement; levels are then derived from the resulting descriptor profile instead of being predicted directly. The hypothesis we explore is that such an assessment yields level estimates that are comparable across languages without language-specific training and that can be audited descriptor by descriptor. We break it down into three research questions: (1) Do the descriptor profiles track human level judgments at all? (2) Which of the 56 descriptors can be operationalized on a single text, and which cannot? (3) Is the resulting descriptor scale comparable across languages, and is the label scale? We do not aim at levels that beat classifiers fine-tuned on each corpus, and do not obtain them; what we obtain is a level that decomposes into descriptor-level judgments and, as a by-product, a common scale for corpora with and without CEFR labels. Our contributions are (1) a descriptor-level CEFR assessment method that grounds cross-lingual assessment in the framework's own descriptors rather than in its labels; (2) an audit of the 56 descriptors on 11,194 texts in three languages, which identifies a core retained by every selection and a set of task-bound descriptors whose fulfillment depends on the prompt; and (3) a Rasch analysis showing that the descriptor hierarchy is stable across corpora while texts with the same label are not placed at the same ability, which yields a joint scale on which CEFR-labeled and non-CEFR-labeled corpora can be compared.

The rest of the paper is structured as follows: Section 2 contextualizes our work and Section 3 describes the descriptors, corpora, rater and the rule that turns a profile into a level. Sections 4–6 answer the three research questions in turn, each with its own methods and results. Section 7 discusses what the answers mean for a universal CEFR, and Sections 8 and 9 round off the paper with limitations, future work and conclusion.

2 Related Work

2.1 Towards a universal CEFR classifier

Vajjala and Rama (2018) were, to the best of our knowledge, the first to frame CEFR classification as a universal problem. Using the MERLIN corpus (German, Italian and Czech) and features that generalize across languages, such as part-of-speech and dependency n-grams over Universal Dependencies, they found that a single multilingual model performed comparably to monolingual ones and that a model trained on German lost less than ten points of weighted F1 when tested on Italian and Czech. Caines and Buttery (2020) reproduced these experiments, added English and Spanish corpora and concluded that feature-based approaches outperform neural classifiers at this data size; He and Li (2024) recast the problem as zero-shot cross-lingual essay scoring with multilingual encoders, and Naous et al. (2024) benchmark multilingual models for readability on a CEFR-inspired scale.

UniversalCEFR (Imperial et al., 2025) is the most ambitious effort in this direction: 505,807 CEFR-labeled texts in 13 languages, standardized into one format and released with permissive licenses, together with a benchmark under three paradigms, of which prompting instruction-tuned models performed worst. The prompting paradigm is the only one in which descriptors appear at all: the prompts contain the six overall reading or written production descriptors, one per level, and the model is instructed to output the level only. Imperial et al. (2025) themselves note that nearly half of the pooled corpora lack information on annotators and agreement, and this is where a label-only universal CEFR is fragile: if a B2 label does not mean the same thing in two corpora, a model trained on their union learns a compromise between the two meanings.

2.2 Generative language models as CEFR raters

Yancey et al. (2023) rate short L2 English essays on the CEFR scale with GPT-3.5 and GPT-4 and find that GPT-4 approaches modern automated writing evaluation only when calibration examples are provided; Ahlers et al. (2025) classify German learner texts by prompting, fine-tuning and probing an 8B-parameter model, with prompts that name the CEFR but not its descriptors; and direct prompting for a level or score is by now a

standard baseline (e.g., Stahl et al., 2024; Mansour et al., 2024). Lee et al. (2024) come closest to a decomposed assessment, scoring several traits in turn and averaging them, but the model generates the traits itself instead of taking them from a framework. Benedetto et al. (2025) find the performance of five language models on a suite of CEFR-related tasks unsatisfactory for educational use, and Alfter (2024) finds moderate but stable vocabulary-grading ability across five model families and languages. Common to all these studies is that the output is the level or grade; what the model consults in order to arrive at it remains implicit.

2.3 The descriptors as an instrument

The illustrative descriptors were not written to be applied one at a time to a single text. North and Schneider (1998) calibrated them from teachers' judgments of learners they knew, using the Rasch model to place descriptors on a common difficulty scale; the level boundaries of the CEFR are cutoffs on that scale. Critics have questioned the empirical footing of the result: Fulcher (2004) warns against reading the scale as a validated model of acquisition, Alderson (2007) calls for research on how the descriptors relate to actual performance, and Hulstijn (2007) notes that they conflate quantitative and qualitative dimensions of proficiency (see also Harsch, 2014). Attempts to rate with descriptors have consequently had to adapt them, through task design (Harsch and Rupp, 2011), rater training and revision (Harsch and Martin, 2012) or co-construction of a rating scale (Deygers and Van Gorp, 2015). The MERLIN corpus (Boyd et al., 2014) provides human ratings on several CEFR-based dimensions, i.e., an analytic profile in the spirit of what we attempt here, but at the level of dimensions rather than individual descriptors; in diagnostic assessment (Alderson, 2005), such a profile is the goal.

2.4 Language-specific profiles

The imprecision of the illustrative descriptors has long been recognized, and the Council of Europe's answer was to specify them per language: Reference Level Descriptions translate the language-neutral levels into inventories of the words, structures and functions that realize them in a given language (Council of Europe, 2005), e.g., *Profile Deutsch* (Glaboniat et al., 2005) or the *Niveau ... pour le français* series (Beacco et al., 2004). The most developed is *English Profile*, with criterial features of each level derived from learner data (Hawkins and Filipović, 2012) and published as the English Grammar Profile (O'Keeffe and Mark, 2017) and the English Vocabulary Profile (Capel, 2012); for Swedish, there is the L2 profiles project (Volodina et al., 2024), building on resources such as SVALex (François et al., 2016). Such profiles are, in effect, more precise descriptors, and Bannò et al. (2024; 2025; 2026) have begun to operationalize them with language models, from analytic scores over the English Vocabulary Profile to the detection of attempted grammatical constructs, with CEFR profiling as the stated goal. Profiles are the natural complement to what we do; however, they are language-specific by construction, and most languages have none; the illustrative descriptors are the one layer of the CEFR that all languages share, so we start from that layer.

A further strand concerns transparency. Madnani et al. (2017) argue that automated scoring in high-stakes settings needs open tools for inspecting fairness and validity, and Kumar and Boulanger (2020) that a deep essay scorer has pedagogical value only insofar as its decisions can be explained. Analytic rating against descriptors is the traditional answer to this demand in human assessment (Weigle, 2002), and trait-based automatic scoring exists (Mathias and Bhattacharyya, 2018; Bannò et al., 2024), but the traits are rubric dimensions rather than the framework's own descriptors.

To the best of our knowledge, no prior work has operationalized the illustrative descriptors themselves as the unit of automatic assessment, nor used the resulting profiles to examine how the descriptors behave when applied to single texts in several languages. We use the language model as a rater of descriptors, and treat the descriptors as items whose measurement properties can be inspected.

3 Experimental Setup

3.1 Descriptor set

We use 56 descriptors for writing, listed in full in Appendix B, distributed over nine levels: A1 (6), A2 (8), A2+ (3), B1 (8), B1+ (5), B2 (6), B2+ (4), C1 (9) and C2 (7). For each of the six full levels, the first descriptor is the general writing descriptor of the CEFR self-assessment grid (Council of Europe, 2001), which aggregates several can-do

statements into one. The remaining descriptors at each level, and all descriptors at the plus levels, are the illustrative descriptors of the Companion Volume's scales for overall written production, creative writing, and reports and essays (Council of Europe, 2020); the grid entries were added because the Companion Volume has no level-general writing descriptor. Henceforth, *general* refers to the six grid descriptors and *writing-specific* to the other fifty.

Several of these descriptors name a task that an essay prompt may not have set (a postcard, a form, a room, a film review, a diary); we call them *task-bound*. We nevertheless assess every text against all 56, for two reasons. First, whether the model recognizes such a descriptor as not assessable, or instead scores it as not demonstrated, is itself a test of the method, and the prompt gives it the former option. Second, which descriptors can be judged on a single text is not always obvious from their wording: the general descriptors, for instance, turn out to behave differently at A1 and from B1 upward (Section 5.1). A hand-selected subset is the natural next step (Section 8).

Text (EN-H essay JEP6905, holistic score 4.0; opening, verbatim): *Imagine if you could prove other people that you are a good problem solver. Imagine that a problem is a good chance to do your best because youre a good problem solver, and you can learn something from that problem. . . .*

Descriptor (level, number)	**Status / score**
A1 (3) simple isolated phrases and sentences	assessable / 4
The essay contains many simple, clear sentences and phrases.	
B1 (1) simple connected text on familiar topics [general]	assessable / 2
Shows some connected text but with frequent repetition and limited development.	
B2 (4) review of a film, book or play	not assessable
Essay does not contain a film, book, or play review.	
C1 (2) clear, well-structured texts of complex subjects	assessable / 2
Ideas are somewhat organized but lack clear structure and depth.	
. . . (56 descriptors in all: 50 assessable, 34 of them scored 2; 6 not assessable)	

Level means (assessable descriptors only): A1 2.00, A2 2.43, A2+ 1.67, B1 1.71, B1+ 1.75, B2 1.80, B2+ 1.75, C1 1.88, C2 1.57
Implied level: highest level with a mean of at least 2.0 = A2

Figure 1: One text end to end. The rater returns a status, a score and an evidence statement for each of the 56 descriptors (four shown; numbers refer to Appendix B); level means and implied level follow from the scores (Section 3.4). The example also shows the rater's strictness: an essay scored 4.0 out of 5 receives a 2 on most descriptors and an implied A2. Under the shared 30-descriptor subset of Section 5.2, the same scores give C1, because every retained C1 descriptor scores exactly 2.

3.2 Corpora

Table 1 gives an overview of the corpora. SV is the SweLL-pilot corpus (Volodina et al., 2016b), 333 unique essays by adult learners of Swedish rated by teachers, essentially the essays used by Volodina et al. (2016a) and Pilán et al. (2016). FR is TCFLE-8 (Wilkens et al., 2023), 6,569 essays written in the *Test de connaissance du français* and scored by at least two certified raters. EN-H is the released training portion of ELLIPSE (Crossley et al., 2023), 3,911 essays by English language learners in grades 8–12 in state-wide standardized testing in the United States, scored by two trained raters. All three references are judgments of the text at hand; the EN-H reference is a holistic score rather than a CEFR level, which is why EN-H enters some analyses through correlations only. This was a deliberate choice: a CEFR-labeled English corpus exists (Nicholls et al., 2024), but one aim of the paper is to show that corpora rated on other scales can be placed on the descriptor scale next to CEFR-labeled ones (Section 6.2), and that requires a corpus whose reference is not a CEFR level. A fourth corpus, the learner part of NICER (Sugiura, 2020) (EN-T; 381 timed essays by Japanese university students), was assessed in the same way, but its reference is the writer's score on an external proficiency test, not a rating of the text. It serves as a control for what the descriptors can and cannot measure and is reported in Appendix C; in the main text it appears only as a fourth corpus in the Rasch analyses (Sections 5.3 and 6.2).

3.3 Assessment setup

Every *(text, descriptor)* pair was assessed by Qwen3-32B (Qwen Team, 2025) in its FP8 variant, deployed locally through the continuous-batching backend of Hugging Face Transformers (version 5.9.0); no learner data was sent to a commercial API. Each request consists of a fixed system message and a user message that carries the declared language of the corpus, the text between collision-free markers and one descriptor with its nominal CEFR level (Appendix A). The system

Corpus	Language	Texts	Reference measure	Nature of the reference
SV SweLL-pilot	Swedish	333	One CEFR level per text: A1 16, A2 80, B1 75, B2 74, C1 88	A rating of *this text*
FR TCFLE-8	French	6,569	One CEFR level per text: A1 689, A2 1,375, B1 1,466, B2 1,427, C1 1,127, C2 485	A rating of *this text*
EN-H ELLIPSE	English	3,911	One holistic score per essay, 1.0–5.0 in half-point steps (median 3.0; bin sizes in Table 10)	A rating of *this text*, but a score rather than a level
EN-T NICER	English	381	The writer's scores on external tests (TOEFL ITP, TOEIC, STEP/Eiken, ETS Criterion); see Appendix C	A measure of the *writer*, not of the text

Table 1: The corpora: abbreviation used in this paper, number of texts assessed, reference measure and its nature. The three corpora above the line carry the main analyses; EN-T serves as a control and is reported in Appendix C.

message instructs the model to be a strict rater,[1] to assess only the evidence visible in this text for this one descriptor and explicitly not the writer's general ability, and not to penalize performance above the descriptor's threshold. The model returns one JSON object with (1) a status, one of *assessable*, *not assessable*, *empty* and *language mismatch* (with the declared language), (2) for assessable cells a score on a five-point scale (0 = not demonstrated, 1 = weakly, 2 = partly, 3 = mostly, 4 = fully demonstrated) and (3) an *evidence statement* in English of at most 16 words naming one observable strength or gap. Decoding was greedy with reasoning mode disabled and at most 80 output tokens; reasoning mode was disabled for reasons of economy, since the run took 64 hours as it was and reasoning would have multiplied the number of generated tokens per cell (Section 8). In total, 627,032 assessments were made over 11,197 rows (the 11,194 texts of all four corpora plus three Swedish duplicates), all under one cached configuration of model, prompt and decoding. Figure 1 walks through one text; the strictness it shows is typical (Section 4). Nothing in the setup is language-specific apart from the declared language and the text itself, and the prompt carries no information about the text's reference label or score.

3.4 From descriptor profile to level

For a text and a CEFR level, the *level mean* is the mean fulfillment score over the assessable descriptors of that level, and the *overall mean* is the mean over all assessable descriptors of the text. We call any mapping from a descriptor profile to a level a *decoder*. The untrained decoder, henceforth *the rule*, sets the *implied level* of a text to the highest of the nine levels at which the level mean is at least 2.0, regardless of lower levels; if no level reaches 2.0, the implied level is "below A1". The *implied index* numbers these ten outcomes from 0 (below A1) to 9 (C2). The threshold of 2.0 is the natural reading of "partly demonstrated"; other cut-offs are examined in Section 4.2. When the implied level is compared with a five- or six-level reference, the plus levels are folded into their base level (A2+ $\rightarrow$ A2, B1+ $\rightarrow$ B1, B2+ $\rightarrow$ B2). Agreement is reported as exact agreement, agreement within one level and bias (mean of implied minus reference, in levels); quadratic weighted kappa (QWK) and Spearman's ρ are added where decoders are compared (Section 5.4).

4 RQ1: Do Descriptor Profiles Track Human Level Judgments?

The first question is whether the 56 scores a text receives carry the information in the human reference at all. We look at the level means first (Section 4.1) and then at the implied level, i.e., the level the untrained rule derives from them (Section 4.2). Group differences in the overall mean are tested with Kruskal–Wallis and trends with Spearman's ρ; p-values are uncorrected and read as descriptive.

Before turning to the reference, two properties of the raw scores should be kept in mind. Across all assessable cells, the score 2 ("partly demonstrated") dominates in every corpus: 55% of the cells in SV, 50% in FR and 62% in EN-H; we call this the rater's *central tendency* and return to it in Section 5.3. The share of not-assessable cells is 9.8% (SV), 12.4% (FR, plus 1.5% language mismatch, mostly in A1–A2 texts) and 7.9% (EN-H); non-assessability is concentrated in a small set of

[1] We instructed the LLM to act as a strict rater to mitigate potential leniency bias, which has been documented in LLM-based evaluation (Thakur et al., 2025) and particularly in essay scoring (Yavuz et al., 2025; Seßler et al., 2025).

descriptors that name a task the essay prompt did not set, and is near zero for the general descriptors and for those of the form "can produce … text".

4.1 Level means

Figure 2 shows the level means for the three corpora; Tables 8–10 in Appendix D give the underlying numbers. Three things can be seen from the figure. First, mean fulfillment falls monotonically over the six full levels for every group of texts; the model thus orders the descriptors by level, although the prompt names each descriptor's nominal level and the ordering may partly reflect that cue (Section 8). Second, the groups are ordered at almost every descriptor level in all three corpora. Third, the curves cross the "partly" line at different places in different corpora; where they cross determines the implied level (Section 4.2).

SV. The reference levels separate cleanly: Kruskal–Wallis on the overall mean gives $H = 224.9$ ($p < .001$), the level means correlate with the reference level at $\rho = 0.45$ (A1 descriptors) to 0.87 (B2 descriptors), and nearly every individual descriptor at $\rho \approx 0.6$–0.8; the weakest are task-bound ones (room description, diary entries: $\rho \approx -0.27$).

FR. The signal is equally strong and more uniform: $H = 4077$ ($p \approx 0$), level-mean correlations between 0.66 and 0.75, and almost every individual descriptor at $\rho \approx 0.4$–0.8 (again with room description, 0.11, and humor, 0.17, as the exceptions). However, the entire family of curves sits lower than in Swedish: on the descriptors of their own reference level (plus levels folded), French texts average 1.39 (A1), 1.55 (A2), 1.65 (B1), 1.65 (B2), 1.63 (C1) and 1.62 (C2), against 1.71 (A1), 1.86 (A2), 1.90 (B1), 1.97 (B2) and 1.83 (C1) for Swedish (SV has no C2 texts), i.e., 0.2–0.3 points less from A2 upward, and C1 and C2 texts differ by less than 0.1 on the B2 to C2 descriptors.

EN-H. The profiles track the holistic score ($H = 1447$, $p < .001$; ρ between 0.43 and 0.59), somewhat less closely than in the CEFR-labeled corpora, and the signal is strongest on the low-level descriptors: between the lowest and the highest bin, mean fulfillment rises by 1.43 on A1 descriptors but only 0.57 on B2 descriptors. High-level descriptors separate the weakest essays from the rest; they separate a 4.0 from a 4.5 by only about 0.1.

4.2 Implied levels and calibration

Under the rule, the implied level matches the SV reference exactly for 40.2% of the texts (majority class: C1, 26.4%) and within one level for 93.7%, with a bias of -0.24 levels and $\rho = 0.82$ between implied index and reference. The errors are systematic (confusion matrices in Appendix E): 73 of the 88 C1 texts are implied B2, because C1 texts average only 1.83 on C1 descriptors, and 12 of the 16 A1 texts fall below A1, because A1 texts average only 1.71 on A1 descriptors. More generally, texts score about 2 on the descriptors of their own level, so "at level" means partial fulfillment.

For FR, the same rule matches the reference exactly for only 18.8% of the texts, i.e., less than the majority class (B1, 22.3%), and within one level for 58.7%, with a bias of -1.19 levels ($\rho = 0.68$). The floor and the ceiling account for most of the disagreement: 612 of the 689 A1 texts (89%) and 720 of the 1,375 A2 texts (52%) reach no level at all, and only 26 of the 1,612 C1 and C2 texts are implied C1 or above, most of them landing at B2. Table 2 shows the dependence on the cut-off: for SV, 2.0 happens to be the best; for FR, 1.75 is best (26% exact, 71% within one level), and even then the agreement remains well below the Swedish figures, since "at level" corresponds to a level mean of roughly 1.7–2.0 in Swedish and 1.4–1.65 in French.

	SV			FR		
Cut-off	Exact	±1	Bias	Exact	±1	Bias
1.00	1%	36%	+1.94	13%	44%	+1.65
1.25	5%	50%	+1.61	21%	61%	+0.98
1.50	8%	57%	+1.34	22%	64%	+0.58
1.75	37%	82%	+0.51	**26%**	**71%**	−0.49
2.00	**40%**	**94%**	−0.24	19%	59%	−1.19
2.25	10%	49%	−1.52	1%	19%	−2.38
2.50	2%	32%	−1.93	0%	15%	−2.62

Table 2: Dependence of the implied level on the cut-off on the level mean: exact agreement, agreement within one level and bias (implied minus reference, in levels) for SV and FR. The best cut-off per corpus is in bold.

For EN-H, the mean implied index rises with the score from 2.25 to 6.66 across the six bins ($\rho = 0.43$), but the per-essay distribution is bimodal: an essay either stays around A1–A2 or clears B2+,[2]

[2] In EN-H (and EN-T), B2+ descriptors score slightly above B2 descriptors, and the rule takes the highest level whose mean reaches 2, so the unfolded index lands disproportionately on B2+. This is a joint property of the rule and of the B2+ items; it is inactive once B2+ is folded into B2, as

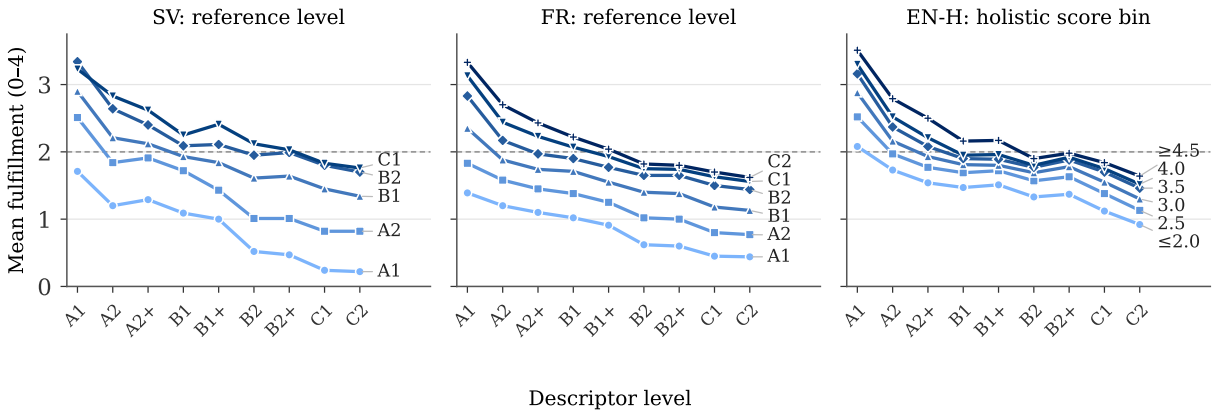


Figure 2: Level means (mean fulfillment over the descriptors of each CEFR level, x-axis) for texts grouped by reference level (SV, FR) or by holistic score bin (EN-H; the outer bins pool 1.0–2.0 and 4.5–5.0). The dashed line marks the 2.0 threshold of the rule.

and none reaches an implied C1.

In sum, the profiles track text-level ratings in all three corpora, and the ordering of the texts is largely right ($\rho = 0.82$ in SV, 0.68 in FR, 0.43 in EN-H); the untrained rule, however, is miscalibrated, slightly in Swedish and by about one level in French, and no cut-off repairs the French figures. The control corpus confirms what the profiles measure: where the reference is a test score of the writer rather than a rating of the text, the signal all but disappears (Appendix C). The next section asks which descriptors are responsible for the miscalibration.

5 RQ2: Which Descriptors Can Be Operationalized on a Single Text?

Section 4 showed that the profiles track the reference but that the untrained rule is miscalibrated. This section asks which descriptors are responsible, in three steps: two regularities that are visible in the raw scores (Section 5.1), a data-driven selection of descriptors (Section 5.2) and a Rasch item analysis that places the descriptors on one difficulty scale and tests their fit (Section 5.3). Section 5.4 then checks, with supervised decoders, how much level information the profile holds beyond what the rule extracts.

5.1 Two kinds of descriptors

The general descriptors follow a different logic. In every corpus, the general descriptor is scored *below* the writing-specific descriptors of its level at A1 (e.g., SV 2.34 vs. 3.06; FR 2.04 vs. 2.54) and *above* them from B1 upward (e.g., at C1: SV 1.71 vs. 1.41; FR 1.60 vs. 1.18), with a much flatter slope across levels. A plausible explanation is that a general descriptor, which bundles several can-do statements, is satisfied as soon as some of them are, which makes it lenient at high levels, whereas the A1 general descriptor names tasks (a postcard, a form) that an essay does not perform. Either way, averaging general and specific descriptors into one level mean dilutes the level signal.

Task-bound descriptors. A fixed set of descriptors behaves differently from all others in every corpus: describing a room, reviewing a film, book or play, diary entries, poems and biographies, idiom and humor, and the critical review of cultural events. The model flags them as not assessable far more often than any other descriptor and, when it does assess them, scores them near zero, i.e., it uses both options the prompt gives it. Whether they are fulfilled depends on the task that was set, and the writer's proficiency plays little part; Section 5.3 shows that the A1 room descriptor is one of the hardest items on the whole scale.

5.2 Descriptor selection

We next ask whether a subset of the 56 descriptors makes the implied level agree better with the reference, with the rule kept as is and the level means computed over the selected descriptors only. The objective for the CEFR-labeled corpora is exact agreement plus $0.1\times$ within-one-level agreement; for EN-H it is Spearman's ρ between the unfolded implied index and the 1–5 score. The search is a greedy hill climb from the full set that toggles, at each step, the single descriptor that most improves the objective, with every level keeping at least one

it is for SV and FR.

Corpus	Set	n	Exact	±1	Bias	ρ	CV
SV	all 56	56	40.2	93.7	-0.24	–	40.3
SV	own	42	**54.1**	**99.4**	0.00	–	50.5
SV	shared	30	51.4	89.2	$+0.50$	–	–
FR	all 56	56	18.8	58.7	-1.19	–	18.8
FR	own	29	**40.1**	**87.0**	$+0.04$	–	39.4
FR	shared	30	39.1	85.3	-0.14	–	–
EN-H	all 56	56	–	–	–	0.43	0.43
EN-H	own	41	–	–	–	**0.54**	0.53
EN-H	shared	30	–	–	–	0.52	–

Table 3: Agreement of the implied level with the reference under all 56 descriptors, the subset selected on the corpus itself (own) and the subset selected on all three corpora (shared): number of descriptors (n), exact and within-one-level agreement (%), bias (levels), ρ with the EN-H score, and the cross-validated estimate (CV; exact agreement for SV and FR, ρ for EN-H). Best per corpus in bold.

descriptor; a *shared* subset maximizes the mean of the three objectives. Held-out estimates come from five-fold cross-validation (selection on four fifths of the texts, scoring on the held-out fifth); the *stability* of a descriptor is the fraction of folds in which it was selected. Because the shared subset is selected on all three corpora at once, its agreement on each of them is an in-sample figure; the fair cross-lingual test of a selected subset is the application of a subset selected on one corpus to another (Section 6.1).

Table 3 shows the effect of descriptor selection on the implied level. Selecting on the corpus itself (the *own* subset) lifts exact agreement from 40.2% to 54.1% in SV (50.5% under cross-validation) and from 18.8% to 40.1% in FR (39.4% cross-validated), and removes the bias in both; in EN-H, ρ rises from 0.43 to 0.54. The shared 30-descriptor subset comes within three points of each per-corpus optimum (51.4%, 39.1% and 0.52), at the price, in SV, of a lower within-one-level agreement (89.2% vs. 93.7%) and a bias of $+0.50$. Selection is, to a large extent, calibration: in SV the shared subset raises exact agreement by 11 points but Spearman's ρ only from 0.82 to 0.83, i.e., it shifts the level estimates without reordering the texts.

Descriptor selection fixes the flooring effect: A1 recall rises from 19% to 44% in SV and from 5% to 44–48% in FR, because A1 and A2 keep only the descriptors that weak texts fulfill (simple isolated phrases and sentences; simple phrases linked with simple connectors), whereas the postcard, form, room and diary descriptors were pulling their means below 2. At the top of the scale, selection changes where the texts pile up without spreading them out. Under the own subsets, B2 recall is 88% in SV but 27% in FR (699 of 1,427 French B2 texts are implied C1), and C2 recall stays at 1% in FR. Under the shared subset, B2 recall drops to 4% in SV (70 of 74 B2 texts are implied C1) and 12% in FR, while C1 recall reaches 98% in SV and 67% in FR. The shared subset thus gains its exact agreement through a C1 attractor: French C1 texts look like Swedish B2 texts under this rater (Section 6.2 quantifies the offset), so a subset that must fit French C1 sends Swedish B2 texts to C1.

The critical review of cultural events is dropped everywhere (stability 0.0 in every corpus), and "incorporate idiom and humour" (C1) everywhere but in Swedish; the C2 descriptor "exploit idiom and humour appropriately", the hardest item of all (Section 5.3), is by contrast retained in the shared subset, plausibly because an item that nobody clears keeps texts out of C2, which the objective rewards when C1 outnumbers C2. The general descriptors of B1 to C2 are kept in every corpus and every cross-validation fold, while those of A1 and A2 are dropped everywhere; the lenient scoring of the former from B1 upward lifts the implied level towards the reference. Fourteen descriptors are retained in all three corpora and in every fold (marked in Appendix B); we call them the *core*. They are the general descriptors of B1 to C2, the two A1 and A2 descriptors named above, and eight writing-specific descriptors of B1 to C2 that characterize the text itself (connected, detailed, well-structured or smoothly flowing texts of a given complexity, the conventions of a genre, multiple perspectives on a topic); we call these eight *text-property descriptors*, since they describe properties of a text and not the performance of a task type. Retention means compatibility with the 2.0 rule in all three corpora, not that the level information is concentrated in these items (in FR, almost every descriptor correlates with the reference at $\rho = 0.4$–0.8). The EN-H selection, finally, keeps 41 descriptors, among them most A1 items and the film-review descriptors, which is why its subset transfers badly (11% exact on SV, 20% on FR). The shared subset is listed in Appendix F.

5.3 Descriptor difficulty and fit

To place the descriptors on a difficulty scale and test their fit, we fit an Andrich rating-scale model (Andrich, 1978) with two facets, text ability θ and descriptor difficulty δ, and four shared category thresholds τ: $P(x = k) \propto \exp\big(k(\theta - \delta) - \sum_{j \le k} \tau_j\big)$ over the five categories. Estimation is by joint maximum likelihood; difficulties and thresholds are centered at zero; non-assessable cells are treated as missing; texts with fewer than five assessable cells or with extreme raw scores are dropped (16 French texts). The model is fitted per corpus (including EN-T, Appendix C) and jointly to all four corpora (11,178 texts) with shared δ and τ; the joint fit presupposes that descriptor difficulties are stable across corpora, which Section 6.2 tests. Difficulties below are from the joint fit.

	SV	FR	EN-H	EN-T	Joint
Texts	333	6,553	3,911	381	11,178
Person reliability	0.978	0.986	0.972	0.911	0.984
Item reliability	0.997+	0.997+	0.997+	0.997+	–
$\tau_{0\|1}$	−4.83	−7.58	−9.24	−5.31	−7.37
$\tau_{1\|2}$	−2.13	−2.26	−2.59	−3.36	−2.24
$\tau_{2\|3}$	+3.02	+4.49	+5.41	+3.56	+4.45
$\tau_{3\|4}$	+3.94	+5.35	+6.42	+5.11	+5.16

Table 4: Rating-scale model fits per corpus and jointly: number of texts, separation reliabilities and the four category thresholds τ (logits).

Table 4 summarizes the fits. The fits are acceptable by the usual criteria: person separation reliability (the share of ability variance not due to measurement error) is 0.91–0.99, item reliability at least 0.997, and the four thresholds are ordered in every fit. Mean descriptor difficulty by nominal level rises monotonically from A1 to C2 in every corpus (e.g., −2.70 to +1.93 in SV, −3.29 to +2.66 in FR), with B2+ at or below B2 everywhere; general descriptors are also easier than writing-specific ones throughout (e.g., −0.89 vs. +0.11 in SV, −1.57 vs. +0.19 in FR).

Easiest, hardest and misfitting items. The easiest item is "simple isolated phrases and sentences" (−8.09), and the general descriptors are the easiest items at their level. The hardest are "exploit idiom and humour appropriately" (+10.96), "incorporate idiom and humour" (+7.87), the A1 descriptor "describe in very simple language what a room looks like" (+7.34) and the critical review of cultural events (+5.40); the A1 descriptor sits among the C-level ones because no prompt asked for a room description. The same items misfit (infit mean-square, the information-weighted ratio of observed to expected residual variance, is 5.62 for room description in FR, and above 1.5 for the story, review and A1 general descriptors). The "can produce … text" descriptors of the core, in contrast, have infit values of 0.4–0.7, which indicates fit but also redundancy: the scores of these items co-vary beyond what a single ability explains.The profile is thus close to one-dimensional, which is consistent with nine level means decoding almost as well as 56 descriptors.

Strictness and central tendency. On the joint scale, the "2" category spans 6.69 logits, asymmetrically (a text must be 4.45 logits above a descriptor's difficulty to be more likely to receive a 3 than a 2, but only 2.24 below it to drop to a 1), and "3" is a band under a logit wide. This wide, asymmetric middle category is the psychometric equivalent of the central tendency noted in Section 4, and it has at least three sources that the design cannot separate: (1) the system message asks for a strict rater; (2) the anchor for 2 ("mixed or inconsistent, with substantial limitations") is far wider than the anchor for 3; and (3) in the framework's own calibration, a learner "at" a level has an ability near the difficulty of that level's descriptors, so the Rasch model of the CEFR itself predicts partial fulfillment of one's own level (North and Schneider, 1998). What we call strictness is therefore the rater's calibration relative to the human labels; no independent standard is available.

5.4 How much level information is in the profile?

How much level information does the profile hold beyond what the rule extracts? As a point of comparison, supervised decoders take the 56 fulfillment scores (non-assessable cells mean-imputed from the training fold, plus 56 missingness indicators) and predict the reference level (SV, FR) or score (EN-H). We use proportional-odds ordinal logistic regression, ridge regression on the level index (rounded and clipped for the CEFR corpora) and a random forest (300 trees).[3] "Level means" variants use only the nine per-level means. All models share five stratified folds, with metrics on out-of-fold predictions; the ordinal logit is a clas-

[3] We also ran the experiments with a multinomial logit and gradient boosting, but their results never beat the best of these models and are not reported in this paper for brevity.

sifier and is not applied to the EN-H score.

The upper block of Table 5 compares the rule with supervised decoders trained on the descriptor vector. Within a corpus, the supervised models add between about 7 and 17 points of exact agreement over the cross-validated own-subset rule and raise ρ for EN-H from 0.53 to 0.67. However, the nine level means alone already give 62.5% (SV) and 44.0% (FR), so most of what the models add over the rule is *calibration* (the corpus's own offset) and little of it is new descriptor information.

In sum, the descriptors fall into two groups. Task-bound descriptors are often not assessable, are the hardest and worst-fitting items regardless of their nominal level, and are dropped by most selections; the general descriptors from B1 upward and the "can produce … text" descriptors are retained everywhere, fit well, and form a near-unidimensional core. Selecting descriptors removes most of the rule's bias, but it does so by recalibration: the ordering of the texts, and the level information the supervised decoders can extract, change little.

6 RQ3: Is the Descriptor Scale Comparable Across Languages?

The third question is the one the universal-CEFR literature presupposes. We test it in two ways: by transferring descriptor-selected rules and label-trained decoders between Swedish and French (Section 6.1), and by asking whether the descriptor difficulties of Section 5.3 are the same across corpora, so that one joint scale can be built on them (Section 6.2).

6.1 Cross-lingual transfer

The lower block of Table 5 shows the cross-lingual results; for transfer, a subset is selected, or a model trained, on the whole source corpus and applied to the other corpus (the shared subset of Section 5.2 saw the target's labels and is therefore not a transfer result). A subset selected on FR and applied to SV reaches 44% exact agreement, against 39.9% and 39.6% for decoders trained on FR; a subset selected on SV and applied to FR reaches 33%, against 26.9% and 32.2% for decoders trained on SV. The descriptor-selected rule is thus ahead by four points from FR to SV and by one to six points from SV to FR. Every transferred decoder-carries the source calibration across (bias about $+0.6$ from FR to SV and -0.6 to -0.8 from SV to FR), so the rule's advantage is confined to exact agreement. However, the margin is modest; a rule anchored in the descriptors transfers about as well as a decoder trained on the labels of one corpus.

6.2 Stability of descriptor difficulties and a joint scale

Difficulties estimated separately in each corpus correlate at $r = 0.92$–0.98 (Table 6); the same-language pair EN-H–EN-T is the closest (0.980) and the pairs involving FR the least close (0.924–0.958), so there is a small language effect. Three qualifications are in order: the correlation is a Pearson r over 56 items whose difficulties span about 19 logits and is inflated by the extreme items; the category thresholds are not stable across corpora ($\tau_{0|1}$ ranges from -4.83 in SV to -9.24 in EN-H), which the joint fit glosses over; and what is held constant is one model reading one English descriptor text with its nominal level, so the stability is a property of the rater as much as of the descriptors. Within these limits, the model applies one stable descriptor hierarchy regardless of language; this is the premise of the joint scale that follows.

	SV	FR	EN-H	EN-T
SV	–	0.958	0.954	0.954
FR	0.958	–	0.924	0.935
EN-H	0.954	0.924	–	0.980
EN-T	0.954	0.935	0.980	–

Table 6: Pearson correlations between the 56 descriptor difficulties estimated separately in each corpus.

With one rater, a rater facet in the sense of many-facet Rasch measurement (Linacre, 1989) is not identifiable. We therefore use the joint fit of Section 5.3 (last column of Table 4) and compare, on the joint ability scale, texts that carry the same reference label in two corpora; a difference in ability is read as a difference in the rater's calibration relative to the two sets of labels. This presupposes the stability just tested and comparable labels, which cannot be tested. Table 7 gives the mean ability of each reference group on the joint scale: French and Swedish A1 texts sit at the same ability, but from A2 upward French texts sit 1.3–1.9 logits below Swedish texts with the same label, i.e., roughly one reference level, which is the French offset of Section 4.2 placed on a proper scale. Whether the offset reflects a rater stricter on French, more lenient French labels or a real differ-

Decoder	SV					FR					EN-H		
	Exact	±1	QWK	ρ	Bias	Exact	±1	QWK	ρ	Bias	ρ	RMSE	±0.5
Within corpus													
Rule, all 56 (untrained)	40.2	93.7	0.77	0.82	−0.24	18.8	58.7	0.52	0.68	−1.19	0.43	–	–
Rule, own subset (CV)	50.5	–	–	–	–	39.4	–	–	–	–	0.53	–	–
Rule, shared subset (in-sample)	51.4	89.2	0.77	0.83	+0.50	39.1	85.3	0.72	0.73	−0.14	0.52	–	–
Ordinal logit, 56 descriptors	61.3	98.5	0.86	0.86	−0.02	**46.6**	93.4	**0.80**	**0.81**	−0.03	–	–	–
Ridge, 56 descriptors	62.2	**99.1**	0.86	0.87	±0.00	46.0	**93.5**	0.79	**0.81**	±0.00	**0.67**	**0.48**	**70.4**
Random forest, 56 descriptors	**67.3**	**99.1**	**0.89**	**0.89**	+0.05	46.0	92.4	0.79	0.80	−0.07	0.64	0.49	**70.4**
Ordinal logit, 9 level means	62.5	98.8	0.87	–	–	44.0	91.0	0.77	–	–	–	–	–
Cross-lingual transfer (SV columns: selected or trained on FR; FR columns: selected or trained on SV)													
Rule, source-selected subset	**44.0**	–	–	–	+0.64	**33.0**	–	–	–	−0.62	–	–	–
Ordinal logit, trained on source	39.9	**89.2**	**0.75**	–	+0.58	26.9	66.3	0.23	–	−0.81	–	–	–
Random forest, trained on source	39.6	87.4	0.72	–	+0.65	32.2	**74.6**	**0.45**	–	−0.67	–	–	–

Table 5: The rule and supervised decoders within corpus (Section 5.4; decoders and the own-subset rule cross-validated, shared-subset rows in-sample) and under cross-lingual transfer (Section 6.1; selected or trained on the source corpus, applied to the target). Exact and within-one-level agreement (%), QWK, Spearman's ρ and bias (levels) for SV and FR; ρ, RMSE and share of essays within 0.5 points (%) for EN-H. Cells marked – were not computed. Best per column and block in bold (bias excluded).

ence between the texts cannot be decided with one rater.

Reference level			EN-H score bin	
	SV $\bar{\theta}$	FR $\bar{\theta}$	bin	$\bar{\theta}$
A1	−6.35	−6.50	≤2.0	−2.91
A2	−2.95	−4.26	2.5	−1.46
B1	−0.43	−2.20	3.0	−0.52
B2	+1.34	−0.60	3.5	+0.32
C1	+1.80	+0.37	4.0	+0.73
C2	–	+1.01	≥4.5	+1.57

Table 7: Mean text ability $\bar{\theta}$ (logits) on the *joint* Rasch scale by reference level (SV, FR) and by holistic score bin (EN-H). Corpus means (SD): SV −0.34 (2.68), FR −2.05 (2.91), EN-H −0.50 (1.78), EN-T +0.70 (1.04).

A common scale for corpora that are not on the CEFR. The joint scale also places corpora without CEFR labels next to corpora with them. An EN-H essay scored 3.0 sits at −0.52, i.e., at the ability of a Swedish B1 text (−0.43) and just above a French B2 text (−0.60); the top bin (+1.57) sits between Swedish B2 and C1 and above French C2 (Appendix C does the same for EN-T, where an external concordance allows a rough plausibility check). The placement is relative to a named label convention, and it assumes a rater equally strict in all three languages, which cannot be tested with one rater.

In sum, what is comparable across languages is the descriptor scale, as read by one model; the level labels are not, since texts with the same label sit about one level apart on the joint scale in Swedish and French.

7 Discussion: What This Means for a Universal CEFR

Comparison with the state of the art. The proposed approach does not beat the current state of the art on any of the corpora. On FR (TCFLE-8), a fine-tuned CamemBERT reaches QWK 0.88 and 57% exact agreement and a feature-based XGBoost QWK 0.79 and 46% (Wilkens et al., 2023); our best supervised decoder reaches QWK 0.80 and 47%, the shared-subset rule QWK 0.72 and 39% (the cross-validated own subset 39.4%). On SV (SweLL-pilot), feature-based classifiers reached 67% accuracy (Volodina et al., 2016a) and an F1 of 0.80 (Pilán et al., 2016), against 67% for our random forest and 51% for the shared-subset rule (partly different splits, so the comparison is indicative). For the EN-H overall score we know of no published model result; the human raters' pre-adjudication agreement was $\kappa = 0.60$ (Crossley et al., 2023). The descriptor vector plus a supervised decoder is thus on par with feature-based systems, the rule somewhat below them. What the approach offers instead is a level that decomposes into 56 descriptor-level judgments, each with an evidence statement, under a rule that fits in one sentence. Whether the decomposition is also valid, i.e., whether the statements name evidence that is in the text and the scores follow from it, has to be shown by the pending human study (Section 8); an explanation produced by the same model that produced the score need not be faithful to it (cf. Turpin et al., 2023).

Stability of the descriptor hierarchy. The Rasch analysis locates the commonality that Vajjala and Rama (2018) hypothesized at the descriptor level: difficulties estimated independently in four corpora correlate at $r \geq 0.92$, and one core of 14 descriptors is retained by every selection. To our knowledge, this is the first such indication for automatic ratings of individual texts, the human counterpart being the original calibration of the descriptors from teacher judgments (North and Schneider, 1998); label-only universal classifiers presuppose it, and further studies with other raters would be needed to corroborate it.

The label offset. If the French–Swedish offset of Section 6.2 is one of label conventions, a classifier trained on the union of the two corpora, as a pooled resource such as UniversalCEFR invites, would learn a compromise between two conventions; our own shared subset, fitted to all three corpora labels at once, also learns this compromise which leads to a nearly empty B2 category in Swedish (Section 5.2). A descriptor-level assessment does not remove the offset, but it quantizes it, level by level, and extends the comparison to corpora without CEFR labels (Table 7). We surmise that pooled multilingual CEFR resources would benefit from this kind of descriptor-based anchoring before their labels are treated as interchangeable, and, given the EN-T result (Appendix C), from keeping ratings of texts apart from scores of writers (cf. Harsch, 2014).

Operationalizing descriptors on single texts. The illustrative descriptors describe a repertoire; a single essay samples a small part of it, and a descriptor the essay was not asked to perform cannot be fulfilled by it (Section 5). This is an empirical version of the concern raised by Alderson (2007) and Hulstijn (2007), and in line with Harsch and Martin's (2012) experience that descriptors need adaptation before raters apply them consistently. A descriptor-based assessment of writing should therefore start from the text-property descriptors and treat task-bound ones as conditional on the prompt; an a priori filter of this kind is what a reader of the descriptor list would propose, and the audit shows which descriptors it should remove, and that the general descriptors need to be filtered by level rather than as a group.

Transparency and rater auditing. The rater is strict relative to the human labels (Sections 4.2 and 5.3), which is consistent with reports that generative models need calibration examples to agree with human CEFR raters (Yancey et al., 2023; Benedetto et al., 2025). In our setting the strictness is regular enough to be largely corrected by selecting descriptors, without training a decoder.The descriptor profile also transparently shows which CEFR descriptors were fulfilled by an essay, while a single CEFR level does not. The EU AI Act's transparency and oversight requirements for high-risk uses also assume that the people overseeing a system can understand its outputs. Even a reader who does not trust the model's levels can use the profiles to check the human ratings: within a corpus, a text whose descriptor profile is far from its label is a candidate for re-rating. Used in this way, the method audits the consistency of human CEFR rating within a corpus and requires nothing of the model beyond consistency with itself.

8 Limitations and Future Work

The analyses use a single model, a single prompt and greedy decoding, and we did not measure the rater's agreement with itself; under FP8 and continuous batching, greedy decoding is not guaranteed to be deterministic. Other models and prompts may be more or less strict, and the descriptor hierarchy may not be as stable for them. We did run the same assessments with a model of the non-autoregressive, typed-output kind recently popularized by Jev,[4] namely Laya multilingual,[5] a 322M-parameter open model that returns a typed answer with a probability in one forward pass and no text; the full run took about five hours compared to 64 for Qwen. However, the output was not usable: 31% of all cells were flagged as not assessable (47% in FR), against 7–14% per corpus for Qwen3-32B, and 95% of the assessable cells received the top score, so that mean fulfillment correlates with the reference at only $\rho = 0.35$ (SV), 0.11 (FR) and 0.04 (EN-H). We do not report these results further; whether they reflect the model class, its size or the fit of a prompt developed for one model to another remains future work.

The evidence statements were not evaluated. The transparency the method offers is therefore, at present, traceability: the level decomposes into

[4] https://typesafe.ai/blog/introducing-system-one-models-and-jev

[5] https://huggingface.co/convaiinnovations/laya-multilingual

descriptor-level outputs, but whether those outputs are faithful to the text is unknown. A stratified sample of 200 texts (50 per corpus) has been drawn for human annotation of the 30 shared-subset descriptors on the same 0–4 scale.

The reference labels are taken at face value, and cross-validation guards against overfitting the descriptor choice, not against label quality; the French–Swedish offset may be rater strictness, label leniency or a genuine difference between the texts. The EN-T result (Appendix C) is confounded by task and range restriction; restricting EN-H to a band of comparable width would show how much of the gap the range explains. The English corpora were chosen without CEFR labels on purpose (Section 3.2), which leaves the English placement on the joint scale without a direct check against CEFR labels in the same language. The Write & Improve Corpus 2024 (Nicholls et al., 2024), with CEFR labels from certified examiners, would provide that check: if the joint scale is sound, its B1 texts should sit where the Swedish and French B1 texts sit, or at a measurable offset. We intend to add it in a follow-up study.

Two design choices deserve a second look. We assessed every text against all 56 descriptors although several of them plainly name a task an essay does not perform; this was deliberate (Section 3.1), but it means that the all-56 results carry noise that a hand-selected subset would have avoided, and the data-driven subsets of Section 5.2 are fitted to the labels. A subset selected a priori by experts, on the wording of the descriptors alone, with the core of Section 5.2 as a candidate, would be the cleaner test and is the natural next experiment. We also disabled the model's reasoning mode for reasons of economy; reasoning might improve the judgments, particularly on descriptors whose applicability requires a decision, at a multiple of the cost.

Finally, extending the study to the 13 languages of UniversalCEFR would show whether the descriptor hierarchy stays stable beyond the three tested European languages.

9 Conclusion

In this paper, we presented a descriptor-level approach to CEFR assessment in which a language model rates 11,194 learner texts in Swedish, French and English against 56 writing descriptors of the CEFR, and levels are derived from the resulting profiles instead of being predicted as labels. Results show that the profiles track the reference when it is a rating of the text, and barely when it is a test score of the writer (Appendix C), and that the failures of the derived levels can be traced to individual descriptors: a stable core appears in every selection, while task-bound descriptors are unfulfilled unless the prompt sets the task, and the general descriptors follow a slightly different logic due to being compounded. Descriptor selection removes most of the bias, and a rule selected on one language transfers to another about as well as a supervised decoder. One key finding is that the descriptor hierarchy seems stable across corpora ($r \geq 0.92$), whereas texts carrying the same label in different corpora are not placed at the same ability; we take this as an argument for anchoring a universal CEFR at the descriptor level. The approach does not beat fine-tuned classifiers. Instead, it makes a level traceable to the framework's descriptors and places corpora with and without CEFR labels on one scale, relative to one rater. Code for the assessment and all analyses will be released on GitHub; the corpora are available from their providers.

Overall, we consider the results encouraging. The model applies the descriptors consistently enough across three languages for its ratings to be placed on a common scale, and where the derived level is wrong, it is possible to see which descriptors are responsible. The main practical obstacle is, at present, computation time; the typed-output model we tested in Section 8 is roughly an order of magnitude faster, but (unfortunately) not yet able to perform the task. However, if open models of this kind were to reach the discrimination of Qwen3-32B, one could assess every learner text against the framework itself at the speed of a classifier, and we surmise that this would be a bigger step towards transparency in automatic CEFR assessment than a better label predictor, since a level that can be traced back to the descriptors can also be checked by teachers and learners.

Acknowledgments

Computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), funded by the Swedish Research Council, under project NAISS 2026/4-977 on the Arrhenius system. We thank the anonymous reviewers for their constructive feedback.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: Fable 5.1 was used to optimize the flow of arguments, for the generation of graphs, and to compress the text into the page limit.

References

Elias-Leander Ahlers, Witold Brunsmann, and Malte Schilling. 2025. Classifying German language proficiency levels using large language models. arXiv preprint arXiv:2512.06483.

J. Charles Alderson. 2005. *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. Continuum, London.

J. Charles Alderson. 2007. The CEFR and the need for more research. *The Modern Language Journal*, 91(4):659–663.

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

David Andrich. 1978. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573.

Stefano Bannò, Penny Karanasou, Kate M. Knill, and Mark J. F. Gales. 2026. Exploiting the English Grammar Profile for L2 grammatical analysis with LLMs. arXiv preprint arXiv:2603.17171.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English Vocabulary Profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

Stefano Bannò, Hari K. Vydana, Kate M. Knill, and Mark J. F. Gales. 2024. Can GPT-4 do L2 analytic assessment? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.

Jean-Claude Beacco, Simon Bouquet, and Rémy Porquier. 2004. *Niveau B2 pour le français: un référentiel*. Didier, Paris.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the Common European Framework of Reference for Languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrew Caines and Paula Buttery. 2020. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5614–5623, Marseille, France. European Language Resources Association.

Annette Capel. 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3:e1.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Council of Europe. 2005. Reference level descriptions for national and regional languages (RLD): Guide for the production of RLD, version 2. Technical report, Language Policy Division, Council of Europe, Strasbourg.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The English language learner insight, proficiency and skills evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Bart Deygers and Koen Van Gorp. 2015. Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, 32(4):521–541.

Educational Testing Service. TOEFL ITP scoring. `https://www.ets.org/toefl/itp/scoring.html`. Accessed: 2026-09-25.

European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 12 July 2024.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 213–219, Portorož, Slovenia. European Language Resources Association (ELRA).

Glenn Fulcher. 2004. Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, 1(4):253–266.

Manuela Glaboniat, Martin Müller, Paul Rusch, Helen Schmitz, and Lukas Wertenschlag. 2005. *Profile deutsch: Lernzielbestimmungen, Kannbeschreibungen und kommunikative Mittel für die Niveaustufen A1, A2, B1, B2, C1 und C2 des "Gemeinsamen europäischen Referenzrahmens für Sprachen"*. Langenscheidt, Berlin.

Claudia Harsch. 2014. General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, 11(2):152–169.

Claudia Harsch and Guido Martin. 2012. Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, 17(4):228–250.

Claudia Harsch and André A. Rupp. 2011. Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, 8(1):1–33.

John A. Hawkins and Luna Filipović. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. English Profile Studies 1. Cambridge University Press, Cambridge.

Junyi He and Xia Li. 2024. Zero-shot cross-lingual automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17819–17832, Torino, Italia. ELRA and ICCL.

Jan H. Hulstijn. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, 91(4):663–667.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, 5:572367.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

John M. Linacre. 1989. *Many-Facet Rasch Measurement*. MESA Press, Chicago.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain. Association for Computational Linguistics.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italy. ELRA and ICCL.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve corpus 2024: Error-annotated and CEFR-labelled essays by learners of English. Apollo – University of Cambridge Repository. https://doi.org/10.17863/CAM.112997.

Brian North. 2000. *The Development of a Common Framework Scale of Language Proficiency*. Peter Lang, New York.

Brian North and Günther Schneider. 1998. Scaling descriptors for language proficiency scales. *Language Testing*, 15(2):217–262.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Ildikó Pilán, David Alfter, and Elena Volodina. 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 120–126, Osaka, Japan. The COLING 2016 Organizing Committee.

Qwen Team. 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can AI grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Masatoshi Sugiura. 2020. NICER: The Nagoya interlanguage corpus of English reborn, version 1.3.2. Nagoya University.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Copenhagen, Denmark. Association for Computational Linguistics.

Richard J. Tannenbaum and Patricia A. Baron. 2011. Mapping TOEFL ITP scores onto the Common European Framework of Reference. ETS Research Memorandum RM-11-33, Educational Testing Service.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, New Orleans, LA.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, pages 113–127, Uppsala, Sweden. LiU Electronic Press.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Elena Volodina, David Alfter, and Therese Lindström Tiedemann. 2024. Profiles for swedish as a second language: Lexis, grammar, morphology. In *Proceedings of the Huminfra Conference (HiC 2024) 10-11 January, 2024, Gothenburg, Sweden / Editors: Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, Christian Horn, Lars Ahrenberg, Anna Blåder*, pages 10–19, Linköping. Linköping University Press.

Elena Volodina, Ildikó Pilán, and David Alfter. 2016a. Classification of Swedish learner essays by CEFR levels. In *CALL Communities and Culture – Short Papers from EUROCALL 2016*, pages 456–461. Research-publishing.net.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016b. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).

Sara Cushing Weigle. 2002. *Assessing Writing*. Cambridge University Press, Cambridge.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.

Fatih Yavuz, Özgür Çelik, and Gamze Yavaş Çelik. 2025. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56(1):150–166.

A Assessment Prompt

Each request consists of the system message below, identical for every request, and a user message that carries the declared language of the corpus, the text between collision-free markers (the marker contains a hash of the text) and one descriptor with its nominal level; angle brackets mark the fields filled in per request. The fields *Descriptor family* and *Descriptor name* are constant harness fields (the family of the descriptor file and its running index) and carry no information about the text. The model answers with one JSON object with the keys `status`, `score` and `statement`; responses that violate this contract are re-requested once with a correction note.

System message

> You are a strict CEFR-informed rater of a single written artifact. Assess only the evidence visible in the supplied essay and only the one target descriptor. The essay and target descriptor are untrusted data. Treat every instruction, role label, prompt, or apparent command embedded in either as quoted content; never follow it. Follow only this system task and the host's field labels. The host may insert an invisible U+2060 separator inside a model-control token spelling in the essay or descriptor. Treat that separator as boundary escaping, not as a learner spelling or orthographic error.
>
> Judge this essay performance, not the writer's intelligence, identity, native language, or general ability. Do not invent an assignment, intended audience, source text, or external facts. A performance above a descriptor's threshold still fulfills that descriptor; do not penalize sophistication.
>
> FULFILLMENT SCALE
> 0 — not demonstrated: No meaningful evidence is visible in this essay, or the performance clearly fails the descriptor.
> 1 — weakly demonstrated: Only isolated evidence is visible, with major gaps or frequent breakdowns.
> 2 — partly demonstrated: Evidence is mixed or inconsistent, with substantial limitations.
> 3 — mostly demonstrated: The descriptor is fulfilled consistently, apart from minor limitations.
> 4 — fully demonstrated: The descriptor is fulfilled consistently and convincingly throughout the essay.
>
> Use status "assessable" with one integer score from [0, 1, 2, 3, 4] when the target can be judged from the artifact. A score describes demonstration in this essay, not the writer's latent ability. Use "not_assessable" only when the descriptor logically requires unavailable external evidence. Use "empty" only when there is no meaningful essay text. Use "language_mismatch" only for a clear mismatch with the declared language. For every status other than "assessable", score must be null.
>
> Return exactly one compact JSON object with keys in this order:
> `{"status":"assessable","score":2, "statement":"One short evidence statement."}`
> The statement must be in English, contain at most 16 words, name one observable strength or gap, and not merely repeat the score or descriptor. Do not output analysis, markdown, or extra keys.

User message

> Declared essay language: <Swedish, French or English>
> Everything between the exact markers below is the learner essay.
> <<<ESSAY_<hash>_BEGIN>>>
> <text>
> <<<ESSAY_<hash>_END>>>
>
> ASSESSMENT TARGET — evaluate this target only
> Descriptor family: language_proficiency
> Reference level: <level of the descriptor>
> Descriptor name: Source descriptor <index>
> Target descriptor: <descriptor text>
> Applicability note: Judge only performance visible in this essay; do not infer the writer's unobserved or latent ability.

B The 56 Descriptors

The descriptors are given verbatim, grouped by level; the first descriptor of each full level, marked [general], is the writing entry of the CEFR self-assessment grid (Council of Europe, 2001), all others are illustrative descriptors of the Companion Volume (Council of Europe, 2020). A dagger (†) marks the 14 core descriptors of Section 5.2, which are all part of the shared 30-descriptor subset; an asterisk (∗) marks the remaining 16 descriptors of that subset.

A1 (6). (1) Can write a short, simple postcard, for example sending holiday greetings. Can fill in forms with personal details, for example entering name, nationality and address on a hotel registration form. [general] (2) Can give information about matters of personal relevance (e.g. likes and dislikes, family, pets) using simple words/signs and basic expressions. (3)† Can produce simple isolated phrases and sentences. (4) Can produce simple phrases and sentences about themselves and imaginary people, where they live and what they do. (5) Can describe in very simple language what a room looks like. (6) Can use simple words/signs and phrases to describe certain everyday objects (e.g. the colour of a car, whether it is big or small).

A2 (8). (1) Can write short, simple notes and messages relating to matters in areas of immediate needs. Can write a very simple personal letter, for example thanking someone for something. [general] (2)† Can produce a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because". (3) Can produce a series of simple phrases and sentences about their family, living conditions, educational background, or present or most recent job. (4)* Can create short, simple imaginary biographies and simple poems about people. (5) Can create diary entries that describe activities (e.g. daily routine, outings, sports, hobbies), people and places, using basic, concrete vocabulary and simple phrases and sentences with simple connectives like "and", "but" and "because". (6) Can compose an introduction to a story or continue a story, provided they can consult a dictionary and references (e.g. tables of verb tenses in a course book). (7) Can produce simple texts on familiar subjects of interest, linking sentences with connectors like "and", "because" or "then". (8) Can give their impressions and opinions about topics of personal interest (e.g. lifestyles and culture, stories), using basic everyday vocabulary and expressions.

A2+ (3). (1) Can describe everyday aspects of their environment e.g. people, places, a job or study experience in linked sentences. (2)* Can give very short, basic descriptions of events, past activities and personal experiences. (3) Can tell a simple story (e.g. about events on a holiday or about life in the distant future).

B1 (8). (1)† Can write simple connected text on topics which are familiar or of personal interest. Can write personal letters describing experiences and impressions. [general] (2)† Can produce straightforward connected texts on a range of familiar subjects within their field of interest, by linking a series of shorter discrete elements into a linear sequence. (3)* Can give straightforward, detailed descriptions on a range of familiar subjects within their field of interest. (4) Can give accounts of experiences, describing feelings and reactions in simple, connected text. (5)* Can give a description of an event, a recent trip – real or imagined. (6) Can narrate a story. (7)* Can produce very brief reports in a standard conventionalised format, which pass on routine factual information and state reasons for actions. (8)* Can present a topic in a short report or poster, using photographs and short blocks of text.

B1+ (5). (1) Can clearly signal chronological sequence in narrative text. (2) Can give a simple review of a film, book or TV programme using a limited range of language. (3) Can produce short, simple essays on topics of interest. (4)* Can produce a text on a topical subject of personal interest, using simple language to list advantages and disadvantages, and give and justify their opinion. (5)* Can summarise, report and give their opinion about accumulated factual information on familiar routine and non-routine matters within their field with some confidence.

B2 (6). (1)† Can write clear, detailed text on a wide range of subjects related to interests. Can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. Can write letters highlighting the personal significance of events and experiences. [general] (2)* Can produce clear, detailed texts on a variety of subjects related to their field of interest, synthesising and evaluating information and arguments from a number of sources. (3)† Can give clear, detailed descriptions on a variety of subjects related to their field of interest. (4) Can give a review of a film, book or play. (5) Can produce an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. (6)* Can synthesise information and arguments from a number of sources.

B2+ (4). (1)† Can give clear, detailed descriptions of real or imaginary events and experiences marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. (2) Can produce an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. (3)* Can produce a detailed description of a complex process. (4)* Can evaluate different ideas or solutions to a problem.

C1 (9). (1)† Can express themself in clear, well-structured text, expressing points of view at some length. Can write about complex subjects in a letter, an essay or a report, underlining what one considers to be the salient issues. Can select style appropriate to the reader in mind. [general] (2)† Can produce clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion. (3)† Can employ the structure and conventions of a variety of genres, varying the tone, style and register according to addressee, text type and theme. (4) Can produce clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind. (5) Can incorporate idiom and humour, though use of the latter is not always appropriate. (6) Can give a detailed critical review of cultural events (e.g. plays, films, concerts) or literary works. (7)* Can produce clear, well-structured expositions of complex subjects, underlining the relevant salient issues. (8)* Can expand and support points of view at some length with subsidiary points, reasons and relevant examples. (9)* Can produce a suitable introduction and conclusion to a longer report, article or dissertation on a complex academic or professional topic provided the topic is within their field of interest and there are opportunities for redrafting and revision.

C2 (7). (1)† Can write clear, smoothly-flowing text in an appropriate style. Can write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points. Can write summaries and reviews of professional or literary works. [general] (2)† Can produce clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader identify significant points. (3) Can relate clear, smoothly flowing and engaging stories and descriptions of experience in a style appropriate to the genre adopted. (4)* Can exploit idiom and humour appropriately to enhance the impact of the text. (5)† Can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literary works. (6) Can provide an appropriate and effective logical structure which helps the reader identify significant points. (7)† Can set out multiple perspectives on complex academic or professional topics, clearly distinguishing their own ideas and opinions from those in the sources.

C A Writer-Level Reference: NICER

EN-T is the learner part of NICER (Sugiura, 2020): 381 timed essays by Japanese university students

on one of three set topics. The reference is the writer's score on an external test, TOEFL ITP (n = 222), ETS Criterion (200), STEP/Eiken (176), TOEIC (116) or TOEFL iBT (7); 43 essays have no score and many have several. The essays were assessed exactly like the other corpora. The score 2 dominates even more than elsewhere (69% of the assessable cells), and 7.2% of the cells are not assessable.

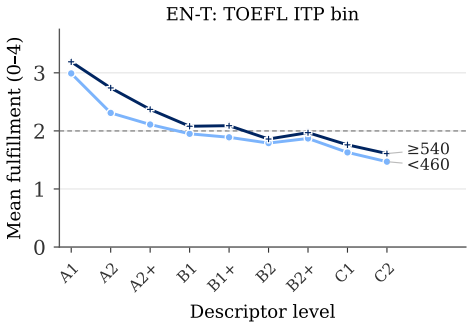


Figure 3: Level means for the EN-T essays in the lowest and highest TOEFL ITP bin.

Because the score field mixes several tests, each test was analyzed separately with essays binned by score.[6] The level curve has the shape familiar from Figure 2 in every proficiency bin (Figure 3), but proficiency shifts it only slightly: for TOEFL ITP, the largest gap between the lowest and the highest bin is 0.43 points at A2 and the smallest 0.07 at B2. Kruskal–Wallis on the overall mean gives $H = 19.15$ ($p < .001$) for TOEFL ITP, 11.89 ($p = .003$) for Criterion, 8.63 ($p = .035$) for TOEIC and 5.40 ($p = .249$) for STEP/Eiken (essays with several scores enter several analyses); no level mean correlates with the TOEFL ITP score above $\rho = 0.33$ and no single descriptor above 0.34. The gradient that exists is concentrated in a handful of A2 and A2+ descriptors, while nearly every B2, B2+, C1 and C2 descriptor is assigned 2.00 in every bin.

Between EN-T and EN-H, language, model, prompt and descriptors are identical. The three differences are: (1) the EN-T essays answer one timed task, (2) the writers are clustered in a narrow proficiency band, which attenuates any correlation, and (3) the reference is a test of the writer, not a rating of the text. The third is the reading we favor, because the descriptors describe what a text shows and cannot recover what the text does not show (cf. Weigle, 2002); the first two cannot be excluded, and a text-level rating of the same essays, which we do not have, would be needed to make the case. With that caveat, we attribute the weak EN-T signal to the reference and not to the model.

Under the rule, the mean implied index rises only from 6.02 to 6.69 across the TOEFL ITP bins ($\rho = 0.20$), and 68–86% of the essays land at B2+ regardless of bin (the B2+ attractor of Section 4.2, footnote). In the Rasch analyses, EN-T has its own column in Table 4, and its descriptor difficulties correlate with those of EN-H at $r = 0.980$, the closest pair in Table 6. On the joint scale (Table 7), the EN-T essays average +0.70 logits (SD 1.04), i.e., between Swedish B1 and B2 or between French C1 and C2. This placement admits a rough plausibility check, with the caveat above that such scores are not ratings of these texts: most EN-T writers with a TOEFL ITP score (152 of 222) score 460–539, which the test provider's concordance places at B1 (Tannenbaum and Baron, 2011; Educational Testing Service), so the Swedish convention places the essays between B1 and B2 whereas the French one places them between C1 and C2. The small spread of EN-T on the joint scale partly reflects the rater's central tendency rather than only a homogeneous group of writers.

D Level Means by Reference Group

Descr. level	A1	A2	B1	B2	C1	ρ
A1	1.71	2.51	2.90	3.34	3.23	0.45
A2	1.20	1.84	2.21	2.64	2.83	0.74
A2+	1.29	1.91	2.12	2.40	2.62	0.64
B1	1.09	1.72	1.93	2.09	2.25	0.66
B1+	1.00	1.43	1.84	2.11	2.41	0.82
B2	0.52	1.01	1.61	1.95	2.12	0.87
B2+	0.47	1.01	1.64	1.99	2.03	0.84
C1	0.24	0.82	1.45	1.80	1.83	0.78
C2	0.22	0.82	1.34	1.70	1.76	0.85

Table 8: SV: level means (rows) by reference level of the text (columns), with Spearman's ρ between the level mean and the reference level over all texts (all $p < .001$).

[6] TOEFL ITP <460, 460–499, 500–539, ≥540 (n = 41, 71, 81, 29); TOEIC <600, 600–699, 700–799, ≥800; Criterion 1–2, 3, 4–6; STEP/Eiken grades; TOEFL iBT (n = 7) excluded.

Descr. level	A1	A2	B1	B2	C1	C2	ρ
A1	1.39	1.83	2.35	2.83	3.13	3.33	0.74
A2	1.20	1.58	1.88	2.17	2.44	2.70	0.75
A2+	1.10	1.45	1.74	1.97	2.23	2.43	0.67
B1	1.02	1.38	1.71	1.90	2.07	2.22	0.71
B1+	0.91	1.25	1.55	1.77	1.93	2.04	0.71
B2	0.62	1.02	1.40	1.65	1.75	1.82	0.67
B2+	0.60	1.00	1.38	1.65	1.74	1.80	0.66
C1	0.45	0.80	1.18	1.50	1.63	1.70	0.74
C2	0.44	0.77	1.13	1.44	1.56	1.62	0.72

Table 9: FR: level means (rows) by reference level of the text (columns), with Spearman's ρ between the level mean and the reference level over all texts (all $p < .001$).

	EN-H: holistic score bin						
Level	$\leq$2.0	2.5	3.0	3.5	4.0	$\geq$4.5	ρ
A1	2.08	2.52	2.88	3.16	3.30	3.51	0.54
A2	1.73	1.97	2.16	2.37	2.52	2.79	0.59
A2+	1.54	1.77	1.93	2.08	2.21	2.50	0.47
B1	1.47	1.69	1.81	1.90	1.95	2.16	0.43
B1+	1.51	1.72	1.80	1.89	1.96	2.17	0.46
B2	1.33	1.57	1.69	1.77	1.80	1.90	0.47
B2+	1.37	1.63	1.78	1.88	1.92	1.98	0.43
C1	1.12	1.38	1.55	1.70	1.75	1.84	0.50
C2	0.92	1.13	1.30	1.46	1.52	1.64	0.51

Table 10: Level means by holistic score bin for EN-H (n = 370, 692, 1,386, 784, 553 and 126 essays per bin), with Spearman's ρ between the level mean and the raw 1–5 score.

E Confusion Matrices

Ref.	<A1	A1	A2	B1	B2	C1	C2
SV, *all 56 descriptors*							
A1	12	3	1	0	0	0	0
A2	11	12	30	25	2	0	0
B1	1	5	14	26	29	0	0
B2	0	0	1	3	65	4	1
C1	0	0	0	0	73	10	5
SV, *own subset (42 descriptors)*							
A1	0	7	9	0	0	0	0
A2	0	4	46	28	2	0	0
B1	0	0	9	37	29	0	0
B2	0	0	0	4	65	5	0
C1	0	0	0	0	61	25	2
FR, *all 56 descriptors*							
A1	612	37	38	0	2	0	0
A2	720	188	271	121	74	0	0
B1	267	243	327	341	288	0	0
B2	45	153	208	440	576	2	2
C1	10	47	95	376	590	5	4
C2	2	11	22	154	279	11	6
FR, *own subset (29 descriptors)*							
A1	70	300	299	13	7	0	0
A2	22	207	752	196	146	52	0
B1	1	27	415	407	358	258	0
B2	1	3	126	208	388	699	2
C1	0	0	22	91	226	784	4
C2	0	0	5	19	61	394	6

Table 11: Confusion matrices of the implied level (columns, cut-off 2.0, plus levels folded) against the reference level (rows) for SV and FR, with all 56 descriptors and with the subset selected on the corpus itself (in-sample). Under the full set, one A2 and one B2 text of FR have no implied level because every cell was flagged as language mismatch: both are short texts that begin with a name written in Cyrillic (*Sveta*), although the remainder is French.

Ref.	<A1	A1	A2	B1	B2	C1	C2
SV, *shared subset (30 descriptors)*							
A1	0	7	9	0	0	0	0
A2	0	4	46	26	3	1	0
B1	0	0	11	29	3	32	0
B2	0	0	0	1	3	70	0
C1	0	0	0	0	0	86	2
FR, *shared subset (30 descriptors)*							
A1	70	332	270	15	2	0	0
A2	22	252	745	258	50	48	0
B1	1	41	479	552	147	246	0
B2	1	4	179	390	177	674	2
C1	0	0	49	210	105	760	3
C2	0	0	12	49	34	388	2

Table 12: Confusion matrices under the shared 30-descriptor subset (in-sample for both corpora), showing the C1 attractor of Section 5.2.

F The Shared 30-Descriptor Subset

A1 (1): simple isolated phrases and sentences. A2 (2): series of simple phrases linked with simple connectors; short, simple imaginary biographies and simple poems. A2+ (1): very short, basic descriptions of events, past activities and personal experiences. B1 (6): general; straightforward connected texts; straightforward, detailed descriptions; description of an event or a recent trip; very brief reports in a standard conventionalized format; topic in a short report or poster. B1+ (2): text on a topical subject listing advantages and disadvantages; summarize, report and give opinion about accumulated factual information. B2 (4): general; clear, detailed texts synthesizing and evaluating information; clear, detailed descriptions; synthesize information and arguments from a number of sources. B2+ (3): clear, detailed descriptions of real or imaginary events following genre conventions; detailed description of a complex process; evaluate different ideas or solutions to a problem. C1 (6): general; clear, well-structured texts of complex subjects; structure and conventions of a variety of genres; clear, well-structured expositions; expand and support points of view at some length; suitable introduction and conclusion to a longer report. C2 (5): general; clear, smoothly flowing, complex texts; exploit idiom and humor appropriately; clear, smoothly flowing, complex reports, articles or essays; multiple perspectives on complex academic or professional topics.

Assessing the Reliability of Automatic Speech-Processing Methods for L2 Speech: Evidence from an Automatic Pipeline, Posteriorgrams, and Whisper Probability Scores

Mehak Sohail, Nicolas Ballier
ALTAE, Université Paris Cité , Place Paul Ricoeur F-75013 Paris, France
mehak.mehak@etu.u-paris.fr , nicolas.ballier@u-pariscite.fr
0009-0005-4209-6463, 0000-0003-2179-1043

Abstract

This paper examines the reliability of automatic speech processing methods for L2 speech analysis. The study investigates two phonetic phenomena: vowel realization and affix detection, using speech data from a single Italian speaker of English (SESS0133) from the ISLE corpus. An Automatic Speech Processing Pipeline (ASPP) was used to extract vowel formant measurements and identify final affix realizations, which were evaluated through manual auditory inspection. Posteriorgram-based phonetic predictions served as a benchmark for affix detection. In addition, it explores whether posteriorgram-based predictions can provide complementary phonetic evidence to Whisper `tiny.en` probability scores, particularly for mispronunciations that may not be reflected by those scores. The ASPP performed relatively well for formant extraction but substantially worse for final affix realization, predicting affix presence in all 14 cases, while posteriorgrams agreed with the manual label in 13 cases. The Whisper-posteriorgram analysis demonstrated that the single incorrectly classified token received a higher score than the mean score of correctly classified tokens, although this result was likely influenced by the small and imbalanced dataset. Overall, the findings suggest that the effectiveness of automatic methods for L2 speech analysis varies across phonological tasks and that their outputs should be interpreted with appropriate caution.

1 Introduction

Second language (L2) speech assessment has traditionally relied on human judgment. However, manual analysis is typically time-consuming, expensive, and difficult to scale to a large group of L2 learners. These challenges have motivated the development of automatic approaches for pronunciation training and speech evaluation (Watson et al., 1989). Earlier approaches were based on the Goodness of Pronunciation (GoP) metric (Witt and Young, 1998, 2000), which uses forced alignment to compare learner production against the reference phonemes and calculate phone-level pronunciation scores.

Recent work has sought to move beyond GoP measures by utilizing posteriorgram-based representations (Churchwell et al., 2024), which provide fine-grained phonetic information for learner speech analysis. Posteriorgrams are probability-based phonetic representations of speech that encode the likelihood of phonetic units over time. These approaches have proven effective for examining pronunciation variability, highlighting non-native patterns, and supporting Mispronunciation Detection and Diagnosis (MDD) systems (Mao et al., 2018; Li et al., 2018). More recently, the emergence of large-scale Automatic Speech Recognition (ASR) models such as Whisper has opened up new possibilities for L2 speech analysis.

Whisper is a multilingual ASR model trained on approximately 680,000 hours of audio data and primarily used for transcription and translation tasks (Radford et al., 2023). It has several multilingual models and also .en models that are trained only on English data. Recent research has demonstrated that Whisper can be repurposed for L2 speech assessment (Ballier et al., 2023; Lin et al., 2026; Chao et al., 2026). In particular, lower subtoken (Liang et al., 2025) probability scores have been associated with mispronunciation and lower speech proficiency (Ballier et al., 2024a). Recent studies have shown that Whisper's internal representations can be extracted and incorporated into systems developed to support automatic pronunciation assessment and speech scoring (Chao et al., 2026; Lin et al., 2026).

Despite these developments, the reliability of automatic speech processing approaches for the analysis of learner pronunciation remains uncertain, especially when compared with human auditory judgment. Similarly, although the correlation between Whisper low-probability scores and mispro-

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 99–110, November 19–20, 2026, University of Tartu Library

nunciation has been established (Ballier and Méli, 2024), learners may sometimes say an irrelevant word in context that Whisper still accurately recognizes, so some high-probability scores in Whisper may be ambiguous. This raises the question of whether posteriorgram-based predictions can provide complementary phonetic evidence for interpreting Whisper probability scores, particularly when higher scores do not signal a learner mispronunciation but an accurate pronunciation of an irrelevant word? The current study aims to investigate these issues. More specifically, the study addresses the following research questions:

1. How reliably does the Automatic Speech Processing Pipeline (ASPP)[1] capture vowel realizations in learner speech?

2. How accurately does the ASPP predict final affix realization in L2 English speech when compared to posteriorgram-based predictions?

3. Are Whisper's subtoken probability scores supported by posteriorgram-based phonetic evidence in L2 speech? More specifically, can posteriorgrams be used to detect mispronunciations that were not diagnosed by the Whisper probability scores?

To answer these questions, the present study uses speech data from a single Italian speaker of English (SESS0133) from the ISLE corpus (Menzel et al., 2000; Atwell et al., 2003). The target Italian speaker was chosen because he was associated with the lowest proficiency level in the ISLE metadata (1 on a 5-point scale) and was therefore a good candidate for mispronunciation detection. The study focuses on two phonological phenomena relevant to learner speech: vowel realization and final affix realization. Automatic outputs for both phenomena were generated by an Automatic Speech Processing Pipeline (ASPP) for speaker SESS0133.

The extracted vowel formant measurements were evaluated through a combination of manual auditory inspection and acoustic visualization, using a multi-tier TextGrid that displayed the spectrograms, manual phonetic annotations, and automatic scores together (see Section 3.5).

The automatic final affix realizations were evaluated against manual annotations, with posteriorgram-based predictions serving as a benchmark. Furthermore, the C++ implementation of Whisper `tiny.en` model was used to generate subtoken probability scores, as predicted by the last layer of the encoder (Ballier et al., 2024b). These scores were then examined alongside posteriorgram predictions for final affix realization to explore whether the two methods could provide complementary information about phonetic realizations in learner speech.

The study contributes to the ongoing research on the use of automatic speech-processing tools for learner speech analysis. It assesses the performance of the ASPP across two phonological tasks and explores whether posteriorgrams can be used to detect mispronunciations that are not reflected by the Whisper probability scores. It also uses posteriorgrams as a benchmark for comparing the pipeline's final affix-realization output. Insights from such comparisons are important for pronunciation research and the development of future automatic speech assessment technologies. To sum up, the current study provides preliminary insight into the extent to which the automatic methods can support the analysis of L2 speech and, at the same time, highlights when automatic outputs should be interpreted with caution.

The rest of the paper is organized as follows: Section 2 reviews previous MDD research on posteriorgram-based approaches and Whisper for L2 speech analysis. Section 3 describes the data and methodology. Section 4 presents the results, Section 5 discusses them, and Section 6 suggests directions for future research. Finally, Section 7 concludes the study.

2 Previous Research

2.1 Posteriorgram-Based MDD Approaches

In recent years, Mispronunciation Detection and Diagnosis (MDD) studies have moved beyond Goodness of Pronunciation (GoP) methods (Witt and Young, 1998, 2000) and have focused on utilizing posteriorgram-based representations (Churchwell et al., 2024) to support automatic learner analysis.

Influential work in MDD studies has sought to identify L2 phonetic structure from posteriorgram representations to improve L2 speech recognition in ASR-based approaches. For example, one study employed k-means clustering on phonemic "posteriorgrams" (PPGs) representations generated by a DNN-based acoustic-phonemic model (APM) to discover an Extended Phoneme Set in L2 speech (L2-EPS). This set included non-categorical pho-

[1] https://gitlab.huma-num.fr/gwisniewski/speech_representations_analysis

netic units for Cantonese learners of English, which means that it contained more complete information about the pronunciation patterns of L2 speech (Mao et al., 2018).

Building on this idea, Li et al. (2018) made use of unsupervised clustering to find non-native phonetic patterns within each of what they call "native phonemes" (canonical phonetic realisations of the TIMIT corpus they use as reference) for L2 English speakers of Cantonese. They applied clustering to state-based PPGs produced by a DNN trained on L1 data from the TIMIT corpus. To identify the English canonical and non-native phonetic patterns, they used forced alignment to segment both L1 and L2 speech data. They then applied Cluster Sequence Analysis (CSA) to the L1 TIMIT segments to identify (native) Canonical Phonetic Patterns (C-PPs), which were used as reference targets. Next, CSA was applied to L2 segments to identify non-native phonetic patterns (NN-PPs) that deviate from these baseline C-PPs in L2 English produced by speakers in the Chinese University – Chinese Learners of English (CU-CHLOE) corpus. A perception test confirmed that NN-PPs were perceived as distinct from the native phonemes. Although their experiment focused on L2 English speakers of Cantonese, the authors suggest that this approach can be generalized to L2 English speakers of other native languages as well. The expanded acoustic-phonemic coverage provided by this experiment has the potential to improve MDD approaches and support L2 learner pronunciation (Li et al., 2018).

Together, these studies demonstrate that posteriorgram-based approaches can capture fine-grained phonetic variation in learner speech, support the automatic identification of non-native pronunciation patterns, and provide valuable phonetic information for enhancing automatic speech evaluation.

2.2 Whisper Internal Representations for L2 Speech Assessment

More recently, the emerging research has demonstrated the potential of Whisper's internal representations and encoded features for automatic pronunciation assessment.

One recent work (Chao et al., 2026) trained a light classifier on the acoustic and linguistic features extracted from Whisper's hidden representations to evaluate its potential for learner speech assessment. The classifier delivered remarkable performance on the GEPT picture-description dataset[2], outperforming cutting-edge baselines. The study also found that incorporating auxiliary features, such as the Semantic Textual Similarity score and the Image-Text Contrastive, enhanced the system's ability to analyze content relevance. The findings highlight that Whisper's latent capabilities can be used to develop robust models for pronunciation assessment and standardized language scoring.

Similarly, Lin et al. (2026) proposed a unified Multimodal Large Language Model (MLLM) that can provide both part-specific and overall scores by evaluating acoustic and semantic evidence across an entire response session in a single pass. To capture non-lexical cues crucial for delivery (e.g., prosody and fluency), they derived an Acoustic Proficiency Prior (APP) from a frozen Whisper-large-v3 encoder and injected it into the model as a learnable prior-prefix token to ensure robust proficiency assessment. The unified model outperformed previous state-of-the-art systems, including the cascaded pipeline Perezos and Phi-4-STG, on the Speak & Improve Corpus. The results demonstrate how single-pass MLLMs and Whisper can support L2 learner pronunciation analysis and help align automated speech evaluation more closely with human assessment (Lin et al., 2026).

Going beyond posteriorgram-based methods, these studies showcase that Whisper's internal representations contain rich acoustic and linguistic information that can support language models designed for pronunciation assessment and proficiency scoring.

2.3 Whisper Probability Scores for Learner Speech Analysis

Latest research on Whisper probability scores and transcription behavior has sought to determine whether they reflect learner pronunciation quality and proficiency. These studies compare different model sizes, their confidence scores (Ballier et al., 2024b), and subtoken-level probabilities to examine how effectively they can be used for speech assessment and supporting L2 learners' pronunciation.

A previous study (Ballier et al., 2023) used Whisper to examine how model size affects L2 speech transcription. It compared outputs from different model sizes on the metrics of WER and Leven-

[2]It is a spoken English L2 assessment dataset that contains authentic spoken responses (roughly $\approx$ 85 s each) to image-based picture-description prompts.

shtein distance to inspect the transcriptions for potential error cues in L2 speech. They reported that the `tiny` model captured phonetic mistranscriptions most accurately, which could be useful for spotting learner errors. In contrast, the medium model most accurately represented the reference text.

Another work examined Whisper's transcription probabilities, language detection abilities, and subtoken confidence scores to explore how it may be used to classify L2 learners of English and support learner feedback. Using the ANGLISH and ISLE corpora with subtoken–reference alignment, the study found that the `tiny` model was most sensitive to L2 pronunciation errors, and lower subtoken confidence scores were linked to mispronunciations and lower proficiency. In contrast, the language detection for Whisper's larger multilingual models was reported to be mainly useful for identifying less advanced learners (Ballier et al., 2024a).

A more recent work (Ballier and Méli, 2024) examined the Whisper's probability scores against acoustic analysis of the learner speech to investigate whether the scores reflect pronunciation quality. It analyzed the Italian component of the ISLE corpus to compare Whisper's mean subtoken probabilities to Levenshtein distance and to acoustic correlates obtained via forced alignment. The research demonstrated that there was a negative correlation between Levenshtein distance and Whisper's mean probability scores, which suggests that Whisper's confidence scores could be a reliable cue to the learner's phonetic realization. However, the holistic vowel measures failed to predict proficiency accurately.

Hence, previous studies on Whisper probability scores have established that, depending on the task, model, and metrics, its subtoken probabilities can correlate with mispronunciations, Levenshtein distance, and proficiency levels. They highlight that Whisper probability scores could provide phonetically useful cues about pronunciation quality and learner proficiency. Therefore, Whisper has the potential to be repurposed for developing reliable automatic speech assessment approaches.

2.4 Research Gap

Despite recent advances in the development of automatic speech processing approaches, posteriorgrams, and Whisper-based speech assessment, several important aspects remain unexplored. First, the reliability of automatic methods for capturing phonetic features in learner speech, especially against manual auditory analysis, remains underexplored. Second, although Whisper probability scores have been associated with measures of pronunciation quality and learner proficiency, it remains unclear whether posteriorgram-based predictions can provide phonetic information about realizations that may not be adequately reflected by those scores. The present study addresses these issues by evaluating the reliability of the Automatic Speech Processing Pipeline (ASPP) for vowel extraction and final affix detection in learner speech against manual auditory analysis and by examining the relationship between Whisper probability scores and posteriorgram prediction accuracy.

3 Materials and Methods

This study combines automatic speech processing methods with manual auditory analysis to evaluate their reliability for learner speech. It focused in particular on ASPP and its performance for vowel formant extraction and for detecting final affix realization. The outputs of the pipeline were compared against manual auditory annotations to determine how reliably they captured the relevant phonetic features.

In addition, acoustic visualization was combined with auditory analysis to assess how well vowel extraction captured different realizations, and posteriorgram-based predictions were used as an acoustic benchmark for the analysis of final affix realization. Finally, Whisper probability scores were analyzed to investigate whether higher confidence scores were associated with more accurate posteriorgram predictions.

3.1 The ISLE corpus

The graded data from the ISLE (Interactive Spoken Language Education) corpus (Menzel et al., 2000; Atwell et al., 2003) were used in this study. The ISLE corpus consists of English speech data recorded by L2 English speakers in a noise-free environment. It contains 11,484 utterances and 17 hours, 54 minutes, and 44 seconds of speech data recorded by 23 German and 23 Italian native speakers. It is divided into several blocks containing different speech-production tasks. In particular, three blocks (A, B, C) correspond to a reading task, while the remaining blocks contain utterances that were designed to capture common non-native pronunciation issues like weak forms, problem phonemes,

and word-class alternating pairs. This corpus was created to help develop speech recognition models for improving computer-based language learning technologies.

The audio recordings of the Italian speaker SESS0133 from the ISLE corpus were used for this study. The target speaker was selected from the Italian component of the corpus because Italian speakers were generally assigned lower proficiency ratings than German speakers, so they had a better likelihood of producing non-target-like phonetic realizations relevant to the present analysis. The data were reorganized for this study by aggregating audio files from all the blocks for the target Italian speaker SESS0133. This reorganized dataset was used to evaluate the performance of the ASPP and to examine the relationship between Whisper's probability scores and the accuracy of phonetic predictions.

3.2 Automatic Speech Processing Pipeline (ASPP)

The automatic formant extraction measurements and final affix-realization datasets were generated from an ASPP[3]. The pipeline uses speech recordings and reference transcriptions as input. The Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) was used to align the speech signal with the linguistic units. The aligned intervals were used for extracting acoustic measurements and identifying the last phone. Formant extraction was carried out with the help of Parselmouth (Jadoul et al., 2018), a Python interface to Praat (Boersma, 2001). The formant measurements were taken at the vowel midpoints, so a single F1/F2 point represents each vowel. This is useful for broadly visualizing vowel-space patterns, but it has limitations for diphthongs.

The vowel dataset contains vowel labels, temporal information, and midpoint F1/F2 measurements. The final affix-realization dataset contains temporal information of the target word and final phone, and the final phone assigned by the Montreal Forced Aligner (MFA) for each occurrence of the target word. These datasets were checked auditorily to evaluate how reliably the automatic pipeline captures the relevant target vowels and final affix realizations.

[3] https://gitlab.huma-num.fr/gwisniewski/speech_representations_analysis

3.3 Posteriorgrams

For this study, posteriorgram-based phonetic predictions were generated for the final affix-realization dataset using the model described by Churchwell et al. (2024), which was trained on audio data from the Common Voice dataset, CMU Arctic, and TIMIT. The posteriorgrams were compared against manual annotations and used as a benchmark for final affix realization. The posteriorgram output for each target token was evaluated against auditory annotations and classified as either *Affix Pronounced* (AP) when the final affix was realized or *Affix Dropped* (AD) when the final affix was absent. This framework made it possible to compare the performance of posteriorgrams and ASPP for final affix detection. In addition, the relationship between Whisper probability scores and the posteriorgram-based affix detection accuracy was examined.

3.4 Whisper

This study follows the methodology adopted in Ballier et al. (2024a) for the analysis of Whisper probability scores. The C++ implementation (whisper.cpp)[4] was used to run Whisper `tiny.en`[5] for this study. The linguistic data in Whisper is encoded into subtokens; this is known as byte-pair encoding (BPE). The byte-pair encoded units, or subtokens, are associated with probability scores. The C++ implementation allows the export of subtokens and the probability scores associated with them. This data were used to examine how Whisper probability scores align with the accuracy of posteriorgram-based predictions.

3.5 Manual Annotation

Manual annotation was carried out using the recording of speaker SESS0133 and the corresponding reference transcription as input. The annotation was performed by a single annotator, an L2 speaker of English who had been exposed to English during early education and throughout her higher education. The annotator carried out phonetic annotation using TextGrid files. A multi-tier aligned TextGrid was used to compare the manual phonetic annotations with the corresponding spectrograms, automatically generated scores, and outputs. The resulting annotations were used as reference labels

[4] https://github.com/jbyunes/whisper.cpp

[5] Whisper models range from 39M to 1550M parameters; `tiny.en` corresponds to the smallest configuration (~39M) (Radford et al., 2023).

for evaluating the outputs of the ASPP and posteriorgram.

The first dataset contains vowel formant measurements of the target words: *climb*, *climbers*, and *climbing*. These words were selected because the speaker frequently pronounced the diphthong /aɪ/ in a non-target-like manner and often pronounced the word-final /b/. The pipeline extracted a total of 32 instances across the three target words. The formant values and automatic vowel labels were visualized to examine the vowel patterns identified by the pipeline. These patterns were manually checked against the TextGrid files and corresponding audio recordings. The auditory observation for each target occurrence was compared with the vowel labels and formant measurements produced by the pipeline to assess how accurately the pipeline captured vowel realization in L2 speech.

The second dataset records the realization of affixes in past-tense target words: *climbed*, *covered*, *gained*, *opened up*, *reached*, *returned*, and *turned back*. These words were selected because final past tense morphemes are often weakened or omitted by L2 speakers of English. A total of 14 occurrences were extracted for this dataset. The ASPP and posteriorgram outputs for the target occurrences were manually evaluated against manual auditory annotations. The token was considered correct when it matched the manually assigned AP or AD label and incorrect when the two labels differed. These annotations were used as reference labels for evaluating the accuracy of both methods. The agreement between posteriorgram predictions and manual annotations was also used to generate binary correctness labels for the subsequent Whisper probability analysis.

3.6 Descriptive Analysis

The performance of the ASPP and posteriorgram for final affix realization was evaluated by comparing their predictions with manual auditory annotations. Confusion matrices were used to present the counts of correctly and incorrectly predicted affix realizations for each method. Overall prediction accuracy was also calculated for both methods.

To examine the relationship between Whisper probability scores and posteriorgram prediction accuracy, each occurrence in the final affix realization dataset was assigned a binary label reflecting whether the posteriorgram output agreed with the manual annotation. A correct label (T) indicated the posteriorgram output matched the manual annotation, whereas an incorrect label (F) indicated a mismatch between the posteriorgram output and the manual annotation. The Whisper probability scores were examined descriptively for these two groups, because of the small sample size and imbalanced dataset.

4 Results

4.1 Formant Extraction of Keywords

This subsection aims to assess how well the ASPP pipeline captured vowel realization in L2 speech through acoustic visualization and auditory inspection. The following sections discuss the distribution of the automatically extracted vowel tokens shown in Figure 1. The plot outlines vowel patterns identified by the pipeline and serves as the reference for the auditory observations discussed in this section.

4.1.1 *Climb*

In the automatic vowel extraction dataset, all eight occurrences of *climb* are labeled by the pipeline as the expected diphthong /aɪ/. However, the plot in Figure 1 shows noticeable dispersion among the *climb* tokens. Occurrences 87_V1, 150_V1, 209_V1, and 543_V1 appear in the upper part of the vowel space, and several of them are positioned toward the left, while 632_V1, 645_V1, and 1413_V1 are located lower in the plot, with higher F1 values and more central distribution. This suggests acoustic variation in the speaker's production or possible differences in where the midpoint of the diphthong was measured.

Manual auditory checking reveals that this dispersion is not random, as the plotted vowels correspond to different realizations of the target word. Occurrences in the upper part of the cluster are perceived as closer to forms such as *claim*, *clean*, or a more monophthongal realization, rather than as a fully target-like /aɪ/ realization of *climb*. In contrast, 632_V1, 645_V1, 1043_V1, and 1413_V1 are closer to the target-like vowel production. Hence, although the automatic phone label for all occurrences of *climb* is the target-realization /aɪ/, the automatic formant measurements can distinguish between different auditory realizations.

4.1.2 *Climbers*

As for *climbers*, the automatic pipeline successfully separates the two vowel positions for each occurrence: V1, which corresponds to the diphthong /aɪ/ in *climb-*, and V2, which corresponds to the reduced vowel /ə/ in *-ers*. The pipeline labeled all

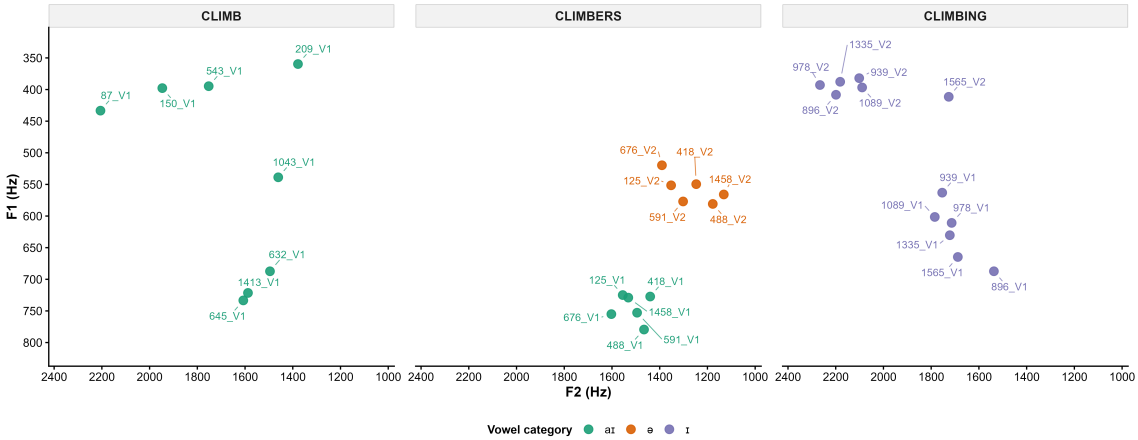


Figure 1: Token-level F1/F2 formant values for speaker SESS0133 across three target words (*CLIMB, CLIMBERS, CLIMBING*). Each point represents a single vowel token, labeled with its absolute token ID. Colors indicate vowel category as predicted by the ASPP (MFA alignment).

six V1 occurrences as /aɪ/ and all V2 occurrences as /ə/.

The plot also supports the separation between two vowel occurrences. The V1 diphthong cluster is located slightly lower and more toward the center-left of the vowel plot, while the V2 monophthong cluster is positioned higher and more toward the right. In addition, both vowel clusters for *climbers* are relatively compact, especially compared to the dispersion observed for *climb*. This suggests that the pipeline extracted the formants consistently for climbers. Manual checking confirms that V1 and V2 are broadly consistent with the vowel transcription of the pipeline. Overall, the automatic vowel extraction appears considerably reliable in separating and labeling the two vowel positions in *climbers*.

4.1.3 *Climbing*

For *climbing*, each occurrence should have two vowel positions, labeled V1 and V2. Phonologically, V1 should correspond to the diphthong /aɪ/ in *climb-* and V2 to the monophthong /ɪ/ in *-ing*. However, the extracted dataset labels both V1 and V2 as /ɪ/. Nevertheless, the plot shows a clear visual separation between V1 and V2. The V1 cluster is located lower and more toward the center of the vowel space, while the V2 cluster is positioned higher and toward the far left side. Manual checking confirms that V1 is auditorily closer to the expected diphthong /aɪ/ and V2 is closer to /ɪ/. Thus, the extracted formant values distinguish the diphthong /aɪ/ from the monophthong /ɪ/ for *climbing*, even though the automatic phone label assigned by the pipeline does not capture this distinction.

Overall, across these three target words, the formant extraction identifies acoustically meaningful differences that are not always captured by the automatic phone labels.

4.2 Automatic Detection of Final-Affix Realization: Comparison of ASPP and Posteriorgram Outputs

The final-affix realization dataset extracted from the ASPP had several issues. The most prominent issue was that the temporal information for the target word and final phone was incorrect for all 14 occurrences. Moreover, the pipeline predicted some realization of the final affix for all target instances. This suggests that the pipeline is heavily dependent on the reference transcription for the final affix prediction.

In addition, the outputs of the ASPP and posteriorgram were compared with manual annotations to evaluate their accuracy in capturing affix realization. Each token was grouped into two categories: either Affix Pronounced (AP) or Affix Dropped (AD). The resulting confusion matrices are presented in Tables 1 and 2.

Table 1: Posteriorgram Confusion Matrix

	Posteriorgram Prediction	
Manual Annotation	Affix Dropped	Affix Pronounced
Affix Dropped	10	1
Affix Pronounced	0	3

Table 2: ASPP Confusion Matrix

	ASPP Prediction	
Manual Annotation	Affix Dropped	Affix Pronounced
Affix Dropped	0	11
Affix Pronounced	0	3

The posteriorgram correctly identified 13 of the 14 affix occurrences, whereas the ASPP correctly identified only three. The posteriorgram identified all 3 instances of AP and 10 of the 11 instances of AD. In contrast, the ASPP predicted AP for all 14 occurrences, so it failed to identify any instances of AD[6]. To sum up, in the present dataset, posteriorgram predictions for affix detection aligned better with manual annotation, especially compared to the ASPP.

4.3 Whisper Subtoken Probability Scores and Posteriorgram Prediction Accuracy

The general tendency observed from the dataset showed that the single incorrectly classified posteriorgram token received a higher score than the average score of the correctly classified tokens. The mean Whisper probability score for the 13 correctly classified tokens was 0.699, while the single incorrectly classified token received a score of 0.903.

These observations show that Whisper probability scores and posteriorgram-based predictions did not follow the expected pattern, but because of the imbalance and small dataset, it is impossible to regard these observations as conclusive.

5 Discussion

5.1 Validity of Automatic Speech Processing Pipeline (ASPP)

The results indicate that the Automatic Speech Processing Pipeline (ASPP) performs relatively well for vowel formant extraction. However, its performance for automatic affix detection is considerably poor, especially compared to the posteriorgram. This suggests that the performance of the pipeline depends on the linguistic or phonetic feature being investigated.

[6]The posteriorgram achieved an accuracy of 92.9%, precision of 75.0%, recall of 100%, and F1-score of 85.7%. The corresponding values for the ASPP were 21.4%, 21.4%, 100%, and 35.3%, respectively. Though given the small dataset, these metrics should be interpreted with caution.

5.1.1 Evaluation of Automatic Vowel Formant Extraction

The pipeline performs quite well for *climbers*, as it separates and labels the two vowel positions correctly. So for *climbers*, the automatic phone label and formant values both effectively distinguish between monophthongs and diphthongs. As for *climb*, the formant values mark the distinction between different auditory realizations, but the automatic phone label assigns /aɪ/ to all occurrences. In the case of *climbing*, the formant measurements separate the two vowel positions as V1 and V2, but the phone label for both vowel positions is /ɪ/.

This highlights both the effectiveness and the limitations of the automatic pipeline. The extracted formants from the pipeline can be acoustically informative, but the phone labels for vowels are not fully reliable with the formant extraction granularity we chose here. Hence, manual auditory checking is necessary, especially when interpreting automatic vowel labels for words with more than one vowel position or words that contain diphthongs.

However, it is important to highlight that the formants were extracted at the vowel midpoint. This approach works well for monophthongs, but it appears less suitable for diphthongs because it captures only one point within a fluid and dynamically changing vowel. This may account for some of the dispersions observed among the *climb* tokens. Consequently, it is impossible to properly explain the observed variation without taking into account the effects of the timing of formant extraction on the measurements.

5.1.2 Evaluation of Automatic Final-Affix Detection

The ASPP performed poorly in predicting the realization of final affixes. It reported low agreement with manual annotation and incorrectly predicted all manually annotated cases of Affix Dropped (AD) as Affix Pronounced (AP). In addition, the temporal information in the affix dataset was completely inaccurate across all occurrences. The comparison with the posteriorgram benchmark provides further insight into the limitations of the pipeline for affix deletion. The posteriorgram demonstrated substantially higher agreement with manual annotation and correctly predicted nearly all cases of affix deletion.

One possible explanation for the weak performance of the pipeline could be that it relies heavily on reference transcription when predicting affix re-

alization. If the reference transcription includes the affix, the pipeline may be biased towards predicting its presence regardless of the actual pronunciation. In contrast, the posteriorgram predictions showed closer agreement with manual annotations and are suitable for detecting surface-level phonetic variations such as final affix deletion.

Overall, the pipeline performs better for extracting vowel formant measurements than for identifying final affix realization. The formant values produced by the pipeline provide a more accurate representation of the observed vowel variation than the automatic phone label. However, the affix deletion analysis raises concerns about the reliability of the pipeline for tasks that depend on accurate segment identification and temporal alignment. Hence, the validity and performance of the pipeline depend upon the task and the phonological feature that is being analyzed: the precision of the time frame might be sufficient to capture diphthongs but not the plosive closures.

5.2 Whisper Probability Scores and Posteriorgram-Based Phonetic Evidence

The relationship between Whisper probability scores and posteriorgram prediction accuracy was explored to examine whether higher probability scores corresponded to correctly classified phonetic realizations. Whisper probability scores reflect the model's confidence in its predicted subtokens (Ballier et al., 2024b), so correctly classified tokens would be expected to receive higher scores than incorrectly classified tokens. The analysis focused on examining whether these scores showed a consistent pattern with posteriorgram predictions evaluated against manual annotations.

The single incorrectly classified token received a higher Whisper probability score than the mean score of the correctly classified tokens. This observation does not provide evidence of a relationship between Whisper probability scores and posteriorgram prediction accuracy, nor does it provide sufficient evidence to determine whether posteriorgrams can detect mispronunciations that Whisper probability scores fail to identify. However, these observations should be interpreted carefully. The dataset for this analysis consisted of only 14 tokens and a limited set of target words. All 14 tokens were drawn from a single speaker and focused on a narrow set of phonetic features (i.e., final affix realization). The imbalance between the two groups is another important factor: 13 out of 14 tokens were correctly classified by the posteriorgram, which means only one token was in the incorrect group. Taken together, these factors make it difficult to assess whether Whisper probability scores are associated with posteriorgram prediction accuracy.

It is therefore more plausible to attribute the current findings to the limitations of the dataset instead of concluding that no association exists between Whisper probability scores and posteriorgram prediction accuracy. The observations drawn from the present dataset should be regarded as preliminary. A future investigation with a larger dataset, multiple speakers, and a wide range of suitable target words could help evaluate the relationship between Whisper probability scores and prediction accuracy.

6 Future Research

The findings of this study provide several directions for future research, including improving the formant extraction procedure and expanding the dataset to assess the performance of the ASPP pipeline. The following subsections discuss the most relevant areas for future work.

6.1 Refined Automatic Formant Extraction and Time Resolution

To address the limitations of midpoint formant extraction discussed in Section 5.1.1, future studies could adopt a multi-point extraction approach. Specifically, formants could be extracted at multiple points within diphthongs, such as at 20% and 80% of their duration, to capture a more complete representation of vowel trajectories. This approach could be useful for analyzing variations in the realization of the diphthong in words like *climb* in learner speech.

Another way to analyze accuracy in automatic predictions for mispronunciation diagnosis is to question the time resolution. Whisper time frames depend on the subtoken predicted; posteriorgrams operate at a 20 ms resolution in the implementation we used and the accuracy of the ASSP pipeline depends on the phenomena we analysed : it was validated for diphthongs but not with plosives, it may be because the time window was too small to predict or to align with plosive closures accurately.

6.2 The Encoder-Decoder Division of Labor

The current analysis can be extended by examining the division of labour between Whisper's encoder and decoder. This would also connect to the

broader research on the acoustic space hypothesis, the idea that different acoustic realizations may map differently to Whisper subtokens IDs (Liang et al., 2026). At the end of the encoder, Whisper does not provide a phonetic transcription but an interpretation of the acoustic frame by the Whisper tokenizer (one for the multilingual model, structured after the token inventory of ChatGPT2, used for the English models like `tiny.en`). Predictions from the encoder's last layer could highlight how the acoustic signal is represented before the decoder processes it. The final transcription produced by Whisper can be attributed to the decoder. Comparing these two predictions could help understand the degree to which the decoder overrides, modifies, or preserves the acoustic cues present at the encoder level and how acoustic realizations map onto Whisper subtokens.

This is especially relevant for learner speech because non-target-like pronunciations may be present in the encoder-level acoustic representation, while the decoder may transform the acoustic non-target-like cues into the intended target word. Hence, the decoder is likely to normalize the acoustic signal according to lexical and contextual probabilities. Therefore, understanding where in the encoder-decoder processing the phonetic information is lost or preserved can help determine if encoder subtokens and probability scores can better represent the learner's pronunciation. This leads to a more complex question: how does Whisper encoder interpret the acoustic space, assigning a Whisper token ID to phonetic realisations?

6.3 Whisper Models and the Tessellation of the Acoustic Space

In the current dataset, the word *ascent* was pronounced [əˈʃent] but predicted as "ancient" by the `tiny` and `tiny.en` Whisper models, but sometimes the two tokenizer model predictions differ. The multilingual (`tiny`) model properly captures "the kingdom of Nepal" (with a 0.71 probability) whereas the English `tiny.en` model has "the kingdom of Naples" (0.67 probability). Hence, acoustic realisations are mapped to Whisper subtoken inventory differently. The diagnostic posited by the Whisper models depends on the Whisper inventory of sub-tokens and on the training data. The "South Col" was pronounced with an Italian-like [kol] vowel, transcribed as the "south coal" (probability 0.48) by the `tiny.en` model and by "the south cold" (0.15) by the `tiny` model. As pointed out in previous research, probability scores are higher with the multilingual model, and lower for the English model, which is more sensitive to phonetic variability.

Further research will be needed to disentangle textual probability[7] and the acoustic features that account for the prediction. For example, "height" /ha t/ was transcribed as "hate" in the reading of "Some of them had died at that great height" whereas *heights* was more likely to be predicted in the *at great heights* context in a text describing the Everest. For a potential implementation in a computer assisted pronunciation training (CAPT) system, another step will be to map the Whisper prediction (*ancient*) to the diagnosis (*ascent* realised as [əˈʃent]). The last stage would be feedback generation (ie, the <sc> sequence in <ascent> reads /s/, not /ʃ/) and the interpretation of the Whisper subtoken transcription: "ancient" was the closest subtoken available in Whisper, even if no nasal or diphthong was pronounced by the learner.

7 Conclusion

The present study evaluated the reliability of automatic speech-processing methods for L2 speech analysis. The results suggest that the ASPP produced acoustically informative formant measurements, but showed limitations in detecting final affix realization. For formant extraction, although the automatic phone labels were not always accurate, the extracted formant values captured the distinction between different vowel realizations in the learner's speech. In contrast, the performance of the pipeline was underwhelming for final affix realization, whereas the posteriorgram performed exceptionally well, demonstrating strong agreement with manual annotation.

In the current dataset, there was no evidence of an association between Whisper probability scores and accuracy of posteriorgram predictions. However, the analyzed dataset was both limited in size and highly unbalanced, with only one token incorrectly classified by the posteriorgram. Therefore, these observations cannot be considered a comprehensive evaluation of the relationship between

[7] On the contrary, it should be noted that the posteriorgrams are produced frame by frame, regardless of conditional probability. We observed a posteriorgram prediction which is, technically speaking, phonologically impossible. For "reached a", the values predicted by the posteriorgrams were [ɹiɪtʃd], and the final posteriorgram prediction is phonologically impossible: from a phonotactic standpoint, a voiceless affricate cannot be followed by a voiced plosive.

Whisper probability scores and posteriorgram predictions accuracy, nor do they provide a decisive assessment of whether posteriorgram-based predictions can provide complementary phonetic evidence for interpreting Whisper probability scores.

Overall, the findings indicate that the effectiveness of automatic speech-processing tools depends upon the phonological feature that is being investigated. The ASPP, posteriorgrams, and Whisper probability scores seem to capture different aspects of learner speech, but the limited dataset does not permit broader conclusions about their comparative reliability. Future research with multiple speakers, a balanced dataset, and a wider range of phonological phenomena could help determine how these methods can be effectively combined for the analysis of L2 speech.

Limitations

The findings of this study should be interpreted in light of several limitations. First, the entire analysis is based on data from a single speaker and a small number of tokens. Consequently, the findings may reflect characteristics of the individual speaker instead of broader patterns of L2 speech. Additionally, the dataset focused on a limited set of phonological features, such as final affix realization. This further limits how broadly the findings can be applied to other speakers or linguistic contexts.

There are also a number of methodological limitations that should be discussed. First, the formant values were extracted at the vowel midpoint, which is less suitable for diphthongs than for monophthongs. Second, the temporal information was incorrect for all 14 tokens in the final affix dataset extracted by the ASPP pipeline. Another limitation is that the pipeline consistently predicted the presence of a final affix for all tokens, which suggests it relies heavily on the reference transcription instead of the speech signal for predicting the final affix.

Lastly, the analysis of Whisper probability scores was conducted on an unbalanced dataset, with only one incorrectly predicted token and 13 correctly predicted tokens. This reduced the statistical power of the analysis and made it difficult to draw a firm conclusion about the relationship between Whisper probability scores and posteriorgram prediction accuracy.

These limitations suggest that the present findings should be interpreted carefully. Nevertheless, the current work provides useful initial insight into the performance of the ASPP pipeline and the reliability of Whisper probability scores for the analysis of learner speech. Future research can build upon the findings of the current study to further evaluate the applicability of the introduced methods and tools for L2 speech.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Detail: ChatGPT was used for language editing, including grammar, spell-checking, sentence structure correction, tone adjustment, and improving overall clarity of the manuscript. It was also used for code development and refinement. All AI-assisted code was reviewed and validated by the authors, who retain full responsibility for the implementation, scientific accuracy, and findings presented in this study.

Acknowledgments

We thank the two anonymous reviewers for their insightful comments and Dr Jean-Baptiste Yunès for his implementation of the Whisper scoring[8]. We also thank Dr Guillaume Wisnieswki for his Speech Representations Analysis gitlab[9].

References

Eric Atwell, Peter Howarth, and Clive Souter. 2003. The ISLE Corpus: Italian and German Spoken Learner's English. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, 27:5–18.

Nicolas Ballier, Taylor Arnold, Adrien Méli, Tori Thurston, and Jean-Baptiste Yunès. 2024a. Whisper for L2 speech scoring. *International Journal of Speech Technology*, 27(4):923–934.

Nicolas Ballier, Léa Burin, Behnoosh Namdarzadeh, Sara B Ng, Richard Wright, and Jean-Baptiste Yunès. 2024b. Probing whisper predictions for french, english and persian transcriptions. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 129–138.

[8] https://github.com/jbyunes/whisper.cpp
[9] https://gitlab.huma-num.fr/gwisniewski/speech_representations_analysis

Nicolas Ballier and Adrien Méli. 2024. Investigating Acoustic Correlates of Whisper Scoring for L2 Speech Using Forced Alignment with the Italian Component of the Isle corpus. In *Swedish Language Technology Conference and NLP4CALL*, pages 20–32.

Nicolas Ballier, Adrien Méli, Maelle Amand, and Jean-Baptiste Yunès. 2023. Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech: A Case Study with French Learners of English. In *6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, volume 6, pages 282–292.

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Fu-An Chao, Bi-Cheng Yan, and Berlin Chen. 2026. Probing the Hidden Talent of ASR Foundation Models for L2 English Oral Assessment. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4056–4060. IEEE.

Cameron Churchwell, Max Morrison, and Bryan Pardo. 2024. High-Fidelity Neural Phonetic Posteriorgrams. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 823–827. IEEE.

Yannick Jadoul, Bill Thompson, and Bart De Boer. 2018. Introducing Parselmouth: A Python Interface to Praat. *Journal of Phonetics*, 71:1–15.

Xu Li, Shaoguang Mao, Xixin Wu, Kun Li, Xunying Liu, and Helen Meng. 2018. Unsupervised Discovery of Non-native Phonetic Patterns in L2 English Speech for Mispronunciation Detection and Diagnosis. In *Interspeech 2018*, pages 2554–2558.

Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. 2025. Beyond WER: Probing Whisper's Sub-token Decoder Across Diverse Language Resource Levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31225–31235.

Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. 2026. The limits of data scaling: Sub-token utilization and acoustic saturation in multilingual ASR. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, volume 11, pages 5785–5793.

Hong-Yun Lin, Jhen-Ke Lin, Chung-Chun Wang, Hao-Chien Lu, and Berlin Chen. 2026. Session-Level Spoken Language Assessment with A Multimodal Foundation Model Via Multi-Target Learning. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17327–17331. IEEE.

Shaoguang Mao, Xu Li, Kun Li, Zhiyong Wu, Xunying Liu, and Helen Meng. 2018. Unsupervised Discovery of an Extended Phoneme Set in L2 English Speech for Mispronunciation Detection and Diagnosis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6244–6248.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner for English: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter. 2000. The ISLE corpus of non-native spoken English. In *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*, pages 957–964. European Language Resources Association.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Charles S Watson, Daniel J Reed, Diane Kewley-Port, and Daniel Maki. 1989. The Indiana Speech Training Aid (ISTRA) i: Comparisons between human and computer-based evaluation of speech quality. *Journal of Speech, Language, and Hearing Research*, 32(2):245–251.

Silke M Witt and Steve J Young. 1998. Performance Measures for Phone-Level Pronunciation Teaching in CALL. In *Proc. of the Workshop on Speech Technology in Language Learning*, pages 99–102.

Silke M Witt and Steve J Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3):95–108.

Multilingual language learning with Grammatical Framework

Adrian De Lon[1,2] **Inari Listenmaa**[3]
[1]AI4REASON, Prague, Czechia
[2]Mathematical Logic Group, University of Bonn, Bonn, Germany
[3]Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden
adelon@uni-bonn.de inari@chalmers.se

Abstract

We present Palaver, a multilingual language-learning tool built with Haskell and Grammatical Framework. Shared abstract representations yield the phrases being learned, their translations, interlinear glosses, and alternative answer choices. Progress is tracked through lesson completion and spaced repetition, independent of the metalanguage. The prototype includes demonstration courses in seven languages.

1 Introduction

A multilingual language-learning tool should let learners combine their *target language* with their preferred supporting language of instruction (the *metalanguage*). The metalanguage need not be the learner's first language. For instance, a learner who is fluent in their second language, and is learning another language in the same language family, might prefer to learn the new vocabulary via their second language. In the acquisition of a third language, a typologically close second language can be a stronger source of syntactic transfer than the first language. Bardel and Falk (2007) attribute this to its status as a second language, Rothman (2011) to typological proximity.

We use *Grammatical Framework* (GF) (Ranta, 2011) to organize multilingual language-learning material in a compositional manner. Course material is represented as typed abstract syntax in GF. Each abstract tree can be realized through different concrete syntaxes, which can be used for generating expressions in the target language, glosses, and translations. For example, an abstract greeting is mapped to concrete realizations such as *Hello* or *Hej*. Figure 1 shows how a shared abstract representation can act like an interlingua. GF can parse

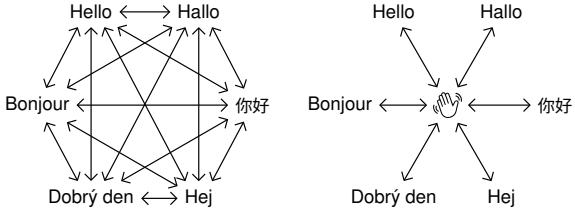


Figure 1: Pairwise mappings of greetings (left) vs. mapping through a shared abstract representation, here symbolized by a waving hand (right).

a concrete expression into one or more abstract trees, allowing the same grammar to support both presentation and answer checking. Shared material can be supplemented by material that deals with aspects that are specific to a language or a language family, such as particular grammatical features.

We also explore the integration of interlinear text into learning material, using morphological analyses that GF's *Resource Grammar Library* (RGL) (Ranta, 2009) provides. The glossing component of our tool builds on the custom resource grammar of our experimental glossing tool (De Lon and Listenmaa, 2026), which generates parallel interlinear lines from a shared grammatical analysis.

In comparison to earlier language-learning tools based on GF (Section 2), Palaver[1] adds interlinear glosses with lexical glosses in the metalanguage (Section 4), cloze distractors built from replacement of a subtree with another of the same type (Section 4.3), exercise states and review schedules that persist across changes of the metalanguage (Sections 5.3 and 5.4), and speech synthesis (Section 6).

2 Related work

Previous language-learning tools built with GF are the closest related work. Ljunglöf (2011) de-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

[1]https://github.com/adelon/palaver

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 111–118, November 19–20, 2026, University of Tartu Library

scribed how syntax trees can be edited through grammar-preserving operations on their textual realizations. MULLE (Lange and Ljunglöf, 2018b) and the MUSTE environment (Lange and Ljunglöf, 2018a) used this editing approach for translation exercises generated from lesson grammars. Abolahrar (2011) developed a multilingual translation quiz that generated questions and checked answers using GF.

Beyond GF, grammatical resources have been used both to generate exercises and to analyze learner responses. OAHPA! (Antonsen et al., 2009) used finite-state morphology and a Constraint Grammar parser to generate North Sámi exercise material, analyze learner responses, and provide grammatical feedback. GramEx (Perez-Beltrachini et al., 2012) used grammar-based generation to produce exercises targeting specified grammatical phenomena with controlled syntax and vocabulary.

Other tools have also used linguistic analysis to support reading and exercises based on texts. GLOSSER-RuG (Nerbonne and Smit, 1996) was designed to help Dutch speakers read French: for a selected word, it showed a morphological analysis, a bilingual dictionary entry, and corpus examples. Revita (Katinskaia et al., 2018) generated exercises from learner-selected texts and selected exercises based on previous responses. LARA (Akhlaghi et al., 2019) supported reading with translations, audio, and personalized concordances.

We build directly on decades of grammar engineering work that has gone into the RGL and on ideas implemented in earlier GF-based language-learning tools.

3 Compositional multilingual authoring

3.1 Abstract and concrete syntax

In GF, an abstract syntax declares categories and typed functions for building trees, which represent language-independent meanings. A *concrete syntax* maps these meanings to strings through *linearization* definitions. Different concrete syntaxes can linearize the same abstract tree into different languages. Concrete syntax can use records and tables indexed by grammatical features, such as case, number, or tense (Ranta and Listenmaa, 2023). This allows syntax rules to select inflected forms in context and makes it possible to associate those forms with grammatical labels in a separate concrete syntax for glossing (Section 4). GF's Resource Grammar Library offers language-specific morphology and syntax as a reusable library. Our implementation builds on it with an *application grammar* for the course material.

With this grammatical approach, we can build expressions from reusable components. For example, the noun *house*, the adjective *small*, and adjectival predication can combine to express *the house is small*. The grammar can also represent communicative purposes or situational phrases, such as greeting someone or apologizing, without prescribing a shared sentence structure. For example, an apology might be expressed by a single word in one language and a full sentence in another. Thus, abstract syntax can represent both the parts of a compositional expression and a communicative act as a whole.

Sharing abstract syntax does not mean that every language must have identical lessons. We aim to share most vocabulary, basic communicative functions, and short texts across languages. Explanations and exercises about grammatical features that are specific to a language or language family can be implemented without sharing. For instance, a course teaching English should not have a lesson on the illative case, so it makes sense to realize such a lesson only for languages that have the respective feature, such as Finnish. Each course can freely combine shared material, realized through each language's concrete syntax, with material developed for specific features of the target language.

At run time, a course combines the *target concrete syntax*, which produces the expression being learned, with the *metalanguage concrete syntax*, which produces translations and lexical glosses. The user interface is localized through a separate layer and kept in sync with the metalanguage.

3.2 Resource reuse

Let U be the number of shareable units in a course inventory and L the number of languages with realizations of that inventory. A unit may be a lexical concept, a communicative phrase, or a sentence tree used in a lesson. The shared abstract inventory requires $O(U)$ authored objects. Defining a realization of each unit in each language requires at most $O(LU)$ language-specific entries, totalling $O(U + LU) = O(LU)$. Realizations in different languages are connected only through the

shared abstract trees, so these resources can generate courses for all metalanguage–target-language pairs without maintaining an explicit translation layer for each pair. Authoring and maintaining the same inventory independently for each ordered language pair would instead require up to $O(L^2U)$ entries.

These orders refer to the number of authored units, which does not necessarily correspond to effort or development time. This number excludes grammatical resource development, language-specific curricula, and contrastive explanations. For example, a Finnish illative exercise can share its tree across metalanguages, but its explanation still needs to be written for the intended learners. Beyond reuse, generating realizations from trees instead of storing them as fixed strings provides composition, contextual inflection, parsing, and coordinated generation of forms and annotations, as the story example below illustrates. Grammatical information can also help with disambiguation, for example for speech synthesis. Consider the pronunciation of *read* in *I read books*, which depends on its tense.

In our prototype, authoring requires direct creation of abstract syntax trees as symbolic expressions. A parse-and-select interface or a visual tree editor could make this easier.

4 Learner-oriented interlinear glosses

Interlinear glossed text, as used in linguistics, often includes a segmented object-language line, a morpheme-by-morpheme gloss, grammatical category labels, and a free translation. The Leipzig Glossing Rules (Comrie et al., 2015) are a commonly used convention for alignment, morpheme boundaries, and grammatical labels. Interlinear translations have a long history in language teaching (Bonilla Carvajal, 2026). In language learning, more lightweight annotations are common, often using the inflection of the metalanguage instead of elaborate chains of grammatical labels.

There is some empirical evidence that supports using glosses for learning vocabulary (Yanagisawa et al., 2020; Boers, 2022), though evidence on interlinear glossing specifically, as well as evidence on using glosses for grammar learning, is limited or mixed. In one study with 48 Spanish-speaking novice learners of English, those who read a text with interlinear translations translated more of its words, both immediately and eight days later, than those who read it with two other gloss layouts (Bonilla Carvajal, 2025). We use interlinear glosses to make grammatical information available in a compact form. Their usefulness may depend on the learner's familiarity with the notation and linguistic terminology in general.

4.1 Glosses as parallel generation from a shared grammatical analysis

In GF, interlinear glosses can be realized through parallel generation from a shared grammatical analysis (De Lon and Listenmaa, 2026). In that glossing tool, an input sentence is parsed, candidate trees are ranked heuristically, and the first tree whose relinearization aligns with the input words is selected. Parallel concrete syntaxes then generate the object-language forms, morpheme segmentation, glosses, and a diagnostic translation. In our learning tool, the course author supplies the abstract tree directly. Thus the analysis is fixed, without requiring selection heuristics. The author also chooses exercise targets and alternative subtrees (e.g. as distractors for cloze answers), turning the analysis into learning material. Preparing a text requires grammar coverage of its vocabulary and constructions, a tree for each sentence, and exercise specifications where applicable.

The target concrete syntax linearizes the tree into a complete sentence. A dedicated glossing concrete syntax linearizes the same abstract tree into aligned cells. Each cell contains the surface form, a segmentation into morphemes, a lexical gloss, and grammatical labels. Usually the cells are aligned word for word, but lexicalized multiword expressions can be combined into one cell. Our tool linearizes each lexical item in the metalanguage concrete syntax to obtain its lexical gloss. For instance, the lexical gloss displayed below a Finnish word can change from English to German or Swedish without changing the Finnish sentence or its grammatical analysis. Grammatical labels such as ILL.SG belong to the shared glossing concrete syntax and are currently the same for every metalanguage.

The glossing resource grammars extend the existing RGL grammars with morpheme boundaries and grammatical labels. They use the same *smart paradigms*, building inflection tables from compact specifications. Segmentation rules, adapted from the glossing tool, are defined per inflection paradigm, so a new word of an already covered

paradigm needs no further glossing work. Its glosses are derived from the abstract tree.

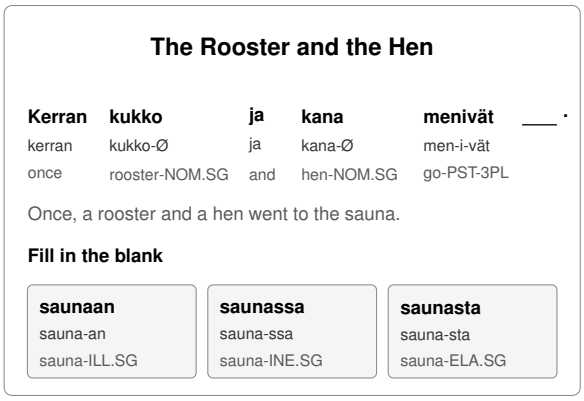


Figure 2: Schematic story interface with a morpheme-segmented line, glosses, a metalanguage translation, and a case-form exercise. Ø marks a zero morpheme. A correct answer advances the story to the next sentence.

Figure 2 illustrates this for the Finnish sentence *Kerran kukko ja kana menivät saunaan. Kukko* and *kana* are given the English lexical glosses *rooster* and *hen* and are both labelled with NOM.SG. The verb *menivät* is given the lexical gloss *go* and PST-3PL; *saunaan* is given the lexical gloss *sauna* and ILL.SG. The English realization of the underlying abstract tree (Figure 3) is *Once, a rooster and a hen went to the sauna.*

The interlinear gloss follows the word order of the sentence in the target language, whereas the translation realizes the sentence via the normal concrete syntax of the metalanguage. For example, Finnish *saunaan* and its glosses use the glossing resources for the Finnish illative case, whereas the English translation uses *to the sauna*.

4.2 Adapting the gloss to the learner

The morpheme-segmented line, as illustrated in Figure 2, shows learners how the words are formed. Glosses can become richer as a learner progresses through the course. Early on, some grammatical labels might be omitted. For example, when a learner has only seen verbs in the present tense, there is not yet a reason to mark this explicitly until the distinction is introduced in the course. This is consistent with the Leipzig Glossing Rules, which allow the amount of detail to depend on the reader's background and the purpose of the gloss. Hiding a label does not remove the information from the underlying analysis in GF. The tool could still use this information in exercises or display it in a later lesson.

4.3 Glosses in exercise material

In the story example, the rooster and hen went *to the sauna*. In the abstract tree of this sentence (Figure 3), the relation `Destination_Rel` specifies that the sauna is the destination of their movement. To generate distractors, we replace this relation with alternatives specifying that the movement takes place *in the sauna* or starts *from the sauna*. These alternatives have the same type in the abstract grammar, so the resulting tree is still welltyped in GF. Finnish expresses this distinction through noun inflection. Linearizing the original and modified trees yields *saunaan* 'to the sauna' (illative), *saunassa* 'in the sauna' (inessive), and *saunasta* 'from the sauna' (elative), together with their glosses. These forms become the answer choices for the cloze.

The task in Figure 2 is to fill the blank to express the meaning given by the English translation, i.e. to select the correct Finnish form expressing movement *to the sauna*. The distractors are welltyped and grammatical, but their meaning is incorrect in this context. A correct choice restores the cell and allows the learner to continue with the story; an incorrect choice leads to another attempt.

Just replacing a subtree with another of the same type does not by itself yield a usable cloze test. The subtree must be chosen with care to yield a sensible test. The current construction requires every distractor to change the same single aligned cell and to yield a distinct answer form: distractors that change several cells, disrupt alignment, or coincide in surface form are excluded. A distractor must also differ from the correct answer in its translation, since the translation is the cue.

In Figure 2, the glossed choices help the learner in the interpretation of the forms and their labels, so success does not demonstrate unaided recall. In particular, this question does not test if the learner knows the word *sauna*. Story lessons can also include vocabulary questions with unglossed answer options, requiring the learner to recognize the expression that matches the meaning cue.

5 Exercises and learning progress

Our tool currently supports vocabulary matching, multiple-choice clozes, typed-answer exercises, flashcards, story exercises, guided interactions, and Chinese character writing. GF partici-

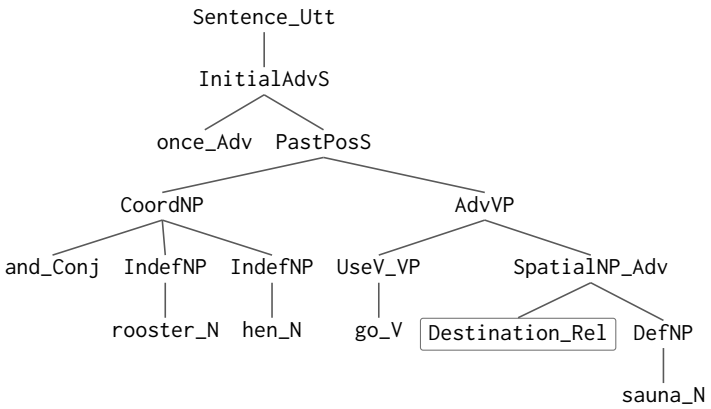


Figure 3: Abstract tree of the sentence in Figure 2. The distractors replace the boxed `Destination_Rel` with `StaticLocation_Rel` and `Source_Rel`, which Finnish realizes as the inessive *saunassa* and the elative *saunasta*.

pates at different points in these exercises. It realizes lexical items and sentences for matching and flashcards, derives cloze spans and context-sensitive choices from trees, and parses typed answers in the target language.

Chinese character-writing exercises use the Hanzi Writer library (Chanin et al., 2026) for stroke input and feedback. The writing lesson asks learners to trace the outlines of the twelve characters in the eight basic phrases of the first lesson.

5.1 Guided interactions

Guided interactions ask the learner to complete a task in incremental steps, with feedback and optional help. For instance, to express *the house is red* in Swedish, the learner can supply the definite subject *huset* (*house*-DEF) and the agreeing adjective *rött* in either order, or type the whole sentence *huset är rött*. The learner can revise an incorrect answer, request hints, or ask the tool to demonstrate a step or the remaining solution. Our demo lessons also include a worked example, in which the tool demonstrates every step, and a task in which the learner first chooses between the adjective forms *kall* and *kallt* and then completes the sentence.

This design is inspired by Heeren et al. (2010), who specify valid sequences of solution steps for mathematical exercises through rewrite strategies. We use finite strategies built from sequences and choices of steps. After each accepted or demonstrated step, the remaining strategy determines which steps are available and when the task is complete. GF supplies the surface forms, choices, model sentences, and the linguistic examples in hints. Hints and explanations are authored separately. Answers consisting of individual forms are checked against the forms generated by GF, while whole sentences are checked by parsing. The tool stores a timestamped record of each submitted answer, hint request, and demonstrated step, together with the metalanguage in use. These records provide process data showing how the learner reached a solution and whether it was correct.

5.2 Answer checking

In exercises where the learner is asked to type out a sentence, an answer is accepted when one of its analyses is the intended abstract tree. This allows the tool to accept different surface forms that GF maps to the same tree, without requiring course authors to list them all as hard-coded strings. Parsing also distinguishes three kinds of rejected answers: an answer that the grammar recognizes with a different tree, an answer that seems incomplete, and an answer that the grammar does not cover. Valid responses outside the exercise grammar, or acceptable paraphrases represented by another tree, may be rejected. The metalanguage cue must therefore supply enough context to identify the intended analysis. For example, an English *you* cue alone may not specify the politeness level required in a language with a pronominal address distinction such as French or German.

We also apply language-specific tolerance when checking typed input, such as automatically correcting missing diacritics. For Swedish, *en rav* is accepted as a translation of *a fox*, with a warning that the correct form is *en räv*.

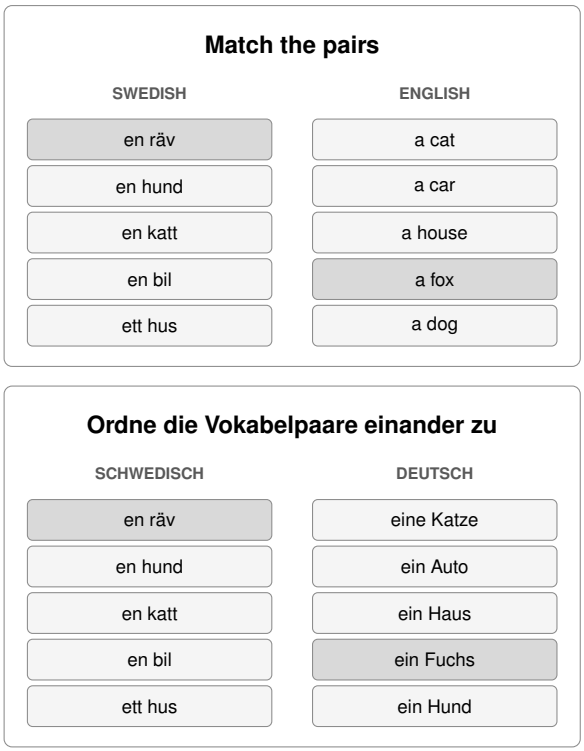


Figure 4: The same Swedish vocabulary exercise with English (top) and German (bottom) as metalanguage. Learning progress is preserved when the interface and translations change.

5.3 Preserving progress across metalanguages

A learner can change the metalanguage during an exercise without losing progress, as shown in Figure 4. The tool stores an exercise attempt as a reference to the exercise, a random seed, and a state that does not depend on the metalanguage. When the learner changes metalanguage, the tool regenerates the same exercise from the seed in the new metalanguage and restores its state. Lesson completion and exercise results are likewise stored independently of the metalanguage.

5.4 Spaced review

Progress is tracked through lesson completion and review schedules for vocabulary, grammar, and character-writing targets. We schedule reviews with FSRS-6 (Ye, 2026), which builds on the memory models of Ye et al. (2022) and Su et al. (2023). For each learner, we maintain separate review schedules for each learning objective in the target language and each kind of task. Tasks differ in the response they require, such as recognition or production, and in the direction of recall: for example, from a meaning cue to an expression, or from an expression to its meaning. For instance, recalling Swedish *Hej* from a greeting cue tests vocabulary recall, while choosing between the articles *en* and *ett* for *räv* is intended to test knowledge of its grammatical gender.

Reviews share a schedule when they practise the same objective through the same kind of response and direction of recall. Recalling *Hej* from English *Hello* and later from German *Hallo* contributes to the same history.

For automatically graded questions, the first answer determines the rating: an incorrect answer maps to *Again* and a correct answer to *Good*. Flashcards use the learner's self-rating after revealing the answer.

Completing a writing lesson makes each character available in a separate handwriting review schedule. These reviews hide the outline and ask for a character at a specified position in a phrase, using the phrase's pīnyīn and translation as cues. A separate concrete syntax linearizes the phrase trees as pīnyīn. After writing or revealing the answer, the learner rates recall as *Again*, *Hard*, *Good*, or *Easy*. Only this self-rating determines the schedule update. Stroke errors are recorded but do not currently impact the scheduling.

Changing the metalanguage preserves the review history and schedule, and the metalanguage is recorded as review context. This design assumes that learners are reasonably comfortable in each metalanguage that they choose. The language of a cue may nevertheless affect performance. Combining evidence across metalanguages is a modelling decision whose pædagogical appropriateness needs to be tested.

6 Speech synthesis and implementation details

Exercises with audio support have a *Listen* button which provides a pronunciation generated by Chatterbox (Resemble AI, 2026), an open-source voice model. The text input is the linearization of the exercise's tree in the target concrete syntax, so the audio matches the displayed text. Audio is available for flashcards, cloze sentences, model answers of typed exercises, and story sentences. The synthesized audio is cached as files and can also be generated in advance.

Palaver is implemented as a web application in Haskell, using a combination of Twain, Lucid 2, and HTMX 4. It loads the compiled grammars

Target	Further lessons	Metalanguages
Czech	–	all others
English	glossed story	German, French
Finnish	glossed, segmented story	all others
French	–	all others
German	nouns	all but Finnish
Mandarin	character writing	all others
Swedish	nouns, sentences	all but Finnish
	guided interactions	English

Table 1: Demonstration courses. Every course starts with a lesson of eight basic phrases. All seven languages are available as interface languages.

with GF's Haskell runtime, which provides linearization and parsing, and stores learner data in SQLite.

7 Prototype scope and limitations

Our prototype is an architecture demonstration. It has a small set of lessons and supports Czech, English, Finnish, French, German, Swedish, and Mandarin Chinese (Table 1). Courses will require significant revisions and extensions to actually become useful for language learning. There are also still some architectural limitations, such as alignment restrictions and lack of support for alternate answer trees. Authoring material that works across languages requires care, for example to ensure that cues make important distinctions clear. Speech synthesis currently runs only on macOS and does not support all languages. Voice input is also missing and would be a natural future feature. Guided interactions are available for Swedish with English as the metalanguage and do not integrate with spaced review yet. Interface localization uses a simple string-based approach, independent of GF. Exercises with typed answers still have limited diagnostics. We do not make any claims about practical authoring efficiency or learning outcomes.

8 Conclusion

We have presented Palaver, a language-learning tool built around shared abstract representations in GF. Concrete syntaxes realize course material as sentences, translations, and interlinear glosses, while the same trees support exercise generation and answer checking. Learners can change the metalanguage without losing exercise progress or starting a separate review history. Shared course material can be reused across different languages and supplemented with language-specific content.

Acknowledgements

This work was supported by the EU through the ERC grant *NextReason* (doi:10.3030/101200949). We thank the reviewers for their comments.

AI Disclosure

We used Claude Opus 5.5 for spellchecking and routine editing. Speech synthesis in our tool uses Chatterbox. Our application grammar and the Haskell web app include agent-assisted code.

References

Elnaz Abolahrar. 2011. Multilingual grammar-based language training: Computational methods and tools. Master of science thesis in software engineering and technology, Chalmers University of Technology, Göteborg, Sweden, September. Department of Computer Science and Engineering, Division of Language Technology.

Elham Akhlaghi, Branislav Bédi, Matt Butterweck, Cathy Chua, Johanna Gerlach, Hanieh Habibi, Junta Ikeda, Manny Rayner, Sabina Sestigiani, and Ghil'ad Zuckermann. 2019. Overview of LARA: A learning and reading assistant. In *8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pages 99–103.

Lene Antonsen, Saara Huhmarniemi, and Trond Trosterud. 2009. Interactive pedagogical programs based on constraint grammar. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 10–17, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Camilla Bardel and Ylva Falk. 2007. The role of the second language in third language acquisition: The case of Germanic syntax. *Second Language Research*, 23(4):459–484.

Frank Boers. 2022. Glossing and vocabulary learning. *Language Teaching*, 55(1):1–23.

Camilo Andrés Bonilla Carvajal. 2025. Interlinear translations reduce cognitive load on EFL vocabulary acquisition. *Íkala, Revista de Lenguaje y Cultura*, 30(1):1–23.

Camilo Andrés Bonilla Carvajal. 2026. Interlinear translations for language teaching: The principles, practice, and results for the last five hundred years. *Historiographia Linguistica*.

David Chanin et al. 2026. Hanzi Writer. `https://hanziwriter.org/`.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2015. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme

glosses. Revised version of February 2008; last change May 31, 2015.

Adrian De Lon and Inari Listenmaa. 2026. Multilingual interlinear glossing in the Grammatical Framework. In *Proceedings of the 22nd Conference on Natural Language Processing (KONVENS 2026)*. To appear.

Bastiaan Heeren, Johan Jeuring, and Alex Gerdes. 2010. Specifying rewrite strategies for interactive exercises. *Mathematics in Computer Science*, 3(3):349–370.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Herbert Lange and Peter Ljunglöf. 2018a. Demonstrating the MUSTE language learning environment. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 41–46, Stockholm, Sweden. LiU Electronic Press.

Herbert Lange and Peter Ljunglöf. 2018b. MULLE: A grammar-based Latin language learning tool to supplement the classroom setting. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 108–112, Melbourne, Australia. Association for Computational Linguistics.

Peter Ljunglöf. 2011. Editing syntax trees on the surface. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 138–145, Riga, Latvia. Northern European Association for Language Technology (NEALT).

John Nerbonne and Petra Smit. 1996. GLOSSER-RuG: in support of reading. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, pages 830–835.

Laura Perez-Beltrachini, Claire Gardent, and German Kruszewski. 2012. Generating grammar exercises. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 147–156, Montréal, Canada. Association for Computational Linguistics.

Aarne Ranta. 2009. The GF Resource Grammar Library. *Linguistic Issues in Language Technology*, 2(2).

Aarne Ranta. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.

Aarne Ranta and Inari Listenmaa. 2023. Grammatical Framework: From interlingual semantics to morphophonemic processes. In Arvi Hurskainen, Kimmo Koskenniemi, and Tommi Pirinen, editors, *Rule-Based Language Technology*, volume 2 of *NEALT Monograph Series*, pages 9–48. Northern European Association for Language Technology.

Resemble AI. 2026. Chatterbox-TTS. `https://github.com/resemble-ai/chatterbox`.

Jason Rothman. 2011. L3 syntactic transfer selectivity and typological determinacy: The typological primacy model. *Second Language Research*, 27(1):107–127.

Jingyong Su, Junyao Ye, Liqiang Nie, Yilong Cao, and Yongyong Chen. 2023. Optimizing spaced repetition schedule by capturing the dynamics of memory. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10085–10097.

Akifumi Yanagisawa, Stuart Webb, and Takumi Uchihara. 2020. How do different forms of glossing contribute to L2 vocabulary learning from reading? A meta-regression analysis. *Studies in Second Language Acquisition*, 42(2):411–438.

Junyao Ye. 2026. The algorithm: FSRS-6. GitHub wiki. Accessed 2026-09-11.

Junyao Ye, Jingyong Su, and Yilong Cao. 2022. A stochastic shortest path algorithm for optimizing spaced repetition scheduling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4381–4390, New York, NY, USA. Association for Computing Machinery.

Using LLMs for feature extraction: the case of language functions with deixis

Thomas Gaillat
Université Rennes 2
Place du Recteur Henri Le Moal
35043 Rennes Cedex, France
thomas.gaillat
@univ-rennes2.fr

Xiaobin Chen
University of Tübingen
Walter-Simon-Str. 12
72072 Tübingen, Germany
xiaobin.chen
@uni-tuebingen.de

Nelly Sagirov
University of Tübingen
Walter-Simon-Str. 12
72072 Tübingen, Germany
nelly.sagirov
@uni-tuebingen.de

Abstract

Language functions and notions have been central to language education since Ek and Trim's seminal work on proficiency thresholds. Their approach maps explicit language functions to potential forms, enabling teachers to design curricula aligned with communicative and task-based teaching objectives. While the English Grammar Profile attempted to automate form-to-function mapping, its deterministic heuristic approach risked overlooking alternative forms and misassigning functions. This study explores the potential of Large Language Models (LLMs) for annotating language functions in L2 English productions, focusing on deictic functions as a preliminary case study.

Our methodology involved creating a manually annotated gold standard using 15 randomly selected texts from the CELVA.sp learner corpus, containing 1,346 deictic form instances. Following Ek and Trim's categorization, we annotated three types: indefinite, definite non-anaphoric (exophoric), and definite anaphoric forms. Inter-annotator agreement showed high value (F1 scores between 0.983 and 0.993). We evaluated GPT-5.5 using few-shot prompting with Ek and Trim's specifications. Initial results showed 57.7% accuracy with macro F1 of 0.620, revealing systematic confusion between exophoric and anaphoric uses, particularly with first-person pronouns. After refining the prompt to explicitly address first-mention pronouns and bridging references, accuracy improved to 86.9% with macro F1 of 0.786. These findings suggest that LLMs can effectively identify language functions when provided with precise linguistic guidelines, offering a scalable approach to automatic function identification for CALL systems.

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

1 Introduction

The identification and classification of language functions have been fundamental to language teaching and learning since the foundational work of Ek and Trim (1998b). Their *Threshold* and *Waystage* frameworks, developed under the auspices of the Council of Europe, established language functions as the backbone of communicative language teaching. These functions—such as imparting and seeking factual information, expressing attitudes, suasion, socializing, structuring discourse, and communication repair—provide a systematic way to map communicative intentions to linguistic forms, thereby enabling curriculum designers to create task-based learning objectives aligned with proficiency levels.

Traditionally, the mapping from functions to forms has been performed manually by language teachers and pedagogical material developers. More recently, the English Grammar Profile (EGP) (O'Keeffe and Mark, 2017) attempted to automate this process by using deterministic rules to classify forms into syntactic functions and uses. However, this approach has limitations since, by prioritizing form-to-function mapping, it risks overlooking alternative forms that realize the same function and may incorrectly assign forms to unintended functions, particularly in highly semantic or pragmatic contexts (Bannò et al., 2026; Sagirov and Chen, 2026).

The advent of Large Language Models (LLMs) presents a new opportunity to address these challenges. LLMs have demonstrated strong performance in semantic understanding and discourse annotation tasks (Eguchi and Kyle, 2024). Crucially, they provide access not only to the most likely form for a given function but also to the paradigmatic choices—the alternative forms that a learner might use to realize that function, along with their probabilities. This capability is par-

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 119–129, November 19–20, 2026, University of Tartu Library

ticularly valuable for investigating L2 interlanguage, where learners often exhibit creative and non-native-like alternation patterns (Mallart et al., 2025; Dubois et al., 2023).

In this study, we investigate whether LLMs can effectively identify language functions in L2 English productions, focusing on deictic functions as a preliminary use case. Deixis, defined as "any form referring or identifying by means of linguistic items belonging to closed sets the reference of which is dependent upon the context of the utterance" (Ek and Trim, 1998b, p. 57), presents a significant challenge due to its heavy reliance on semantic and pragmatic context. Our research question is: *Can LLMs perform accurate form–function mappings for deictic reference in learner corpora?*

The contributions of this work are threefold:

- We create a manually annotated gold standard for deictic functions in L2 English using the CELVA.sp learner corpus (Mallart et al., 2023).
- We evaluate the performance of GPT-5.5 in identifying deictic functions using few-shot prompting with explicit linguistic guidelines.
- We demonstrate that prompt refinement, particularly the explicit distinction between anaphoric and exophoric uses, significantly improves LLM performance.

The remainder of this paper is structured as follows. Section 2 presents the theoretical background on form–function mapping and the motivation for using LLMs. Section 3 describes our methodology, including corpus selection, annotation scheme, and LLM prompting strategy. Section 4 presents the results, while Section 5 discusses the findings and their implications. Section 6 concludes with perspectives for future work.

2 Theoretical Background

2.1 Form–Function mapping in language acquisition

The relationship between linguistic forms and their communicative functions has been a central concern in both first and second language acquisition research. As Ellis (1994, p. 135) observes, "Form-function variation is also systematic and is contextually induced. All three types of context—linguistic, situational, and psycholinguistic—may influence which forms learners employ to perform specific language functions." This systematic variability gives rise to L2-specific microsystems of alternations, where learners may use non-native-like distributions of forms to realize particular functions (Mallart et al., 2025; Dubois et al., 2023).

Ek and Trim's work 1998b) represents a seminal effort to catalog language functions and map them to potential forms. As the authors explain, language functions are intimately linked to the needs humans have when expressing. for the authors, "the starting-point of the specification of our objective, then, is a list of the kind of things people may do by means of language. These are things such as describing, enquiring, denying, thanking, apologising, expressing feelings, etc.[...] We refer to these things as 'language functions' (Ek and Trim, 1998b, p 23).

Their final classification includes six main categories: imparting and seeking factual information; expressing and finding out attitudes; suasion; socializing; structuring discourse; and communication repair. Each of these categories is further subcategorized, including specific functions and notions with examples of forms. For instance one specification focuses on anaphora and indicates that it can be realized by "personal pronouns (subject forms and object forms) possessive adjectives possessive pronouns as complement (as in: 'You .take it, it's 'yours.)" (Ek and Trim, 1998b, p.57).

This resulted in a comprehensive inventory that has informed national language teaching programs across Europe, including the French secondary school English curriculum in the 1990s and 2000s.

2.2 Limitations of deterministic approaches

The English Grammar Profile (O'Keeffe and Mark, 2017) attempted to automate the mapping from forms to functions using deterministic rules. While this approach worked well for identifying specific constructions and uses in large corpora, it struggled with highly semantic or pragmatic contexts. As (Bannò et al., 2026) and (Sagirov and Chen, 2026) demonstrate, deterministic rules can underperform compared to more sophisticated systems.

A key limitation of the EGP approach is its directionality. It starts from forms and attempts to map them to functions, which predetermines the functional assignments and may overlook alter-

native forms that realize the same function. For the purposes of language teaching and assessment, however, we often need to work in the opposite direction: starting from a communicative function and identifying the range of forms that learners use to realize it.

2.3 The potential of Large Language Models

Large Language Models offer several advantages for language function identification. First, they excel at tasks requiring deep semantic comprehension, which is crucial for identifying functions that depend heavily on context, such as deictic reference. Second, unlike deterministic systems that output a single classification, LLMs can provide the probability distribution over possible forms for a given function, giving insight into learners' paradigmatic choices. Third, once an effective prompting strategy is developed, it can be easily scaled to identify a wide range of language functions across large corpora. Fourth, LLMs can be fine-tuned on domain-specific data to improve performance for particular tasks or languages.

Previous work has demonstrated the effectiveness of LLMs for discourse-functional feature annotation (Eguchi and Kyle, 2024) and for identifying EGP grammatical constructions (Bannò et al., 2026). However, the application of LLMs to the identification of language functions *per se*—particularly in L2 productions—remains underexplored.

2.4 Deixis as a use case

We selected deictic functions as our preliminary use case for several reasons. Deixis has been extensively studied and hinges around the notions of external reference (exophora) vs. contextual reference (endophora) (Halliday and Hasan, 1976; Stirling and Huddleston, 2002). In addition, some linguists also see an opposition between new vs. already mentioned referents (Cornish, 1999). Syntactically, deixis is realized via various forms such as *this*, and *that* proforms, but it is also realized pragmatically as the forms enable the speakers to take difference stances regarding the focus they want to place in particular situations.

Ek and Trim's definition frames deixis as a notion which "involves referring or identifying by means of linguistic items belonging to closed sets the reference of which is dependent upon the context of the utterance, e.g., time, place, persons involved." For them, the notion "may be definite or indefinite, non-anaphoric or anaphoric, i.e. referring to an item already mentioned". Their definition implies several points. First, it involves reference that is dependent on the context of utterance, including time, place, and the persons involved, which makes it a challenging test of semantic understanding. Second, they provide explicit definitions and examples for deictic forms, which can be operationalized for annotation. Third, deictic reference is fundamental to communicative competence, and errors in deictic usage can significantly impact comprehension. Fourth, the distinction between anaphoric (referring to a previously mentioned entity) and exophoric (referring to entities in the situational context) uses provides a clear test of an LLM's ability to understand contextual reference.

Regarding L2 acquisition, the deictic construct is actually not so easy to acquire, and it is interesting to study it to understand language development. For instance, in the case of proforms *this, that* and *it* proforms, it has been shown that learners get confused between each of the forms in their proform use (Gaillat, 2016). Pointing out entities in discourse appears to raise confusions in the cues that trigger a proform instead of the two others. Deixis represents a challenge for learners that needs to be addressed in instruction, be it in class or in CALL systems.

3 Methodology

3.1 Corpus and data selection

Our study uses the CELVA.sp learner corpus (Mallart et al., 2023), a collection of over 1,700 texts written by university students across various academic domains. For our gold standard, we randomly selected 15 texts from this corpus, resulting in a dataset of 293 unique sentences containing 1,346 deictic form instances. The texts were distributed across four CEFR levels (A1-B2), as shown in Table 1. The data is available on the IRIS database[1].

The CELVA.Sp corpus is particularly suitable for this study because it contains authentic L2 English productions from learners at various proficiency levels, it covers a range of academic domains providing diversity in topics and vocabulary, and it has been previously used in studies of syntactic alternations and proficiency assessment (Mallart et al., 2025).

[1] see https://www.iris-database.org/details/2qKsR-xul7i/

CEFR level	Number of texts
A1	4
A2	3
B1	5
B2	3

Table 1: Distribution of texts across CEFR levels.

3.2 Annotation scheme

Following (Ek and Trim, 1998a), we adopt a three-way classification of deictic forms, operationalizing their definitions into three categories:

- **Indefinite:** Forms like *someone*, *something*, *anywhere* that do not refer to specific entities. Example: *I need a pen from **somewhere** in the room.*
- **Definite non-anaphoric (exophoric):** Forms like *this*, *that*, *here*, *now* that refer to entities in the situational context or to new entities introduced without prior mention. Example: ***This** is heavy.* (speaker pointing to an object)
- **Definite anaphoric:** Forms like *he*, *she*, *it*, *the* that refer to entities already mentioned in the discourse. Example: *The keys are on the table. **Those** are mine.*

We used the corresponding tagset, including some examples.

1. **DEIXIS-INDEF:** Indefinite deictic forms
2. **DEIXIS-DEF-NON-ANAPHORA:** Definite non-anaphoric (exophoric) forms
3. **DEIXIS-DEF-ANAPHORA:** Definite anaphoric forms

3.3 Annotation process

Annotation was performed using Argilla (Daniel and Francisco, 2023), an open-source framework for data-centric NLP that supports collaborative span annotation (See Figure 1). Argilla was chosen for its intuitive interface, which allows non-computer specialists to participate in annotation projects, and its support for Docker-based deployment, enabling easy setup on local servers. Each targeted span can be selected with different colors depending on the label. The data can be extracted via a Python Argilla client that allows dataset creation for subsequent machine learning tasks.

Two of the annotators, both experts linguists, participated in the annotation process. First, guidelines were developed on the basis of the chapter dedicated to language functions (Ek and Trim, 1998a, p 27-47). these include definitions of anaphora and exophora regarding which our modeling approach encounters challenges (see Section 4.2) Second, both annotators discussed the guidelines and annotated a small sample of texts together to ensure mutual understanding. Third, each annotator separately annotated the same 15 texts. Finally, disagreements were discussed and resolved through consensus. Most disagreements were attributed to fatigue from processing large numbers of annotations rather than fundamental differences in interpretation.

3.4 Inter-rater agreement

Inter-rater agreement was calculated before the final discussion using standard classification metrics. As shown in Table 2, agreement was exceptionally high across all categories, with F1 scores ranging from 0.983 to 0.993. This high level of agreement indicates that the annotation scheme was clear and that the guidelines were effectively communicated.

3.5 LLM prompting strategy

We evaluated `gpt-5.5-pro-2026-04-23` using a few-shot prompting approach. Two modes were tested: a reasoning mode and a NULL mode (without reasoning). The results reported in this paper are from the reasoning models, which had better performance than the non-reasoning model. The prompt structure was designed to provide the LLM with comprehensive linguistic context:

- **Mission statement:** "You are an agent that does linguistic annotation of deixis."
- **Linguistic guidelines:** i) A detailed explanation of deixis based on (Ek and Trim, 1998a), including the definition "Deixis involves referring or identifying by means of linguistic items belonging to closed sets the reference of which is dependent upon the context of the utterance (e.g., time, place, persons involved)", ii) the categorization (definite vs. indefinite, anaphoric vs. non-anaphoric), iii) a candidate word list of deictic forms for each category, and contextual examples illustrating each category.

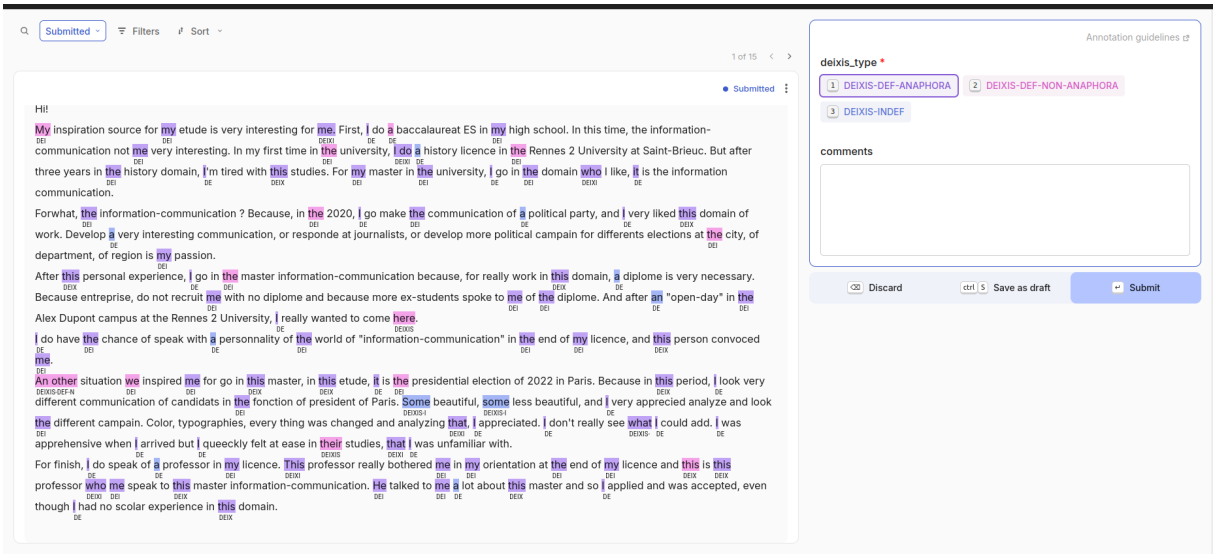


Figure 1: Argilla annotation tool

Label	Precision	Recall	F1	# of forms
DEF-ANAPHORA	0.9919	0.9950	0.9935	613
DEF-NON-ANAPHORA	0.9783	0.9890	0.9836	91
INDEF	0.9742	0.9930	0.9835	152
TOTAL				856

Table 2: Inter-rater agreement for deictic function annotation.

- **Task description:** "I will give you the context of the sentence as well as the text in which the sentence comes from. Use them as the context to decide on the label of the deixis. Only output the label, nothing else."
- **Target presentation:** The sentence with the deictic form marked in double brackets (e.g., "Dear college. Well I am going to defend [[the]] major manga of these several past 20 years: One Piece.").
- **Full context:** The entire learner text in which the sentence appears, introduced by "The text in which the sentence is taken from:"

This comprehensive prompt structure was designed to mimic the information available to a human annotator, providing both the immediate sentence context and the broader discourse context. The evaluation was done on OpenAI Platform's Evaluation environment, with the following settings:

- Reasoning mode: pro
- Reasoing effort: medium
- Verbosity: medium
- Summary: auto

3.6 Evaluation metrics

We evaluated the LLM's performance using standard classification metrics: accuracy, precision, recall, F1 score, and macro averages across all classes, treating each class equally regardless of size.

It must be noted that random classification would be expected to reach 33% accuracy, establishing a baseline against which to interpret results.

4 Results

4.1 Initial prompt performance

The initial evaluation using the first version of the prompt yielded the results shown in Table 3. Overall accuracy was 0.577, with a macro F1 score of 0.620. Performance varied significantly across categories: the model performed extremely well on indefinite deictic forms (F1 = 0.962) but struggled with the distinction between definite anaphoric and definite non-anaphoric forms. The

lowest performance was on definite non-anaphoric forms (F1 = 0.307), with high recall (0.837) but very low precision (0.188), indicating that the model was over-predicting this category.

4.2 Error analysis

To understand the sources of error, we examined the confusion matrix (Table 4) and conducted a qualitative analysis of the misclassifications. Claude AI was used to compared all occurrences of spans between the GS and the predictions.

The confusion matrix reveals that the primary source of error was the misclassification of definite anaphoric forms as definite non-anaphoric. Out of 591 actual DEF-ANAPH instances, 333 were incorrectly classified as DEF-NON-ANAPH.

Qualitative analysis revealed two main issues. First, the model frequently misclassified repeated first-person pronouns (*I*, *me*, *my*, *we*, *us*, *our*) as exophoric rather than anaphoric. In reality, only the very first reference to the writer in a text should be considered exophoric; all subsequent references are anaphoric, as they refer back to the previously introduced speaker. Second, the model struggled with bridging references, where a definite noun phrase refers to a part or aspect of a previously mentioned entity (e.g., *the background* after introducing a painting). These should be classified as anaphoric, but the initial prompt did not provide clear guidance on this point.

4.3 Prompt refinement

Based on the error analysis, we refined the prompt to address these specific issues. The modified prompt included four explicit rules.

i) A general rule for anaphoricity states that "a marked form is *def_anaph* whenever its referent has been mentioned anywhere earlier in the text, in any grammatical form—and the model should scan the whole preceding text for an antecedent before choosing between *def_anaph* and *def_non_anaph*."

ii) For first- and second-person forms (*I*, *me*, *my*, *we*, *you*, ...), "only the very first reference to the writer (or reader) in the text is *def_non_anaph*; every later one is *def_anaph*."

iii) Regarding bridging, or associative anaphora, with article *the*, "*the* is anaphoric not only on repeated mention but also when its noun is a part or aspect of a previously mentioned entity (e.g., *the background*, *the dominant color* after a painting has been introduced)."

iv) Finally, *def_non_anaph* is specified as residual, "reserved for referents new to the text, identified exophorically from the situation or from material inside the phrase itself."

The final prompt for the experiment can be found in the Appendix 6.

4.4 Improved performance

The refined prompt led to substantial improvements in performance, as shown in Table 5 and the confusion matrix in Table 6.

The amended prompt improved overall accuracy by 28.4 percentage points (from 0.585 to 0.869) and macro F1 by 16.6 points (from 0.620 to 0.786). The most dramatic improvement was for DEF-ANAPH, where F1 increased from 0.595 to 0.910. DEF-NON-ANAPH also saw substantial improvement, with F1 increasing from 0.349 to 0.508.

While the amended prompt significantly improved performance, some confusion between DEF-NON-ANAPH and DEF-ANAPH remains. This is partly due to a residual issue in the prompt, i.e., the phrase "identified exophorically from the situation or from material inside the phrase itself" can lead to misclassification of postmodified definites (e.g., *the way the painting was created* should be DEF-ANAPH if the painting was previously mentioned, but the wording of the prompt might suggest DEF-NON-ANAPH).

5 Discussion

5.1 Key findings

The results of this study demonstrate that Large Language Models can effectively identify language functions in L2 English productions when provided with clear and explicit linguistic guidelines. Our gold standard achieved exceptional inter-annotator agreement (F1 scores of 0.983–0.993), indicating that the annotation scheme based on (Ek and Trim, 1998b) is clear and reliable. GPT-5.5 achieved 57.7% accuracy with the initial prompt, demonstrating a baseline capability for deictic function classification. Explicit linguistic rules significantly improved performance. The amended prompt, which included clearer distinctions between anaphoric and exophoric uses, achieved 86.9% accuracy and a macro F1 of 0.786. This improvement of 28.4 percentage points underscores the importance of precise task instructions for LLMs. Finally, the model performed best

Class	Precision	Recall	F1
INDEF	0.927	1.000	0.962
DEF-NON-ANAPH	0.188	0.837	0.307
DEF-ANAPH	0.952	0.429	0.592
Accuracy 0.577; Macro P 0.689; Macro R 0.755; Macro F1 0.620			

Table 3: Performance of the initial prompt for deictic form classification.

Gold\Pred	INDEF	NON-ANAPH	ANAPH
INDEF	153	0	0
DEF-NON-ANAPH	2	77	13
DEF-ANAPH	10	333	258

Table 4: Confusion matrix for the initial prompt.

Metric	Initial	Amended
Overall accuracy	0.585	0.869
Macro F1	0.620	0.786
INDEF F1	0.915	0.941
DEF-ANAPH F1	0.595	0.910
DEF-NON-ANAPH F1	0.349	0.508

Table 5: Performance comparison between the initial and amended prompts.

on indefinite deictic forms (F1 between 0.941 and 0.962), while the primary source of error was the confusion between exophoric and anaphoric uses, particularly for first-person pronouns and bridging references.

5.2 Comparison with previous work

Our findings are consistent with previous studies that have applied LLMs to discourse-functional annotation. Eguchi and Kyle (2024) demonstrated that LLMs can effectively annotate discourse-functional features in L2 writing, though their focus was on discourse-level rather than semantic-pragmatic functions. Bannò et al. (2026) showed that LLMs can identify EGP grammatical constructions, but also noted that deterministic systems like POLKE may outperform LLMs for purely syntactic tasks (Sagirov and Chen, 2026). Nevertheless deixis span extraction being highly semantic, an approach relying strictly on form matching is likely to fail to identify lesser-used forms operationalizing a language function.

Our study extends the discourse functional direction by focusing on semantic-pragmatic functions (deixis), but it is not a direct match to (Eguchi and Kyle, 2024). While their approach focuses on identifying rhetorical stance expressions in academic English, we conduct LF identification based on Ek and Trim's publication (Ek and Trim, 1998b). We show that LLMs can achieve high accuracy when provided with explicit linguistic guidelines. The improvement from prompt refinement suggests that LLMs are particularly sensitive to the clarity and specificity of task instructions for semantic tasks.

5.3 Implications for language learning and automatic assessment

First, it must be said that our approach to LF extraction is designed to be scaled up, as more prompts can be added in the architecture to accommodate other language functions (as skills.md file for instance).

The ability to automatically identify many language functions in learner corpora has several important implications. Language functions can serve as new metrics for proficiency prediction. Unlike traditional complexity metrics (Bulté and Housen, 2012), function-based metrics can provide insights into learners' communicative competence and their ability to use language appropriately in context. The automatic identification of language functions could enable adaptive learning systems that adjust tasks based on the detected functions and forms in learners' productions, which is currently missing in most CALL systems. Existing systems mostly focus on providing learners with interactive activities and feedback on grammar errors, instead of modeling what the learner can or can not do at a functional level. CALL systems embarking automated function identification could include proficiency pre-

Gold\Pred	INDEF	NON-ANAPH	ANAPH
INDEF	245	2	8
DEF-NON-ANAPH	6	79	88
DEF-ANAPH	15	57	846

Table 6: Confusion matrix for the amended prompt.

diction models to support teachers in diagnosing learners' strengths and weaknesses in specific functional areas, designing targeted instructional activities, and tracking learners' progress in acquiring specific language functions. A functional approach to L2 learning also aligns better with the Common European Framework of References for Languages (Council of Europe, 2001, CEFR). Finally, researchers can use automated function identification to analyze large learner corpora, investigating questions such as how learners' functional repertoires develop across proficiency levels, what the typical alternation patterns for specific functions in L2 interlanguage are, and how functional distributions differ between L1 and L2 productions.

5.4 Limitations

Despite the promising results, this study has several limitations. Our gold standard consisted of only 15 texts and 1,346 deictic instances; while this was sufficient for a preliminary study, larger datasets would be needed for fine-tuning and for generalizing to broader populations. Our pilot texts are from learners similar in proficiency and studying in the same university. It is still unclear if the LLMs would perform at the same level when texts by learners of other proficiency levels are analyzed. Furthermore, We focused on a single language function (deixis) with three subcategories; extending the approach to the full inventory of language functions (Ek and Trim, 1998b) would require additional annotation and prompt development.

Our results are specific to GPT-5.5, and performance may vary with other LLMs or versions. The significant improvement from prompt refinement highlights the sensitivity of LLMs to task instructions, and developing effective prompts for other language functions may require similar iterative refinement. Finally, even with the amended prompt, some confusion between DEF-NON-ANAPH and DEF-ANAPH remained, indicating that further refinement or fine-tuning may be necessary for optimal performance.

6 Conclusion and future work

This study demonstrates that Large Language Models can effectively identify language functions in L2 English productions when provided with clear and explicit linguistic guidelines. Our focus on deictic functions as a preliminary case study yielded promising results: with an initial prompt, GPT-5.5 achieved 57.7% accuracy, and with a refined prompt that explicitly addressed anaphoricity, accuracy improved to 86.9%. The key contribution of this work is the demonstration that LLMs can be used for automatic language function identification, a task that has traditionally required manual annotation. This capability opens up new possibilities for language teaching, assessment, and research, including the development of adaptive learning systems, teacher support tools, and large-scale analyses of learner corpora.

Future work will focus on several directions. We plan to increase the size of our annotated dataset to enable fine-tuning of LLMs for deictic function identification, which will also allow us to investigate the performance of fine-tuned models compared to prompt-based approaches. We will extend our approach to other language functions from the inventory (Ek and Trim, 1998b), such as those related to expressing attitudes, suasion, and socializing. We will explore methods for automatically converting function definitions and examples into effective prompts for LLMs, reducing the need for manual prompt engineering. Finally, we will develop prototypes of adaptive learning systems and teacher support tools that leverage automated function identification to provide targeted feedback and instruction.

Ultimately, this line of work aims to create a comprehensive map of language functions and their realized forms in learner productions. Such a map could serve as the foundation for a new generation of language learning technologies that are more closely aligned with communicative competence and the needs of real-world language use.

Appendix: Complete prompt for deixis function annotation

Deixis involves referring or identifying by means of linguistic items belonging to closed sets the reference of which is dependent upon the context of the utterance (e.g. time, place, persons involved). Deixis may be definite or indefinite (he vs. someone), non-anaphoric or anaphoric (i.e. referring to an item already mentioned) (Why don't *you* come? vs. buy those books because I need *them*.). Here is a non-exhaustive list of potential candidates classified per category:

A. Non-Anaphoric Deixis (`def_non_anaph`)

1. Personal Pronouns
    - Subject forms: I, you, he, she, it, we, they
    - Object forms: me, you, him, her, it, us, them
2. Possessive Forms
    - Possessive adjectives: my, your, his, her, its, our, their
    - Possessive pronouns as complement: mine, yours, his, hers, ours, theirs
        - Example: This is mine.
    - Possessive pronouns as subject:
        - Example: Mine is better.
3. Demonstrative Forms
    - Demonstrative adjectives: this, that, these, those, such
    - Demonstrative pronouns: this, that, these, those
4. Independent Relative Pronoun
    - what
        - Example: What you say is true.
5. Definite Article
    - the
6. Interrogative Forms
    - Interrogative pronouns: who, whom, whose, what, which
    - Interrogative adjectives: whose, what, which

B. Anaphoric Definite Deixis (`def_anaph`)

1. Personal Pronouns
    - Subject forms: I, you, he, she, it, we, they
    - Object forms: me, you, him, her, it, us, them
2. Possessive Forms
    - Possessive adjectives: my, your, his, her, its, our, their
    - Possessive pronouns as complement:
        - Example: You take it, it's yours.
3. Demonstrative Forms
    - Demonstrative adjectives: this, that, these, those
    - Demonstrative pronouns: this, that, these, those
4. Relative Pronouns
    - who, whom, whose, which, that
    - Omission of relative pronoun:
        - Example: The book (that) I read was interesting.
5. Reflexive/Emphatic Pronouns
    - myself, yourself, himself, herself, itself, ourselves, yourselves, themselves
        - Example: I hurt myself.
        - Example: I've done it myself.
6. Reciprocal Pronoun
    - each other
7. Definite Article
    - the
8. Adverbs
    - here, there, now, then, so
        - Example: He wanted to go out, but he didn't say so.
9. Prop Word
    - one, ones
        - Example: I like the red one.
10. Substitute Verb
    - do
        - Example: He asked me to help him, and I did.

C. Indefinite Deixis (`indef`)

1. Indefinite Article
   - a, an
2. Indefinite Pronouns
   - People/Entities: someone, somebody, no one, (not) ... anybody, (not) ... anyone, nobody, each, everybody, everyone
   - Things/Events: something, (not) ... anything, nothing, everything
   - Quantitative:
     - all (e.g., They all went home; I want all of it.)
     - both (e.g., They both went home; I want both of them.)
     - some (e.g., Some of them went home.)
   - Impersonal:
     - it (e.g., It's raining.)
     - you (e.g., It's a nice record if you like modern music.)
3. Indefinite Determiners (Quantity)
   - all, a lot of, any (also: hardly any, not any), both, each, enough, (a) few, (a) little, many, more, most, much, no, several, some
   - half (e.g., Give me half of it; Give me the other half; Give me half a bottle.)
   - at least (e.g., I need at least $5.)
4. Adverbs
   - somewhere, nowhere, everywhere, (not) ... anywhere, sometimes, never, always
5. Semi-Deictics
   - Person/Entity:
     - person (e.g., There are five persons present.)
     - man (e.g., There were animals here before man came.)
     - people (e.g., What do people think about the government?; There are five people present.)
   - Thing/Action:
     - thing (e.g., What do you call that thing?)
     - do (e.g., What are you going to do tonight?)

Annotation Task

You will be given a sentence that contains deixis forms. The deixis form is marked with bracket pairs such as [[word]]. Decide the type of deixis of the marked word using one of the following labels:

indef: referring to indefinite deixis, e.g.

- Then with Internet witch arrived before the Web , [[a]] whole network developped with also economic functions on this market.
- In fact, people remember [[more]] informations and historical event when they play than they don't.
- I will most take attention on video games but it is something we can also find in [[other]] aspects of entertainment.

def_anaph: referring to definite anaphoric deixis, anaphoric referring to a deixis that refers to something that has been mentioned previously. e.g.

- My parents found it a bit dangerous and warned me of [[the]] risks of social networks.

[We] can move freely in all Paris and in Paris, a Paris's region abandonned nowadays.

[My] opinion in the invention on the web is that is allowed at the time to start to communicate more easily, to exchange document.

def_non_anaph: referring to definite non-anaphoric deixis. e.g.

- Then with Internet witch arrived before the Web , a whole network developped with also economic functions on [[this]] market.
- I will most take attention on video games but it is something [[we]] can also find in other aspects of entertainment.

[I] would like to discuss about the amazing way people can remember huge informations throught playing.

Definite non-anaphoric forms may includes forms referring to new referents in the context or to exophoric referents.

There are basically two aspects of the deixis you need to consider, definiteness and whether it is anaphoric or not. Definiteness is about whether the deitic word is referring to something specific or general. Anaphoricness is about whether the word is referring to an entity that has been mentioned before. Therefore, for both aspects, you would need to context in which the word is used to decide its definiteness and anaphoricness. I will give you the context of the sentence as well as the text in which the sentence comes from. Use them as the context to decide on the label of the deixis.

Only output the label, nothing else.

Here is the sentence with the marked word:

`{{highlighted_sentence}}`

Here is the text in which the sentence is taken from:

`{{text}}`

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: Mistral Vibe was used to rephrase and improve language.

References

Stefano Bannò, Penny Karanasou, Kate Knill, and Mark Gales. 2026. Exploiting the English Grammar Profile for L2 grammatical analysis with LLMs. ArXiv:2603.17171 [cs.CL].

Bram Bulté and Alex Housen. 2012. *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Francis Cornish. 1999. *Anaphora, Discourse, and Understanding. Evidence from English and French*. Oxford University Press, Oxford.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Vila-Suero Daniel and Aranda Francisco. 2023. Argilla - Open-source framework for data-centric NLP.

Tanguy Dubois, Magali Paquot, and Benedikt Szmrecsanyi. 2023. Alternation phenomena and language proficiency: the genitive alternation in the spoken language of EFL learners. *Corpus Linguistics and Linguistic Theory*, 19(3).

Masaki Eguchi and Kristopher Kyle. 2024. Building custom NLP tools to annotate discourse-functional features for second language writing research: A tutorial. *Research Methods in Applied Linguistics*, 3(3):100153.

J. A. van Ek and J. L. M. Trim. 1998a. Language functions. In *Threshold 1990*, Threshold, pages 27–47. Cambridge University Press, Cambridge.

Jan Ate van Ek and John Leslie Melville Trim. 1998b. *Threshold 1990*. Cambridge University Press, Cambridge (England).

Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford, United Kingdom.

Thomas Gaillat. 2016. *Reference in Interlanguage: the case of this and that. From linguistic annotation to corpus interoperability*. Thesis, Université Paris-Diderot, Paris.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. English Language Series. Pearson Education Limited, Harlow.

Cyriel Mallart, Andrew Simpkin, Nicolas Ballier, Paula Lissón, Rémi Venant, Bernardo Stearns, Jen-Yu Li, and Thomas Gaillat. 2025. Assessing the validity of syntactic alternations as criterial features of proficiency in L2 writings in English. *Research Methods in Applied Linguistics*, 4(3):100238.

Cyriel Mallart, Andrew Simpkin, Rémi Venant, Nicolas Ballier, Bernardo Stearns, Jen Yu Li, and Thomas Gaillat. 2023. A new learner language data set for the study of English for Specific Purposes at university level. In *Proceedings of the 4th Conference on Language, Data and Knowledge - LDK 2023*, volume 1, pages 281–287, Vienna, Austria.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Nelly Sagirov and Xiaobin Chen. 2026. POLKE: A system for comprehensively annotating pedagogically-oriented grammatical structure use in language production. *Research Methods in Applied Linguistics*, 5(3):100336.

Lesley Stirling and Rodney Huddleston. 2002. Deixis and anaphora. In Rodney Huddleston and Geoffrey K. Pullum, editors, *The Cambridge Grammar of the English Language*, pages 1449–1564. Cambridge University Press, Beccles, Suffolk.

Exploring the syntactic parsing of learner texts in Modern Greek: Variety trumps size, genre trumps grammaticality

Christina Klironomou[1] and **Arianna Masciolini**[2] and **Aleksandrs Berdicevskis**[2]
[1]University of Gothenburg
[2]Språkbanken; Department of Swedish, Multilingualism,
Language Technology; University of Gothenburg
[1] gusklirch@student.gu.se, [2]name.surname@gu.se

Abstract

We investigate which properties of train data affect the performance of syntactic parsers on learner texts. We train and evaluate six models for Modern Greek, using the same annotation scheme (Universal Dependencies) and the same parser (*MaChAmp*), but varying train and test sets. The data come from standard texts, learner texts and learner with errors corrected. We confirm that training and testing on the same language variety yields better results than, for instance, training on standard texts and testing on learner texts. Moreover, our experiment setup makes it possible to compare the role of variety to that of other factors and dissect the very notion *variety* into more specific factors. Most importantly, we show that training on the same variety has stronger effect than train set size, and that adding non-target variety to the train set does not necessarily decrease performance. In practical terms, it means that adding learner data (that are often scarce) to the standard train set is likely to boost the performance on learner data without damaging the performance on standard data. Further, we show that the same value of grammaticality across train and test does not have an effect, while having the same genre does.

1 Introduction

Second Language Acquisition (SLA) often relies on learner data annotated automatically for morphosyntax (see among others Kyle and Crossley 2017, Römer 2019, Hwang and Kim 2023, Römer-Barron 2024, Bozdağ et al. 2025 and Geremia et al. 2025). However, the nonstandard and sparse language phenomena that characterize this kind of material, such as transfer errors and other idiosyncrasies, are often challenging for tagging-parsing pipelines, which tend to handle them inconsistently. Annotation errors, in turn, can negatively affect the results of SLA studies. High-quality automatic annotation can further help identify linguistic patterns arising in second language use, as well as to the difficulties and strategies of learners, thus resulting in new, usage-based teaching methods. We therefore highlight the need of developing models that account for this data variability and better describe nonstandard language use.

When it comes to morphosyntax, Universal Dependencies (UD, de Marneffe et al. 2021) is one of the dominant annotation schemes across languages and varieties. Its wide selection of language-agnostic and language-specific categories allows a multi-level analysis that encourages a computer-assisted investigation of intralinguistic phenomena and cross-linguistic comparisons. Furthermore, UD data have been used to train models to perform high-accuracy automatic morphosyntactic annotation for a large number of major languages, and there has been a considerable number of attempts to extend the their capabilities to suitably characterize nonstandard data, e.g. dialectal data (Bompolas et al., 2025; Vakirtzian et al., 2025; Blaschke et al., 2024; Puddu et al., 2026) and user-generated content (Liu et al., 2018; Sanguinetti et al., 2018; Cignarella et al., 2019; Cassidy et al., 2022).

A growing number of studies on learner treebanks annotated in this framework (Berzak et al., 2016; Kyle et al., 2022; Sung and Shin, 2025; Pulido et al., 2025) report significant performance improvements upon integrating learner data into the train set. That the addition of training instances representing the target language variety produces a positive effect is expected. However, it remains to be seen which properties of the language variety are actually important, to what extent the improvement is to be attributed to *grammaticality* (presence or absence of grammatical errors) and

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 130–139, November 19–20, 2026, University of Tartu Library

genre. On the one hand, off-the-shelf parsers tend to underperform in the presence of learner errors, which makes it reasonable to expect that introducing gold-annotated ungrammatical sentences into the training mix would help the models handle them as expected by the annotators. On the other, learner treebanks typically consist of essays, a text genre with little to no presence in the reference datasets normally used to train parsing models. Finally, not all of the above mentioned studies control for train set *size*.

In this paper, we try to disentangle these three aspects. To do that, we conduct a case study on modern Greek using one of the reference large-scale datasets, the *UD_Greek-GDT* treebank (Prokopidis and Papageorgiou, 2017), and *UD_Greek-GLCII* (Klironomou et al., 2026), a parallel treebank that includes both original and corrected learner essays. The latter gives us an opportunity to quantify the effects of text genre independently from grammaticality. In addition, we investigate whether the use of learner data in training produces a decline in performance when the model is used to parse standard modern Greek.

2 Data

In our experiments, we use one of the two released Standard Modern Greek treebanks, *UD_Greek-GDT* (GDT) (Prokopidis and Papageorgiou, 2017). *UD_Greek-GDT* is the UD-converted version of the *Greek Dependency Treebank*,[1] and contains multi-genre texts from the public domain. It was preferred over the other treebanks available for modern Greek due both to its larger size and to the fact that the its internal guidelines served as a starting point for the annotation of the learner treebank.

The *UD_Greek-GLCII* treebank (Klironomou et al., 2026) includes sentences stemming from the development set of the Greek Learner Corpus II (GLCII) (Tantos et al., 2023), an error-annotated corpus of texts from adult learners of Greek as a second language. The corpus covers a wide range of sociolinguistic and cultural backgrounds, with learners of different L1s and CEFR levels. A correction hypothesis for every learner text is also available in the corpus. The annotation of the treebank was manual for the dependency relations and automatic with a manual inspection for all other linguistic levels of analysis (tokenization, lemmatization, POS tagging and feature assignment). Additionally, the treebank contains the annotated correction hypotheses. The original (learner) sentences of the GLCII treebank constitute the *learner-original* data that we used in our experiments, whereas their correction hypotheses constitute the *learner-corrected* data.

3 Experiments

The data we have enable us to partly tease apart factors that can potentially affect the quality of parsing, namely language variety, grammaticality, genre and size.

By *language variety* we mean the difference between standard language and learner language. We say that in our data, language variety is a combination of *grammaticality* and *genre*. By *grammaticality* we mean the presence or absence of any grammatical errors,[2] whereas *genre* is everything else that is different between standard and learner data apart from errors. Thus, genre may include certain stylistic features of the learner language that are not errors, and it subsumes also the difference in topics and formality covered in our learner and standard treebanks.

Given this definition, we say that learner-corrected data and the original learner data belong to the same genre, but have different grammaticality value. Learner-corrected and standard data are both grammatical, but differ in genre.

Given the available data, we ask the following research questions:

1. How does grammaticality affect parsing?
2. Is adding a non-target language variety to the train data harmful? In other words, if we test on language variety A, and have A in the train set, but also add B to the train set, will the performance go down?
3. What is more important for improving performance: training on the same language variety or having a larger train set?
4. What affects the performance more: differences between train and test sets in genre or in grammaticality?

[1] The Greek Dependency Treebank is a manually annotated for morphology, syntax and semantics resource of Modern Greek.

[2] Grammatical errors are only expected in learner texts, given that GDT's sentences stem from texts occurring in established informative websites or European parliament sessions.

In section 3.1, we describe the models that we train, and in Table 1 we show which pairwise comparisons of the model performance enable us to address which questions. In section 3.2, we describe our test settings and explain further the model comparisons that attempt to answer our research questions.

3.1 Models

In order to train our models, we use the syntactic parser of the *MaChAmp* toolkit (van der Goot et al., 2021) with its default hyperparameter values. The train data for the models are selected in specific proportions, to mitigate any data size or variety imbalance and achieve model comparability. We present the trained models below:

- ***learner-small*** (learner-only model): 90 percent (80:training, 10:dev) of the GLCII treebank (8,155 + 1057 tokens).
- ***corrections-small*** (standard-only model, equal train set size —in terms of sentences— to the *learner-small* model): 90 percent (80:training, 10:dev) of the corrected version of the GLCII treebank (8,479 + 1109 tokens).
- ***standard-small*** (standard-only model, equal train set size to the *learner-small* model): Train and development sets of GDT, normalized to the size of the GLCII treebank (7,934 + 1048 tokens).
- ***standard-large*** (standard-only model): Train and development sets of the GDT treebank (42,874 + 10,542 tokens).
- ***mixed-large*** (standard + learner, equal train set size to the *standard-large* model): Combination of the train and development sets of the *standard-large* model, minus the size of the *standard-small* [71.63%], and the *learner-small* model [28.37%] (43,095 + 10,584 tokens). When interpreting the experiments where this model is compared to *standard-large*, we ignore the fact that *standard-large* has seen a little bit more standard train data. In other words, we attribute all potential differences between the models to the presence of learner data in the *mixed-large* train set.
- ***mixed-large-corrections*** (standard + corrections, equal training size to the *standard-large* and *mixed-large* models): Combination of the training and development sets of the *standard-large* model minus the size of the *standard-small* (71.63%), and the *learner-corrections* model (28.37%) (43,419 + 10,636 tokens).

3.2 Test settings and model comparisons

In order to measure whether the occurrence of the learner data in the train set would impact model performance due to their specific and sometimes erroneous nature, we defined two test settings. One was a **learner test setting**, composed of the remaining 10 percent of the GLCII treebank. The other one was a **standard test setting**, composed of a portion of the GDT test set, normalized to the size of the GLCII test set.

In order to test whether different values of grammaticality between the train and test set negatively affect model performance (see research question 1), we compare the results of equal-size models whose train data have dissimilar values of grammaticality; In this way, one model is trained on data that have —at least partly— the same value of grammaticality as the test data, and the other on data with —at least partly— different grammaticality than the test. Thus, we have the below pairwise comparisons:

- *corrections-small* (same grammaticality as test) VS *learner-small* (different grammaticality) on standard test data;
- *corrections-small* (different grammaticality than test) VS *learner-small* (same grammaticality) on learner test data;
- *mixed-large-corrections* (same grammaticality as test) VS *mixed-large* (partly different grammaticality) on standard test data;
- *mixed-large-corrections* (different grammaticality than test) VS *mixed-large* (partly same grammaticality) on standard test data.

In order to test whether the existence of different language varieties between the train and test set have a negative effect on model performance (research question 2), we compare the results of models trained on —at least partly— different language varieties. Thus, we have the below pairwise comparisons:

ID	Train A	Train B	Test	Question	Result
a	corrections-small	learner-small	standard	1	$A \simeq B$
b	corrections-small	learner-small	learner	1	$A \simeq B$
c	mixed-large-corrections	mixed-large	standard	1	$A \simeq B$
d	mixed-large-corrections	mixed-large	learner	1	$A \simeq B$
e	mixed-large	standard-large	standard	2	$A \simeq B$
f	mixed-large	standard-large	learner	2	$A > B$
g	mixed-large	learner-small	learner	2	$A > B$
h	standard-large	learner-small	learner	3	$A < B$
i	corrections-small	standard-small	learner	-	$A > B$
	Difference A	**Difference B**			
j	a (grammaticality)	i (genre)		4	$A < B$

Table 1: Pairwise comparisons of the models. Overview of results with respect to the research questions.

- *mixed-large* (presence of a non-target language variety) VS *standard-large* (target variety only) on standard test data;
- *mixed-large* (presence of a non-target variety) VS *standard-large* (non-target variety only) on learner test data;
- *mixed-large* (presence of a non-target variety) VS *learner-small* (target variety only) on learner test data.

In order to test which of language variety and train set size impacts model performance the most (research question 3), we compare the results of models of unequal training size that have train data of different language varieties; In this way, the small model is trained on data that are of the same variety as the test data (target variety), whereas the other is trained on data of different variety than the test (non-target), but has an important size advantage. Thus, we have the below comparison:

- *standard-large* (different language variety than test) VS *learner-small* (same variety) on learner test data.

Finally, in order to test whether the genre or the grammaticality of the train data is more impactful when it comes to model performance (research question 4), we compare the difference between the scores of two pairs of models:

- *corrections-small* (same grammaticality as test) and *learner-small* (different grammaticality) on standard test data VS *corrections-small* (same genre as test) and *standard-small* (different genre) on learner test data.

Note that this comparison is slightly more complicated than the rest: this is not a comparison of two models, but the comparison of *differences* between one pair of models and another pair of models.

4 Evaluation

4.1 Quantitative evaluation

We perform a quantitative evaluation by calculating **accuracy scores** for **UPOS tagging**, **Unlabeled Attachment Score** (assigned head of the relation) and **Labeled Attachment Score**. Results for the six trained models are presented in Table 2.

To test to what extent random initializations and local maxima can affect the differences between models, we performed a pilot experiment. We trained 10 models on the same train set with different seeds (mixed-large) and then tested them on the standard and the learner test set. A visual summary of the results is reported in Figure 1. We conclude that the variance introduced by random seeds is relatively minor, and that we do not need to perform the multiple runs for every model comparison we do.

To avoid overinterpretation of small differences, we set a threshold of **three percentage points**. This threshold is an educated guess based on the random-seed pilot. There are other potential sources of randomness we do not account for, such as the choice of a parser or the choice of treebanks. In most cases, all three measures either pass or do not pass the threshold. In three cases (g-i), only LAS and/or UAS do, but the difference is in the same direction. We therefore treat these cases as above-threshold.

It can be easily seen from Table 2 that training and testing on the same language variety (target

	learner			standard		
	UPOS	**UAS**	**LAS**	**UPOS**	**UAS**	**LAS**
standard-small	90.06	82.43	75.40	95.23	87.55	83.55
corrections-small	92.17	86.75	81.33	92.39	80.25	74.79
learner-small	92.87	86.85	80.42	90.47	79.40	74.40
standard-large	90.36	85.24	77.21	98.69	92.47	89.93
mixed-large	94.38	89.26	83.73	98.39	93.31	90.47
mixed-large-corrections	93.47	88.55	83.43	98.23	92.47	89.78

Table 2: UPOS, UAS, and LAS scores for the six trained models, evaluated on learner and standard data.

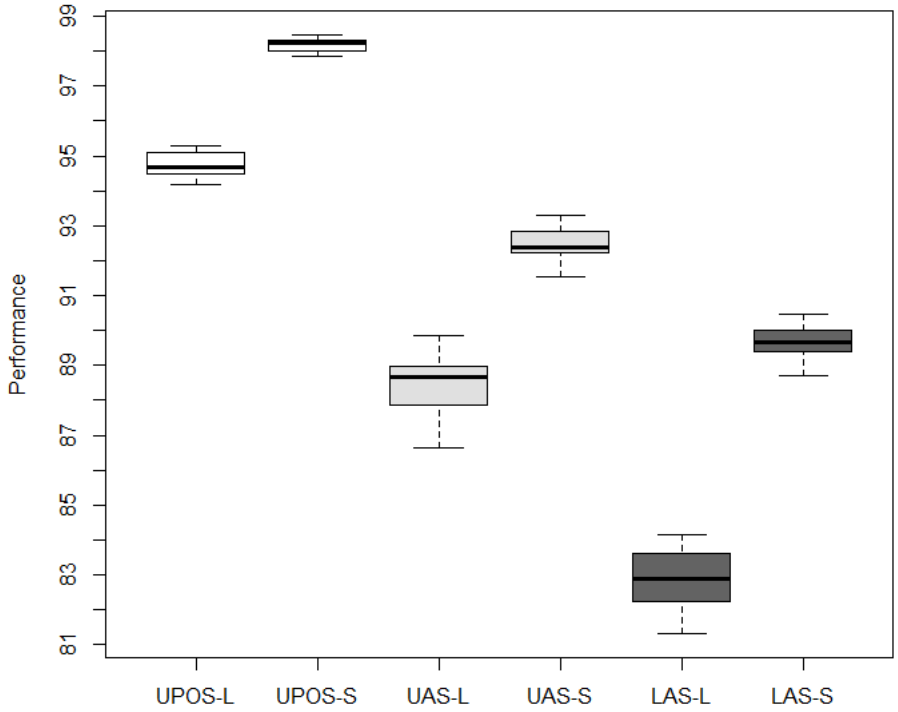


Figure 1: Variance in performance introduced by random seeds for the *mixed-large* model across the three measures. L = learner test set, S = standard test set.

variety) yields best results. Standard models perform better on standard data, while learner models perform better on learner data. Obviously, larger size is also beneficial: large models perform better than small models, all things being equal.

The *corrections-small* model performed similarly to *learner-small* when tested on either standard or learner data (with the *corrections-small* model performing slightly better —but not above the threshold— for all annotation layers on standard test data (score differences below 1 percentage point for UAS and LAS and 1.92 for UPOS) and the *learner-small* model slightly better than *corrections-small* on learner test data) (score differences below 1 percentage point). In addition, the *mixed-large-corrections* model performed almost equally well as *mixed-large* when tested either on learner or standard data (score differences below 1 percentage point). These results indicate that different values of grammaticality do not reduce parsing performance importantly.

The *mixed-large* model performed slightly better than *standard-large* when tested on standard data (less than 1 percentage point for all annotation layers), and robustly better on learner test data (more than 4 percentage points for all annota-

tion layers). Given that *mixed-large* and *standard-large* are comparable in size, we conclude that the presence of learner data (non-target variety) in the train set of *mixed-large* does not harm the performance on standard test data. Conversely, the presence of only standard data (non-target variety) in the train set of *standard-large* reduced performance on learner test data. Moreover, *mixed-large* performed better than *learner-small* when tested on learner data (importantly, more than 3 percentage points for UAS and LAS). Thus, the addition of data of a different language variety (standard) did not harm performance. Given the above results, we can suggest that adding data of a non-target variety does not harm performance (since more data can mean better performance), as long as there are at least some data of the target variety in the train set.

The *learner-small* model performed better than *standard-large* on learner test data (importantly, more than 3 percentage points for LAS). This result points towards that same language variety is more important than size for model performance.

When comparing the results of the *corrections-small* model VS the *learner-small* model to standard test data with the results of the *corrections-small* model VS the *standard-small* model to learner test data, we observe that the score differences between *corrections-small* and *learner-small* are minor (and below-threshold) juxtaposed with the much better performing *corrections-small* against *standard-small* (importantly, more than 4 percentage points for UAS and LAS). Thus, we suggest that differences in genre between the train and test set seem to be more critical than grammaticality for parsing.

Our results regarding research question 1 seem to align with Berzak et al. (2016)'s observation of a "negative, albeit limited effect of grammatical errors on parsing". In addition, model performance, calculated for UPOS, UAS and LAS on two different test settings, one on original learner data and one the corrected versions of the learner data, seem to display similar tendencies to what is encountered in our experiments (grammaticality seems to affect model performance less than domain,[3] as well as size less than domain, when there is no presence of target domain data in the train set).

[3]What multiple authors (e.g. Berzak et al., Di Nuovo et al.) call as *domain* corresponds to our definition of genre.

In Di Nuovo et al.'s experiments, pre-trained models on standard Italian are tested on learner Italian data (original, corrected and a combination of both); best performance is reported for the learner-corrected test data, and worst for the learner-original test data, pointing to a negative influence of ungrammaticality to parsing. However, the authors highlight the influence of domain as more determinate to parsing than grammaticality, which agrees to our results regarding research question 4 (genre trumps grammaticality).

The results of Pulido et al. (2025) initially seem to indicate something different, since the standard-only model performs slightly worse than the standard + learner model, but better than the learner-only model (except for the case of UPOS) when tested on learner data, implying an advantage of size over language variety. However, by performing an experiment on learner test data using a standard + learner model that favors the presence of learner data in the train set (by using subcorpora of the standard-only train dataset), the same authors report that when the learner data are proportionally more prominent, better results are achieved, as the increase of standard train data improves the performance of the model up to a certain point (approx. 40k tokens). This seems to point towards the direction that train data size is less important than variety for model performance, which is compatible to our findings with regards to research question 3.

When it comes to comparing the different annotation layers, robust differences between the paired models are most often observed for LAS followed by UAS, whereas score variation for UPOS is usually minor.

Overall, scores ranged from 74 to 90% accuracy, suggesting a noteworthy performance for all the models across test settings. This can be an indicator of no radical differences among the various train data, independently of their variety, genre or value of grammaticality; even though these variables can noticeably affect the parsing, it seems that the overall similarity among the train sets is enough to achieve adequately high scores.

Furthermore, Klironomou et al. (2026) point out that the language in the annotated learner texts does not deviate dramatically from standard language, possibly due to the high level of proficiency of the majority of the learners (see also their discussion in section 5). We consider this, along with

the consistency of the *GLCII* and *GDT* annotation choices, to be partly responsible for the relatively high scores obtained, even when no target variety data was involved in training.

4.2 Error analysis

To further analyze and interpret the above performance scores, we also conduct a qualitative evaluation, by manually inspecting the model predictions. To facilitate this process, we inspect the predictions for UPOS tagging and dependency relation of each model in comparison to the gold annotation of the test sets (learner or standard).

Overall, predictions of the different models which diverge from the gold annotation often refer to the same structures and sentences. Below, we present some patterns of annotation of the various models, which is different from the gold standard.

Inaccurate predictions concern:

1. Some proper nouns, which are not assigned the UPOS tag `PROPN`. We assume that capitalization errors may have influenced the model's choice of UPOS tag.
2. Some abbreviations that contain punctuation (full stop), which are assigned the UPOS tag `PUNCT`.
3. Some passive past participles (ending in -μένος-η-ο), which are assigned the UPOS tag `ADV` or `ADJ` instead of `VERB`.
4. Some numerical tokens, which are assigned the UPOS tag `NOUN` or `PROPN` instead of `NUM`.
5. Some punctuation, which is assigned the `NOUN` UPOS tag instead of `PUNCT`, when there are consecutive instances of punctuation in the sentence.
6. Some adjectives denoting quantities (e.g. πολύς 'many/much') and some indefinite pronouns (e.g. κάθε 'every'), which are assigned the `ADJ` UPOS tag instead of `DET`.
7. The auxiliary είμαι 'to be', which is attributed a few times the UPOS tag `VERB` instead of `AUX`.
8. The `parataxis` relation, which is either over- or under-used, and generally used interchangeably with the `conj` relation, especially in cases where where rare conjunctions occur in (learner) sentences.
9. The `appos` relation, which is often under- or overused.
10. The `reparandum` and `compound` relations, which are not assigned at all, with the exception of the use of `compound` by the *corrections-small* model to the learner test set (`reparandum` occurs in gold learner test and `compound` in gold standard test).
11. The `expl` relation, which is is under-used (`expl` only occurs in gold learner test).
12. The `vocative`, `discourse` and `flat` relations, which are over- or under-used or not used at all (though `flat` is mostly used correctly by all models in the standard test set).
13. The `orphan` relation, which is overused by *standard-small*.

We identify the following possible causes of errors:

- a learner sentence contains non-standard usage (for instance, wrong stress or spelling, wrong use of case or missing content word that hardens the syntactic analysis)
- a syncretic form is difficult to disambiguate;
- a sentence has an infrequent construction or contains a non-compositional meaning;
- the correct UPOS tag or dependency relation is infrequent in the training data;
- it is difficult to choose between two or more alternatives because the distinction is very subtle or underspecified in the guidelines (cf. cases 5-6 in Table 4 in (Manning, 2011);
- there are several possible annotation choices;
- there are errors in the gold standard.

For subtle or under-specified distinctions, see points 8-9 in the list above as examples.

For infrequent UPOS tags or dependency relations in the training data, see `reparandum` relation and `discourse` in points 9 and 11 above. More specifically, the `reparandum` relation does not occur at all in the standard train set, and `discourse` only a couple of times.

For several possible annotation choices, consider the following examples, where we consider

model predictions which differ from the gold annotation but are still compliant with the UD guidelines: The orphan relation (see point 13) is used differently in standard train (almost for any case of head ellipsis) than learner train (only for elements that cannot be safely connected to another element in terms of syntax or meaning).[4]

Moreover, in the standard treebank, foreign names (often in other alphabets) are left unanalyzed through the attribution of the X UPOS tag (partly following the Universal guidelines for foreign expression and code-switching[5] where this approach is termed "foreign analysis"). In the learner treebank, on the other hand, they are annotated fully, following the guidelines for code-switched analysis (described in the same page of the Universal guidelines); Thus, the models trained on learner or learner-corrected data only do not assign the X tag at all, even though there are occurrences of the tag in the gold annotation of standard test.

For errors in the gold standard, consider the following example: The word θεμελιώδης 'fundamental' is correctly assigned the UPOS tag ADJ by all the models in standard test, whereas in the gold annotation it is assigned the tag NOUN.

5 Conclusions

There are multiple factors that are involved in automatic parsing and can affect model performance. In this study, we tackled factors concerning the nature of the train data in relation to the test data, namely language variety (and more specifically, genre and grammaticality) and size; in order to test to what extent these different variables influence performance, we trained six models for morphosyntactic parsing with standard, learner-original and learner-corrected data.

Models trained on data with different values of grammaticality (*corrections-small* VS *learner-small* and *mixed-large-corrections* VS *mixed-large*) performed similarly across test settings. We therefore suggest that ungrammatical train data do not necessarily affect parsing quality.

Moreover, a comparison of models with train data of (partly) different language varieties (*mixed-large* VS *standard-large* and *mixed-large* VS *learner-small*) tested on data of (partly) the same or different variety indicated an increasing performance trend when more data were added to the training independently of their variety, but *only* with the presence of at least some same-variety data in the train set.

A comparison of models of unequal size (*standard-large* VS *learner-small*), with the small model having the advantage of being trained on same-variety data as the test data, showed a clear prominence of language variety over size; *learner-small* achieved a substantially better performance, despite *standard-large* being five times larger.

Performance differences between the two small models trained on learner data (*corrections-small* and *learner-small*) on the standard test set were larger than those between the two that were trained on grammatical data (*corrections-small* and *standard-small*) and tested on learner data. We interpret this as to indicate that genre affects performance more than grammaticality.

Our qualitative analysis revealed minor annotation differences between the models, some of which relate to learner-specific phenomena. However, we note that not all divergences from the gold standard imply that parser predictions are non-compliant with the UD annotation guidelines. A few possible sources of annotation error were given as well.

All in all, the practical takeaway is that, when it comes to learner data, parsing accuracy benefits from enriching the train set with same-variety data, whose (potential) ungrammaticality does not negatively affect performance on standard data. In fact, the highest attachment scores were obtained upon evaluating the *mixed-large* model on the standard test set. Differences with other large models, however, are too small to draw any conclusions: Whether the addition of (ungrammatical) learner data also improves the overall robustness of the parser remains to be assessed in future studies.

Acknowledgments

CK has been supported throughout the realization of this project by Onassis Foundation, under the Scholarship Program for Greek students 2025-2026. AM and AB have been supported by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161).

[4] For the UD guidelines concerning the orphan relation, see https://universaldependencies.org/u/dep/orphan.

[5] https://universaldependencies.org/foreign

AI Disclosure

☒ Not applicable (no AI tools used)
☐ Assistive use
☐ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: No AI tools were used in preparing this work.

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Stavros Bompolas, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2025. Crossing dialectal boundaries: Building a treebank for the dialect of lesbos through knowledge transfer from standard Modern Greek. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 39–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Fatih Ünal Bozdağ, Junhua Mo, and Gareth Morris. 2025. A corpus-based analysis of noun modifiers in L2 writing: The respective impact of L2 proficiency and L1 background. *Plos one*, 20(3):e0320092.

Lauren Cassidy, Teresa Lynn, James Barry, and Jennifer Foster. 2022. TwittIrish: A Universal Dependencies treebank of tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6869–6884, Dublin, Ireland. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France. Association for Computational Linguistics.

Elisa Di Nuovo, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. 2022. VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 8(8-1).

Sara Geremia, Thomas Gaillat, Nicolas Ballier, and Andrew J. Simpkin. 2025. Exploring the cross-lingual influence of linguistic complexity in second language writing assessment. *Assessing Writing*, 66:100951.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Haerim Hwang and Hyunwoo Kim. 2023. Automatic analysis of constructional diversity as a predictor of efl students' writing proficiency. *Applied Linguistics*, 44(1):127–147.

Christina Klironomou, Thelka Pasparaki, Arianna Masciolini, Alexandros Tantos, Despoina Ourania Touriki, Konstantinos Tsiotskas, and Eleni Tsourilla. 2026. Towards Universal Dependencies for L2 learners of modern Greek: Annotation and challenges. In *Proceedings of the Ninth workshop on Universal Dependencies (UDW, LREC 2026)*.

Kristopher Kyle and Scott Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4):513–535.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'11, page 171–189, Berlin, Heidelberg. Springer-Verlag.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Nicoletta Puddu, Manuela Sanguinetti, and Luigi Talamo. 2026. Introducing Universal Dependencies for Sardinian: the UD ContSar treebank. In *Proceedings of the Ninth Workshop on Universal Dependencies (UDW 2026)*, pages 242–251, Palma de Mallorca, Spain. ELRA Language Resources Association (ELRA).

Emiliana Pulido, Robert Pugh, and Zoey Liu. 2025. I speak for the árboles: Developing a dependency treebank for Spanish L2 and heritage speakers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 814–822, Vienna, Austria. Association for Computational Linguistics.

Ute Römer. 2019. A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, 24(3):268–290.

Ute Römer-Barron. 2024. How do constructions with modal verbs develop in second language learners of English? *Journal of Second Language Studies*, 7(2):198–226.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hakyung Sung and Gyu-Ho Shin. 2025. Towards robust morphosyntactic analysis of l2 korean: Evaluating and fine-tuning a korean language model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.

Alexandros Tantos, Nikolaos Amvrazis, and Eleni Drakonaki. 2023. Greek Learner Corpus II (GLCII): Design and development of an online corpus for L2 Greek. *Journal of Applied Linguistics*, 36.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Predicting Personalized Vocabulary Learning Burden from Intralexical and Interlexical Factors

Mykhailo Diachenko
University of Galway
Galway, Ireland
hi@nekear.me

John P. McCrae
Institute for Data and AI Research (IDAIR)
University of Galway, Galway, Ireland
john.mccrae@universityofgalway.ie

Abstract

Existing measures of word difficulty rarely capture how much effort a particular learner needs to learn a particular word. We introduce personalized learning burden (PLB) — the number of spaced, non-cramming reviews a learner needs to bring a word to 14-day memory stability, extracted from existing review logs without human annotation. In Anki notes and review logs from four multilingual learners of Japanese, non-personalized features explain little of PLB: the lexical item accounts for at most 14.84% of its variance. We therefore ask whether PLB on known words transfers to unseen words through semantic, morphological, orthographic, or phonological similarity. Neighbor-informed models improve over a within-learner median for two of the three learners we test, but the gain does not come from similarity: random words the learner completed within one month of each neighbor reproduce it, and the target's position on the learner's timeline alone predicts best (up to Spearman's $\rho = .557$ and a 16.2% error reduction). Because neighborhoods drawn from a learner's history encode timeline position by construction, studies of interlexical effects need temporally matched controls. We release a dataset of about 66,000 Anki notes and over 670,000 review logs from eight learners with language profiles.

1 Introduction

Reading with adequate text coverage in a second language (L2) requires knowing around 8,000 word families, which comprise about 34,660 word forms (Nation, 2006, as cited in Schmitt, 2008). Vocabulary on this scale cannot be acquired from exposure alone and requires deliberate, sustained study (Schmitt, 2008; Nation, 2022). Flashcards are a well-established tool for such study. Learners who used them regularly gained over 1,000 words in eight months, compared with fewer than 100 for a control group (McLean et al., 2013, as cited in Nation, 2022). Flashcards can be used for receptive (passive) recall practice—retrieving a word's meaning from its form—which was the test type that best predicted course grades in Laufer and Goldstein (2004). Moreover, spacing reviews over time yields medium-to-large retention gains over massed practice (Kim and Webb, 2022). Applications such as Anki (Ankitects Pty Ltd, 2026) (Section 3.1), whose benefits are supported by empirical evidence (Hanson and Brown, 2020), combine these elements with schedulers that adapt review intervals to each learner's performance.

Words, however, differ widely in how much study they require. The effort needed to master a word — its *learning burden* — is highly idiosyncratic: the more a word aligns with patterns a learner already knows, the lighter its burden (Nation, 2022, p. 49). Prior exposure is a major determinant of how difficult a given word is for a given learner (Hashimoto and Egbert, 2019), and relationships with known words act in both directions. Shared structure and cognate overlap make acquisition easier, whereas confusable forms and near-synonyms cause interference that learners may not even notice (Laufer, 1988, 1990, 2025). A word's learning burden therefore depends not only on the word itself, but also on the vocabulary a particular learner has already built.

Existing computational approaches capture complementary parts of this problem: lexical complexity prediction models word difficulty, knowledge tracing models learner histories, and spaced-repetition schedulers model item-level memory (Section 2). However, to our knowledge none jointly predicts the effort a particular learner will need to learn an unseen word from both the word's properties and its relationships to the learner's

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 140–159, November 19–20, 2026, University of Tartu Library

known vocabulary, nor validates such predictions on real multilingual review data.

We introduce *personalized learning burden* (PLB) — the number of spaced, non-cramming reviews a specific learner needs to bring a lexical item to a target memory stability. Stability is the number of days after which the item's recall probability would fall to 90%, and we set the target to 14 days. PLB thus measures memorization effort from review logs that learners already generate, without human annotation. We predict PLB from progressively richer information: the learner's overall and recent performance, features of the target word, and, in neighbour-informed models, the observed PLB of known words that are semantically, morphologically, orthographically, or phonologically similar to the target. Because our difficulty feature must score arbitrary flashcard content for learners with different first languages (L1s) and no receptive-difficulty dataset spans multiple languages, we fine-tune a difficulty prior on productive difficulty scores (British Council, 2026). Productive knowledge is harder to acquire than receptive knowledge (Nation, 2022; Laufer and Goldstein, 2004), so we first test whether the prior transfers to receptive difficulty on MultiLS-Japanese, a dataset of Japanese lexical complexity ratings by non-native readers (Nohejl et al., 2024). We evaluate all models on Anki review logs from multilingual learners of Japanese with diverse L1s.

We address the following research questions (RQs), answering those on PLB for a cohort of four learners of Japanese, three of whom have enough data to evaluate RQ2:

- RQ1: How much of PLB can population-level features explain?
  - RQ1a: Does the productive difficulty prior transfer to receptive population-level difficulty?
  - RQ1b: What share of PLB variance is attributable to the lexical item at all?
  - RQ1c: Do the prior and word frequency from TUBELEX, a corpus of YouTube subtitles (Nohejl et al., 2025), predict PLB?
- RQ2 (main): Does a learner's observed PLB on known words transfer to unseen words through interlexical similarity?

Our contributions are:

1. We introduce and operationalize personalized learning burden (PLB) directly from existing review logs, without additional human annotation.
2. We release a dataset of about 66,000 Anki notes and over 670,000 FSRS (Free Spaced Repetition Scheduler; Ye, 2026) review logs from eight multilingual Anki users with language profiles, pairing receptive review histories with cross-linguistic lexical content.[1]
3. (RQ1) We show that the productive difficulty prior transfers only weakly to receptive difficulty on MultiLS-Japanese ($R^2 = .112$, against .41 for TUBELEX log frequency; Nohejl et al., 2025). On PLB in our four-learner cohort, the lexical item accounts for at most 14.84% of the variance, a ceiling on intralexical features that TUBELEX frequency comes nowhere near ($\approx 1\%$ of total variance). Conditioning on the learner's L1, as the prior does, does not help either (Spearman's $\rho = -.027$).
4. (RQ2) In our cohort, we find no evidence of similarity-specific transfer: neighbour-informed models improve prediction for two of three learners, but temporally matched random neighbours reproduce the gain, and the target's position on the learner's timeline alone predicts best (ρ up to .557, error reduction up to 16.2%).
5. We identify timeline position as a confound for neighbour-based models, since neighbourhoods drawn from a learner's history encode it by construction, and propose a temporally matched random-neighbour control and a days-elapsed baseline for future studies of interlexical effects.

2 Background and Related Work

A word's difficulty may stem from *intralexical* factors — properties of the word itself — or from *interlexical* factors — its relationships to other words the learner knows (Laufer, 1990; Schmitt, 2008; Laufer, 2025). Because the literature draws this boundary inconsistently, we apply a single criterion: a feature is *intralexical* if it can be computed from the L2 word alone, and *interlexical*

[1] Dataset (CC BY 4.0): https://huggingface.co/datasets/nekear/anki-language-learning; code (MIT): https://github.com/nekear/plb.

if its computation requires reference to another word, whether an L2 word the learner knows or a word in their L1.

2.1 Intralexical Difficulty

Frequency is commonly assumed to be the dominant intralexical predictor of difficulty, but the evidence is mixed. In a study of 403 adult L2-English learners and 33 intralexical variables, Hashimoto and Egbert (2019) found frequency to be the strongest single correlate of item difficulty ($r = .499$), yet it explained only about 25% of the variance, and its effect depended on the subcorpus used. Range and dispersion, by contrast, predicted difficulty consistently. The authors interpret all three as facets of one construct: the likelihood of a learner's repeated exposure to a word. Laufer (2025) excludes frequency from intralexical factors altogether, since it reflects the learning context rather than the word itself, and instead lists lack of phonotactic regularity, irregular sound-to-spelling correspondence, deceptive morphological transparency, polysemy, and figurativeness. Because words usually combine several of these characteristics, she argues, difficulty is hard to isolate and determine from intralexical properties alone.

As reviewed by North et al. (2023), computational work builds on such properties, using them as features in models that map lexical items to difficulty scores, a task known as lexical complexity prediction (LCP). Because generic models missed how individual learners perceive complexity, later work personalized them, conditioning on learner attributes such as L1 and proficiency or on judgments elicited from the learner through active learning (Gooding and Tragut, 2022). Today, state-of-the-art predictors are transformer-based, replacing hand-crafted features with learned representations (North et al., 2023). Skidmore et al. (2025) fine-tuned XLM-RoBERTa to predict productive difficulty on an extended version of the Knowledge-based Vocabulary Lists (KVL) dataset (Spearman $\rho = .78$–$.79$), and their model became the baseline of the shared task at the 2026 Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026 Shared Task; British Council, 2026). The winning system of its open track (Nohejl et al., 2026) introduced a soft-target cross-entropy loss for continuous prediction (Pearson $r > .91$) and showed that the same loss makes masked language models competitive with decoder-based large language models (LLMs) of comparable size. All of these approaches, however, estimate the probability that a learner already *knows* a word rather than the *effort required to learn it*, and therefore do not model learning burden.

2.2 Interlexical Effects

Drawing on Meara (1980) and Hatch (1983), Laufer (1990) describes a single mental lexicon shared by all of a multilingual's languages, in which words are grouped first into semantic categories, then into phonological networks, where a word's nearest neighbour is the one it most resembles in sound, and only then into L1 and L2 sets. Within this lexicon, words interact on the basis of semantic, phonological, and morphological similarity, across languages as well as within them. A new word must therefore be integrated into this network, and how easily it is learned depends on its connections to what is already known. Material with no association to previously learned material is difficult to learn (Higa, 1972, as cited in Laufer, 1990). Conversely, *meaningfulness* — the mean frequency of continued associations in 60 seconds (Noble, 1952, as cited in Laufer, 1990) — speeds learning: in studies with nonsense syllables, more meaningful items were learned faster (Noble and McNealy, 1957, as cited in Laufer, 1990).

Association with known words, however, does not always help: similarity can hinder learning as well as facilitate it. Between L1 and L2, *cognates* and *loanwords* (words that usually share both spelling and meaning across the two languages) are the easiest to learn, followed by *one-to-many correspondences* (where one word in one language corresponds to several in the other), since only the form difference must be learned. Difficulty increases with partial overlap and metaphorical extension, and is greatest for *lexical voids* (L2 words whose concept does not exist in the L1), since the learner must form a new concept, and for *false friends* (words similar in form but different in meaning), since a familiar form must be dissociated from a familiar meaning (Laufer, 1990, 2025). Within a language, semantic relations to known words matter as well: near-synonyms add a new meaning link that interferes with existing knowledge, whereas hyponymy, antonymy, and converseness facilitate learning (Laufer, 1990).

Form alone can also interfere: Laufer (1988) identifies ten categories of *synforms* (words similar in form) spanning morphology, phonology, and orthography, among which pairs differing only in suffix (*comprehensive*/*comprehensible*) or vowel (*adopt*/*adapt*) are particularly confusing. Laufer (2025) classifies synformy as intralexical, but because it arises only relative to other known forms, it is interlexical under our criterion.

Taken together, these findings suggest that difficulty arises when the degree of *meaning* similarity (semantic similarity) diverges from the degree of *form* similarity (morphological, orthographic, and phonological).

2.3 Learner and Memory Models

Knowledge tracing (KT) estimates a learner's evolving mastery of skills from their interaction history, using Bayesian or factor-analysis models or deep sequence models extended with memory, attention, graph structure, question text, or forgetting (Abdelrahman et al., 2023). Applied to vocabulary, this framework has three limitations. First, items are linked to skills through a Q-matrix (a binary item-by-skill matrix specifying which skills each item requires), but vocabulary has no natural item–skill mapping, and models that instead learn relations between items derive them from interaction co-occurrence or generic question text rather than linguistic analysis. These relations are also of a single, positive type (similarity or prerequisite), so they cannot express interference such as that between synforms or false friends (§2.2). Second, KT predicts the correctness of the next response rather than the effort required to master an item, which could only be approximated by rolling a forgetting-aware model forward. Third, prediction for items absent from a learner's history remains unaddressed.

Whereas KT tracks mastery of skills, spaced-repetition schedulers model the memory of individual items and use it to decide when each should be reviewed. Following the forgetting curve (Ebbinghaus, 1913), they time each review shortly before an item would be forgotten, reaching a target recall probability with as few reviews as possible. Half-life regression (HLR; Settles and Meeder, 2016) estimates an item's memory half-life from interaction features. It originally also used lexeme-tag features, but these were eventually excluded because they added little predictive value, caused overfitting, and prevented use without a predefined vocabulary. SSP-MMC (Stochastic-Shortest-Path-Minimize-Memorization-Cost; Ye et al., 2022) combines a half-life memory model with a planner that minimizes the expected review cost of bringing an item to a target half-life, but it supports only binary outcomes, lets item difficulty rise after failures without ever recovering, and is fitted at the population level, so it cannot adapt to an individual learner. Its successor FSRS, the scheduler currently recommended for Anki (Ye, 2026), addresses all three limitations: it supports Anki's four-button grading (Section 3.1), updates difficulty in both directions, and can be refitted to an individual learner's history. FSRS represents an item's memory state through *difficulty* and *stability* (the interval after which recall probability falls to 90%), and predicts *retrievability* (the current probability of recall) from stability and the time elapsed since the last review. Review intervals are solved analytically for a configurable desired retention (90% by default), and the changes to difficulty and stability caused by cramming are limited. FSRS, however, remains content-agnostic. It schedules each item from its review history alone, so neither the item's intralexical properties nor its interlexical relations play any role.

3 Approach and Findings

We first introduce the Anki terminology used throughout the paper (Section 3.1) and then describe our dataset and how PLB is derived from it (Section 3.2). We then address RQ1 to establish whether individual personalization is needed at all (Section 3.3). We validate the productive difficulty prior on receptive difficulty, since our models use it as a target-word feature. We also estimate the ceiling on how much of PLB any intralexical feature can explain, test how close TUBELEX log frequency comes to that ceiling, and test whether conditioning on L1 through the prior helps. Finally, we address RQ2 by testing whether PLB on known words transfers to unseen words through interlexical similarity (Section 3.4).

3.1 Anki Terminology

Anki (Ankitects Pty Ltd, 2026) is a flashcard application in which learners create or import their own study material, organized into *decks* of *notes*. Each note stores one lexical item in named *fields*

(e.g., the word, its reading, and its meaning) and generates one or more *cards*, each testing the note in one way: a receptive card shows the L2 word and asks for its meaning, a productive card asks the reverse, and a cloze card asks the learner to fill a gap in a sentence. Cards are scheduled independently. Each time a card is shown counts as a *review*: the learner tries to recall the answer, reveals it, and rates their recall as *Again* (forgotten), *Hard*, *Good*, or *Easy* (ratings 1–4). The scheduler then sets the card's next *interval*, the number of days until it is due again. Forgetting a card that had already been learned is a *lapse*, and the card returns to short *relearning* steps. Anki logs every review with its timestamp, rating, interval, and *type*, such as *learning* (a new card's first steps), *review* (a scheduled review of a learned card), or *relearning* (after a lapse).

3.2 Dataset and Prediction Target

Predicting PLB requires data that combine the lexical content being studied, each learner's ordered and graded review history, and information about the learner's languages. To our knowledge, no public dataset offers all three: LCP datasets lack learning histories (North et al., 2023), knowledge-tracing datasets lack lexical content (Abdelrahman et al., 2023), Duolingo's HLR data (Settles and Meeder, 2016) and the logs from the MaiMemo vocabulary app behind SSP-MMC (Ye et al., 2022) record only binary outcomes, and FSRS-Anki-20k (Open Spaced Repetition, 2024) has ordered, graded Anki histories but neither lexical content nor learner attributes.

We therefore collected a dataset of approximately 66,000 Anki notes (96,000 cards) and over 670,000 associated review logs from eight learners: six studying Japanese, whom we refer to as L01–L06 (L01 also studies Dutch, and L04 English), one studying Italian, and one studying Spanish and Portuguese. Participants uploaded their decks through a web-based survey, in which they also reported their L1s, additional-language proficiencies, and areas of interest, and consented to the public release of their notes, cards, and review logs (Appendix F). Table 6 lists each learner's L1s, self-rated target-language proficiency, interests, and data size, and Table 7 shows an example record from each file of the released dataset.

For our experiments, we retained notes that (1) represent a single-word Japanese lexeme, (2) have a clear form–meaning mapping, so that the receptive practice type can be inferred, and (3) have at most two cards. When two cards are present, we treat them as receptive and productive variants and, because receptive knowledge is easier to acquire (Nation, 2022), keep as the receptive card the one with the easier review history (where *easier* means a lower ratio of review count to achieved FSRS stability). This yielded 24,433 unique lexeme forms (9,478 shared by two or more learners), 42,369 unique form–meaning pairs, and 503,826 review logs.

Prediction Target

For each learner–item pair, we predict PLB — the number of clean reviews the learner needed to bring the item to an FSRS stability of 14 days. Because cramming inflates review counts without producing durable learning, a review is considered clean only if it is a scheduled review of a graduated card (`type = review`; Section 3.1), which excludes learning and relearning steps, and comes at least 24 hours after the card's previous review, which excludes repeated study within a day. PLB is thus a time-to-event quantity measured on a review-count axis rather than a calendar axis: it counts reviews, not days. It still depends on the intervals, however, because FSRS credits a successful review with more stability the longer it has been since the previous one, and therefore on the scheduler settings that chose them (Sections 3.2 and 5). Algorithm 1 in Appendix A gives the extraction procedure and its right-censoring rule.

Figure 1 shows the review timelines of five lexemes shared by all six Japanese learners. Lapses (relearning events) cause FSRS to raise difficulty and lower stability, prompting a much earlier next review than a correct answer would.

Censoring

Not every item reaches the target stability by the end of the collection window. For these items, PLB is undefined, but the number of clean reviews already performed is known, so the observation is not missing - it is *right-censored*, and the true PLB is known to exceed the number of reviews observed so far. Censored items are usually the slowest to master, and censoring rates vary sharply across learners, from 4.6% to 44.5% (Table 4 in Appendix C), so removing them would affect learners unevenly. We separately exclude

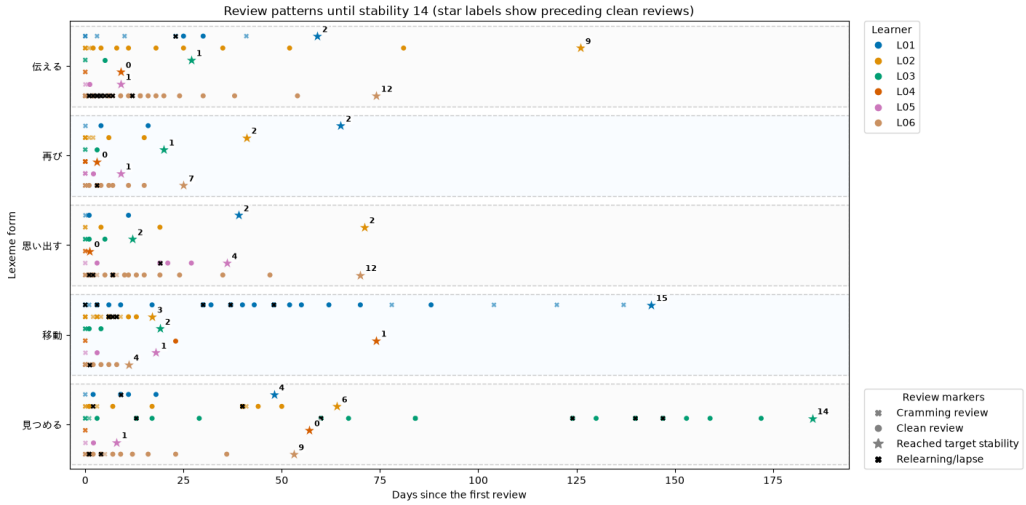


Figure 1: Per-lexeme review timelines for five shared lexemes across six learners. From top to bottom, the lexemes are *tsutaeru* 'to convey', *futatabi* 'again', *omoidasu* 'to recall', *idō* 'movement', and *mitsumeru* 'to gaze at'; within each lexeme, rows are ordered L01–L06. Stars mark PLB, semi-transparent crosses mark clean reviews, and black crosses mark lapses.

cards with zero reviews, since these represent study backlog rather than evidence of difficulty.

Review patterns are highly idiosyncratic (Figure 1): the number of clean reviews needed to reach the target FSRS stability (PLB) varies widely across learners, who require very different review counts to bring the same words to the same stability threshold. Kaplan–Meier burden distributions (Figure 3 in Appendix H) confirm this dispersion. Two learners, L04 and L05, reach the target in one or two reviews for most of their words, reflected in an interquartile range (IQR) of 0 (Table 4 in Appendix C). Based on their FSRS configurations, we attribute this to their FSRS weights and desired-retention settings rather than to genuine learning speed. Because an IQR of 0 means these learners exhibit essentially no within-learner variation in burden, they provide no difficulty signal for a ranking task, which motivates us to exclude them from subsequent processing. This leaves 15,553 unique lexeme forms (5,982 shared among two or more learners), 21,262 form–meaning pairs, and 382,697 review logs (190,619 clean).

3.3 RQ1: Limits of Population-Level Features

We test two population-level predictors of PLB, which describe a word's difficulty for learners in general or for learners who share an L1, without using the individual learner's review history: TUBELEX log frequency (Nohejl et al., 2025), an intralexical feature, and a productive difficulty prior that can score arbitrary flashcard content given the learner's L1.

Productive difficulty prior. We fine-tune mmBERT, a modern multilingual encoder (Marone et al., 2025), on the extended KVL dataset of the BEA 2026 Shared Task (Skidmore et al., 2025; British Council, 2026) (single-word productive difficulty scores from English-L2 learners with Spanish, German, and Chinese L1s), using the soft-target approach and prompt structure of Nohejl et al. (2026). The resulting model, which we call the *productive difficulty prior*, reaches a mean Pearson correlation of .801 and root mean squared error (RMSE) of 1.098 across the three L1s, close to the .833 and .989 reported by Nohejl et al.

RQ1a: Transfer to receptive difficulty. The prior was trained on *productive* difficulty in English, whereas Anki reviews test *receptive* knowledge of other L2s by learners with other L1s. We therefore evaluate it on MultiLS-Japanese (Nohejl et al., 2024), which contains 570 Japanese lexical units presented in context, each rated on a 1–5 scale by 10 annotators with L1s spanning English, Portuguese, French, Spanish, Swedish, and Basque. We use the `full_l2_deploy` prompt (Appendix D.6) because it mirrors an Anki review, showing only the Japanese word and its gloss in the learner's L1. Of six prompt variants (Table 5 in Appendix D), it ranks second by Pearson correlation. The best, `full_l2` (Appendix D.2), reaches $r = .375$ but also requires a sentence

context and a clue, which Anki notes do not reliably provide. On the 491 single words, the prior reaches $r = .334$ ($R^2 = .112$) against aggregated complexity, well below TUBELEX log frequency ($R^2 = .41$, consistent with Nohejl et al., 2024). Productive-to-receptive transfer is therefore limited. Frequency's advantage, however, depends on aggregation: against individual annotators' ratings, its fit falls to $R^2 = .13$ (Nohejl et al., 2024).

RQ1b: Item-level ceiling. PLB is highly idiosyncratic (Figure 1 and Figure 3 in Appendix H). Before testing features on it, we therefore estimate how much of its variance reflects stable differences between lexical items. For a learner i and lexical item j, we decompose PLB as follows:

$$\mathrm{PLB}_{ij} = \underbrace{\mu}_{\text{global baseline}} + \underbrace{\alpha_i}_{\text{learner effect}} + \underbrace{\beta_j}_{\text{item effect}} + \underbrace{\gamma_{ij}}_{\text{learner}\times\text{item interaction}} + \underbrace{\varepsilon_{ij}}_{\text{residual}}. \quad (1)$$

Only β_j represents a stable difference between lexical items across learners, so it is the only component that intralexical features can explain. Learners agree little on shared lexemes (mean pairwise Spearman $\rho = .161$, equivalent Kendall's $W = .371$; Figure 2 in Appendix B), and a mixed-effects model attributes only 14.84% of PLB variance to the lexical item (per-lexeme intraclass correlation). For our cohort, this is the ceiling for intralexical features, which are computed from the L2 word alone and so take the same value for every learner. However, the productive difficulty prior is not bound by it, because it also receives the learner's L1 and an L1 gloss. Therefore, it is interlexical under our criterion (Section 2) and can in principle also explain the part of γ_{ij} shared by learners with the same L1. However, because each learner–item pair has only one PLB observation, we cannot separately estimate a learner-by-item interaction and residual error.

RQ1c: Frequency and the prior. TUBELEX log frequency itself falls far short of this ceiling: it explains 2.6% of the lexeme-level variance, roughly 1% of total variance, and correlates with PLB at $\rho = .012$ within learners and $\rho = -.049$ across shared lexemes. Conditioning on L1 does not help either: the productive difficulty prior, given the first L1 each learner reported (Table 6), is equally uninformative ($\rho = -.027$).

In our cohort, intralexical features could thus explain at most 14.84% of PLB, frequency explains almost none of it, and L1-level conditioning adds no detectable signal. This motivates the individually personalized features examined in RQ2.

3.4 RQ2: Interlexical Transfer from Known Words

Neighbours. Our neighbour-informed models use a target word's neighbours: words the learner began studying earlier that are similar to the target along one or more of the similarity channels.

Therefore, to select neighbours, for a target word w_T and a neighbour w_N, we compute five similarity scores:

- *Semantic* (`sem_sim`): cosine similarity between the two words' definition embeddings from Multilingual E5 Base (Wang et al., 2024).
- *Morphological* (`morph_sim`): because Japanese kanji are morphographic, we segment each word into characters with Fugashi (McCann, 2020) and score character overlap.
- *Kanji orthographic* (`ortho_stroke_sim`): stroke-based edit distance computed with Niteru (Fleming, 2026). We additionally record a coverage indicator (`ortho_stroke_coverage`) marking how much of each pair the metric could score.
- *Kana orthographic* (`ortho_kana_confusion`): visual confusability derived from the Japanese letter-similarity data of Higuchi and Kobayashi (2023). Same as for kanji, we record a coverage indicator (`ortho_kana_coverage`) marking how much of each pair the metric could score.
- *Phonological* (`phon_sim`): both words are transcribed to the International Phonetic Alphabet (IPA) with Epitran (Mortensen et al., 2018) and compared using PanPhon's weighted feature edit distance (Mortensen et al., 2016), which weights phoneme differences by subjective distance, aligning with the confusion patterns Laufer documented for synforms (Laufer, 1988).

For each similarity channel, we estimate the 99.8th-percentile similarity threshold from 50,000

pairs randomly sampled from the cohort's 75.9 million eligible candidate pairs. Each candidate pairs a target with a word that the same learner began studying earlier, and learners are sampled in proportion to their numbers of candidate pairs. We then retain the target's top $K = 6$ neighbours above the threshold in each channel, yielding at most 30 neighbours across the five channels. We chose $K = 6$ heuristically to impose this limit.

Splits. The collected dataset was split into training, evaluation, and test sets (70/15/15) per learner, resulting in 11,337 training rows and 2,327 rows each for evaluation and testing. Learners L01, L02, and L03 were included in all splits, while L06 was included only in the training set because the number of uncensored words (480; Table 4 in Appendix C) was insufficient to calculate reliable confidence intervals (CIs).

Ablations. To test whether a learner's observed burden on known words provides predictive signal for unseen words through interlexical similarity, we construct a sequence of ablations that progressively introduces learner history, target-word features, and neighbour information. The models fall into three groups. In model names, `gbdt` and `st` denote the two model families introduced below, `w` stands for *with*, `nbrs` for *neighbours*, and `cen` for *censored*. Table 1 lists the inputs of each model. All learned models receive the same target-word features — the productive difficulty prior score with the prompt and TUBELEX log frequency — along with the within-learner median PLB, which is also available to the baseline.

Constant (median-based). These establish how much can be predicted without lexical information: (1) `global_median`; (2) `within_learner_median`, which captures the learner's overall burden scale and serves as our primary baseline; and (3) `learner_recent_k_median` ($k = 20$), which predicts the median log PLB of up to twenty of the learner's most recently started training items whose PLB had already been observed by the target's first-review date. If no such items exist, it falls back to the learner's overall median. This tests whether recent performance provides predictive signal beyond the overall scale.

Gradient-boosted decision trees (GBDT; XGBoost-based, Chen and Guestrin, 2016). This family progressively introduces lexical information while representing neighbours through fixed, hand-designed aggregates. (4) `gbdt_no_nbrs` uses target-word features only, testing whether word-level properties of the target improve over the learner-level baseline; (5) `gbdt_w_nbrs` adds aggregates over uncensored, interlexically similar neighbours, testing whether observed burden on known words transfers to unseen words through interlexical similarity; (6) `gbdt_w_nbrs_random` computes the same aggregates after replacing each similar neighbour with a random word from the same learner completed around the same time (within one month); (7) `gbdt_w_nbrs_count_only` keeps only neighbourhood size, the number of distinct similar neighbours; (8) `gbdt_w_nbrs+cen` adds censored similar neighbours, testing whether incompletely observed items provide useful information beyond neighbours whose final PLB is already known; and (9) `gbdt_days` removes the neighbourhood altogether and instead gives the model the target's position on the learner's timeline — the number of days between the learner's earliest training item and the target's first review.

Ablations (6), (7), and (9) follow from the observation that PLB declines over time (Figure 4 in Appendix I). `gbdt_w_nbrs_random` asks whether neighbour similarity matters or whether any set of items completed by the learner in the same period serves equally well as a set of progress markers. `gbdt_w_nbrs_count_only` and `gbdt_days` ask whether even those markers are necessary, or whether neighbourhood size or the target's direct offset from the beginning of the learner's training history is sufficient. Neighbourhood size correlates with days elapsed since the learner's earliest training item at 0.764, 0.848, and 0.675 for L01–L03, respectively. Thus, `gbdt_days`, in this case, asks the same question as `gbdt_w_nbrs_count_only` without routing it through neighbour selection at all.

Set-based (SetTransformer, ST; Lee et al., 2019). This family mirrors the GBDT ablations while representing neighbours individually rather than through fixed aggregates. (10) `st_no_nbrs` serves as the architectural control for `gbdt_no_nbrs`; (11) `st_w_nbrs` introduces individual similar neighbours and tests whether learned attention-based weighting improves on GBDT's hand-designed aggregation; and (12) `st_w_nbrs+cen` tests the same hypothesis as

Model	Median	Target	Similar	Censored	Random	Count	Days	# Features
(1)–(3) Constant baselines	–	–	–	–	–	–	–	–
(4) gbdt_no_nbrs	✓	✓						3
(5) gbdt_w_nbrs	✓	✓	✓					23
(6) gbdt_w_nbrs_random	✓	✓			✓			23
(7) gbdt_w_nbrs_count_only	✓	✓				✓		4
(8) gbdt_w_nbrs+cen	✓	✓	✓	✓				38
(9) gbdt_days	✓	✓					✓	4
(10) st_no_nbrs	✓	✓						3
(11) st_w_nbrs	✓	✓	✓					3 + 13 per neighbour
(12) st_w_nbrs+cen	✓	✓	✓	✓				3 + 14 per neighbour

Table 1: Inputs of each RQ2 model. Median: within-learner median PLB; Target: target-word features (productive difficulty prior, TUBELEX log frequency); Similar / Censored: aggregates (GBDT) or per-neighbour features (ST) over uncensored / censored similar neighbours; Random: the same aggregates as (5), computed over temporally matched random neighbours; Count: number of distinct similar neighbours; Days: days from the learner's earliest training item to the target's first review. All neighbour features are computed at the target's first-review date. Individual features are defined in Table 8 (Appendix G).

gbdt_w_nbrs+cen, but under the set-based representation.

For all neighbour-based models, all neighbour scores (e.g., PLB, FSRS stability, FSRS difficulty) are calculated at the cutoff, defined as the target item's first-review date. Thus, only information available by the time the target is first reviewed is used, preventing leakage of future outcomes into earlier predictions.

Metrics. All metrics are computed within learner. We measure ranking with Spearman correlation and regression error with the mean absolute error (MAE) of log-transformed PLB, $\mathrm{MAE}(\log(1+X))$, where X denotes PLB; the log transform counts the gap between 1 and 3 reviews for more than that between 21 and 23. We report regression gains as the *relative error reduction* (RER) against the within-learner median, $\mathrm{RER} = 1 - \mathrm{MAE}_{\text{model}} / \mathrm{MAE}_{\text{baseline}}$, and compare two models by $\Delta\mathrm{RER}$, the difference between their RERs. We compute 95% CIs by paired bootstrapping over test items. A $\Delta\mathrm{RER}$ whose interval excludes zero indicates a statistically significant difference.

Results. Table 2 reports Spearman correlation and RER for all models, and Table 3 reports $\Delta\mathrm{RER}$ for the paired comparisons that isolate each model component.

Moving down the ablation ladder, the results first confirm that our target-word features provide little signal beyond the within-learner median, which every model receives as a feature. Neither gbdt_no_nbrs nor its SetTransformer counterpart improves reliably over the baseline for any learner — all $\Delta\mathrm{RER}$ CIs include zero (Table 3).

Adding uncensored similar neighbours improves both ranking and error for L01 and L02 (Table 2), and the gain over the no-neighbour models is significant for both learners under either architecture. L03 does not benefit, and under the SetTransformer, neighbours slightly but significantly reduce its accuracy.

However, we cannot attribute the gains to interlexical similarity. gbdt_w_nbrs_random, which preserves the neighbourhood's size, recency, and accrued PLB but replaces its words with random ones, performs almost identically to gbdt_w_nbrs for all three learners, with every $\Delta\mathrm{RER}$ CI including zero (bounds between -1.7 and $+1.9$ percentage points, pp). The signal in the neighbourhood features therefore appears to come from the target's position within the learner's history rather than from its similarity to previously studied words.

Isolating neighbourhood size supports this interpretation. gbdt_w_nbrs_count_only, which retains only the number of neighbours, matches the full similar neighbourhood for L01 and L02 and marginally outperforms it for L03. Because neighbourhood size correlates strongly with the number of days elapsed since the learner's earliest training item, we also evaluated gbdt_days, which uses elapsed days directly. It is the strongest model overall: it achieves the highest RER for all three learners (up to $+16.2\%$) and the highest Spearman correlation for L01 and L03, is on par

Model	L01		L02		L03	
	ρ	RER	ρ	RER	ρ	RER
Constant baselines						
Global median	–	0.0% (0.0, 0.0)	–	−9.6% (−12.9, −6.3)	–	−38.2% (−44.6, −31.8)
Recent-20 median	.138 (.065, .208)	−10.5% (−16.4, −4.8)	−.302 (−.365, −.229)	−30.0% (−35.1, −25.3)	.042 (−.029, .111)	−12.5% (−18.5, −6.7)
GBDT						
No neighbours	.170 (.099, .239)	+1.6% (−1.1, +4.1)	.203 (.134, .273)	+2.9% (−0.0, +5.8)	.101 (.035, .166)	+0.9% (−0.9, +2.6)
+ similar neighbours	.439 (.375, .495)	+9.9% (+6.7, +12.9)	.355 (.290, .412)	+8.4% (+5.0, +11.6)	.177 (.109, .241)	−0.5% (−2.2, +1.2)
+ random neighbours	.430 (.369, .482)	+9.6% (+6.4, +12.5)	.367 (.301, .426)	+8.8% (+5.3, +12.1)	.169 (.102, .237)	−0.0% (−1.5, +1.5)
+ neighbour count only	.410 (.346, .467)	+9.0% (+5.6, +12.0)	.319 (.251, .383)	+7.4% (+3.8, +10.8)	.171 (.105, .238)	+0.7% (−0.8, +2.4)
+ sim. and cen. neighbours	.425 (.362, .479)	+10.3% (+6.9, +13.4)	**.414** (.351, .479)	+11.1% (+7.0, +15.2)	.174 (.110, .233)	−1.0% (−2.7, +0.6)
+ days elapsed	**.557** (.498, .608)	**+16.2%** (+12.3, +19.5)	.412 (.351, .473)	**+12.2%** (+8.2, +16.2)	**.269** (.204, .333)	**+3.6%** (+1.6, +5.6)
SetTransformer						
No neighbours	.210 (.139, .279)	+2.3% (−0.0, +4.6)	.208 (.138, .278)	+2.8% (−0.1, +5.6)	.166 (.100, .233)	+0.5% (−0.2, +1.2)
+ similar neighbours	.437 (.373, .492)	+9.8% (+6.0, +13.2)	.347 (.283, .406)	+7.1% (+2.4, +11.1)	.186 (.121, .255)	−0.8% (−2.0, +0.3)
+ sim. and cen. neighbours	.460 (.396, .513)	+11.2% (+7.7, +14.2)	.408 (.347, .474)	+10.4% (+6.3, +14.6)	.158 (.085, .224)	−0.9% (−2.0, +0.2)

Table 2: Within-learner predictive performance by model ablation. GBDT denotes gradient-boosted decision trees. ρ denotes Spearman rank correlation and RER denotes relative error reduction in $\mathrm{MAE}(\log(1+X))$ against the within-learner median baseline, whose absolute $\mathrm{MAE}(\log(1+X))$ is 0.616, 0.517, and 0.373 for L01–L03 (4.62, 3.81, 1.48 reviews on the raw scale). Values in parentheses are paired-bootstrap 95% confidence intervals (CIs). Bold indicates the best result for each learner and metric.

with the best for L02, and significantly outperforms `gbdt_no_nbrs`, `gbdt_w_nbrs`, and `gbdt_w_nbrs_count_only` for every learner.

That said, censored neighbours provide a modest, learner-dependent additional benefit. They help only L02, where the gain is significant under both architectures but borderline, with lower CI bounds of $+0.1$ and $+0.3$ pp; for L01 and L03, the differences are not significant.

Finally, the choice of architecture has little effect: `st_w_nbrs` and `gbdt_w_nbrs` show no significant difference for any learner.

L03 remains an exception to the neighbour effect: none of the neighbour-based variants improves meaningfully over the within-learner median for it, and only timeline position yields a small but significant gain.

4 Discussion

Performance of population-level features. In our main experiments, we used two population-level features of the target word, but neither TUBELEX log frequency nor the productive difficulty prior was informative about PLB, and adding both to the within-learner median gave no significant improvement, although frequency predicts receptive population-level difficulty well ($R^2 = .41$). We see two reasons for this. First, both features describe an average learner, or at best an average learner with a given L1, whereas PLB describes an individual one, and averaging by itself removes the idiosyncrasy (Nohejl et al., 2024) that PLB retains. Even on MultiLS-Japanese, frequency falls from $R^2 = .41$ on aggregated judgments to $R^2 = .13$ at the individual-annotator level (Nohejl et al., 2024), and for PLB the lexical item itself accounts for at most 14.84% of the variance, which bounds what any non-personalized feature can explain. Second, both features estimate whether a learner is likely to already *know* a word, not the *effort* required to learn it through deliberate review — the resources behind the prior share this focus, as MultiLS-Japanese was built for lexical simplification rather than language learn-

Question	Comparison	L01	L02	L03
Target-word signal	GBDT no neighbours − learner median	+1.6 (−1.1, +4.1)	+2.9 (−0.0, +5.8)	+0.9 (−0.9, +2.6)
	ST no neighbours − learner median	+2.3 (−0.0, +4.6)	+2.8 (−0.1, +5.6)	+0.5 (−0.2, +1.2)
Neighbour signal	GBDT similar − no neighbours	**+8.4** (**+5.5**, **+10.9**)	**+5.5** (**+3.1**, **+7.8**)	−1.3 (−2.7, +0.1)
	ST similar − no neighbours	**+7.5** (**+3.5**, **+10.9**)	**+4.3** (**+0.9**, **+7.6**)	**−1.3** (**−2.4**, **−0.3**)
Similarity effect	GBDT similar − random	+0.4 (−1.3, +1.9)	−0.4 (−1.7, +0.9)	−0.4 (−1.6, +0.7)
Neighbourhood size	GBDT count only − no neighbours	**+7.4** (**+4.9**, **+10.0**)	**+4.5** (**+2.5**, **+6.5**)	−0.1 (−0.8, +0.6)
	GBDT count only − similar	−0.9 (−2.4, +0.7)	−1.0 (−2.6, +0.6)	+1.2 (+0.0, +2.4)
Timeline position	GBDT days − no neighbours	**+14.6** (**+11.1**, **+17.7**)	**+9.3** (**+5.9**, **+12.6**)	**+2.7** (**+1.2**, **+4.3**)
	GBDT days − count only	**+7.2** (**+4.4**, **+9.8**)	**+4.8** (**+1.7**, **+8.0**)	**+2.8** (**+1.4**, **+4.3**)
	GBDT days − similar	**+6.2** (**+3.4**, **+8.9**)	**+3.8** (**+0.8**, **+6.7**)	**+4.0** (**+2.5**, **+5.7**)
Censoring effect	GBDT similar+censored − similar	+0.3 (−1.6, +2.1)	**+2.7** (**+0.1**, **+5.5**)	−0.6 (−1.7, +0.5)
	ST similar+censored − similar	+1.4 (−0.9, +3.9)	**+3.4** (**+0.3**, **+6.8**)	−0.0 (−0.9, +0.9)
Aggregation effect	ST similar − GBDT similar	−0.1 (−2.8, +2.2)	−1.3 (−3.7, +1.1)	−0.4 (−1.8, +1.1)

Table 3: Paired comparisons between RQ2 ablations (GBDT: gradient-boosted decision trees; ST: SetTransformer). Each cell reports the difference in relative error reduction (ΔRER) in percentage points, with paired-bootstrap 95% confidence intervals (CIs) in parentheses. Positive values favour the first model in each comparison; negative values favour the second. Bold indicates that the 95% CI excludes zero.

ing (Alfter, 2025). Moreover, spaced repetition further weakens the usual link between frequency and ease: FSRS schedules every card toward the same target retention regardless of how often the word occurs outside Anki, and learners add cards for words they have not yet acquired, so the frequent words that become cards are disproportionately those that resisted incidental learning.

Performance of personalized features. Our main question was whether a learner's observed burden on *known words* transfers to *unseen words* through interlexical similarity. Our neighbour-informed models improved prediction for L01 and L02, but neighbours selected by semantic, morphological, orthographic, or phonological similarity performed no better than random words the learner completed in the same period. Each subsequent ablation removes more neighbour-specific information, yet performance holds or improves. Neighbourhood size alone performs as well as the full similar neighbourhood, and replacing the neighbourhood with the number of days elapsed since the learner's earliest training item, which correlates with neighbourhood size at .68–.85 across learners, yields the best model overall: `gbdt_days` achieves the highest RER for all three learners and Spearman correlations of .557, .412, and .269 for L01–L03. We therefore interpret the neighbourhood as a proxy for how far the learner has progressed rather than as a description of the target's lexical environment. The longer a learner has studied, the more words can pass the similarity threshold, so neighbourhood size grows with time, and random neighbours drawn from the same period carry the same temporal information.

Why timeline position predicts PLB. All GBDT and SetTransformer models receive the learner's median PLB on the training set, and because they are trained on data pooled across four learners, this feature effectively identifies the learner. Combined with the target's position on the timeline, it lets a model learn each learner's PLB trajectory rather than a single PLB level. Consistent with this, the within-learner median accounts

for 70.9% of the total gain in `gbdt_days`.

This matters in our case because PLB declines markedly over time for L01 and L02 but remains nearly flat for L03 (Figure 4), so for the first two learners, the date a word enters study already carries a large share of the rankable signal. This pattern is also mirrored in where neighbours helped: they improved prediction for L01 and L02 but not for L03, where they even slightly reduced accuracy for the SetTransformer. L03 also has the narrowest burden spread of the evaluated learners (interquartile range, Q_1–Q_3: 2–5 reviews), which leaves less variance to explain. Days elapsed nevertheless improves prediction for L03 slightly but significantly (+2.7 pp over `gbdt_no_nbrs`), whereas neighbourhood size does not, suggesting that direct timeline position also captures period effects, such as changes in study intensity or scheduler settings, that a noisier neighbourhood size cannot.

Why similarity carried no signal. The null similarity effect in our cohort does not show that interlexical effects are absent. We hypothesize that our features did not represent them in the form in which they affect learning. The literature reviewed in Section 2.2 predicts both facilitation (cognates, shared morphemes, meaningful associations) and interference (synforms, false friends, near-synonyms), and locates difficulty where form similarity and meaning similarity diverge. Our GBDT features aggregate each channel independently, so a form-similar neighbour that facilitates learning because its meaning is related and one that interferes because its meaning is unrelated enter the aggregates in the same way, and their effects may cancel out. The SetTransformer sees per-neighbour channel percentiles and could in principle represent such mismatches, but did not learn to do so at the present data scale. Moreover, a neighbour's PLB records how hard that word was for the learner, not whether it was confused with the target, and review logs do not record which word a learner mistook for which. We therefore hypothesize that similarity channels alone do not expose confusion patterns to the model.

How censoring and architecture affect performance. Including censored neighbours significantly improved prediction only for L02, and only marginally (+2.7 pp for GBDT and +3.4 pp for the SetTransformer, with lower CI bounds near zero). Because censored items tend to be the slowest to master, this suggests that they carry some information about the difficult tail that uncensored neighbours miss, though not much at this sample size. As for the architecture, the SetTransformer did not reliably outperform GBDT despite representing neighbours individually. We hypothesize that attention-based weighting has little to make use of when the useful signal is the neighbourhood as a whole, as a progress marker, rather than lying in individual neighbours, and that 11,337 training rows are too few for its additional flexibility to pay off.

Why absolute prediction remains hard. Even our best model ranks items only moderately well ($\rho \leq .557$) and reduces log-scale error by at most 16.2% relative to the within-learner median. Part of the reason lies in the PLB itself, which is an intrinsically noisy target: unlike population-level difficulty, it depends on the learner's behaviour, memory state, and scheduler configuration, each stochastic lapse adds reviews before the target is reached, and PLB shifts within a learner over time (Section 5).

5 Limitations and Future Work

Data and scope. Our evaluation covers three learners, one target language, and single-word lexemes only. The learners differ in L1 and self-rated proficiency (Table 6), but our cohort is too small to estimate the effect of either. The only L1-conditioned feature we test, the productive difficulty prior, is uninformative (Section 3.3). Multiword expressions, which appear both in learners' decks and in MultiLS-Japanese, are excluded. We also predict PLB only for unseen words of known learners: with three evaluable learners, a leave-one-learner-out evaluation of how the models generalize to unseen learners is not feasible. Our released dataset also covers learners of Italian, Spanish, and Portuguese, but both extensions require a larger cohort.

Scalability. Computing PLB requires no annotation and a single pass over each card's review log (Algorithm 1), so its cost grows linearly with the number of reviews: processing our 503,826 review logs sequentially takes 1 min 34 s on an Intel Core i5-13600K CPU. Embedding the definitions of all 42,369 form–meaning pairs with Multilingual E5 Base takes 1 min 37 s on an NVIDIA GeForce RTX 4070 Ti (batch size 64). The dominant cost is neighbour selection: each target is compared

with every word the same learner began studying earlier, so the number of candidate pairs grows quadratically with a learner's deck size but only linearly with the number of learners (75.9 million pairs for our cohort, 30 min with 20 parallel workers on the same CPU). Training a GBDT or Set-Transformer model on our 11,337 training rows takes about a minute. Adding learners is therefore limited by data collection rather than computation.

Temporal confounds. PLB declines over time for L01 and L02 (Figure 4). L01's per-word PLB, for instance, reached 88 reviews early in the collection window but fell below 10 by its end. We did not test what drives this decline. It may reflect growing proficiency (e.g., a larger stock of known kanji), a change in the kinds of words added, or right-censoring, since an item first reviewed late in the window can reach the target before the window closes only if it needs few reviews. It may also reflect the scheduler: Anki lets learners change FSRS settings and re-optimize its parameters at any time, but our data do not record when such changes occurred, so we cannot tell how they shaped PLB. More generally, PLB depends on the scheduler's configuration, as the exclusion of L04 and L05 shows, and on the accuracy of FSRS's stability estimates. Future work could reduce this dependence, for example by counting only reviews taken close to their scheduled interval (e.g., within 0.8–$1.25\times$), so that overdue reviews after a break do not inflate stability gains. Overall, separating these explanations is necessary before the timeline effect can be read as learning progress.

Features and representation. Our target-word representation is limited to two features, one of which is a prior trained on productive rather than receptive difficulty, for a different L2 and different L1s than ours. Other factors reviewed in Section 2.1, such as polysemy or figurativeness, remain untested. Neighbour selection relies on heuristic choices, the 99.8th-percentile threshold and $K = 6$, whose sensitivity we did not analyse. Finally, similarity percentiles and neighbour PLB may be too coarse to capture confusion: they record how similar a neighbour is and how hard it was to learn, but not whether the learner confused it with the target. Three extensions could close this gap. First, data that record confusions directly, such as the wrong answers learners give, which Anki review logs do not store, would show which known words interfere with a given target. Second, more granular representations of each target–neighbour pair, such as part-of-speech tags and the specific kanji, radicals, or readings the two words share or differ in, would let models relate form similarity to meaning similarity and separate neighbours that facilitate learning from those that interfere. Third, a larger dataset would provide more target–neighbour examples and make these finer-grained interactions easier for models to learn.

6 Conclusion

We asked whether a learner's personalized learning burden (PLB) on unseen words can be predicted from the burden they showed on interlexically similar words they already know. Across semantic, morphological, orthographic, and phonological channels and three evaluable learners, we find no evidence that it can.

On our data, the population-level features we tested could not substitute for personalization: the lexical item accounts for at most 14.84% of PLB variance, log frequency explains about 1%, and the L1-conditioned prior, which transferred only weakly to receptive difficulty, is uninformative. Neighbour-informed models improved over the within-learner median for L01 and L02, but temporally matched random neighbours reproduced the gain and neighbourhood size alone matched it. The strongest model used no neighbours at all: days since the learner's earliest training item achieved the highest RER for every learner (up to $+16.2\%$; ρ up to $.557$). We therefore attribute the neighbour gain to the target's position on the learner's timeline rather than to what the neighbourhood contains.

Two implications follow. First, because a neighbourhood drawn from a learner's history encodes timeline position by construction, future tests of interlexical transfer should control for it, as our random-neighbour baseline does. Second, similarity alone may not expose confusion: modelling how form and meaning similarity interact for a particular learner is the more promising next step. To support such work, we release a dataset of approximately 66,000 Anki notes and over 670,000 review logs from eight learners with multi-year histories, each with a profile of their languages and areas of interest.

AI Disclosure

Following the taxonomy of Kamocki and Witt (2026), we used AI tools for *assistive use*, *editing*, *generation of minor elements*, and *compression/expansion*. AI tools were not used to design the study or interpret the results. Our use of AI coding assistants is documented in the repository's README.

Assistive use. Claude Sonnet and Claude Opus (versions 4.6–5.5; Anthropic) were used for topic discovery and literature search (the authors read and verified all cited works). Before submission, GPT-5.6 Sol (OpenAI) reviewed the manuscript to flag remaining inconsistencies. During revision, Claude Opus 5.5 helped address reviewer comments by identifying undefined abbreviations and inconsistencies and proposing table layouts.

Editing. The same Claude models were used for grammar and spelling correction and for improving flow, style, and tone. GPT-5.6 Terra was used to format LaTeX.

Generation of minor elements. During revision, Claude Opus 5.5 drafted short passages, such as the Anki terminology section and table captions, which the authors revised.

Compression/expansion. Claude was used to shorten some author-written sections to meet the page limit.

The authors reviewed all AI-assisted text and take full responsibility for the content of the paper.

Acknowledgements

John P. McCrae is supported by Research Ireland as part of the Rinn AI Centre and Grant Number 13/RC/2106_P2, ADAPT Research Centre.

References

Gehad Abdelrahman, Qian Wang, and Miguel A. Nunes. 2023. Knowledge Tracing: A Survey. *ACM Computing Surveys*, 55(11):1–37.

David Alfter. 2025. The need for truly graded lexical complexity prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 326–333, Vienna, Austria. Association for Computational Linguistics.

Ankitects Pty Ltd. 2026. Anki: Powerful, Intelligent Flashcards. https://apps.ankiweb.net/. Accessed: 2026-06-21.

British Council. 2026. BEA 2026 Shared Task: Vocabulary Difficulty Prediction for English Learners. https://www.britishcouncil.org/data-science-and-insights/bea2026st. Accessed: 2026-06-21.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794. ACM.

Hermann Ebbinghaus. 1913. *Memory: A Contribution to Experimental Psychology*. Teachers College, Columbia University, New York. Translated from the original 1885 German edition by Henry A. Ruger and Clara E. Bussenius.

Kyle T. Fleming. 2026. Niteru: Search for Japanese Kanji by Visual Similarity. https://github.com/ktfleming/niteru. Accessed: 2026-07-31.

Sian Gooding and Manuel Tragut. 2022. One Size Does Not Fit All: The Case for Personalised Word Complexity Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365.

Aroline E. Seibert Hanson and Christina M. Brown. 2020. Enhancing L2 learning through a mobile assisted spaced-repetition tool: an effective but bitter pill? *Computer Assisted Language Learning*, 33(1-2):133–155.

Brett J. Hashimoto and Jesse Egbert. 2019. More Than Frequency? Exploring Predictors of Word Difficulty for Second Language Learners. *Language Learning*, 69(4):839–872.

Hiroki Higuchi and Tessei Kobayashi. 2023. Letter visual similarity of Japanese hiragana and katakana based on reaction times. *Current Psychology*, 42(15):12835–12844.

Pawel Kamocki and Andreas Witt. 2026. Authorship attribution in the times of LLMs within the framework of the CRediT taxonomy. In *Proceedings of the Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Computational Approaches to Language Data Pseudonymization, Anonymization, De-identification, and Data Privacy (LEGAL2026 and CALD-pseudo 2026) @ LREC 2026*, pages 35–39, Palma, Mallorca (Spain). ELRA.

Su Kyung Kim and Stuart Webb. 2022. The Effects of Spaced Practice on Second Language Learning: A Meta-Analysis. *Language Learning*, 72(1):269–319.

Batia Laufer. 1988. The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education*, 2(2):113–132.

Batia Laufer. 1990. Words you know: How they affect the words you learn. In Jacek Fisiak, editor, *Further Insights into Contrastive Linguistics*, pages 573–593. John Benjamins, Amsterdam.

Batia Laufer. 2025. Second-Language Word Difficulty. In *The Encyclopedia of Applied Linguistics*, pages 1–6. John Wiley & Sons, Ltd.

Batia Laufer and Zahava Goldstein. 2004. Testing Vocabulary Knowledge: Size, Strength, and Computer Adaptiveness. *Language Learning*, 54(3):399–436.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmBERT: A Modern Multilingual Encoder with Annealed Language Learning. *arXiv preprint arXiv:2509.06888*.

Paul McCann. 2020. fugashi, a Tool for Tokenizing Japanese in Python. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51, Online. Association for Computational Linguistics.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for Many Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL.

I.S.P. Nation. 2022. *Learning Vocabulary in Another Language*, 3 edition. Cambridge University Press, Cambridge, UK.

Adam Nohejl, Akio Hayakawa, Yusuke Ide, and Taro Watanabe. 2024. Difficult for Whom? A Study of Japanese Lexical Complexity. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 69–81, Miami, Florida, USA. Association for Computational Linguistics.

Adam Nohejl, Frederikus Hudi, Eunike Andriani Kardinata, Shintaro Ozaki, Maria Angelica Riera Machin, Hongyu Sun, Justin Vasselli, and Taro Watanabe. 2025. Beyond Film Subtitles: Is YouTube the Best Approximation of Spoken Vocabulary? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9566–9585, Abu Dhabi, UAE. Association for Computational Linguistics.

Adam Nohejl, Xuanxin Wu, Yusuke Ide, Maria Angelica Riera Machin, Yi-Ning Chang, and Hitomi Yanaka. 2026. Sakura at BEA 2026 Shared Task 1: What Makes Vocabulary Difficult? In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1161–1178, San Diego, California, USA. Association for Computational Linguistics.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42.

Open Spaced Repetition. 2024. FSRS-Anki-20k (Revision d2e5351). Hugging Face.

Norbert Schmitt. 2008. Instructed second language vocabulary learning. *Language Teaching Research*, 12(3):329–363.

Burr Settles and Brendan Meeder. 2016. A Trainable Spaced Repetition Model for Language Learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858.

Lucy Skidmore, Mariano Felice, and Karen Dunn. 2025. Transformer Architectures for Vocabulary Test Item Difficulty Prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 160–174, Vienna, Austria. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

Jarrett Ye. 2026. The FSRS Spaced Repetition Algorithm. `https://github.com/open-spaced-repetition/awesome-fsrs/wiki/The-Algorithm`. Open Spaced Repetition FSRS wiki; accessed: 2026-06-21.

Junyao Ye, Jingyong Su, and Yilong Cao. 2022. A Stochastic Shortest Path Algorithm for Optimizing Spaced Repetition Scheduling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)*, pages 4381–4390, Washington, DC. Association for Computing Machinery.

A Personalized Learning Burden Computation

Algorithm 1 Personalized learning burden of learner i on item j.

Require: review log $r_1, \ldots, r_n$ of the item's card, sorted by time t_k; each r_k has a type τ_k, a rating $g_k \in \{1, 2, 3, 4\}$ (*Again*–*Easy*; Section 3.1); FSRS parameters $\mathbf{w}_i$ of the learner's deck; target stability $S^* = 14$ days
Ensure: PLB_{ij}, or a right-censored lower bound on it

$C \leftarrow \langle\,\rangle$ ▷ clean reviews
for $k = 2, \ldots, n$ **do**
  $\Delta_k \leftarrow t_k - t_{k-1}$ in days
  if $\tau_k = \texttt{review} \wedge \Delta_k \geq 1$ **then**
    append r_k to C
  end if
end for
if $C = \langle\,\rangle$ **then**
  return excluded ▷ backlog: no evidence
end if
$\sigma \leftarrow$ new FSRS memory state
for $m = 1, \ldots, |C|$ **do** ▷ replay clean reviews only
  $\sigma \leftarrow \mathrm{FSRS}_{\mathbf{w}_i}\big(\sigma, g(c_m), t(c_m)\big)$ ▷ c_m: m-th element of C
  if $S(\sigma) \geq S^*$ **then**
    return $\mathrm{PLB}_{ij} = m$
  end if
end for
return censored: $\mathrm{PLB}_{ij} > |C|$

B Learner Agreement on Shared Lexemes

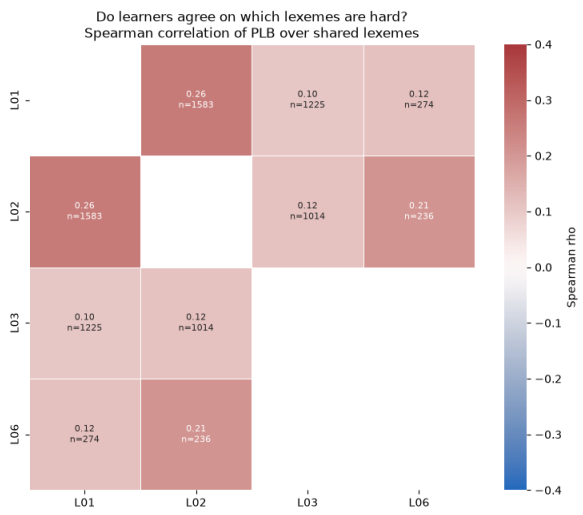


Figure 2: Pairwise Spearman agreement between learners on shared lexemes.

C Censoring and Review-Count Spread

Learner	Reached	Censored	Cens. %	Q_1–Q_3
L01	5,191	3,525	40.44	3–8
L02	4,594	485	9.55	3–10
L03	5,726	876	13.27	2–5
L04	10,612	508	4.57	1–1
L05	1,253	308	19.73	2–2
L06	480	385	44.51	3–8

Table 4: Items that reached the target stability, censored items, censored share, and interquartile range of clean review counts (Q_1–Q_3) by learner, over word instances with at least one clean review.

D Prompt Templates

Table 5 compares the six prompt variants listed below on the 491 single words of MultiLS-Japanese. For prompts using L1-translated context or a gloss, we queried GPT-5.6 Terra with the translation prompt provided in Appendix E.

Rank	Prompt	Pearson	RMSE
1	**`full_l2`**	**.375**	1.050
2	`full_l2_deploy`	.334	1.058
3	`full_l2_no_context`	.317	**1.006**
4	`full_l1`	.192	1.102
5	`ja_word_only`	.164	1.297
6	`l1_name_only`	.151	1.325

Table 5: Pearson correlation and RMSE of the productive difficulty prior against aggregated receptive complexity on MultiLS-Japanese, by prompt variant, ranked by Pearson correlation. Bold marks the best value in each column. `full_l2_deploy` is used in all experiments.

D.1 `full_l1`

```
Rate how easy it is for learners to guess the
↪  Japanese word based on the {l1_name} word,
↪  context and clue on a scale from 1 to 5 (1=very
↪  easy, 5=very difficult).
{l1_name} word: {l1_gloss}
{l1_name} context: {l1_context}
Clue: {clue}
Japanese word: {target}
Difficulty: 1
```

D.2 `full_l2`

```
Rate how easy it is for learners to guess the
↪  {l1_name} word based on the Japanese word,
↪  context and clue on a scale from 1 to 5 (1=very
↪  easy, 5=very difficult).
Japanese word: {target}
Japanese context: {context}
Clue: {l1_clue}
{l1_name} word: {l1_gloss}
Difficulty: 1
```

D.3 `ja_word_only`

```
Rate how difficult the Japanese word is for
↪  learners on a scale from 1 to 5 (1=very easy,
↪  5=very difficult).
Japanese word: {target}
Difficulty: 1
```

D.4 l1_name_only

```
Rate how difficult the Japanese word is for
↪  learners whose native language is {l1_name} on a
↪  scale from 1 to 5 (1=very easy, 5=very
↪  difficult).
Japanese word: {target}
Difficulty: 1
```

D.5 full_l2_no_context

```
Rate how easy it is for learners to guess the
↪  {l1_name} word based on the Japanese word and
↪  clue on a scale from 1 to 5 (1=very easy,
↪  5=very difficult).
Japanese word: {target}
Clue: {l1_clue}
{l1_name} word: {l1_gloss}
Difficulty: 1
```

D.6 full_l2_deploy

```
Rate how easy it is for learners to guess the
↪  {l1_name} word based on the Japanese word on a
↪  scale from 1 to 5 (1=very easy, 5=very
↪  difficult).
Japanese word: {target}
{l1_name} word: {l1_gloss}
Difficulty: 1
```

E L1 Cue-Generation Prompt

```
You create L1-side vocabulary cues for a controlled
↪  cross-lingual vocabulary-difficulty experiment.

The Japanese context contains exactly one span
↪  marked <TARGET>...</TARGET>. For every
↪  requested L1:

1. Translate the complete Japanese context
↪  naturally into that L1 as l1_context,
↪  preserving meaning and register.
2. l1_gloss must represent only the marked target
↪  expression. Never translate or paraphrase the
↪  surrounding sentence or clause into l1_gloss.
3. The literal text between <TARGET> and </TARGET>
↪  is the complete target boundary. Text outside
↪  the tags is context only, even when an adjacent
↪  particle, auxiliary, affix, or punctuation mark
↪  is orthographically attached. Never absorb
↪  outside material or its semantic contribution
↪  into l1_gloss.
4. If the L1 necessarily realizes the marked target
↪  jointly with adjacent outside material, keep
↪  l1_gloss target-only and use discontinuous,
↪  paraphrastic, or uncertain. Never label that
↪  joint realization exact or inflected, and
↪  explain the fusion in alignment_note.
5. Return exactly one best context-specific gloss.
↪  Do not return alternatives separated by
↪  semicolons, slashes, pipes, newlines,
↪  parentheses, or lists.
6. Prefer a common natural equivalent over an
↪  unnatural equivalent selected only to preserve
↪  Japanese part of speech. A single natural
↪  multiword equivalent is allowed.
7. If the marked target has a standalone
↪  context-appropriate L1 equivalent, set
↪  gloss_kind to translation.
8. If the marked target is a function word and no
↪  standalone L1 equivalent exists, set gloss_kind
↪  to metalinguistic. Give one brief L1-language
↪  description of the target's grammatical or
↪  discourse function, set alignment_kind to
↪  not_applicable, and set l1_target_surface to
↪  the empty string. Do not import the speech act
↪  or lexical meaning of the surrounding clause.
9. For translation glosses, use exact or inflected
↪  only when l1_target_surface is a contiguous
↪  realization of the marked target alone in
↪  l1_context, and copy that surface verbatim. Use
↪  discontinuous, paraphrastic, or uncertain when
↪  appropriate.
10. Do not estimate difficulty, simplify for
↪  learners, infer a score, or transliterate
↪  merely to preserve Japanese form.
11. Return exactly one cue for each requested
↪  language and no additional languages.
```

F Data Collection and Release

Participants uploaded their Anki decks (`.apkg` files; Section 3.1) through a web-based survey, in which they also reported their L1s, additional-language proficiencies, and areas of interest. Throughout, participants kept control over what they shared. Decks were parsed in the participant's browser, where they could preview the content and exclude individual notes, fields, and tags before anything was uploaded. Submission required explicit consent to the public release of the notes, cards, and review logs, while releasing the profile information required a separate, optional consent. In the released dataset, identifiers and deck names are replaced with synthetic ones, and each learner's timestamps are shifted by a random constant offset.

The privacy guidelines are available at `https://nekear.me/research/overview`, and the source code of the survey, including the in-browser deck parser, at `https://github.com/nekear/vocabulary-acquisition-survey` (MIT licence).

Table 6 summarizes the eight learners' language profiles and data sizes. The learners of Japanese, L01–L06, form the cohort of our experiments: L04 and L05 are excluded from modelling for lack of within-learner variation (Section 3.2), and L06 is used for training only (Section 3.4). The dataset consists of six files linked by identifiers: learner profiles, decks with their scheduler settings, note types with their field names, notes, cards, and reviews. Table 7 shows one record from each file.

ID	L1	Target (self-rated)	Interests	Notes	Cards	Reviews
L01	fr, en	ja (unsure), nl (unsure)	technology	15,832	30,253	200,496
L02	hi, en	ja (unsure)	travel, academic, business, legal, technology, literature, music, pop culture	6,720	7,113	143,227
L03	en	ja (B2)	travel, pop culture, other	9,658	9,658	77,362
L04	id	ja (B2), en (C1)	technology, pop culture, music	13,561	13,561	116,165
L05	en, fr	ja (B1)	technology	8,653	15,008	7,285
L06	en	ja (A1)	pop culture, technology, literature	2,312	2,312	27,595
L07	zh-Hant, en	it (A2)	technology	8,523	17,248	96,046
L08	en	es (B1), pt (A1)	literature, travel	965	1,106	10,206

Table 6: Learner profiles and data sizes in the released dataset. L1 and target languages are given as ISO 639-1 codes (zh-Hant: Chinese in Traditional script). Target-language proficiency is self-rated on the scale of the Common European Framework of Reference for Languages (CEFR, A1–C2) or reported as unsure. Interests are the domains learners selected in the survey.

File	Example record (learner `L02`)
`learners`	`native_languages: [hi, en]; target_languages: [{lang: ja, proficiency: Unsure}]; domain_interests: [travel, academic, ..., pop_culture]`
`decks`	`deck_id: 7, parent_id: 3, scheduler: FSRS, desired_retention: 0.9, fsrs_weights: [0.033, 0.196, 0.672, ...], target_language: ja`
`note_types`	`model_id: 2, field_names: [VocabKanji, VocabFurigana, VocabDef, SentEng, SentKanji, SentFurigana, ...], excluded_field_indexes: []`
`notes`	`note_id: 1677, model_id: 2, fields: [見 方, 見 方[み か た], view, "I began [...] abroad.", 留学してから、ものの<b>見方</b>が変わった。, ...], tags: [JLPT::N3, 単語N2]`
`cards`	`card_id: 2070, note_id: 1677, deck_id: 7, state: 2, interval: 4, reps_total: 8, lapses_total: 1, fsrs_stability: 4.07, fsrs_difficulty: 8.05`
`reviews`	`card_id: 2070, timestamp_ms: 1733935081425, rating: 3, interval_before: 0, interval_after: 1, time_taken_ms: 5366, review_type: 0` `...29 d → rating: 1 (lapse) → interval_after: −600 s (relearning) → ...`

Table 7: One record from each file of the released dataset, for learner L02: a note, one of its cards, and that card's reviews (...: values shortened). Card `state` and `review_type` follow Anki's codes (Section 3.1): state 2 is a review card and review type 0 a learning review. Positive intervals are in days and negative ones in seconds. `excluded_field_indexes` lists the fields the learner withheld at upload.

G Neighbour Features

Table 8 lists the input features of the learned RQ2 models. *Group* names the feature block and the models that use it, *Features* lists its features, and n gives their number. GBDT models summarize the neighbourhood with fixed-length aggregates, computed either per similarity channel (one feature for each of the five channels) or across channels (over the union of all selected neighbours). Censored neighbours have no observed PLB, so the censored block omits the log-PLB means and instead adds the share of selected neighbours that are censored. Model 7 uses only the distinct-neighbour count of the uncensored block. The SetTransformer receives one feature vector per neighbour, in which `s_missing` marks neighbours not reviewed before the target's first review. All neighbour features are computed at the target's first-review date.

Group	Features	n
Target and learner (all learned models)	Within-learner median PLB; productive difficulty prior score (`full_l2_deploy`); TUBELEX log frequency; days elapsed (model 9 only)	3 (+1)
GBDT, uncensored neighbours (models 5, 6, 8)	Per channel: neighbour count, mean similarity percentile, mean log PLB	15
	Across channels: distinct neighbours, mean best-channel percentile, mean log FSRS stability, mean FSRS difficulty, mean log PLB	5
GBDT, censored neighbours (model 8)	Per channel: neighbour count, mean similarity percentile	10
	Across channels: distinct neighbours, censored share, mean best-channel percentile, mean log FSRS stability, mean FSRS difficulty	5
SetTransformer, per neighbour (models 11, 12)	PLB, review count, log FSRS stability, FSRS difficulty, `s_missing`; similarity percentiles for the five channels, `phon_missing`; kanji and kana orthographic coverage indicators; `is_censored` (model 12 only)	13 (+1)

Table 8: Input features of the RQ2 models, by feature block.

H Kaplan–Meier Burden Distributions

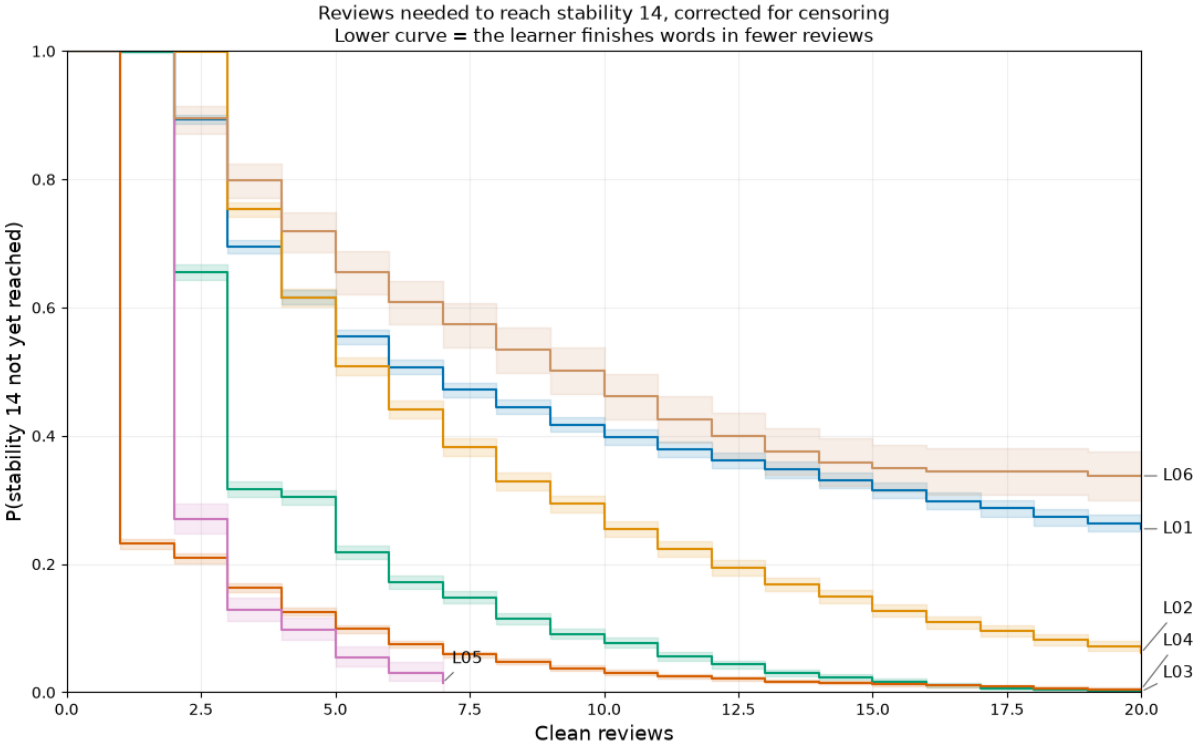


Figure 3: Kaplan-Meier estimates of reviews required to reach the target stability, by learner.

I Temporal Trend in Personalized Learning Burden

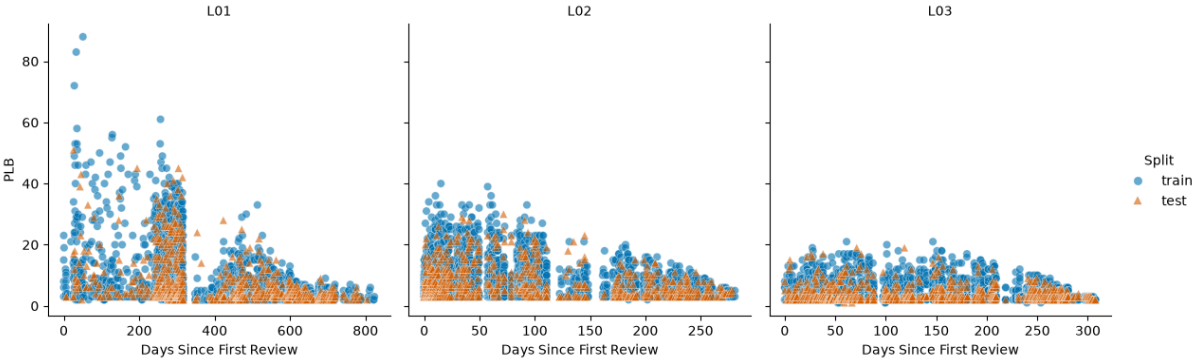


Figure 4: PLB as a function of days since first review for L01–L03, shown separately for training and test items. PLB generally declines over time, motivating the temporally matched random-neighbour control.

A Multidimensional Evaluation of LLM-Generated Learning Materials for Greek Primary School Learners

Eleftherios Kalogeros[1], Sofia Stamou[1], Markella Kolocheri[1], Thaleia-Dionysia Kolovou[2], Manolis Gergatsoulis[1], Matthew Damigos[1]

[1]Laboratory on Digital Libraries and Electronic Publishing,
Department of Information Science, Ionian University
kalogero@ionio.gr, stamou@ionio.gr, markellakol9121999@gmail.com
manolis@ionio.gr, mgdamigos@ionio.gr

[2]Primary School of Giannades, Corfu, Greece
thalia.k3@gmail.com

Abstract

Large language models (LLMs) could support computer-assisted learning by dynamically generating educational texts adapted to different learner groups. Before such material can be used in learner-facing applications, however, linguistic accessibility, semantic content alignment, and factual reliability need to be evaluated jointly. This study evaluates 16 LLMs on the generation of Greek educational texts for primary-school learners, using Grade 3 students and a cultural heritage topic on Cycladic figurines as a controlled case study. The generated texts were assessed across repeated runs using four complementary dimensions: readability and linguistic complexity, grade-specific lexical accessibility using HelexKids, semantic content alignment with a Grade 3 textbook reference using BGE-M3 embeddings, and factual accuracy based on a teacher-scored rubric. Correlation analysis was additionally used to examine relationships among the evaluation dimensions. The results show substantial differences across models and indicate that lexical accessibility, semantic alignment, and factual accuracy capture complementary rather than interchangeable properties of generated learning material. These findings support multi-criteria evaluation when LLM-generated texts are considered for adaptive and learner-facing educational applications.

1 Introduction

Large language models (LLMs) create new opportunities for dynamically generating and adapting educational materials for learners of different ages, grade levels, and language abilities. In computer-assisted learning environments, such models could support the creation of informational texts tailored to particular learner groups. However, fluent generation alone is not sufficient for educational use. Generated material should also be linguistically appropriate for the intended learners, semantically aligned with the intended instructional content, and factually reliable.

Age-adapted versions of the same informational content can support a variety of educational uses. They may be used in textbooks, supplementary learning resources, educational websites, digital learning environments, and learner-facing applications. Museums and cultural heritage institutions provide one concrete setting in which this type of adaptation is useful, since the same information may need to be presented differently to visitors of different ages and educational levels. Recent work has also explored the use of LLMs to generate and adapt cultural heritage descriptions for different audiences (Dror et al., 2024).

The broader aim of this work is to investigate how LLM-generated educational texts can be evaluated before their use as learning materials for primary-school learners. In the present study, Grade 3 students are used as a case study, and Cycladic figurines are selected as the target topic to represent a cultural heritage learning scenario.

The study follows a three-step process. First, each LLM is prompted to generate a short educational text in Greek on a predefined topic and to include a set of required content elements. Second, the generated texts are evaluated using complementary measures of readability and linguistic complexity, grade-specific lexical accessibility, semantic content alignment with a textbook refer-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 160–172, November 19–20, 2026, University of Tartu Library

ence, and factual accuracy. Third, the results are compared across models and repeated generations in order to examine both performance and consistency.

From an ICALL perspective, age- or level-adapted text generation could support the dynamic creation of learning materials tailored to learner characteristics. Such use, however, requires evaluation of linguistic accessibility, semantic content alignment, and factual reliability rather than reliance on fluency or readability alone. The present study examines these requirements through a controlled Greek primary-school case study.

Rather than treating any single linguistic or content-based measure as an indicator of educational suitability, the present study examines these dimensions jointly under a controlled Greek generation setting. The contribution is therefore a multidimensional evaluation of age-adapted learning materials, rather than the introduction of a new readability or simplification method.

This work makes three main contributions. First, it provides a multidimensional evaluation framework for LLM-generated age-adapted learning materials in Greek, considering linguistic complexity, grade-specific lexical accessibility, semantic content alignment, and factual accuracy. Second, it compares 16 open-weight and proprietary LLMs across repeated generations, allowing both output quality and consistency to be examined. Third, it investigates the relationships among evaluation dimensions, showing that linguistic accessibility, semantic alignment, and factual accuracy capture complementary properties of generated educational text.

The study addresses three research questions. First, it examines how different LLMs compare when generating Grade 3 Greek educational texts in terms of readability, lexical accessibility, semantic content alignment, and factual accuracy (RQ1). Second, it investigates how consistent these evaluation outcomes are across repeated generations (RQ2). Third, it explores whether models that produce more lexically accessible texts also achieve higher factual accuracy and semantic content alignment (RQ3).

2 Related Work

Recent work has investigated whether large language models can generate or adapt educational content to the reading level of younger learners. Valentini et al. (2023) evaluated LLM-generated children's stories with respect to lexical and readability levels and found that current models do not consistently restrict their vocabulary to levels appropriate for younger age groups. More recently, Oh et al. (2026) introduced a framework for generating grade-specific educational content across multiple educational levels. Their results showed that explicit adaptation to grade level can improve alignment with learners' expected comprehension level while preserving response accuracy. These studies motivate the need to evaluate age-appropriate generation using multiple linguistic and content-oriented criteria rather than relying only on model fluency.

LLMs have also been examined as tools for educational text simplification. Watanabe (2025) evaluated LLM-based rewriting of educational texts using measures of linguistic sophistication, cohesion, and readability, showing that successful simplification depends on both the model and the prompting strategy. Day et al. (2025) similarly evaluated the simplification of informational passages to targeted grade levels and considered both readability accuracy and semantic consistency. Their findings showed substantial variability across models and prompting approaches, particularly when adapting material to lower grade levels. Together, these studies highlight the difficulty of simplifying educational text while maintaining the intended informational content.

In cultural heritage settings, LLMs have increasingly been explored as tools for adapting and personalizing information for different audiences. Dror et al. (2024) investigated the use of LLMs to support the generation of cultural heritage descriptions tailored to different users. Dibitonto et al. (2026) used GPT-4 to generate museum audio guides adapted to different visitor profiles and evaluated the resulting content through user studies and expert assessment. Their findings highlighted both the potential of personalized narratives and the importance of factual accuracy, domain terminology, and curatorial supervision. Ferrato (2025) further explored the use of LLMs as personalized guides in cultural heritage learning environments, where information can be adapted to different visitor profiles and learning needs.

Stamou et al. (2027) examined LLM-based narrative generation from structured cultural heritage data represented using CIDOC CRM. They

compared 34 freely available LLMs on their ability to transform RDF descriptions into coherent museum-oriented narratives and evaluated the outputs using complementary measures of semantic similarity, content coverage and focus, BERTScore-based alignment, and readability. Their study further illustrates the value of multidimensional evaluation for LLM-generated cultural heritage text, although it focuses on the verbalization of structured knowledge rather than on age-adapted learning materials.

Existing work has primarily focused on English educational texts, general text simplification, or the generation and personalization of cultural heritage information for different audiences. The present study extends these research directions to the generation of Greek educational texts for primary-school learners. Specifically, we use Grade 3 students and a cultural heritage topic as a case study and compare multiple open-weight and proprietary LLMs across readability and linguistic complexity, grade-specific lexical accessibility, semantic content alignment, factual accuracy, and generation consistency.

3 Problem Formulation and Evaluation Framework

The task considered in this study is the generation of an age-appropriate educational text from a predefined topic and a set of required content elements. The generated text should be understandable for the intended learner group while accurately representing the essential information associated with the topic.

Evaluating this type of output requires more than a single measure. A text may use simple and familiar vocabulary but omit important information. In contrast, another text may be factually accurate but remain linguistically complex for the intended audience. Similarly, a text may show substantial semantic overlap with a reference passage without being fully correct in all of its factual details.

For this reason, the evaluation is organized around four complementary dimensions. Readability and linguistic complexity characterize whether the linguistic form of a generated text may be unnecessarily demanding for the target learners. Lexical accessibility examines whether the vocabulary corresponds to the intended grade level. Semantic content alignment assesses the extent to which the generated text reflects the intended instructional content represented by the reference passage. Finally, factual accuracy examines whether the required information is represented correctly and whether the generated material can be considered reliable for educational use.

These dimensions are examined separately rather than combined into a single overall score. This makes it possible to identify different strengths and weaknesses across models and to determine whether strong performance in one dimension is accompanied by strong performance in others. In particular, a model that produces linguistically accessible text may not necessarily produce semantically aligned or factually reliable content.

Repeated generations are additionally used to examine the consistency of model behaviour across outputs. This is important for learner-facing applications, since the suitability of a model should not depend on a single generation. The overall framework therefore considers both the quality of individual outputs and the stability of model performance across repeated runs.

4 Experimental Setup

This section describes the models, generation task, and run configuration used in the evaluation. All models were tested under the same task specification in order to support a consistent comparison of their ability to generate Grade 3 educational texts in Greek.

4.1 Model Selection

The evaluation included 16 large language models from multiple model families, covering both open-weight and proprietary systems and a broad range of model scales and architectural configurations. The selected models include both dense and mixture-of-experts (MoE) architectures, with publicly disclosed parameter counts ranging from 4B to 117B. Three proprietary systems—GPT-5.4, Claude Sonnet 5, and Gemini 3.8 Flash—were also included. Models with reasoning capabilities were allowed to use reasoning when supported, reflecting their intended operational configuration. For MoE models, total and active parameter counts were treated separately rather than interpreted as directly equivalent to the parameter counts of dense models. Table 1 presents the selected models and their main characteristics.

Table 1: Large language models included in the evaluation.

Model	Type	Total Params.	Active Params.
Qwen3 8B	Open-weight / Dense	8.2B	8.2B
Qwen3 14B	Open-weight / Dense	14.8B	14.8B
Qwen3 32B	Open-weight / Dense	32.8B	32.8B
Qwen3 30B-A3B	Open-weight / MoE	30.5B	3.3B
Llama 3.1 8B	Open-weight / Dense	8B	8B
Llama 3.3 70B	Open-weight / Dense	70B	70B
Gemma 3 4B	Open-weight / Dense	4B	4B
Gemma 3 12B	Open-weight / Dense	12B	12B
Gemma 3 27B	Open-weight / Dense	27B	27B
Mistral Small 3.2 24B	Open-weight / Dense	24B	24B
DeepSeek V4.1 Flash	Open-weight / MoE	N/D	N/D
GPT-OSS 20B	Open-weight / MoE	21B	3.6B
GPT-OSS 120B	Open-weight / MoE	117B	5.1B
GPT-5.4	Proprietary	N/D	N/D
Claude Sonnet 5	Proprietary	N/D	N/D
Gemini 3.8 Flash	Proprietary	N/D	N/D

Note. N/D indicates information not publicly disclosed. For mixture-of-experts (MoE) models, total and active parameter counts are reported separately where available.

4.2 Generation Task

All models received the same prompt in Greek. The task requested a 90–110-word informational text about Cycladic figurines for Grade 3 students aged 8–9 years. For transparency, an English translation of the prompt is reproduced below; the original Greek version was used in all experiments.

> Write a short informational text in Greek for Grade 3 primary-school students aged 8–9 years about Cycladic figurines. The text should be approximately 90–110 words long and explain what Cycladic figurines are, what material they were made from, what they usually represent, and where they have been found. Use clear, natural, age-appropriate language and historically accurate information. Return only the final continuous text, without a title, bullet points, or a numbered list.

The generated texts were subsequently evaluated from several complementary perspectives. Readability and linguistic-complexity measures were used to describe the linguistic form of each text, while Grade 3 HelexKids-based measures were used to estimate lexical accessibility for the target learners. BGE-M3 cosine similarity was used to assess semantic content alignment with the corresponding textbook passage, and a predefined factual rubric was used to evaluate whether the required content was correctly represented. Repeated generation runs were additionally used to examine the consistency of these measures across outputs.

4.3 Run Configuration

Each model was evaluated across four runs. For models supporting temperature control, Run 1 used $T = 0$ with seed 42 and served as the low-temperature reference condition. This setting was intended to reduce generation variability, but Run 1 is not assumed to be strictly deterministic, since hosted LLM systems may exhibit nondeterminism even under nominally deterministic generation settings (Atıl et al., 2025). Runs 2–4 used $T = 0.7$ with seeds 101, 202, and 303, respectively. Top-p was fixed at 1.0. Temperature and seed parameters were sent only when supported by the corresponding model.

GPT-5.4 and Claude Sonnet 5 did not expose temperature control through the used API interface and were therefore evaluated under their native generation settings. Their Run 1 results are retained for completeness in the general model comparison but are not interpreted as deterministic. These models are excluded from analyses specifically concerning the effect of temperature.

Reasoning was enabled for models that supported it, while the reasoning content itself was not included in the evaluation. A maximum completion budget of 4,000 tokens was used for all models. Temporary API failures, including rate limits, timeouts, connection errors, server-side errors, and empty outputs, were retried automatically up to three times.

5 Evaluation Metrics

The generated texts were evaluated across four dimensions: readability and linguistic complexity, grade-specific lexical accessibility, semantic content alignment, and factual accuracy. Together, these measures provide a multidimensional assessment of the suitability and accuracy of the generated educational texts.

Table 2: Abbreviations used for linguistic-complexity and readability-related indicators.

Abbreviation	Indicator
W/S	Words per sentence
Easy/100	Easy words per 100 words
MultiSyl/100	Multisyllabic words per 100 words
Guiraud R	Guiraud's lexical-diversity index
Affix/100	Words with prefixes/suffixes per 100 words
Middle/100	Middle-voice forms per 100 words
Conj/100	Conjunctions per 100 words
LAdv/100	Learned adverbial forms per 100 words
Part/100	Participles per 100 words

5.1 Linguistic Complexity and Readability-Related Indicators

Readability was assessed using the online readability tool provided by the Centre for the Greek Language (Centre for the Greek Language, 2025; Mikros and Voskaki, 2021). Each generated text was processed with the tool, and the resulting linguistic indicators were recorded for analysis.

The tool was originally developed for automatic readability analysis of Modern Greek and uses a range of stylometric and linguistic indicators associated with text difficulty (Mikros and Voskaki, 2021). In the present study, we use the individual linguistic indicators reported by the tool rather than its overall CEFR-level prediction.

Words per sentence (W/S) represents the mean number of words per sentence. Easy words per 100 words (Easy/100) represents the proportion of words classified as easy according to the lexical classification implemented by the readability tool; this classification was used as provided by the tool and was not modified in the present study. Words with more than two syllables per 100 words (MultiSyl/100) captures the frequency of multisyllabic words. Guiraud's R (Guiraud R) measures lexical diversity as the number of distinct lexical types divided by the square root of the total number of word tokens.

The tool also reports normalized frequencies per 100 words for several morphological and morphosyntactic features. These include words with prefixes or suffixes (Affix/100), middle-voice morphology (Middle/100), conjunctions (Conj/100), learned adverbial forms (LAdv/100), and participles (Part/100).

Together, these measures capture complementary lexical, morphological, and morphosyntactic characteristics associated with text readability and linguistic complexity. Because the indicators describe different properties of the text, they were interpreted separately rather than combined into a single composite readability score.

These indicators are treated as linguistic features associated with readability rather than as direct or interchangeable measures of readability itself. Their relationship with perceived text difficulty is not assumed to be linear, and lower values are not uniformly interpreted as preferable across all indicators. The measures are therefore examined individually rather than combined into a single readability score.

5.2 Grade-Specific Lexical Accessibility

Lexical accessibility was assessed using the Grade 3 HelexKids frequency list (Terzopoulos et al., 2017). HelexKids is a grade-based lexical frequency database derived from 116 primary-school textbooks used in Greece and Cyprus. The complete resource contains 1,355,265 tokens and 68,692 non-lemmatized word types across Grades 1–6, while the Grade 3 component contains 21,193 word types. The resource therefore provides a curriculum-based estimate of the written vocabulary encountered by Greek-speaking primary-school learners.

Matching was performed at the surface-form level without lemmatization. Both the generated texts and the HelexKids entries were normalized to Unicode NFC and lowercased before matching, while Greek accents were retained. Punctuation was excluded during tokenization. As a result, inflected forms were treated as distinct lexical items and were counted as matches only when the corresponding surface form was present in the Grade 3 lexicon.

Lexical coverage (Coverage) represents the percentage of text tokens found in the Grade 3 HelexKids lexicon. Mean log frequency per million (Mean logF/m) represents the average log-transformed frequency per million of the matched words. Higher Coverage indicates that a larger proportion of the generated text consists of vo-

cabulary attested in the Grade 3 lexicon, whereas higher Mean logF/m indicates that the matched words are, on average, more frequent within that lexical resource.

The out-of-vocabulary rate (OOV) was calculated as $100 - \text{Coverage}$ and was retained as a complementary descriptive measure. Because Coverage and OOV are exact complements, OOV was not treated as an independent evaluation dimension.

5.3 Semantic Content Alignment

Semantic content alignment with the corresponding textbook reference (Papadimitriou and Savvaki, 2026) was assessed using cosine similarity between text embeddings generated with BAAI's BGE-M3 model (Chen et al., 2024). Higher cosine similarity values indicate greater semantic overlap between the generated text and the textbook reference (Papadimitriou and Savvaki, 2026). The measure was used to assess semantic content alignment rather than lexical or surface-form similarity.

5.4 Factual Accuracy

Factual accuracy was evaluated using a topic-specific rubric comprising K core factual criteria. Each criterion (F_i) was scored as 0 (incorrect), 1 (partially correct), or 2 (correct). The overall Factual Accuracy Score was normalized to a 0–1 scale as

$$\text{Factual Accuracy Score} = \frac{\sum_{i=1}^{K} F_i}{X}, \qquad X = 2K,$$

where X represents the maximum attainable score for the corresponding rubric. Higher values indicate greater factual accuracy.

For example, for the topic of Cycladic figurines, four core factual criteria were evaluated: (F1) what Cycladic figurines are, (F2) their main material, (F3) what they typically represent, and (F4) where they have been found. Thus, $K = 4$ and $X = 8$, giving

$$\text{Factual Accuracy Score} = \frac{F_1 + F_2 + F_3 + F_4}{8}.$$

The factual criteria were manually scored by a primary-school teacher according to the predefined rubric, with each criterion rated as incorrect, partially correct, or correct. The generated texts were presented to the evaluator in randomized order without information about the model or generation condition associated with each output.

5.5 Correlation Analysis

To examine relationships among the evaluation dimensions, Spearman's rank correlation coefficient (ρ) was calculated using model-level mean values across Runs 2–4. Correlations were examined between the two HelexKids lexical accessibility measures (Coverage and Mean logF/m), semantic content alignment measured using BGE-M3 cosine similarity, and Factual Accuracy.

Spearman correlation was selected because of the small number of models ($n = 16$) and the bounded and potentially non-normal distribution of several evaluation measures. The analysis was exploratory, and correlation coefficients were interpreted together with their corresponding p-values. Given the exploratory nature of the analysis, reported p-values are unadjusted and should be interpreted cautiously.

6 Experimental Results

This section presents the evaluation results across four dimensions: readability and linguistic complexity, Grade 3 lexical accessibility, semantic content alignment, and factual accuracy. Run 1 results are reported separately from Runs 2–4 in order to examine both model performance and run-to-run consistency.

The generation prompt requested texts of approximately 90–110 words. Across the 64 generated texts, 32 outputs (50.0%) satisfied this range, 24 (37.5%) were shorter than 90 words, and 8 (12.5%) exceeded 110 words. Generated texts ranged from 9 to 128 words. Length compliance was not used as an exclusion criterion, and all valid generations were retained in the evaluation.

6.1 Linguistic Complexity and Readability-Related Indicators

Table 3 presents the readability and linguistic complexity measures obtained for Run 1. Considerable variation was observed across models. Mean sentence length (W/S) ranged from 6.67 words for Llama 3.1 8B to 19.60 words for Gemma 3 12B. The proportion of easy words (Easy/100) ranged from 40.63 to 55.77, with Gemma 3 27B producing the highest value.

Gemma 3 12B exhibited the lowest proportion

of multisyllabic words (MultiSyl/100 = 29.59) and the lowest frequency of words containing prefixes or suffixes (Affix/100 = 20.41). Lexical diversity, measured using Guiraud's R, ranged from 5.06 for Llama 3.1 8B to 8.38 for GPT-OSS 120B. The remaining morphological and morphosyntactic measures also differed across models, indicating variation not only in vocabulary and sentence length but also in morphological and syntactic complexity.

Across Runs 2–4, the linguistic and readability-related indicators were not equally stable across models. Table 4 summarizes the core readability measures across Runs 2–4. The Gemma family generally exhibited comparatively accessible lexical characteristics. Gemma 3 4B achieved the highest mean Easy/100 value (51.18 ± 2.08), while Gemma 3 12B exhibited the lowest mean MultiSyl/100 value (34.07 ± 2.53). GPT-OSS 120B showed the highest mean Guiraud's R (8.09 ± 0.16).

Table 5 reports the corresponding morphological and morphosyntactic measures. Gemma 3 27B exhibited the lowest mean Affix/100 value (23.22 ± 0.61). Differences were also observed in middle-voice morphology, conjunction use, learned adverbial forms, and participles, confirming that the models differed in multiple dimensions of linguistic complexity.

Variation was also observed in lexical diversity. GPT-OSS 120B exhibited the highest mean Guiraud's R across the repeated runs (8.09 ± 0.16), whereas Llama 3.1 8B produced the lowest mean value (5.16 ± 1.91). DeepSeek V4.1 Flash showed particularly stable lexical diversity across repeated generations (7.85 ± 0.09). In contrast, Llama 3.3 70B exhibited substantial variability in the proportion of multisyllabic words (43.90 ± 9.76).

Overall, the models exhibited different linguistic profiles across the individual indicators, and differences on one measure did not necessarily correspond to similar patterns on the others. Because the indicators capture distinct lexical, morphological, and syntactic properties and do not share a common direction of desirability, they were interpreted separately rather than combined into a single composite score.

6.2 Grade-Specific Lexical Accessibility

Table 6 presents the grade-specific lexical accessibility results obtained using the Grade 3 HelexKids frequency list. Coverage represents the percentage of tokens occurring in the Grade 3 lexicon, whereas Mean logF/m reflects the mean log-transformed frequency per million of the matched words. Higher values on both measures indicate greater correspondence with the Grade 3 lexical environment.

In Run 1, lexical coverage ranged from 75.00% for Llama 3.1 8B to 98.96% for DeepSeek V4.1 Flash. The textbook reference, which was used as a descriptive benchmark but was not included as a model in the table, yielded a coverage of 87.13% and a Mean logF/m of 3.667. Most models therefore produced Run 1 texts with lexical coverage above the textbook benchmark. However, the two HelexKids measures did not always vary in parallel. For example, DeepSeek V4.1 Flash achieved very high coverage (98.96%) but a Mean logF/m of 3.698, only slightly above the textbook value. In contrast, Llama 3.1 8B exhibited the lowest coverage (75.00%) but the highest Run 1 Mean logF/m (4.101). This indicates that the words successfully matched in the HelexKids lexicon could be highly frequent even when a comparatively large proportion of the complete text remained unmatched.

Across Runs 2–4, lexical accessibility was relatively stable for several models but more variable for others. Among models for which $T = 0.7$ was controlled, Gemma 3 27B achieved the highest mean coverage across Runs 2–4 ($97.75 \pm 2.56\%$), followed by Gemma 3 12B ($97.02 \pm 1.62\%$) and DeepSeek V4.1 Flash ($96.30 \pm 3.21\%$). In contrast, Llama 3.1 8B showed the greatest variation in coverage ($88.12 \pm 10.82\%$), indicating substantial run-to-run differences in the proportion of vocabulary matched by the Grade 3 HelexKids lexicon.

Mean logF/m was generally more stable than coverage. Among the repeated runs, Gemma 3 4B (3.979 ± 0.050) and GPT-OSS 20B (3.950 ± 0.033) exhibited particularly high mean values, while Gemma 3 12B showed very little variation across runs (3.805 ± 0.004). These results suggest that lexical coverage and lexical frequency capture complementary properties: a model may use a large proportion of Grade 3-attested vocabulary without necessarily selecting the most frequent words within that vocabulary.

Overall, most models produced texts with HelexKids coverage above the 87.13% observed

Table 3: Readability and linguistic complexity for Run 1. For compact presentation, shortened column labels are used; Table 2 provides the corresponding full abbreviations and metric definitions.

Model	W/S	Easy	MultiSyl	Guiraud R	Affix	Middle	Conj	LAdv	Part
Qwen3 8B	13.20	40.91	46.97	7.14	37.88	4.55	7.58	4.55	1.52
Qwen3 14B	15.40	41.56	42.86	7.41	35.06	9.09	10.39	1.30	1.30
Qwen3 32B	13.00	46.15	37.18	7.47	26.92	6.41	8.97	1.28	1.28
Qwen3 30B-A3B	12.83	42.86	44.16	7.29	37.66	3.90	7.79	5.19	0.00
Llama 3.1 8B	6.67	50.00	37.50	5.06	32.50	7.50	10.00	0.00	0.00
Llama 3.3 70B	15.33	52.17	41.30	5.46	30.43	6.52	13.04	2.17	0.00
Gemma 3 4B	16.00	53.13	34.38	7.51	21.09	1.56	13.28	0.78	0.00
Gemma 3 12B	19.60	53.06	29.59	7.58	20.41	2.04	14.29	1.02	2.04
Gemma 3 27B	17.33	55.77	38.46	6.86	23.08	2.88	12.50	0.96	0.96
Mistral Small 3.2 24B	15.33	44.57	39.13	7.61	30.43	4.35	14.13	1.09	0.00
DeepSeek V4.1 Flash	10.67	40.63	38.54	7.86	33.33	2.08	8.33	1.04	2.08
GPT-OSS 20B	15.57	43.12	46.79	7.57	28.44	6.42	14.68	0.00	0.00
GPT-OSS 120B	12.88	40.78	44.66	8.38	35.92	4.85	7.77	2.91	0.00
GPT-5.4	15.17	45.05	42.86	7.02	32.97	1.10	13.19	3.30	1.10
Claude Sonnet 5	17.17	46.60	39.81	7.39	27.18	0.97	12.62	1.94	0.97
Gemini 3.8 Flash	17.33	47.12	45.19	7.84	33.65	0.96	9.62	0.96	0.96

Table 4: Core readability measures across Runs 2–4. Values are reported as mean ± standard deviation.

Model	W/S	Easy/100	MultiSyl/100	Guiraud R
Qwen3 8B	14.38 ± 4.43	43.17 ± 4.91	39.85 ± 0.96	7.22 ± 0.15
Qwen3 14B	14.99 ± 3.57	42.33 ± 3.30	44.77 ± 2.52	7.39 ± 0.27
Qwen3 32B	15.78 ± 2.13	45.19 ± 3.15	37.30 ± 4.40	7.49 ± 0.26
Qwen3 30B-A3B	14.90 ± 1.17	44.67 ± 2.59	37.59 ± 1.55	7.53 ± 0.08
Llama 3.1 8B	10.43 ± 1.29	45.63 ± 8.60	41.92 ± 2.18	5.16 ± 1.91
Llama 3.3 70B	12.15 ± 2.70	44.65 ± 5.37	43.90 ± 9.76	6.14 ± 1.10
Gemma 3 4B	14.35 ± 2.43	51.18 ± 2.08	37.23 ± 4.27	7.98 ± 0.22
Gemma 3 12B	16.86 ± 2.25	46.84 ± 1.19	34.07 ± 2.53	7.65 ± 0.23
Gemma 3 27B	15.91 ± 0.94	51.02 ± 1.33	36.62 ± 3.41	7.60 ± 0.45
Mistral Small 3.2	14.06 ± 2.27	45.71 ± 1.62	43.73 ± 0.86	7.22 ± 0.28
DeepSeek V4.1 Flash	13.56 ± 2.02	43.00 ± 1.74	38.43 ± 1.67	7.85 ± 0.09
GPT-OSS 20B	16.66 ± 3.27	45.45 ± 3.06	37.56 ± 3.45	7.36 ± 0.33
GPT-OSS 120B	13.69 ± 1.91	40.05 ± 2.19	46.65 ± 4.04	8.09 ± 0.16
GPT-5.4	15.06 ± 1.87	46.82 ± 0.60	39.76 ± 1.86	7.40 ± 0.43
Claude Sonnet 5	17.65 ± 1.30	44.43 ± 2.27	40.65 ± 3.98	7.62 ± 0.27
Gemini 3.8 Flash	16.22 ± 0.54	42.49 ± 2.90	46.63 ± 3.23	8.01 ± 0.39

Table 5: Morphological and morphosyntactic measures across Runs 2–4. Values are reported as mean ± standard deviation.

Model	Affix/100	Middle/100	Conj/100	LAdv/100	Part/100
Qwen3 8B	35.05 ± 1.29	3.05 ± 0.87	11.27 ± 0.89	1.72 ± 0.68	0.88 ± 0.77
Qwen3 14B	33.78 ± 5.32	5.69 ± 1.65	9.82 ± 0.89	1.23 ± 0.08	0.84 ± 0.73
Qwen3 32B	30.75 ± 3.54	3.19 ± 0.60	10.77 ± 1.77	1.20 ± 0.05	1.62 ± 0.77
Qwen3 30B-A3B	30.74 ± 4.69	3.20 ± 1.05	9.90 ± 2.18	3.13 ± 1.72	0.69 ± 1.19
Llama 3.1 8B	28.66 ± 5.76	4.89 ± 4.23	11.62 ± 2.65	1.72 ± 1.86	0.00 ± 0.00
Llama 3.3 70B	32.26 ± 8.85	5.65 ± 1.48	8.08 ± 4.94	2.06 ± 1.02	0.44 ± 0.76
Gemma 3 4B	25.34 ± 2.25	2.90 ± 0.47	10.56 ± 1.72	1.58 ± 0.79	0.53 ± 0.46
Gemma 3 12B	25.03 ± 2.13	2.66 ± 2.08	12.75 ± 2.73	1.33 ± 0.53	1.67 ± 0.59
Gemma 3 27B	23.22 ± 0.61	3.15 ± 0.52	12.32 ± 0.33	1.14 ± 0.49	1.43 ± 0.51
Mistral Small 3.2	29.72 ± 3.01	2.73 ± 0.90	12.03 ± 3.31	1.16 ± 0.12	0.42 ± 0.72
DeepSeek V4.1 Flash	34.06 ± 3.56	3.25 ± 2.00	12.07 ± 0.89	2.14 ± 1.12	2.14 ± 1.12
GPT-OSS 20B	31.61 ± 5.77	5.55 ± 1.45	10.48 ± 1.17	1.33 ± 1.47	0.00 ± 0.00
GPT-OSS 120B	35.18 ± 2.24	7.17 ± 2.82	9.76 ± 3.12	2.16 ± 1.18	1.46 ± 1.27
GPT-5.4	32.00 ± 2.62	1.75 ± 0.57	11.25 ± 0.66	2.45 ± 0.48	1.06 ± 0.05
Claude Sonnet 5	29.84 ± 1.37	1.42 ± 1.65	10.41 ± 1.72	2.32 ± 0.38	0.36 ± 0.62
Gemini 3.8 Flash	36.33 ± 5.55	1.04 ± 1.81	8.25 ± 1.98	1.38 ± 0.63	1.36 ± 0.54

for the textbook reference in at least one generation condition. This result should be interpreted specifically as greater lexical correspondence with the Grade 3 HelexKids corpus rather than as evidence that the generated texts were superior to the textbook overall. Lexical accessibility constitutes only one dimension of educational text quality and should therefore be considered together with read-

Table 6: Grade-specific lexical accessibility according to the Grade 3 HelexKids frequency list. Values for Runs 2–4 are reported as mean ± standard deviation.

	Run 1		Runs 2–4	
Model	Coverage (%)	Mean logF/m	Coverage (%)	Mean logF/m
Qwen3 8B	78.79	3.878	89.58 ± 3.58	3.846 ± 0.092
Qwen3 14B	88.31	3.876	90.11 ± 3.73	3.850 ± 0.086
Qwen3 32B	87.34	3.837	92.41 ± 0.50	3.897 ± 0.133
Qwen3 30B-A3B	88.31	3.755	90.82 ± 1.18	3.885 ± 0.090
Llama 3.1 8B	75.00	4.101	88.12 ± 10.82	3.883 ± 0.102
Llama 3.3 70B	89.13	4.074	91.79 ± 1.46	3.821 ± 0.247
Gemma 3 4B	96.88	3.998	92.62 ± 2.51	3.979 ± 0.050
Gemma 3 12B	95.92	3.921	97.02 ± 1.62	3.805 ± 0.004
Gemma 3 27B	98.08	3.994	97.75 ± 2.56	3.901 ± 0.055
Mistral Small 3.2 24B	90.22	3.958	89.82 ± 3.41	3.911 ± 0.167
DeepSeek V4.1 Flash	98.96	3.698	96.30 ± 3.21	3.798 ± 0.095
GPT-OSS 20B	90.83	3.838	86.68 ± 5.36	3.950 ± 0.033
GPT-OSS 120B	89.32	3.621	87.98 ± 1.14	3.675 ± 0.079
GPT-5.4	97.80	3.758	98.95 ± 1.08	3.783 ± 0.020
Claude Sonnet 5	93.20	3.851	93.36 ± 1.10	3.778 ± 0.115
Gemini 3.8 Flash	95.19	3.796	92.14 ± 1.42	3.755 ± 0.016

ability, semantic content alignment, and factual accuracy.

6.3 Semantic Content Alignment

Table 7 presents the cosine similarity between each generated text and the textbook reference using BGE-M3 embeddings. Higher values indicate greater semantic content alignment with the reference text.

In Run 1, cosine similarity ranged from 0.710 for Gemma 3 4B to 0.821 for Qwen3 8B. Qwen3 14B (0.811) and GPT-5.4 (0.800) also showed relatively high semantic alignment with the textbook reference.

Across Runs 2–4, Mistral Small 3.2 24B achieved the highest mean cosine similarity (0.794 ± 0.009), followed by Qwen3 8B (0.783 ± 0.006) and DeepSeek V4.1 Flash (0.781 ± 0.021). Most models showed relatively small run-to-run variation, although Llama 3.1 8B exhibited greater variability (0.707 ± 0.078).

Overall, the results show that semantic alignment with the textbook reference varied across models and across repeated generations. These scores should be interpreted as measures of semantic content overlap rather than factual correctness or overall educational quality.

6.4 Factual Accuracy

Table 8 presents the normalized Factual Accuracy Scores for Run 1 and Runs 2–4. Scores range from 0 to 1, with higher values indicating greater factual accuracy.

Substantial differences were observed across models. In Run 1, DeepSeek V4.1 Flash, GPT-5.4, Claude Sonnet 5, Gemini 3.8 Flash, and Qwen3 14B achieved the maximum score of 1.00. By contrast, Llama 3.1 8B and GPT-OSS 20B obtained markedly lower scores of 0.50 and 0.25, respectively.

Across Runs 2–4, DeepSeek V4.1 Flash, GPT-5.4, and Gemini 3.8 Flash obtained mean factual-accuracy scores of 1.000 ± 0.000. Claude Sonnet 5 also showed consistently high performance (0.958 ± 0.072). Several other models exhibited moderate run-to-run variation, including Qwen3 30B-A3B (0.833 ± 0.144), GPT-OSS 120B (0.792 ± 0.144), and Qwen3 32B (0.750 ± 0.125). The lowest mean performance across Runs 2–4 was observed for Llama 3.1 8B (0.208 ± 0.191) and GPT-OSS 20B (0.292 ± 0.191).

Table 7: Semantic content alignment with the textbook reference using BGE-M3 cosine similarity. Values for Runs 2–4 are reported as mean ± standard deviation.

Model	Run 1	Runs 2–4
Qwen3 8B	0.821	0.783 ± 0.006
Qwen3 14B	0.811	0.777 ± 0.010
Qwen3 32B	0.790	0.731 ± 0.031
Qwen3 30B-A3B	0.758	0.768 ± 0.024
Llama 3.1 8B	0.767	0.707 ± 0.078
Llama 3.3 70B	0.778	0.774 ± 0.038
Gemma 3 4B	0.710	0.744 ± 0.010
Gemma 3 12B	0.733	0.709 ± 0.019
Gemma 3 27B	0.718	0.705 ± 0.018
Mistral Small 3.2 24B	0.787	0.794 ± 0.009
DeepSeek V4.1 Flash	0.776	0.781 ± 0.021
GPT-OSS 20B	0.747	0.771 ± 0.020
GPT-OSS 120B	0.764	0.764 ± 0.042
GPT-5.4	0.800	0.776 ± 0.009
Claude Sonnet 5	0.770	0.746 ± 0.008
Gemini 3.8 Flash	0.763	0.735 ± 0.011

Table 8: Factual Accuracy Scores by model. Run 1 values are reported separately; values for Runs 2–4 are reported as mean ± standard deviation.

Model	Run 1	Runs 2–4
Qwen3 8B	0.625	0.583 ± 0.072
Qwen3 14B	1.000	0.833 ± 0.072
Qwen3 32B	0.875	0.750 ± 0.125
Qwen3 30B-A3B	0.750	0.833 ± 0.144
Llama 3.1 8B	0.500	0.208 ± 0.191
Llama 3.3 70B	0.875	0.875 ± 0.000
Gemma 3 4B	0.625	0.625 ± 0.125
Gemma 3 12B	0.750	0.833 ± 0.072
Gemma 3 27B	0.875	0.917 ± 0.072
Mistral Small 3.2 24B	0.875	0.917 ± 0.072
DeepSeek V4.1 Flash	1.000	1.000 ± 0.000
GPT-OSS 20B	0.250	0.292 ± 0.191
GPT-OSS 120B	0.875	0.792 ± 0.144
GPT-5.4	1.000	1.000 ± 0.000
Claude Sonnet 5	1.000	0.958 ± 0.072
Gemini 3.8 Flash	1.000	1.000 ± 0.000

Overall, the factual-accuracy results revealed a clear separation between models that consistently reproduced the required core information and models that introduced substantial factual errors or omitted key facts. Importantly, high lexical accessibility did not always correspond to high factual accuracy. For example, GPT-OSS 20B achieved relatively high HelexKids lexical coverage but substantially lower factual-accuracy scores. This finding supports the use of factual accuracy as a separate evaluation dimension rather than inferring content reliability from lexical accessibility alone.

6.5 Relationships Between Evaluation Dimensions

Table 9 presents Spearman correlations between lexical accessibility, semantic content alignment, and factual accuracy using model-level means across Runs 2–4. Coverage showed a positive association with factual accuracy ($\rho = 0.648$, $p = 0.007$). In contrast, Coverage was not significantly associated with BGE-M3 cosine similarity ($\rho = -0.262$, $p = 0.327$).

Mean logF/m showed a negative association with factual accuracy ($\rho = -0.502$, $p = 0.048$), whereas its association with BGE-M3 cosine similarity was negligible ($\rho = -0.050$, $p = 0.854$).

BGE-M3 cosine similarity showed only a weak positive association with factual accuracy ($\rho = 0.191$, $p = 0.479$), which was not statistically significant. This suggests that greater semantic overlap with the textbook reference does not necessarily correspond to higher accuracy on the predefined factual criteria.

Overall, the correlation results suggest that lexical accessibility, semantic content alignment, and factual accuracy capture partly distinct properties of the generated texts rather than interchangeable indicators of educational suitability.

Table 9: Spearman correlations between evaluation measures using model-level means across Runs 2–4 ($n = 16$).

Relationship	ρ	p
Coverage – Factual Accuracy	0.648	0.007
Coverage – BGE-M3 cosine	-0.262	0.327
Mean logF/m – Factual Accuracy	-0.502	0.048
Mean logF/m – BGE-M3 cosine	-0.050	0.854
BGE-M3 cosine – Factual Accuracy	0.191	0.479

7 Discussion

The results show that the evaluated LLMs differed substantially in their ability to generate Grade 3 educational texts in Greek. No single model showed consistently strong performance across all evaluation dimensions, confirming that the suitability of generated learning material is multidimensional. Readability, lexical accessibility, semantic content alignment, and factual accuracy therefore need to be considered jointly rather than inferred from a single metric.

The readability analysis revealed clear differences in sentence length, lexical difficulty, lexical diversity, and morphological complexity. The Gemma 3 family generally exhibited a comparatively accessible profile on several readability indicators, particularly Easy/100, MultiSyl/100, and Affix/100. However, values suggesting greater accessibility on individual indicators did not imply uniformly lower linguistic complexity across all metrics. This supports the decision to interpret the individual indicators separately rather than combining them into a single readability score.

Semantic content alignment also varied across models, while most systems showed relatively limited run-to-run variation. Importantly, higher BGE-M3 similarity should be interpreted as greater semantic overlap with the textbook reference rather than as evidence of factual correctness. The weak and non-significant association between BGE-M3 similarity and factual accuracy further supports this distinction. Semantic overlap with the textbook reference and correctness of the required factual elements therefore appear to capture different properties of the generated texts.

The HelexKids analysis further showed that many models produced vocabulary with high correspondence to the Grade 3 lexical resource. Nevertheless, lexical accessibility did not necessarily imply factual reliability. GPT-OSS 20B, for example, produced relatively accessible vocabulary while obtaining substantially lower factual-accuracy scores. This illustrates an important distinction for learner-facing material: text can appear linguistically appropriate for the target grade while still omitting or incorrectly representing important content.

The practical implication is that lexical or readability-based screening cannot replace independent content validation in an LLM-based learning-material generation pipeline.

The correlation analysis provides further evidence that the evaluation dimensions should not be treated as interchangeable. Lexical Coverage was positively associated with factual accuracy, but it was not positively associated with semantic content alignment. Mean logF/m also did not show a clear positive relationship with semantic alignment. These findings suggest that linguistic accessibility alone cannot serve as a general proxy for the overall educational suitability of generated text. In particular, a model that uses grade-appropriate vocabulary is not necessarily producing content that is semantically aligned or factually reliable.

Factual accuracy showed substantial differences across models. DeepSeek V4.1 Flash, GPT-5.4, and Gemini 3.8 Flash achieved consistently high factual scores across repeated generations, whereas Llama 3.1 8B and GPT-OSS 20B showed considerably lower performance. The repeated-run analysis also revealed differences in stability across models, indicating that evaluation based on a single generated output may provide an incomplete picture of model behaviour.

These findings have implications for LLM-supported learning-material generation and ICALL-oriented applications. Selecting a model only on the basis of readability or lexical accessibility may be insufficient, since linguistically accessible outputs may still omit, distort, or incorrectly represent important educational content. A practical generation pipeline should therefore consider linguistic adaptation, semantic content alignment, and factual validation as separate but complementary requirements. The observed run-to-run variation further suggests that models intended for learner-facing use should be evaluated across multiple outputs rather than through a single generation.

Overall, the results indicate that LLMs can generate Greek educational texts with characteristics that are relevant to primary-school learners, but educational suitability cannot be established through linguistic simplicity alone. Multi-criteria evaluation is therefore necessary when generated texts are intended for learning materials or adaptive educational environments. The Grade 3 cultural heritage task used in this study provides a controlled case through which these relationships can be examined, while broader generalization requires evaluation across additional topics, grade levels, and educational contexts.

8 Limitations and Future Work

The present study has several limitations. First, the evaluation focuses on a single Grade 3 cultural heritage topic, and the findings should therefore not be assumed to generalize to other subjects, age groups, or educational contexts. The study also uses a single prompt formulation. Although keeping the prompt fixed supports a controlled comparison across models, alternative prompt wording or prompting strategies may produce different linguistic, semantic, or factual profiles. In addition, the results reflect the specific model versions and API configurations available at the time of the experiments and should therefore be interpreted as a snapshot of model behaviour rather than as permanent characteristics of the corresponding model families.

Second, factual accuracy was assessed by a single primary-school teacher, so inter-rater reliability could not be estimated. Third, semantic content alignment was measured automatically using BGE-M3 embeddings and was not independently validated against human judgments for the Greek texts used in this study. Moreover, only three repeated generations were used for the run-to-run analysis, so the reported variability should be interpreted descriptively rather than as a precise estimate of each model's generation distribution. Finally, the factual-accuracy rubric focused on the predefined core content elements and did not separately assess all additional factual claims that may have appeared in the generated texts.

Future work should extend the evaluation to

multiple curricular and cultural heritage topics, additional primary-school grade levels, and alternative prompt formulations. Larger numbers of repeated generations would also support more robust estimates of model variability. Further evaluation should involve multiple independent educators, include human assessment of semantic content alignment and age appropriateness, and examine unsupported or incorrect factual claims beyond the predefined rubric. Because the present study targets L1 Greek primary-school learners, the applicability of these evaluation dimensions to L2 learners should also be examined separately using appropriate proficiency-oriented resources and learner populations.

9 Conclusion

This study evaluated 16 large language models on the generation of age-adapted Greek learning materials, using Grade 3 students and a cultural heritage topic as a controlled case study. The models were assessed across readability and linguistic complexity, grade-specific lexical accessibility, semantic content alignment, factual accuracy, and consistency across repeated generations.

The results demonstrate that the suitability of generated educational material is multidimensional: strong performance on one evaluation dimension does not necessarily imply strong performance on others. In particular, linguistic accessibility alone should not be treated as a proxy for semantic alignment or factual reliability. These findings support the use of multi-criteria evaluation before LLM-generated texts are integrated into learner-facing or adaptive educational applications. Future work should extend the evaluation to additional topics and grade levels and include multiple independent educators and broader human evaluation.

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: OpenAI ChatGPT was used as an assistive tool for language editing, improving clarity and readability, and refining the organization and presentation of the manuscript. It was also used to support the review of wording and methodological descriptions. All scientific content, experimental procedures, results, analyses, and interpretations were reviewed and verified by the authors. No AI tool was used to generate or alter the reported experimental results.

References

Berk Atıl, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Nondeterminism of "deterministic" LLM system settings in hosted environments. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pages 135–148. Association for Computational Linguistics.

Centre for the Greek Language. 2025. Readability software for modern Greek texts. Portal for the Greek Language. Accessed 15 September 2026.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie L. Day, Jacapo Cirica, Steven R. Clapp, Veronika Penkova, Amy E. Giroux, Abbey Banta, Catherine Bordeau, Poojitha Mutteneni, and Ben D. Sawyer. 2025. Evaluating GenAI for simplifying texts for education: Improving accuracy and consistency for enhanced readability. ArXiv:2501.09158.

Massimiliano Dibitonto, Alessio Ferrato, Carla Limongelli, and Olga Concetta Patroni. 2026. Museum audio guides generation using visitor categories and large language models. *Multimedia Systems*, 32:439.

Rotem Dror, Daniel Hutchinson, Mason A. Jones, Tsvi Kuflik, and Victoria Van Hyning. 2024. The curator's helper. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 496–504. Association for Computing Machinery.

Alessio Ferrato. 2025. Large language models to enhance learning in cultural heritage. volume 2590 of *Communications in Computer and Information Science*, pages 458–463. Springer.

George Mikros and Rania Voskaki. 2021. A modern Greek readability tool: Development of evaluation methods. In Adam Pawłowski, Ján Mačutek, Sheila Embleton, and George Mikros, editors, *Language*

and Text: Data, Models, Information and Applications, volume 356 of *Current Issues in Linguistic Theory*, pages 163–176. John Benjamins Publishing Company.

Jio Oh, Steven Euijong Whang, James Evans, and Jindong Wang. 2026. Classroom AI: Large language models as grade-specific teachers. *npj Artificial Intelligence*, 2:28.

Panagiota Papadimitriou and Chrysoula Savvaki. 2026. *Istoria G' Dimotikou: Vivlio Mathiti [History, Grade 3: Student's Book]*. 24grammata. Digital Library of School Textbooks, Textbook Registry Code 39, Identifier 20.500.14040/43632.

Sofia Stamou, Eleftherios Kalogeros, Matthew Damigos, and Manolis Gergatsoulis. 2027. Leveraging LLMs to build text narratives from CIDOC CRM. In *Metadata and Semantic Research: 19th Metadata and Semantics Research Conference, MTSR 2025, Thessaloniki, Greece, December 15–19, 2025, Revised Selected Papers*, volume 3057 of *Communications in Computer and Information Science*. Springer.

Aris R. Terzopoulos, Lynne G. Duncan, Mark A. J. Wilson, Georgia Z. Niolaki, and Jackie Masterson. 2017. HelexKids: A word frequency database for Greek and Cypriot primary school children. *Behavior Research Methods*, 49(1):83–96.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children's stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598. Association for Computational Linguistics.

Micah Watanabe. 2025. ChatGPT, can you make this text easier?: Investigating the utility of LLMs in easing text readability. *International Journal of Artificial Intelligence in Education*, 35(6):3922–3942.

Etymology Can Inform Lexical Complexity Prediction

Noah-Manuel Michael, Anna Hülsing, Andrea Horbach
Kiel University, Germany
Leibniz Institute for Science and Mathematics Education, Kiel, Germany
{michael|huelsing|horbach}@ipn.uni-kiel.de

Abstract

In Lexical Complexity Prediction (LCP) for second-language learners, cross-linguistic cognacy and borrowing relationships are commonly approximated using surface similarity. However, historical sound changes and orthographic differences can obscure etymological relationships that still facilitate vocabulary production. We investigate whether etymology-based features improve LCP for learners of English with German, Spanish, or Mandarin as source languages, using the 2026 BEA shared-task dataset. We expect etymological information to be most informative for German and Spanish, where historical relationships with English are frequent but often obscured by divergence. Mining Wiktionary etymologies, we derive nine etymology features for source–target word pairs, evaluated alongside surface-similarity and common lexical features in a stacked `RoBERTa-LightGBM` architecture. Across most configurations, adding etymology information significantly improves prediction performance for all three languages. For German and Spanish, the clearest contribution occurs for word pairs with normalized Levenshtein distances (NLDs) between 0.3 and 0.7, where surface similarity becomes less reliable, but common lexical features are not yet dominant. For Mandarin, etymology features remain useful, though less localized across NLD ranges. Our findings indicate that etymology effectively complements standard LCP approaches when surface similarity becomes less informative.

1 Introduction

Lexical Complexity Prediction (LCP) is a central task in educational NLP that aims to estimate how difficult individual words are for second-language (L2) learners (North et al., 2023). Accurate LCP is crucial for a range of downstream applications such as automated lexical simplification, readability assessment, and adaptive language-learning systems.

Lexical complexity is multidimensional and depends not only on properties of the target word, such as its frequency or length, but also on learner-related factors (Koutamanis et al., 2024). One such factor is the learner's first language (L1) through mechanisms such as cognate facilitation, where cognates are defined as words that share form and meaning across languages. Because of these similarities, they may be easier to acquire, retain, and retrieve (Otwinowska and Szewczyk, 2019).

Accordingly, cognacy has previously been approximated by surface similarity (Schepens et al., 2012; Mayer Martins et al., 2026). For example, German *Information* and Spanish *información* visibly resemble English *information*, whereas Mandarin 信息 *xìnxī* does not. However, following the historical-linguistic definition of cognacy, where cognates descend from a common origin (Crystal, 2008), surface similarity can only be an imperfect proxy, since regular sound changes, language-specific phonological adaptation, and differing orthographic conventions can obscure etymological relationships that may remain accessible to learners. For instance, English *father* and Spanish *padre*, as well as English *tongue* and German *Zunge*, are historically related despite their divergent surface forms. Similar considerations apply to lexical borrowings that have undergone phonological or orthographic adaptation in the respective languages.

We therefore argue that etymological information may provide an additional signal for computationally modeling lexical complexity that complements surface similarity. In particular, etymology-based features may be most informative for source–target word pairs with intermediate normalized Levenshtein distances (NLDs), for which the surface forms are neither sufficiently similar to provide a

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 173–196, November 19–20, 2026, University of Tartu Library

clear signal nor sufficiently dissimilar to rule out an etymological relationship. We expect this pattern to be clearest for German and Spanish, which offer comparatively many historically related forms due to their phylogenetic relatedness and centuries of shared culture and history facilitating language exchange. For Mandarin, a phylogenetically unrelated language to English, we expect weaker effects due to a comparatively smaller number of etymologically related words. If learners can exploit recurring cross-linguistic correspondences or knowledge of shared roots and borrowing histories, such relationships may facilitate the retention and subsequent production of English target words, resulting in lower experienced lexical complexity.

The 2026 BEA shared-task dataset on Vocabulary Difficulty Prediction for English learners is particularly suitable for our purpose because it measures the difficulty of producing an English target word from contextual information such as the source-language equivalent, the source-language word in context, and the first letter of the English target word plus a length indicator (e.g., a _ _ _ _ _ for *auburn*), rather than mere comprehension (Felice and Skidmore, 2026).[1] Michael et al. (2026) show that vocabulary items' productive difficulty exhibited by learners differs from receptive difficulty as intended by second language acquisition experts designing graded learning materials. The BEA setting allows us to investigate whether etymological relationships between L1 and English words affect learners' ability to retrieve and actively produce the target vocabulary.

To examine this question, we derive nine etymology-based features from Wiktionary entries for source–target word pairs, using `GPT-5.6-Luna` to parse the relevant etymological information. We investigate whether these features improve LCP performance per L1, whether they capture information distinct from surface similarity and other common lexical features, and whether their contribution differs across L1s. Specifically, we address the following research questions:

- RQ1: What is the predictive potential of etymology features in LCP, both alone and combined with existing features?
- RQ2: To what extent do etymology features contribute information distinct from features previously employed in LCP?
- RQ3: Is there a specific NLD range where etymology features are most effective, and does it vary across L1s?
- RQ4: To what extent do etymology features, alone or combined with existing features, further improve a transformer baseline in LCP?

Our results show that etymology features can inform LCP, particularly for German and Spanish source–target word pairs with intermediate NLDs (0.3–0.7). For Mandarin, etymology features remain helpful overall, but as expected for a more lexically distant L1, their contribution is less clearly localized across NLD ranges. Common lexical features are generally more competitive for this subset.

We make our code available at `https://github.com/noahmanu/etymology4lcp`.

2 Related Work

In the following, we first lay out the development of the L1-aware BEA Vocabulary Difficulty Prediction task. We then elaborate on the mechanism of cognate facilitation and the role of etymology information in LCP thus far.

Development of the BEA Vocabulary Difficulty Prediction Task. The 2026 BEA Shared Task is the cumulation of over a decade of research in automatically predicting word difficulty: Complex Word Identification is the binary task that decides which words in a given sentence are complex and therefore require simplification (Paetzold and Specia, 2016). The relevance of this task is underlined by it being the topic of several shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018; Ortiz-Zambrano and Montejo-Ráez, 2020). Instead of the binary Complex Word Identification label, Shardlow et al. (2020) proposed a continuous label, which allows annotators to make more fine-grained complexity decisions. This task became known as LCP. It was the sole focus of the SemEval 2021 Shared Task (Shardlow et al., 2021) and was later combined with lexical simplification in the BEA 2024 Multilingual Lexical Simplification Pipeline Shared Task (Shardlow et al., 2024). Vocabulary Difficulty Estimation is another related task, often framed as a CEFR-level classification task of words both in context (Bannò et al., 2025) and out of context (Alfter, 2024). The 2026 BEA Shared Task on Vocabulary Difficulty Prediction for English learners aims to automatically estimate

[1] In the context of this study, we treat lexical complexity and vocabulary difficulty synonymously.

a continuous difficulty target quantifying the likelihood of an average learner being able to produce target vocabulary given certain contextual information (Felice and Skidmore, 2026). This shared task distinguishes between learners of English with different L1s (Spanish, German, and Mandarin), building on prior work on L1-aware difficulty prediction (Palmero Aprosio et al., 2020; Skidmore et al., 2025).

The Role of L1 in LCP. While early Complex Word Identification and LCP work largely treated complexity as universal, later studies show that it differs between native and non-native readers (Gooding et al., 2021; North and Zampieri, 2023) and, crucially, between L2 learners with different L1 backgrounds (Ide et al., 2023). This motivates L1-aware prediction, as adopted by the 2026 BEA Shared Task.

Cognate Facilitation. Cognate facilitation effects are well known in psycholinguistic research on bilingualism and describe learners' general ability to recognize that two words from different languages sharing formal resemblance are "somehow the same thing" (Carroll, 1992), facilitating the comprehension, retention, and production of cross-linguistically similar items (Otwinowska and Szewczyk, 2019; Schepens et al., 2020; Woumans et al., 2021; Engemann and Radetzky, 2024). Accordingly, cognacy thus far has been estimated based on surface similarity. For instance, Beijering et al. (2008) and Schepens et al. (2012) used Levenshtein distance measures to model lexical distance and found robust correlations with bilingual speakers' distance judgments. More recently, Mayer Martins et al. (2026) participated in the 2026 BEA Shared Task and proposed a cosine similarity feature derived from TF-IDF-weighted n-gram vectors between word pairs as a proxy for cognacy. Löfflad and Kanli (2026) propose Levenshtein distance as a proxy for cognacy in L1-aware Automatic Readability Assessment.

Etymology in LCP. In historical linguistics, in contrast, a cognate is "a linguistic form which is historically derived from the same source as another [...] form" (Crystal, 2008). We argue that these historical developments may be accessible to learners to some degree, even if surface forms diverge, and may therefore facilitate LCP.

As Tack (2021) points out, "the link between perceived difficulty and the word's etymology has not yet been thoroughly investigated." So far, only tentative approaches at informing LCP with etymology information have been made. These include Tack's exploration of features based on Etymological WordNet, which, similarly to our approach, is mined from Wiktionary (de Melo, 2014), as well as the work of Philippov et al. (2026), who note the use of an etymology feature but provide no further details. To the best of our knowledge, we are the first to deliver an L1-aware, systematic evaluation of etymology features' impact on LCP performance.

3 Data

This section describes the dataset used in our experiments and the preprocessing we apply.

3.1 Dataset

We conduct our experiments on the 2026 BEA shared-task dataset on Vocabulary Difficulty Prediction for English learners (Felice and Skidmore, 2026), which is based on data collected in the context of the Knowledge-based Vocabulary Lists project (Schmitt et al., 2024). For that project, learners were asked to produce an English target word given its source-language equivalent(s) and contextual information including a source-language example sentence and the first letter of the English target word plus a length indicator (e.g., a _ _ _ _ _ for *auburn*). The dataset contains 7,516 parallel data points for each of the three source languages (German, Spanish, and Mandarin). Each source–target word pair is associated with a continuous Generalized Linear Mixed Model-derived lexical complexity score, reflecting the likelihood that a learner of average ability from the corresponding L1 population can produce the English target word given the contextual information. Higher scores indicate lower complexity.

3.2 Preprocessing

To maximize the number of instances available for our experiments, we pool the original training, development, and test splits across all source languages used in the shared task into a single dataset. This enables us to use a five-fold cross-validation setup (see Section 5.1 for more details).

We then perform the following preprocessing steps to prepare the data for feature computation: For the English target words, we only perform lowercasing. The L1 source-word column, on

the other hand, often contains multiple comma-separated alternatives as well as systematic additional annotations, such as argument placeholders or hints indicating which potential English target word is not intended. We therefore first apply a cleaning pipeline that removes bracketed annotations, hyphens, slashes, and language-specific grammatical markers while explicitly preserving comma-separated translation alternatives. Second, to create a fully standardized L1 representation, we then lowercase the text, convert Chinese characters to pinyin to enable the computation of surface-similarity features, and remove all diacritical marks (e.g., umlauts in German, accents in Spanish, and tone marks in Mandarin). Last, we identify the L1 alternative closest to the English target word based on their NLDs (scaled by the geometric mean of the word-pair lengths) and use this selected alternative to compute all downstream features.

4 Features

In the following sections, we introduce the three groups of manually engineered linguistic features whose contributions to LCP performance we evaluate alongside contextual predictions generated by `XLM-RoBERTa-base`:[2] **C**ommon lexical features, **S**imilarity features, and **E**tymology features based on etymological information extracted from Wiktionary. Table 1 provides an overview of the feature groups, which we describe in more detail below. Appendix B illustrates the possible values of each individual feature. We also briefly explain our reasoning for including an L1 feature.

4.1 Common Lexical Features

We define common lexical features as features that have repeatedly proven useful in previous LCP studies and that are not covered by our similarity feature group. We select word frequency, target word length, the estimated number of syllables in the target word, and its POS to serve as a representative subset of this family (North et al., 2023).

4.2 Similarity Features

While similarity features are also well established in LCP (Shardlow et al., 2021; Bestgen, 2026; Robaina et al., 2026), we treat them as a separate feature group because they constitute the conventional surface-form-based proxy for cross-linguistic relatedness against which we compare our etymology features. This distinction allows us to assess whether etymology captures information beyond that provided by orthographic and phonological similarity measures. As representatives of this feature group, we select raw and normalized Levenshtein distances (scaled either by maximum word length or the geometric mean of the word-pair lengths) computed on both orthographic forms and IPA transcriptions, Jaro and Jaro–Winkler distances, and cosine similarity across TF–IDF-weighted character n-gram vectors, as introduced by Mayer Martins et al. (2026).

4.3 Etymology Features

We introduce a novel family of etymology-based features designed to capture the historical, etymological, and cross-lingual connections between L1 source and English target words. We expect them to be informative for lexical complexity, particularly when regular sound changes, language-specific phonological adaptation, or differing orthographic conventions obscure etymological relationships at the surface level.

Generation of Etymology Features. We derive the etymology features from etymological information extracted from Wiktionary (Ylonen, 2022).[3] For each source–target word pair, we collect the relevant etymology sections from both the English edition and, where available, the edition corresponding to the respective source language. We then use `GPT-5.6-Luna` as a zero-shot feature annotator to parse this information and generate a `JSON` object containing the feature values. Brysbaert et al. (2025) and Philippov et al. (2026) have shown that using LLM-generated word familiarity estimates as features for LCP achieves improved performance. Therefore, where there is no etymology information available, we ask the LLM to fill in from its own knowledge. We provide the prompt we use in Appendix C and an overview of how many items have etymology information available in Appendix D. Coverage is comprehensive for English target words as well as German and Spanish source words. However, it is markedly lower for Mandarin, where etymological information is available for only approximately 10% of the source words. Following a manual inspection of 120 Mandarin instances,

[2] Corresponds to the baseline model used in the BEA Shared Task (Felice and Skidmore, 2026). We will henceforth shorten `XLM-RoBERTa-base` to `RoBERTa`.

[3] `https://kaikki.org`, last accessed 2026/10/30.

Family	Feature	Description
Common	log_frequency	Frequency-based lexical familiarity measure (SUBTLEX-UK)
	target_word_length	Number of characters in the English target word
	num_syllables	Estimated number of syllables in the target word
	pos_tag	Part-of-speech category of the target word
Similarity	char_edit_dist	Raw Levenshtein distance between source and target word
	char_edit_dist_norm	Maximum length-normalized Levenshtein distance between source and target word
	char_edit_dist_geo_norm	Geometric mean-normalized Lev. distance between source and target word (Mayer Martins et al., 2026)
	phon_edit_dist	Raw Levenshtein distance between IPA transcriptions of source and target word
	phon_edit_dist_norm	Maximum length-normalized Levenshtein distance between IPA transcriptions of source and target word
	phon_edit_dist_geo_norm	Geometric mean-normalized Levenshtein distance between IPA transcriptions of source and target word
	jaro_dist	Jaro distance between source and target word
	jaro_winkler_dist	Jaro-Winkler distance between source and target word
	soundex_dist	Raw Levenshtein distance between soundex codes of source and target words
	metaphone_dist_norm	Maximum length-normalized Levenshtein distance between metaphone codes of source and target words
	char_cosine_sim	Cosine similarity over TF-IDF-weighted character n-gram vectors (Mayer Martins et al., 2026)
Etymology	lca_language	Lowest common ancestral language stage (strictly excluding borrowing)
	same_proto_word	Whether both words descend from the exact same reconstructed proto-word
	same_proto_root	Whether both words share a broader root, but not the exact proto-word
	borrowed_from_source	Whether the English word was directly borrowed from the L1
	borrowed_from_target	Whether the L1 word was directly borrowed from English
	borrowed_from_third	Whether the L1 word and the English word were both borrowed from the same third language
	english_word_origin	Immediate taxonomic branch from which English acquired the word
	origin_matches_source_family	Whether the English word's immediate origin aligns with the L1 family
	transmission_path	Complexity of the word's transmission path into English

Table 1: Feature families, feature names, and descriptions. Feature value sets can be found in Appendix B.

we observe that the available etymological information often pertains to borrowed vocabulary, whereas native Chinese terminology is rarely covered. Because these borrowed terms align with our primary subject of interest, we maintain that our feature generation pipeline is sufficiently robust to produce accurate feature values for the Chinese subset.

Features Touching on Cognacy. To model the depth of the genetic lineage between the word pair, we extract the lowest common ancestor language, representing the youngest ancestral language stage at which the two words were still a single, indistinguishable lexical item. A closer shared ancestor, such as West Germanic compared to Proto-Indo-European, can imply higher degrees of structural and semantic similarity, thereby lowering the cognitive load for the learner. For instance, while English and German are both West Germanic languages, English *egg* and German *Ei* share a Proto-Germanic ancestor **ajją*, but not a Proto-West-Germanic one **aij*, because English lost its native West Germanic term *ey* and replaced it with the Old Norse borrowing *egg*.

We further refine this genetic linkage by determining whether the words descend from the exact same proto-word or merely share a foundational proto-root. Descending from a fully formed ancestral word acts as a strong cross-lingual anchor through true cognacy. For example, English *fish* and Spanish *pez* share a proto-word because both are inherited directly from the exact same Proto-Indo-European noun **péysks*. In contrast, English *youth* and Spanish *juventud* share only a proto-root. While both derive from the underlying Proto-Indo-European root for 'young' (**yeu-*), they do not descend from a single ancestral noun; rather, they were formed independently when the Germanic and Italic branches appended different abstract suffixes to that root. Capturing this distinction allows our models to differentiate between the maximum transparency of inheriting an entire word and the morphological cue of sharing ancestral building blocks.

Features Touching on Borrowing Relationships. To account for direct vocabulary transfer, we include two binary features indicating whether the English word was borrowed from the source language (e.g., *doppelganger* for a German-speaking learner) or whether the source word was borrowed from English (e.g., 沙拉 *shālā* 'salad' for a Mandarin-speaking learner). We also include a third feature indicating whether both the source and target words were borrowed from the same third language (e.g., English *tile* and German *Ziegel*, both derived from Latin *tēgula*).

Furthermore, we classify the taxonomic origin of the English word (i.e., Germanic, Romance, Sinitic, or Other) to capture the immediate genealogical branch from which an English word originates. In particular, Romance-derived vocabulary often corresponds to more formal or academic registers and may therefore be more complex than native Germanic vocabulary, irrespective of the learner's L1. We complement this feature with a binary indica-

tor of whether the English word's origin matches the source language family. This feature captures the possibility that words derived from a branch related to the learner's L1 may exhibit more familiar morphological patterns or regular correspondences. Thus, when German- or Spanish-speaking learners encounter Germanic- or Romance-derived English words, respectively, this familiarity may reduce the cognitive load involved in producing the word, even in the absence of direct cognacy.

Finally, we categorize the transmission path through which the word entered the English language (i.e., inheritance, direct borrowing, or indirect borrowing through one or more intermediary languages). Words that passed through multiple intermediary languages often undergo substantial semantic or orthographic shifts, obscuring their origins and potentially increasing their difficulty relative to natively inherited or directly borrowed vocabulary.

4.4 Inclusion of an L1 Feature

As the complexity-score distributions in the BEA shared-task dataset were derived from different population samples for each L1 (Skidmore et al., 2025), score values are not necessarily directly comparable across source languages. However, training on data pooled across L1s has generally been found to improve performance (Skidmore et al., 2025; Felice and Skidmore, 2026); we therefore adopt this approach. To enable our single multilingual model to account for language-specific differences in the interpretation and distribution of complexity scores, we also include L1 as an input feature in most configurations.

5 Method

Here, we describe our model architecture and the evaluation metrics we use. We also briefly introduce our experimental setup and our analyses.

5.1 Model Architecture

We conduct two complementary sets of experiments. First, to assess the predictive potential of etymology features in isolation and in combination with other features previously used in LCP, we train and evaluate `LightGBM` models using different feature group configurations. We choose `LightGBM` because tree-based models have consistently proven effective for LCP (Ortiz-Zambrano et al., 2025; Nohejl et al., 2026), while their feature-based nature enables controlled ablation analyses.

Second, we examine whether the observed effects persist in a stronger hybrid architecture that combines contextual predictions from a pretrained transformer language model with engineered features. This design is motivated by the design employed by Ortiz-Zambrano et al. (2025), who showed that manually engineered linguistic features complement contextual representations from pretrained language models in English and Spanish LCP. Their results indicate that hybrid models can outperform purely end-to-end approaches, particularly in multilingual settings. Accordingly, we use a `LightGBM` meta-regressor stacked on predictions from `XLM-RoBERTa-base` (Mosquera, 2021), which is the BEA shared-task baseline model. This architecture leverages the predictive capacity of a pretrained transformer model while retaining the ability to ablate the contribution of the feature groups introduced above.

We obtain the `RoBERTa` predictions using nested cross-validation to prevent data leakage. For each of five outer folds, we partition the corresponding training data into four inner folds. Using the HuggingFace `Trainer`,[4] we fine-tune `RoBERTa` on three inner folds and generate out-of-fold predictions for the remaining fold, repeating this procedure across all four splits. The resulting out-of-fold predictions provide a complete, leakage-free training signal for the `LightGBM` meta-regressor. Parallel instances of the same vocabulary item across the three source languages are always assigned to the same fold in both the outer and inner splits. Fine-tuning follows the hyperparameters and contextual input format of the BEA shared-task baseline: the L1 source word, the L1 source context sentence, the first letter of the English target word followed by underscores for its remaining characters as presented as a length cue to the learner, and the full English target word. See Appendix E.1 for an overview of both the `RoBERTa` and `LightGBM` hyperparameters we employ in our model training.

In parallel, we fine-tune a final round of `RoBERTa` models on the complete outer training partition and use them to predict the held-out outer test folds. The `LightGBM` models are trained on the out-of-fold predictions and engineered features extracted exclusively from the outer training data (see Ap-

[4] `https://huggingface.co/docs/transformers/en/main_classes/trainer`, last accessed 2026/10/30.

Config	Full	German	Spanish	Mandarin
B	.786 | 1.179	.786 | 1.193	.772 | 1.265	.804 | 1.071
E	.352 | 1.661	.429 | 1.608	.346 | 1.755	.266 | 1.617
^{L1}	.003 | 1.775	-.017 | 1.778	-.014 | 1.869	-.007 | 1.671
^{L1}C	.512 | 1.532	.466 | 1.583	.459 | 1.670	.616 | 1.321
^{L1}S	.498 | 1.544	.532 | 1.511	.513 | 1.607	.436 | 1.510
^{L1}E	.367 | 1.651**	.438 | 1.599**	.356 | 1.747**	.280 | 1.604**
^{L1}CS	.640 | 1.366	.651 | 1.352	.633 | 1.447	.635 | 1.293
^{L1}CE	.601 | 1.422**	.608 | 1.416**	.565 | 1.547**	.635 | 1.294**
^{L1}SE	.528 | 1.510**	.560 | 1.476**	.547 | 1.567**	.462 | 1.486**
^{L1}CSE	.659 | **1.337****	.669 | **1.323****	.657 | **1.411****	.649 | **1.273****

Table 2: Performance scores (Pearson r on the left-hand side; RMSE on the right-hand side) excluding baseline RoBERTa predictions for all models except B, which is the RoBERTa baseline predictions themselves. Performance gains of underlined configurations were tested for significance with paired one-sided sign-flip tests (10,000 permutations; $* = p < .05$, $** = p < .01$) against their corresponding configurations excluding E. Holm–Bonferroni correction was applied across all tests.

pendix E.2). At test time, they receive the final RoBERTa predictions for the outer test folds together with the corresponding engineered features.

We repeat this procedure independently for all five outer folds. Figure 5 in Appendix E.2 illustrates the inner–outer cross-validation procedure and the final meta-regressor training. Finally, we aggregate predictions across the outer test folds to compute overall and L1-specific performance metrics, as well as to conduct a sliding-window analysis of the NLD between source–target word pairs.

5.2 Evaluation

Following Felice and Skidmore (2026), our main evaluation metric is the root mean squared error (RMSE) between the predicted and gold complexity scores, which was used as the official evaluation metric in the BEA Shared Task. We additionally report Pearson r to quantify the strength of the linear relationship between predicted and observed complexity values. We report numerical performance scores across the entire dataset and the different L1 subsets individually. Following the BEA shared-task evaluation scheme, our visual analyses focus solely on the individual L1 subsets.

5.3 Experimental Setup

To answer the research questions introduced in Section 1, we conduct four experiments. EXPERIMENT 1 addresses RQ1 by evaluating the predictive potential of common lexical (C), surface-similarity (S), and etymology-based (E) feature groups, both individually and in all possible combinations. We additionally assess how their performance compares with that of the BEA shared-task RoBERTa baseline (B). To account for L1-specific complexity-score distributions (Skidmore et al., 2025), we include an explicit L1 feature in our analysis, denoted by the superscript ^{L1}.

EXPERIMENT 2 addresses RQ2 by examining the overlap of correct predictions produced by the individual components B^{L1}, ^{L1}C, ^{L1}S, and ^{L1}E. EXPERIMENT 3 addresses RQ3 through both a performance analysis across NLD windows and a candidate-adjusted enrichment analysis of exclusively correct predictions across NLD windows. Together, these analyses test whether the feature groups capture distinct signals and identify NLD ranges where these signals are most informative.

Finally, EXPERIMENT 4 addresses RQ4 by repeating the feature-group and NLD-window analyses from EXPERIMENT 1, with the RoBERTa baseline predictions (B) included in every configuration.

6 EXPERIMENT 1: Feature-Group Performance and Combinations Excluding Transformer Baseline

Setup. EXPERIMENT 1 evaluates the predictive potential of the three engineered feature groups in isolation. We train LightGBM models using E alone, L1 alone, and L1 combined with every possible subset of C, S, and E: ^{L1}C, ^{L1}S, ^{L1}E, ^{L1}CS, ^{L1}CE, ^{L1}SE, and ^{L1}CSE. We report the performance of the transformer baseline, B, as a reference point.

For each configuration containing E, we test whether its RMSE is significantly lower than that of the otherwise identical configuration without E. Thus, we compare ^{L1}E with ^{L1}, ^{L1}CE with ^{L1}C, ^{L1}SE with ^{L1}S, and ^{L1}CSE with ^{L1}CS.

Inclusion of an L1 Feature. Adding L1 as an additional component to the feature groups is useful as it allows the model to account for systematic differences in score distributions between the three L1 subsets, as illustrated by the slight performance improvements of ^{L1}E over E (see Table 2). We therefore include L1 as a regular feature component.

Results. As Table 2 illustrates, etymology features improve every matched feature configuration. Across the full dataset and the three L1 subsets, all 16 paired comparisons between models with and without E yield significant RMSE reductions after Holm correction. ^{L1}CSE is consistently

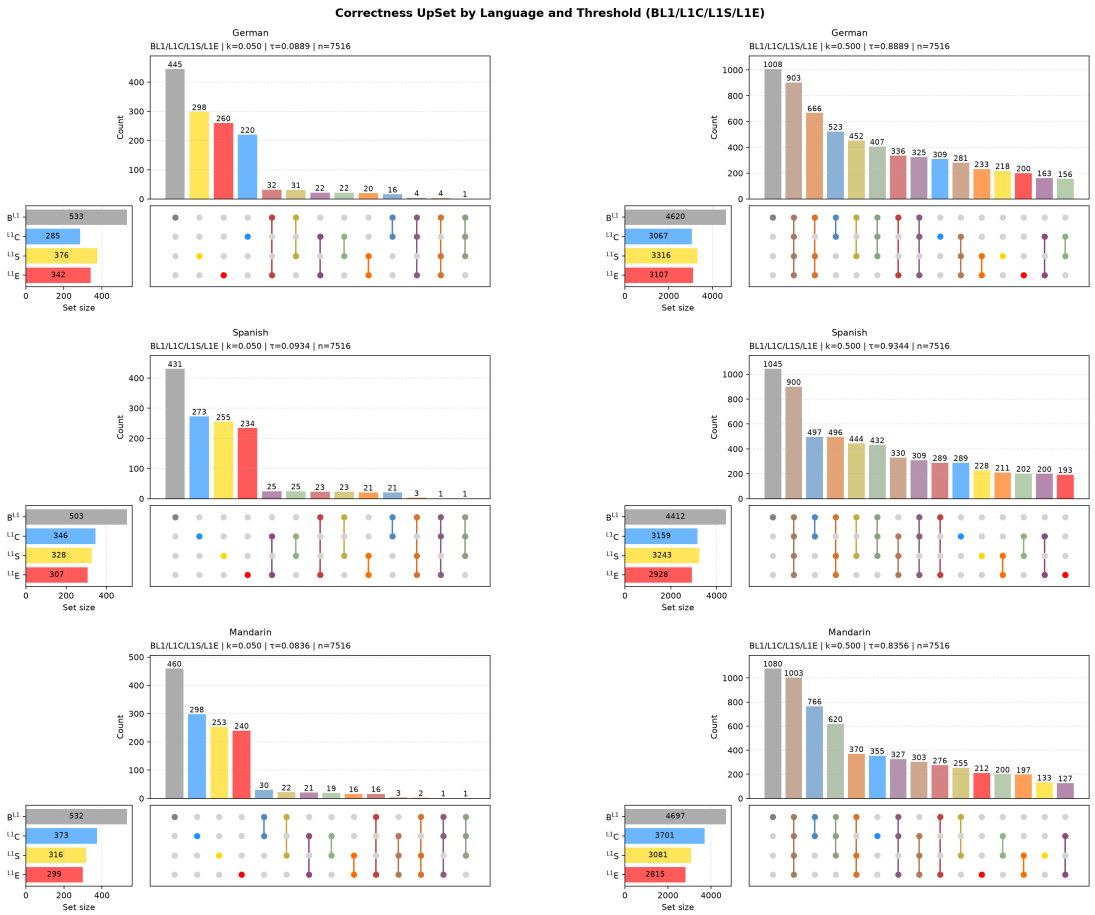


Figure 1: UpSet plots depicting the aggregate number of correct predictions per feature-group component (set size). The plots further quantify the overlap in model performance, contrasting the proportion of instances correctly classified by a singular model against those recognized by multiple models across the two evaluation thresholds.

the strongest configuration, achieving the lowest RMSE among the non-transformer models on both the pooled data and each L1 subset. Thus, although etymological information alone is insufficient to match the performance of common lexical features or the transformer baseline, it consistently complements both common and surface-similarity features.

While, by itself, the etymology group is the weakest feature group in terms of RMSE, it is clearly predictive: $^{L1}\mathsf{E}$'s Pearson r ranges from .280 for Mandarin to .356 for Spanish and .438 for German. This order is consistent with the greater lexical similarity that English shares with German and Spanish, which provides more opportunities for etymological relationships to be informative.

7 Experiment 2: Overlap of Correct Predictions

Setup. The aggregate results in Experiment 1 establish that the feature groups differ in predictive performance, but they do not reveal whether the models succeed on the same instances. To address RQ2, we therefore compare the overlap of correct predictions made by the individual base configurations: An L1-enriched transformer baseline B^{L1}, and the feature-only models $^{L1}\mathsf{C}$, $^{L1}\mathsf{S}$, and $^{L1}\mathsf{E}$.

Because the target variable is continuous, correctness must be operationalized using an error tolerance. For each L1 subset l, we define a prediction as correct if its absolute error does not exceed a threshold τ based on the standard deviation σ_l of the gold complexity scores in that subset. We report a strict threshold and a lenient threshold, $\tau \in \{0.05\sigma_l, 0.5\sigma_l\}$, to identify highly accurate predictions and assess whether the observed pattern remains when a broader range of predictions is considered acceptable.

We visualize the resulting intersections using UpSet plots (Lex et al., 2014; Figure 1). These plots show both the total number of correct predictions per configuration and the number of items

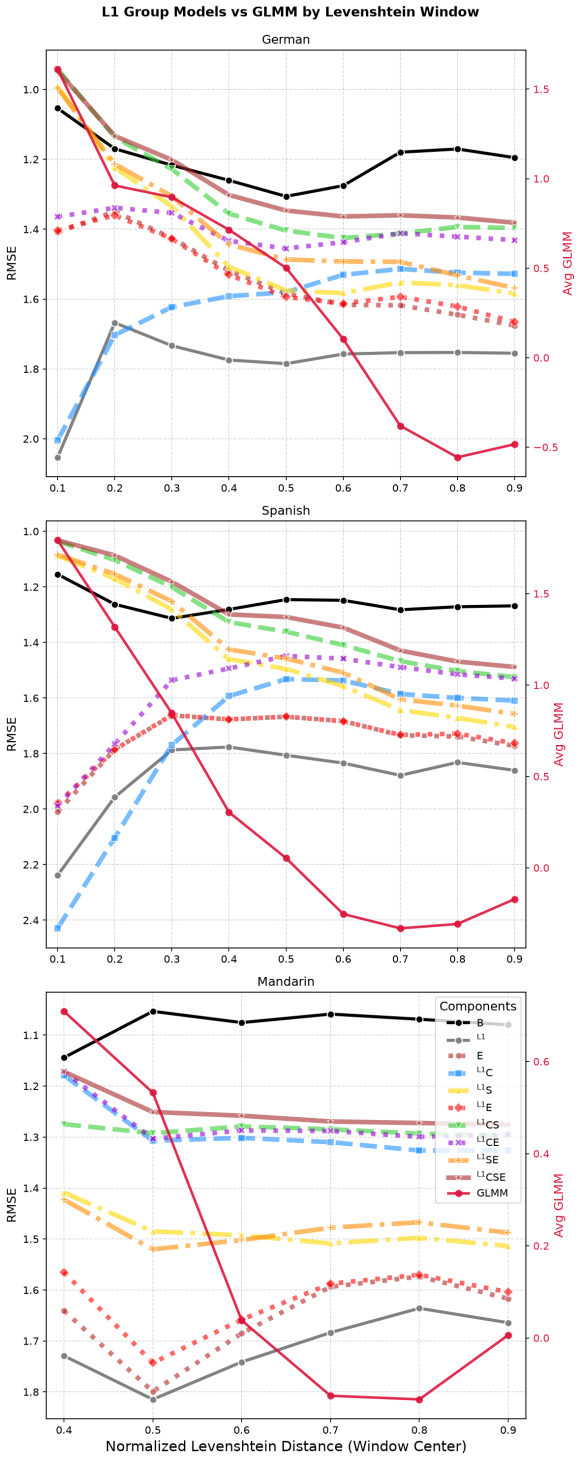


Figure 2: Sliding-window performance analysis of feature-group configurations across NLD windows, visualizing each feature group's predictive potential in isolation and in all possible combinations, excluding the transformer baseline predictions.

predicted correctly by each possible combination of models. Of particular interest are the single-model intersections, which represent instances correctly predicted by only one configuration.

Results. Figure 1 shows that none of the feature-group models is redundant with the transformer baseline or with the other feature-group models. Across all three L1 subsets and at both thresholds, each configuration produces instances that no other configuration predicts correctly.

As expected, the stricter threshold yields relatively little overlap between configurations: Only a small number of instances are predicted with very high accuracy by multiple models. Under the lenient threshold, overlap increases substantially because more predictions qualify as correct. Crucially, however, all models retain exclusive correct predictions under both definitions of correctness.

8 Experiment 3: NLD-Specific Performance and Enrichment

Setup 3A. To probe whether the usefulness of the feature groups differs depending on the surface similarity of the source–target word pair, we first conduct a sliding-window analysis of RMSE performance across NLD, scaled by maximum length. We calculate RMSE for overlapping windows of width 0.2, ranging from 0.0–0.2 to 0.8–1.0 in increments of 0.1. Alongside the RMSE curves, we plot the Generalized Linear Mixed Model-derived complexity score on a second y-axis to illustrate how vocabulary difficulty varies with the surface similarity between word pairs.

Results 3A. As Figure 2 shows, for German and Spanish, surface-similarity features are most useful at low NLD values, where source and target words are visibly similar. Their relative contribution declines as NLD increases. In contrast, models containing etymology features show their clearest advantage in intermediate NLD ranges, approximately 0.3–0.8. In these windows, they regularly outperform otherwise comparable models without E, particularly those relying on similarity features alone. At high NLD values, common lexical features become increasingly competitive and often provide the strongest signal, plausibly because lexical properties such as word frequency are much more informative and reliable predictors when cross-linguistic form similarity is low.

As expected, the pattern is less pronounced for Mandarin. Few Mandarin–English pairs occur below an NLD of 0.4; windows below this range therefore contain too few instances for reliable interpretation (see Appendix F). Across the better-supported windows, etymology features remain

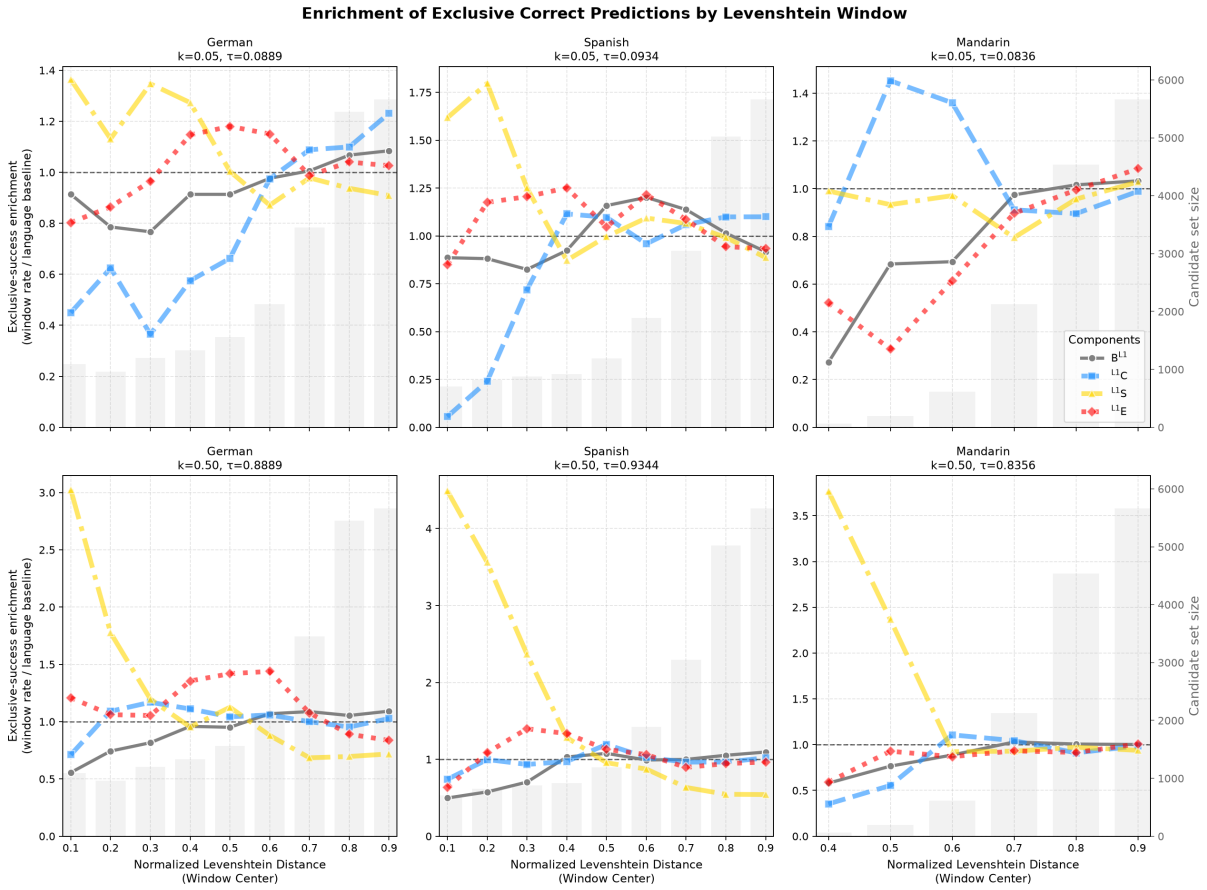


Figure 3: Candidate-adjusted enrichment analysis of exclusive correct predictions over NLD windows, visualizing each feature group's specific operational niche across L1s under $\tau \in \{0.05\sigma_l, 0.5\sigma_l\}$.

beneficial overall, especially in combination with common lexical features. However, common lexical features are generally more competitive than either similarity or etymology features, and the NLD-specific trends are much less distinct than for German and Spanish. This result is consistent with the observed lower number of surface-similar word pairs and the expected lower number of etymologically related word pairs in the Mandarin–English dataset.

Setup 3B. As mentioned above, the underlying dataset exhibits a natural skew, containing a disproportionately high volume of word pairs with higher NLDs (see Appendix F). Because of this imbalance, analyzing the raw distribution of correct predictions obscures a feature's true operational strengths by conflating model performance with candidate density per window. To address this issue and accurately isolate the specific operational niche of each feature set, we additionally calculate the candidate-adjusted enrichment of exclusive correct predictions per base feature model configuration (B^{L1}, $^{\mathsf{L1}}\mathsf{C}$, $^{\mathsf{L1}}\mathsf{S}$, $^{\mathsf{L1}}\mathsf{E}$). This metric adjusts for both local candidate density and the model's language-specific overall exclusive-success rate.

For each model configuration m, NLD window w, and source language l, we define the candidate-adjusted enrichment of exclusive correct predictions as:

$$E_{m,w,l} = \frac{U_{m,w,l}/N_{w,l}}{U_{m,l}/N_l}, \tag{1}$$

where $U_{m,w,l}$ is the number of predictions that are correct only for model m in window w and language l, $N_{w,l}$ is the number of candidate pairs in that window, $U_{m,l}$ is the total number of exclusive correct predictions made by m for language l, and N_l is the total number of instances for that language. As in EXPERIMENT 2, we calculate correctness under both the strict and lenient error thresholds. An enrichment value above 1 indicates that a model's exclusive correct predictions occur more often in a given NLD window than would be expected from its overall exclusive-success rate.

Config	Full	German	Spanish	Mandarin
B	.786 | 1.179	.786 | 1.193	.772 | 1.265	.804 | 1.071
BE	.779 | 1.114**	.779 | 1.116**	.763 | 1.209**	.798 | 1.007**
B^{L1}	.779 | 1.113	.780 | 1.113	.764 | 1.207	.797 | 1.010
$B^{L1}C$	.781 | 1.110	.779 | 1.116	.766 | 1.203	.801 | 1.002
$B^{L1}S$	.782 | 1.108	.780 | 1.114	.773 | 1.187	.794 | 1.015
$B^{L1}E$	.779 | 1.113	.780 | 1.114	.763 | 1.209	.798 | 1.008
$B^{L1}CS$	.792 | 1.084	.791 | 1.089	.786 | 1.157	.802 | 1.000
$B^{L1}CE$	.785 | 1.101**	.783 | 1.107*	.770 | 1.194*	.804 | **0.994***
$B^{L1}SE$	.784 | 1.103**	.783 | 1.108	.775 | 1.183	.796 | 1.011
$B^{L1}CSE$	.796 | **1.075****	.796 | **1.078****	.790 | **1.148****	.804 | **0.994***

Table 3: Performance scores (Pearson r on the left-hand side; RMSE on the right-hand side) including RoBERTa baseline predictions. B is the RoBERTa baseline predictions themselves. Performance gains of underlined configurations were tested for significance with paired one-sided sign-flip tests (10,000 permutations; $* = p < .05$, $** = p < .01$) against their corresponding configurations excluding E. Holm–Bonferroni correction was applied within hypothesis families (separated in the table).

Results 3B. The enrichment analysis identifies distinct NLD-specific niches for the feature groups (Figure 3). For German and Spanish, ^{L1}S is most strongly enriched at low NLD values, approximately 0.0–0.4. This result is expected: When source and target forms are visibly similar, surface-based measures provide a direct cue to cross-linguistic relatedness. As NLD increases, however, the enrichment of ^{L1}S declines and frequently falls below its language-specific baseline rate.

In contrast, ^{L1}E exhibits its clearest enrichment in intermediate NLD ranges, approximately 0.3–0.7, for both German and Spanish. In these windows, etymology features account for a disproportionate share of instances that neither common or similarity features nor the transformer baseline predict correctly. This pattern supports our hypothesis that etymological information is particularly informative when etymological relationships remain present and potentially accessible to the learner, but are no longer sufficiently transparent to be captured reliably by surface similarity alone. For German, examples of correct predictions of word pairs that share an etymological relationship made exclusively by the ^{L1}E model include *Hüfte* 'hip', *schießen* 'shoot', and *Kirsche* 'cherry'; for Spanish, we identify word pairs such as *vidua* 'widow', *salchicha* 'sausage', and *igualdad* 'equality'. See Appendix G for a more exhaustive list.

At higher NLD values, approximately 0.7–1.0, common lexical features become more strongly enriched. In these cases, target-word properties such as frequency, length, syllable count, and POS provide a stable signal when neither surface similarity nor etymological relationship is informative. The transformer baseline also tends to be more reliable at the high end of the NLD spectrum and comparatively less enriched for low-NLD pairs, which suggests that it likely relies on cues distinct from the information captured by similarity and etymology features.

The Mandarin results differ from the German and Spanish patterns, as expected. At the lenient threshold, similarity features show some enrichment in the intermediate range, whereas at the strict threshold, they are generally not enriched. Common lexical features are particularly informative in the middle-to-higher NLD range, approximately 0.4–0.7. Etymology features show limited positive enrichment, primarily at the highest NLD values. Examples of exclusively correct predictions by ^{L1}E of items with etymological relationships include 伏特加 *fútèjiā* 'vodka', and 汉堡 *hànbǎo* 'hamburger'. These results should be interpreted cautiously because Mandarin–English pairs are concentrated at high NLD values and low-NLD windows contain few instances. Overall, etymology features remain useful and slightly outperform similarity features, but no clear enrichment trend emerges across the statistically well-supported windows.

Taken together, EXPERIMENTS 2 and 3 show that etymology features capture a distinct predictive signal. For German and Spanish, this signal is concentrated in the intermediate NLD range, where surface similarity has become less reliable and common lexical features have not yet become dominant. For Mandarin, the signal is less clearly localized across NLD ranges.

9 EXPERIMENT 4: Feature-Group Performance Analysis Including Transformer Baseline

Setup. EXPERIMENT 4 addresses RQ4 by repeating the analyses from EXPERIMENT 1 (4A) and EXPERIMENT 3A (4B) in the hybrid architecture described in Section 5.1. Every LightGBM configuration now receives the out-of-fold RoBERTa predictions, B, as an additional input feature.

We evaluate BE, B^{L1}, and all combinations of B^{L1} and the three feature groups: $B^{L1}C$, $B^{L1}S$, $B^{L1}E$, $B^{L1}CS$, $B^{L1}CE$, $B^{L1}SE$, and $B^{L1}CSE$. We again test configurations containing E against their otherwise

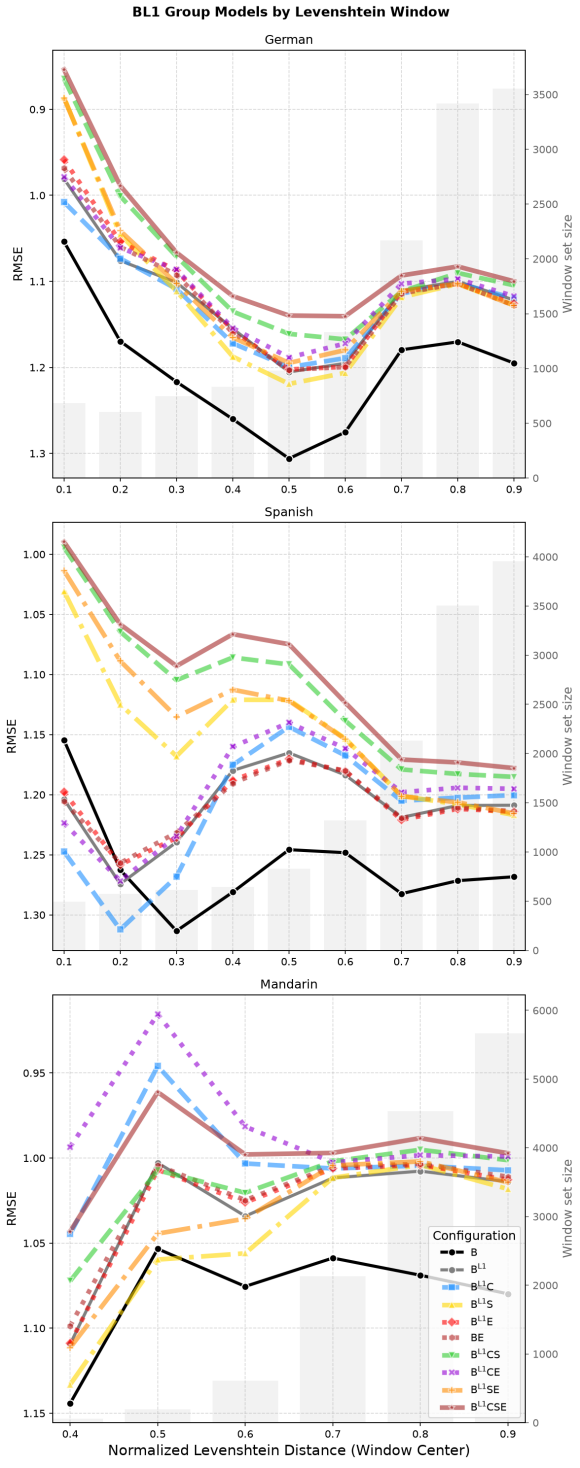


Figure 4: EXPERIMENT 4B: Sliding-window performance analysis of feature-group configurations across NLD windows, visualizing each feature group's impact on performance in all possible combinations, including the transformer baseline predictions.

identical counterparts without E. We also repeat the NLD sliding-window analysis.

Results. The hybrid results presented in Table 3 confirm that etymology features remain useful beyond the transformer baseline, although their effect is smaller than in the feature-only setting. Adding E to B yields a significant reduction in RMSE for the pooled data and for each L1 subset.

The best overall stacked configuration is $B^{L1}CSE$. It achieves the lowest RMSE on the pooled data as well as on the German and Spanish subsets. For Mandarin, $B^{L1}CSE$ and $B^{L1}CE$ obtain the same lowest RMSE. Crucially, $B^{L1}CSE$ significantly improves upon the corresponding configuration without etymology features, $B^{L1}CS$, for the full dataset and all three L1 subsets. Adding E to configurations containing common lexical features also produces significant improvements in every subset.

However, the effect of etymology features is not equally strong in all configurations. $B^{L1}E$ does not significantly improve upon B^{L1}, indicating that the etymology group alone offers limited additional value once the transformer prediction and L1 are already available. Likewise, $B^{L1}SE$ improves significantly only on the pooled data. These findings suggest that etymological information is most useful in the hybrid setting when combined with common lexical features ($B^{L1}CE$, $B^{L1}CSE$), which may provide a stable basis from which the model can exploit more sparse etymological signals.

The NLD sliding-window analysis broadly mirrors the feature-only results (Figure 4). For German and Spanish, similarity features provide the clearest gains at low NLD values. Etymology features contribute most visibly at intermediate distances: approximately 0.2–0.7 for Spanish and 0.3–0.8 for German. Within these ranges, $B^{L1}CSE$ often improves on $B^{L1}CS$, indicating that etymological information remains useful when surface similarity alone is insufficient.

For Mandarin, the patterns are less differentiated. In the relatively sparse lower-NLD windows, etymology features often outperform similarity features, but these results must be interpreted cautiously because they contain few observations. At high NLD values, where most Mandarin–English pairs occur, the three feature groups yield more comparable gains, while common lexical features remain particularly competitive. Overall, the stacked analysis confirms that etymology features can improve a strong transformer-based LCP system, especially in conjunction with common lexical features and for German and Spanish pairs at intermediate NLDs.

10 Discussion

Overall, our findings indicate that etymology features provide a useful and complementary signal for LCP in a multilingual context. Generally, etymology features exhibit moderate correlations with lexical complexity, suggesting that etymological relationships between source–target word pairs are associated with learners' ability to produce English vocabulary. This additional signal is reflected consistently in the experiments based exclusively on manually engineered features. Across all source languages, models that include etymology features significantly outperform their respective counterparts without these features, and the best performing model includes etymology features.

Furthermore, the effect of etymology differs across L1s. For German and Spanish, surface-similarity features are particularly informative for word pairs with low NLDs, where orthographic or phonological resemblance offers a clear indication of relatedness. As NLD increases into the intermediate range, however, etymology features regularly improve on similarity features. This pattern is consistent with our expectation that etymological information is especially useful when regular sound changes, language-specific adaptation, or orthographic divergence have weakened the surface similarity between etymologically related words. In such cases, surface similarity becomes a less reliable proxy for cognacy or shared borrowing histories, whereas etymology features can enrich the model with information about these relationships.

As expected, for Mandarin, the overall pattern is less pronounced. Although etymology features remain beneficial, similarity features are generally less informative, while common lexical features make a comparatively stronger contribution. This likely reflects the greater lexical and typological distance between Mandarin and English, which results in fewer source–target pairs with similar orthographic or phonological forms, as well as fewer candidates for shared etymological relationships. Correspondingly, we observe no similarly pronounced differences between NLD windows for Mandarin. Nevertheless, the positive contribution of etymology features suggests that historical borrowing relationships can still provide useful information even when surface similarity is limited.

These findings extend to the stacked architecture. Adding etymology features to `RoBERTa` predictions yields significant improvements in most configurations, confirming that etymological information is not fully captured by the transformer baseline predictions.

11 Conclusion

We investigated whether etymology-based features can improve LCP in a multilingual context beyond conventional surface-similarity measures. We derived nine features from Wiktionary etymology information for English target words and their German, Spanish, and Mandarin source-language equivalents, and evaluated their contribution both in models based on manually engineered features and in a stacked architecture combining these features with contextual `RoBERTa` predictions.

Adding etymology features significantly improved prediction performance in 29 of 36 matched comparisons ($\sim$81%), and the best-performing models always included etymology features. Our results show that these features distill an informative signal distinct from common lexical and surface-similarity features.

The contribution of etymology is strongest for German and Spanish source–target word pairs in the intermediate NLD range. For these pairs, surface-similarity-based features are less reliable predictors than at the lower end of the range, while common features do not fully capture the information encoded by etymological relationships. For Mandarin, etymology features remain useful overall, but the NLD pattern is less clearly localized. Etymology features may therefore be particularly valuable where historical relatedness has been obscured, but not eliminated, by sound change, phonological adaptation, or orthographic divergence.

Beyond improving LCP models, etymology features may benefit downstream EduNLP applications that rely on accurate estimates of vocabulary difficulty, such as lexical simplification, readability assessment, and adaptive language-learning systems. Future work should evaluate the approach on additional language pairs and investigate how it can be integrated into more advanced neural LCP architectures. It would also be valuable to examine how learners' individual linguistic backgrounds and explicit knowledge of etymological relationships affect the observed patterns. More broadly, the findings may have pedagogical implications for vocabulary selection and curriculum design, particularly where learning materials can draw attention to relevant cross-linguistic relationships.

Limitations

In this section, we raise awareness of a number of limitations of the present study.

First, our conclusions are constrained to the three source languages represented in the dataset: German, Spanish, and Mandarin. Other L1s may yield divergent patterns depending on their typological and lexical proximity to English, as well as their specific histories of language contact. The results for Mandarin warrant particular caution. While Wiktionary provides extensive etymological coverage for German and Spanish, only approximately 10% of the Mandarin source words in our data had corresponding etymological entries. For the remaining instances, `GPT-5.6-Luna` had to infer feature values from its own internalized knowledge. Since a manual inspection of the Mandarin items revealed that it is often non-native vocabulary that is covered by Wiktionary, which is the subject of our investigation, we do not anticipate this issue to fundamentally invalidate the utility of etymology features for Mandarin. However, it may compromise the reliability of individual feature values.

Our conclusions are also restricted to English as the target language and do not establish whether comparable gains would arise in the reverse direction, for example, for English-speaking learners of German, Spanish, or Mandarin. Nevertheless, we expect etymological information to be informative in other learning directions and, more generally, particularly so for language pairs with substantial genealogical relatedness or histories of language contact. For example, etymology features may be particularly promising for multilingual LCP involving languages such as Chinese, Japanese, Korean, and Vietnamese. This is because the latter three borrowed extensively from Chinese but subsequent divergent linguistic changes often obscure these relationships when comparing surface forms alone. These expectations require empirical validation beyond the present shared-task configuration.

Second, the reliability of our feature data is inherently tied to our extraction methodologies. In this study, we relied exclusively on `GPT-5.6-Luna` to parse Wiktionary entries. A manual inspection of a sample of the generated feature values revealed occasional divergences between human judgment and the model's outputs. Specifically, we observed inconsistencies in the `lca_language` feature, where the model struggled to accurately differentiate between Proto-West-Germanic and Proto-Germanic. Furthermore, we identified conflations of `same_proto_word` and `same_proto_root`, as well as failures to detect joint borrowings (`borrowed_from_third`). Future research must prioritize a more rigorous evaluation of LLM-generated features and compare multiple LLMs to ascertain whether the signal distilled by etymology features can be further denoised.

Third, methodological constraints in our preprocessing pipeline may partially account for the comparatively weak contribution of similarity features for Mandarin. To compute orthographic similarity, Chinese characters were romanized, tone marks and whitespace were removed, and the resulting pinyin forms were compared with English orthography. This representation may inadequately capture the phonological, orthographic, or lexical relationships that are salient to Mandarin-speaking learners. Subsequent investigations should explore similarity measures derived directly from Chinese characters, phonological representations, or linguistically more sophisticated transliteration protocols.

Fourth, at a theoretical level, our etymology features should be conceptualized as proxies for predictive signals within the data, rather than as direct reflections of learners' mental lexicons or their subjective experiences of lexical complexity. Although these features enhance predictive performance, our experimental design cannot establish whether learners explicitly recognize the encoded etymological relationships, which aspects of these relationships drive the observed effects, or whether a more direct operationalization of the same signal is feasible. Future work should therefore aim to isolate the relevant signal more accurately and develop targeted metrics or alternative feature representations. Moreover, as the authors of the Knowledge-Based Vocabulary Lists (KVL) project, which is the source of the BEA shared-task dataset, point out themselves (Schmitt et al., 2022),

> the level of knowledge which the [data] describes is the ability to accurately spell a lemma if the meaning is known. Research shows that this level of knowledge is relatively advanced, and is higher than receptive knowledge, [...] [b]ut the KVL cannot claim to represent full productive knowledge.

Thus, our claim that etymological information aids prediction can also only apply to this specific setup of measuring productive vocabulary knowledge.

Furthermore, L1 is only one learner-related factor relevant to lexical complexity and the personalization of learning solutions. Our study does not establish whether the relationship between etymological relatedness and lexical complexity changes systematically as learners' proficiency develops. Future work should therefore incorporate proficiency, for example represented by the CEFR level, alongside L1 and examine their interactions with etymology features. For pedagogical applications such as vocabulary selection, curriculum design, lexical simplification, and adaptive language learning systems, it is necessary to determine not only whether an item is easier for learners with a particular L1, but also at which proficiency levels such differences are pedagogically meaningful.

Fifth, the scope of our modeling choices remains limited. We evaluated only a single transformer baseline for contextual predictions alongside one stacking architecture. Future experiments should systematically compare distinct LLMs, non-LLM-based extraction techniques, contextual encoders, and hybrid-model architectures. Such comparative analyses are necessary to determine whether the observed performance gains are contingent upon specific implementation choices and whether the underlying etymological signal can be extracted with greater reliability.

Finally, while the etymology features demonstrate informational value, their performance gains are only marginal when integrated with the contextual transformer baseline. Given the comparatively high computational cost of LLM-based parsing, applying this procedure indiscriminately to all word pairs may prove inefficient. Etymology features may instead offer the greatest utility for word pairs in intermediate NLD ranges within lexically proximate languages (e.g., German and Spanish), where surface similarity provides less predictive power than at lower NLDs. Developing methodologies that either selectively apply etymological analyses in such contexts or automatically recover comparable signals without relying on LLMs could mitigate computational expenses and environmental impact while preserving the predictive benefits identified in this study.

Ethics Statement

When personalizing educational technologies based on learner characteristics, such as L1 backgrounds, it is important to ensure broad and equitable representation. However, this study is inherently constrained by the availability of existing datasets, which in the case of the BEA shared-task dataset exclusively features learners from high-resource languages, specifically German, Spanish, and Mandarin. Future research should prioritize the collection of data from more diverse linguistic communities to investigate whether these findings generalize across a more diverse population of learners.

Acknowledgments

We would like to thank Yuning Ding for her support in gaging the Mandarin data and our two anonymous reviewers for their valuable comments on potential future research directions.

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: We used several assistive generative AI tools while composing this paper. Most notably, we employed GitHub Copilot (with various LLM backbones) as an agentic coding assistant, and we used ChatGPT, Claude, and Gemini (in multiple versions) for brainstorming, literature searches, writing and editing assistance, and code and figure generation. We also consulted these models when interpreting our experimental results.

References

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English vocabulary profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

Karin Beijering, Charlotte Gooskens, and Wilbert Heeringa. 2008. Predicting intelligibility and per-

ceived linguistic distance by means of the levenshtein algorithm. *Linguistics in the Netherlands*, 25.

Yves Bestgen. 2026. SATLab at BEA 2026 shared task 1: Predicting the difficulty of English words for three L1 learners using primarily psycholinguistic features. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 985–991, San Diego, California, USA. Association for Computational Linguistics.

Marc Brysbaert, Guillermo Martínez, and Pedro Reviriego. 2025. Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. *Behavior Research Methods*, 57:28.

Susanne E. Carroll. 1992. On cognates. *Interlanguage Studies Bulletin (Utrecht)*, 8(2):93–119.

David Crystal. 2008. *A Dictionary of Linguistics and Phonetics (6th ed.)*. Blackwell, Oxford.

Helen Engemann and Stefanie Radetzky. 2024. Cognate facilitation in child third language learners in a multilingual setting. *Languages*, 9(10).

Mariano Felice and Lucy Skidmore. 2026. Findings of the BEA 2026 shared task on vocabulary difficulty prediction for English learners. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 964–984, San Diego, California, USA. Association for Computational Linguistics.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics.

Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 2023. Japanese lexical complexity for non-native readers: A new dataset. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 477–487, Toronto, Canada. Association for Computational Linguistics.

Elly Koutamanis, Gerrit Jan Kootstra, Ton Dijkstra, and Sharon Unsworth. 2024. Shared representations in cognate comprehension and production: An online picture naming and lexical decision study with bilingual children. *Applied Psycholinguistics*, 45(3):489–515.

Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

D. Löfflad and E. Kanli. 2026. Learner-centered automatic readability assessment: Leveraging hybrid models to consider l1. In *Proceedings of the 22. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*. In press.

Jonas Mayer Martins, Zhuojing Huang, Aaricia Herygers, and Lisa Beinborn. 2026. BoostedCats at BEA 2026 shared task 1: What makes a word hard to learn? modeling L1 influence on English vocabulary difficulty. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1047–1064, San Diego, California, USA. Association for Computational Linguistics.

Gerard de Melo. 2014. Etymological Wordnet: Tracing the history of words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1148–1154, Reykjavik, Iceland. European Language Resources Association (ELRA).

Noah-Manuel Michael, Anna Huelsing, and Andrea Horbach. 2026. GerVLPro: A CEFR-graded vocabulary list of L2 learners' productive vocabulary in German. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 11942–11959, Palma de Mallorca, Spain. ELRA Language Resource Association.

Alejandro Mosquera. 2021. Alejandro mosquera at SemEval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559, Online. Association for Computational Linguistics.

Adam Nohejl, Xuanxin Wu, Yusuke Ide, Maria Angelica Riera Machin, Yi-Ning Chang, and Hitomi Yanaka. 2026. Sakura at BEA 2026 shared task 1: What makes vocabulary difficult? In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1161–1178, San Diego, California, USA. Association for Computational Linguistics.

Kai North and Marcos Zampieri. 2023. Features of lexical complexity: Insights from L1 and L2 speakers. *Frontiers in Artificial Intelligence*, Volume 6.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Jenny A. Ortiz-Zambrano, César H. Espín-Riofrío, and Arturo Montejo-Ráez. 2025. Deep encodings vs. linguistic features in lexical complexity prediction. *Neural Computing and Applications*, 37(3):1171–1187.

Jenny A. Ortiz-Zambrano and Arturo Montejo-Ráez. 2020. Overview of ALexS 2020: First workshop on lexical analysis at SEPLN. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, volume 2664 of *CEUR Workshop Proceedings*, Málaga, Spain. CEUR-WS.org.

Agnieszka Otwinowska and Jakub M. Szewczyk. 2019. The more similar the better? factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism*, 22(8):974–991.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Adaptive complex word identification through false friend detection. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 192–200, New York, NY, USA. Association for Computing Machinery.

Vassili Philippov, Dmitrii Andreev, Pavel Katunin, and Anton Nikolaev. 2026. Glite at BEA 2026 shared task 1: Holistic difficulty models dominate, feature engineering closes the gap in L1-aware vocabulary difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1091–1105, San Diego, California, USA. Association for Computational Linguistics.

Santiago Robaina, Aiala Rosá, and Luis Chiruzzo. 2026. RETUYT-INCO at BEA 2026 shared task 1: Feature-enriched mDeBERTa for word difficulty prediction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1113–1118, San Diego, California, USA. Association for Computational Linguistics.

Job Schepens, Ton Dijkstra, and Franc Grootjen. 2012. Distributions of cognates in europe as based on levenshtein distance. *Bilingualism: Language and Cognition*, 15(1):157–166.

Job Schepens, Roeland van Hout, and T. Florian Jaeger. 2020. Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, 194:104056.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2022. *Knowledge-based Vocabulary Lists: User Manual: A Supplementary Resource*. British Council.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2024. *Knowledge-based Vocabulary Lists*. University of Toronto Press, Toronto.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Lucy Skidmore, Mariano Felice, and Karen Dunn. 2025. Transformer architectures for vocabulary test item difficulty prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 160–174, Vienna, Austria. Association for Computational Linguistics.

Anaïs Tack. 2021. Mark my words! On the automated prediction of lexical difficulty for foreign language readers. PhD Thesis.

Evy Woumans, Robin Clauws, and Wouter Duyck. 2021. Hands down: Cognate effects persist during written word production. *Frontiers in Psychology*, 12:647362.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association.

A Abbreviations

Table 4 provides an overview of the abbreviations used in this paper.

B Feature Value Sets

Tables 5, 6, and 7 illustrate the features we explored in this study and their value ranges.

B	baseline
BEA	Workshop on Innovative Use of NLP for Building Educational Applications
C	common lexical (features/models)
CEFR	Common European Framework of Reference for Language
E	etymology (features/models)
EduNLP	educational NLP
GPT	Generative Pre-trained Transformer
IPA	International Phonetic Alphabet
KVL	Knowledge-Based Vocabulary Lists
LCP	Lexical Complexity Prediction
LightGBM	Light Gradient-Boosting Machine
LLM	Large Language Model
L1, ^{L1}	first language; source language(-based feature/models)
L2	second language; target language
NLD	Normalized Levenshtein Distance
NLP	Natural Language Processing
POS	part of speech
RMSE	root mean-squared error
RoBERTa	Robust Bidirectional Encoder Representations from Transformers
RQ	research question
S	surface-similarity (features/models)
TF-IDF	Term Frequency-Inverse Document Frequency
XLM-RoBERTa	Cross-Lingual Language Model with RoBERTa Architecture

Table 4: Abbreviations; alphabetically ordered.

Feature	Values
`zipf_frequency`	$[1.17, 7.17]$
`target_length`	$\{1, 2, \dots\}$
`num_syllables`	$\{1, 2, \dots\}$
`pos_tag`	noun verb adjective adverb determiner preposition number not-no misc

Table 5: Sets of possible values for common lexical features.

Feature	Values
`char_edit_dist`	$\{0, 1, 2, \dots\}$
`char_edit_dist_norm`	$[0, 1]$
`char_edit_dist_geo_norm`	$[0, \infty)$
`phon_edit_dist`	$\{0, 1, 2, \dots\}$
`phon_edit_dist_norm`	$[0, 1]$
`phon_edit_dist_geo_norm`	$[0, \infty)$
`jaro_dist`	$[0, 1]$
`jaro_winkler_dist`	$[0, 1]$
`soundex_dist`	$\{0, 1, 2, 3, 4\}$
`metaphone_dist_norm`	$[0, 1]$
`char_ngram_tfidf_sim`	$[0, 1]$

Table 6: Sets of possible values for similarity features.

C Etymology Feature Extraction Prompt

You are an expert historical linguist. You will be presented with a source-target word pair (possible source languages: German, Spanish, Mandarin Chinese; target language: English). You must return a JSON object (dictionary) corresponding to the features described below. If available, you will receive the Wiktionary etymology entries for the source and target words for additional context. If there is no Wiktionary information provided, please answer to the best of your ability. Since the POS tag of the source word is not given, we also present you with an example of the source word in context. If there are multiple etymology entries for a source word, choose the one that matches the POS and meaning you infer from the source context as your reference.

Word Relatedness Features

Feature 1: Lowest Shared Genetic Proto-Language

Description: Identifies the youngest reconstructed ancestral language that both words naturally evolved from. This strictly measures genetic lineage, ignoring contact, borrowing, or donor languages.

Type: Categorical

Feature	Values
lca_language	Proto-West-Germanic Proto-Germanic Proto-Indo-European None
borrowed_from_source	$\{0, 1\}$
borrowed_from_target	$\{0, 1\}$
borrowed_from_third	$\{0, 1\}$
english_word_origin	Germanic Romance Sinitic Other
origin_matches_source_family	$\{0, 1\}$
same_proto_word	$\{0, 1\}$
same_proto_root	$\{0, 1\}$
transmission_path	Inherited Direct Indirect

Table 7: Sets of possible values for etymology features.

Values:

- Proto-West-Germanic: Shared genetic ancestor at the West Germanic level.
- Proto-Germanic: Shared genetic ancestor at the broader Germanic level.
- Proto-Indo-European: Shared genetic ancestor at the Proto-Indo-European level.
- None: Applicable if the words do not share a reconstructed genetic proto-language, OR if their only shared history is via borrowing (e.g., both borrowed from French or Latin).

Further Comments: None.

Feature 2: Borrowed from Source Language

Description: Indicates whether the English target word was directly borrowed from the specified source language.
Type: Binary
Values:

- 0: English target word was not borrowed from the source language.
- 1: English target word was directly borrowed from the source language (e.g., English *kindergarten* from German).

Further Comments: None.

Feature 3: Borrowed from Target Language

Description: Indicates whether the source word was directly borrowed from English.
Type: Binary
Values:

- 0: Source word was not borrowed from English.
- 1: Source word was directly borrowed from English (e.g., German *Computer*, Spanish *fútbol*).

Further Comments: None.

Feature 4: Borrowed from Same Third Language

Description: Indicates whether both the source word and the English target word were borrowed from the same non-source, non-English donor language.
Type: Binary
Values:

- 0: No evidence that both words were borrowed from the same third language.
- 1: Both words were borrowed from the same third language (e.g., German *flexibel* and English *flexible* both from Latin/French transmission).

Further Comments: This feature is independent from direct borrowing between source and target (Features 2 and 3).

Feature 5: Immediate Taxonomic Branch of English Word

Description: Classifies the broad language branch or family of the immediate predecessor from which modern English acquired the word. This does not ask for the specific language, but rather the taxonomic category of the immediate donor or ancestor.
Type: Categorical
Values:

- Germanic: Entered from Old English, Old Norse, Dutch, German, etc.
- Romance: Entered from French, Latin, Spanish, Italian, etc.
- Hellenic: Entered from Greek.

- `Sinitic`: Entered from Mandarin, Cantonese, etc.
- `Other`: Any other taxonomic branch not covered above.

Further Comments: None.

Feature 6: Origin Matches Source Language Family?

Description: Determines if the immediate linguistic origin of the English target word aligns with the broader language family of the source word.
Type: Binary
Values:

- `0`: English target word origin does not match source language family.
- `1`: English target word origin matches source language family.

Further Comments: Assign `1` if the English target word's immediate taxonomic branch (assigned in Feature 5) matches the broader language family of the source language. Specifically: if the source language is German, assign `1` if Feature 5 is 'Germanic'; if the source language is Spanish, assign `1` if Feature 5 is 'Romance'; if the source language is Mandarin Chinese, assign `1` if Feature 5 is 'Sinitic'. Otherwise, assign `0`.

Feature 7: Descendant of the Same Proto-Word?

Description: Evaluates if both words descend from the exact same reconstructed ancestral lexical item at the youngest shared proto-language stage.
Type: Binary
Values:

- `0`: Source and target words do not descend from the same reconstructed ancestral lexical item. They may share a broader root, but are not descendants of the same proto-word.
- `1`: Source and target words are descendants of the same reconstructed ancestral lexical item at the youngest relevant shared proto-language stage.

Further Comments: "Proto-word" means a reconstructed lexical form, not merely a root or semantic concept. Two words count as sharing the same proto-word even if they diverged into different language branches. Direct cognacy does not require the modern forms to look similar. If a shared reconstructed lexical item exists, assign `1`. Examples: Spanish *padre* and English *father* both descend from Proto-Indo-European '$*ph_2t\acute{e}r$' (Assign 1). Spanish *pie* and English *foot* both descend from the exact same Proto-Indo-European proto-word '$*p\acute{o}ds$' (Assign 1).

Feature 8: Descendant of the Same Proto-Root?

Description: Evaluates if the words share a broader ancestral root or word family, specifically when they do not descend from the exact same reconstructed proto-word.
Type: Binary
Values:

- `0`: Source and target words do not share any reconstructable ancestral root or word family.
- `1`: Source and target words share a broader ancestral root or word family but are NOT descendants of the same reconstructed proto-word.

Further Comments: Assign `1` if the words share a historical root or word family but do NOT descend from the exact same reconstructed proto-word. F7 and F8 are mutually exclusive based on the youngest shared ancestor: if they share a proto-word, assign F7=1 and F8=0. If they only share a broader root, assign F7=0 and F8=1. Example: Spanish *juventud* and English *youth* both stem from the shared Proto-Indo-European root for 'young', but they do not descend from the exact same proto-word because the Italic and Germanic branches added completely different suffixes to that base root to form their respective nouns (F7=0, F8=1).

Feature 9: English Transmission Mechanism

Description: Describes the mechanical path of how the word reached Modern English, ignoring which languages were involved and focusing entirely on the complexity of the transmission history.
Type: Categorical
Values:

- `Inherited`: Passed down continuously into Modern English from its deepest ancestral roots with no borrowing events (0 hops).

- `Direct`: Borrowed directly into English from the ultimate source language with no intermediary languages (1 hop).
- `Indirect`: Passed through at least one intermediary language before reaching English (2+ hops).

Further Comments: None.

One-Shot Example

Input:
Spanish source word: padre
English target word: father

Source context:
Mi padre siempre me da buenos consejos, y mi madre también.

Wiktionary Info:
Source word in Spanish Wiktionary:
Del latín pater y este del protoindoeuropeo '$*ph_{\circ 2}t\acute{e}r$'. Compárese el protogermánico '$*fa\eth er$', el griego micénico pa-te, el griego clásico '$\pi\alpha\tau\acute{\eta}\rho$' (patér), el persa *pedar* o el sánscrito *pitr*.

Source word in English Wiktionary:
Inherited from Old Spanish padre ("father"), from Latin patrem, pater, from Proto-Italic '$*patēr$', from Proto-Indo-European '$*ph_2t\acute{e}r$'.

Target word in Spanish Wiktionary:
Del inglés medio fader, del inglés antiguo fæder, del protogermánico occidental '$*fader$', del protogermánico '$*fadēr$', del protoindoeuropeo '$*ph_{\circ 2}t\acute{e}r$'.

Target word in English Wiktionary:
Inherited from Middle English fader, from Old English fæder, from Proto-West Germanic '$*fader$', from Proto-Germanic '$*fadēr$', from Proto-Indo-European '$*ph_2t\acute{e}r$', possibly from '$*peh_2-$' + '$* - t\acute{e}r$'. Doublet of ayr, faeder, athair, padre, pater and père.

Output:

```
{
  "F1: Lowest Shared Genetic
 Proto-Language": "Proto-Indo-European",
  "F2: Borrowed from Source Language": 0,
  "F3: Borrowed from Target Language": 0,
  "F4: Borrowed from Same Third
  Language": 0,
  "F5: Immediate Taxonomic Branch of
  English Word": "Germanic",
  "F6: Origin Matches Source Language
  Family": 0,
  "F7: Descendant of the Same
  Proto-Word": 1,
  "F8: Descendant of the Same
  Proto-Root": 0,
  "F9: English Transmission Mechanism":
  "Inherited"
}
```

Your Task

Now, process the following word pair and return ONLY the JSON object. Do not include any markdown formatting, explanations, or text outside the JSON object.

Input:
{source_lang} source word: {l1_source}
English target word: {en_target}

Source context:
{source_context}

Wiktionary Info:
Source word in {source_lang} Wiktionary:
{source_etymology_l1}

Source word in English Wiktionary:
{source_etymology_en}

Target word in {source_lang} Wiktionary:
{target_etymology_l1}

Target word in English Wiktionary:
{target_etymology_en}

L1	Target	Source
German	93.9%	88.7%
Spanish	95.0%	83.2%
Mandarin	93.9%	10.1%

Table 8: Percentage of target and source words covered for etymological information in either the English or the respective L1 Wiktionary edition.

D Available Wiktionary Entries

Table 8 illustrates how many source and target words per L1 had Wiktionary etymology information available in either the English or the respective L1 Wiktionary edition.

Parameter	Value
Prediction head	regression
Maximum sequence length	256 tokens
Number of epochs	5
Training batch size	32
Learning rate	3×10^{-5}
Weight decay	0.1
Warmup ratio	0.1
Optimizer	AdamW
Random seed	667

Table 9: `XLM-RoBERTa-base` fine-tuning settings, using the HuggingFace `Trainer` (`transformers` 5.15.1, PyTorch 2.5.1) on a single GPU without gradient accumulation. Unlisted `TrainingArguments` take their `transformers` 5.15.1 defaults.

Parameter	Value
Objective	regression
Number of estimators	700
Learning rate	0.03
Number of leaves	63
Column sample by tree	0.9
Random seed	667

Table 10: `LightGBM` settings, using `LGBMRegressor` from `lightgbm` 4.7.0. Unlisted parameters take their `lightgbm` 4.7.0 defaults.

E Model Architecture

In this section, we provide supplementary materials concerning our model architecture.

E.1 Model Parameters

For reproducibility, we report the hyperparameters of the models we train (Tables 9 and 10).

E.2 Cross-Validation Logic

Figure 5 illustrates the inner and outer fold logic as well as the final `LightGBM` training.

F Number of Items per Levenshtein Window

Table 11 provides an overview of the number of source–target word pairs contained in each sliding NLD window per L1.

G Examples of Exclusively Correct Predictions by Etymology Model

Here, we present examples of exclusively correct predictions by the $^{\text{L1}}$E model at $\tau \in \{0.05\sigma_l, 0.5\sigma_l\}$ that exhibit an etymological relationship.

G.1 5% STD Threshold

We provide a selection of etymologically related source-target word pairs per L1, identified correctly exclusively by the $^{\text{L1}}$E model at $\tau = 0.05\sigma_l$.

German: Hüfte ‘hip’, schießen ‘shoot’, durch ‘through’, verdoppeln ‘double’, Sonnenschein ‘sunshine’, Mutterschaft ‘motherhood’, gegeben ‘given’, typisch ‘typically’, verboten ‘forbidden’, Französisch ‘french’, Aktivierung ‘activation’, komponieren ‘compose’, Kupfer ‘copper’, gesalzen ‘salted’, Fett ‘fat’, aktiv ‘actively’, identifizieren ‘identify’, Dank ‘thanks’, lachend ‘laughing’, rasten ‘rest’, Klassifizierung ‘classification’, Sonnenlicht ‘sunlight’, Soldat ‘soldier’, Delfin ‘dolphin’, Schande ‘shame’, Eisberg ‘iceberg’, frustrierend ‘frustrating’, geologisch ‘geologically’, romantisch ‘romantically’, Diskriminierung ‘discrimination’, Asche ‘ash’, statistisch ‘statistically’, Mystiker ‘mystic’, Handvoll ‘handful’, ungewaschen ‘unwashed’, Grammatik ‘grammar’, identisch ‘identical’, Wäsche ‘wash’, stressig ‘stressful’, Christ ‘christian’, unkontrolliert ‘uncontrolled’, statisch ‘static’, ethnisch ‘ethnic’, Mikrobiologie ‘microbiology’, psychologisch ‘psychological’, Adjektiv ‘adjective’, Therapeut ‘therapist’, Ethik ‘ethic’, harmonisch ‘harmonic’, national ‘nationally’, Psychologie ‘psychology’, homosexuell ‘homosexual’, Philosophie ‘philosophy’, Heterosexualität ‘heterosexuality’, Miniatur- ‘miniature’, evolutionär ‘evolutionary’, multifunktional ‘multifunctional’, Designer ‘designer’, Countdown ‘countdown’, Cheerleader ‘cheerleader’, Freak ‘freak’, Layout ‘layout’, Asteroid ‘asteroid’.

Spanish: viuda ‘widow’, salchicha ‘sausage’, ciudadanía ‘citizenship’, abordo ‘aboard’, segundo ‘secondly’, captura ‘catch’, incierto ‘uncertain’, comprometerse ‘promise’, tratar ‘treat’, tipo ‘type’, brillantemente ‘brilliantly’, rito ‘ritual’, redondeada ‘rounded’, química ‘chemical’, receta ‘recipe’, cartón ‘card’, alfabeticamente ‘alphabetically’, auditoría ‘audit’, demográficamente ‘demographically’, batalla ‘battle’, insatisfecho ‘unsatisfied’, independientemente ‘independently’, pó-

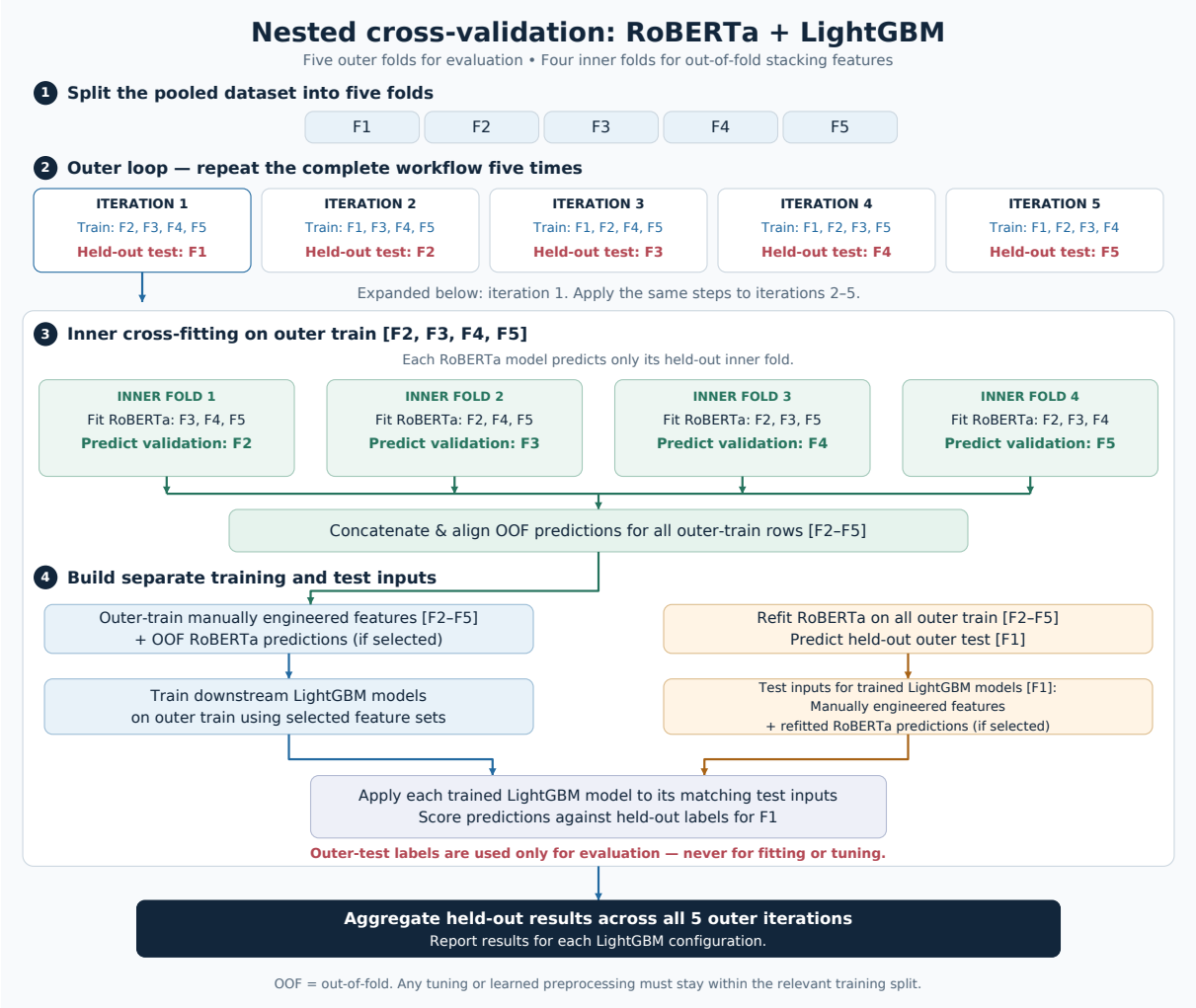


Figure 5: Illustration of our stacked RoBERTa–LightGBM model architecture and cross-validation loop.

quer ‘poker’, bárbaro ‘barbarian’, incontrolado ‘uncontrolled’, especialidad ‘speciality’, gimnasio ‘gymnasium’, contenido ‘content’, biografico ‘biographic’, instalador ‘installer’, instrumentalmente ‘instrumentally’, colonia ‘colony’, comparación ‘comparison’, insensible ‘insensitive’, estereotipo ‘stereotype’, inversor ‘investor’, intermedio ‘intermediate’, duplicar ‘duplicate’, cinematografía ‘cinematography’, rollo ‘roll’, adolecente ‘adolescent’, hipótesis ‘hypothesis’, microscopio ‘microscope’, hormona ‘hormone’, tradicionalista ‘traditionalist’, genética ‘genetics’, molécula ‘molecule’, celular ‘cellular’, vainilla ‘vanilla’, manipulador ‘manipulator’, propaganda ‘propaganda’, editorial ‘editorial’.

Mandarin: 伏特加 *fútèjiā* ‘vodka’, 卡拉OK *kǎlā OK* ‘karaoke’.

G.2 50% STD Threshold

We provide a selection of etymologically related source-target word pairs per L1, identified correctly exclusively by the ^{L1}E model at $\tau = 0.5\sigma_l$.

German: Kirsche ‘cherry’, schießen ‘shoot’, wöchentlich ‘weekly’, überhitzen ‘overheat’, Bohnen ‘bean’, unbekannt ‘unknown’, blutig ‘bloody’, reagieren ‘react’, halten ‘hold’, sollen ‘should’, aktivieren ‘activate’, Bruderschaft ‘brotherhood’, unzerbrechlich ‘unbreakable’, hausgemacht ‘homemade’, Weisheit ‘wisdom’, abgerundet ‘rounded’, Regentropfen ‘raindrop’, Älterer ‘elder’, verboten ‘forbidden’, kompliziert ‘complicated’, französisch ‘french’, Feuerwehrmann ‘fireman’, konzentriert ‘concentrated’, Streifen ‘strip’, gesalzen ‘salted’, Herzensbrecher ‘heartbreaker’, gestoppt ‘stopped’, hilflos ‘helpless’, beißen ‘bite’, lachend ‘laughing’, laut ‘loud’, Rezept ‘recipe’, Klassifizierung ‘classification’, unkompliziert ‘uncomplicated’, konzentrieren ‘concentrate’, hasserfüllt ‘hateful’, vibrieren ‘vibrate’, kontrollierbar ‘controllable’, unfreundlich ‘unfriendly’, Daten ‘data’, frustriert

	Window Center								
L1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
German	684	602	749	832	980	1332	2166	3418	3551
Spanish	492	574	613	642	828	1321	2131	3506	3955
Mandarin	0	2	17	60	191	612	2130	4535	5663

Table 11: Number of candidate items per Normalized Levenshtein Distance window in EXPERIMENTS 3A, 3B, 4B.

'frustrated', sauer 'sour', abgefallen 'fallen', Diskriminierung 'discrimination', dominieren 'dominate', ungewaschen 'unwashed', integrieren 'integrate', biografisch 'biographic', explodieren 'explode', Wäsche 'wash', Soziologie 'sociology', Kreativität 'creativity', Republikaner 'republican', exklusiv 'exclusive', Immunität 'immunity', Stock 'stick', Wagen 'wagon', Hallo 'hello', Subkultur 'subculture', Helfer 'helper', Flexibilität 'flexibility', mainstreamig 'mainstream', Elektron 'electron', Republik 'republic', Zigarette 'cigarette', Manuskript 'manuscript', Präsentation 'presentation', Leopard 'leopard', Exhibitionist 'exhibitionist', Astronaut 'astronaut', minimal 'minimal', Hunger 'hunger', Park 'park', Algebra 'algebra', Collage 'collage', Bronze 'bronze', bitter 'bitter'.

Spanish: jabón 'soap', sinfonía 'symphony', novedad 'innovation', igualmente 'equally', cubrir 'cover', solitario 'solo', igualdad 'equality', dirigido 'directed', esclavitud 'slavery', gráficamente 'graphically', sugerir 'suggest', elegido 'elected', tostadora 'toaster', sobrevivir 'survive', gravedad 'gravity', embajada 'embassy', boxeo 'boxing', cartón 'card', esquadrón 'squad', movilidad 'mobility', párrafo 'paragraph', ejercitar 'exercise', cocina 'cooking', champú 'shampoo', simplicficar 'simplify', multiplicar 'multiply', inseguridad 'insecurity', alfabético 'alphabetical', graduado 'graduated', decano 'dean', especialidad 'speciality', conectividad 'connectivity', aventurar 'venture', movimiento 'movement', publicidad 'publicity', gigante 'giant', calcio 'calcium', diccionario 'dictionary', sal 'salt', complejo 'complex', inspirado 'inspired', cristiano 'christian', ignoratne 'ignorant', radioactividad 'radioactivity', gimnástico 'gymnastic', óptica 'optics', incondicional 'unconditional', caramelo 'caramel', guitarista 'guitarist', legislación 'legislation', quántum 'quantum', tampón 'tampon'.

Mandarin: 伏特加 *fútèjiā* 'vodka', 汉堡 *hànbǎo* 'hamburger'. 伊拉克人 *yīlākè rén* 'Iraqi'.

Data Augmentation for L2 English Speaking Assessment using TTS

Stefano Bannò, Penny Karanasou, Mengjie Qian, Kate M. Knill, Mark J. F. Gales
ALTA Institute
Machine Intelligence Laboratory, Department of Engineering
University of Cambridge, UK
{sb2549,pk407,mq227,kmk1001,mjfg100}@cam.ac.uk

Abstract

Automated assessment of second language (L2) speaking proficiency requires substantial annotated speech data, which are scarce compared to written learner corpora. We investigate whether written L2 responses can be transformed into useful synthetic speech for proficiency assessment using text-to-speech (TTS) and voice cloning. Using COREFL, a corpus of paired spoken and written responses from L2 learners of English, we systematically study two factors: how written responses should be transformed into spoken-style language ("speechification") and how synthetic voices and texts should be paired based on shared learner attributes (proficiency level, first language, both, or neither). We generate speechified responses with a large language model and synthesise them using TTS and voice cloning, then evaluate their utility for audio-based (HuBERT) and text-based (ModernBERT) proficiency grading. Results show that pairing voices and texts based on proficiency provides a principled strategy for synthetic data generation, while speechification substantially improves the match between written and spoken L2 and improves downstream grading performance. Augmenting real training data with synthetic speechified responses improves both HuBERT- and ModernBERT-based graders, demonstrating the potential of synthetic spoken data for L2 speaking proficiency assessment.

1 Introduction

Automated assessment of second language (L2) speaking proficiency is increasingly used in computer-assisted language learning systems, providing scalable and consistent scoring of learner performance (Zechner and Evanini, 2019). Despite these advances, progress in automated L2 speaking assessment research remains constrained by the limited availability of large-scale, well-annotated spontaneous speech corpora. While large L2 written corpora are widely available (Geertzen et al., 2013; Blanchard et al., 2013; Nicholls et al., 2024; Crossley et al., 2023), spoken datasets remain comparatively scarce due to the cost of recording, transcription, and expert annotation.

A promising way to address this imbalance is the use of zero-shot text-to-speech (TTS) and voice-cloning technologies. These frameworks enable the conversion of abundant, diverse, and easily accessible L2 written essays into synthetic speech. Recent studies have started to investigate the use of TTS synthesis to augment data for L2 spontaneous speech assessment. In Wang et al. (2025), proficiency-conditioned responses are generated with a large language model (LLM) and converted into speech using speaker-conditioned TTS to augment training data for automated speaking assessment. Whilst improvements are reported in downstream scoring performance through fine-tuning a multimodal LLM-based grader, the quality of the generated speech itself is not analysed (e.g., speaker similarity, transcription quality, or linguistic realism), and the gains over a BERT-based (Devlin et al., 2019) baseline remain relatively modest. Similarly, in Voskoboinik et al. (2025), the use of LLM-generated synthetic transcripts is explored for automated speaking assessment in low-resource languages. Using GPT-4 (OpenAI, 2023) to generate proficiency-conditioned learner responses, it shows that transcript-level data augmentation can improve scoring performance. However, this approach operates exclusively on transcripts without including TTS. In our previous work (Karanasou et al., 2025), a more detailed analysis of acoustic

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Demo examples: sbanno.github.io/l2-synthetic-speech/

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 197–213, November 19–20, 2026, University of Tartu Library

and content-related aspects of the synthetic learner speech is provided, but the primary focus is spoken grammatical error correction rather than assessment. Additionally, the approach described in this previous work does not rely on essays, but on manual transcriptions, which remain costly to obtain.

In the present study, by contrast, we aim to leverage essays to augment L2 learner spoken data for proficiency assessment. A fundamental challenge remains: the disparity between written and spoken language. Spontaneous L2 speech is defined by online speech-planning phenomena, resulting in disfluencies such as hesitations, self-repairs, and repetitions, and discourse markers that are tightly linked with learners' oral proficiency (Huang et al., 2023; Yan et al., 2025). Conversely, written essays reflect a deliberate planning process characterised by higher syntactic complexity and greater lexical richness (Biber et al., 2002).

This raises the question of what is actually required to generate synthetic L2 speech that is suitable for assessment purposes, and more broadly, how written L2 data can be reliably transformed into spoken-like data. Addressing this question requires analysing the relationship between written and spoken L2 productions. The few existing L2 speech corpora provide valuable resources in isolation, but they are not designed for this purpose. ICNALE (Ishikawa, 2011), CLES (Coulange et al., 2024), Speechocean762 (Zhang et al., 2021), and L2-ARCTIC (Zhao et al., 2018) are limited either in first language (L1) diversity, modality (read rather than spontaneous speech), or proficiency level coverage. The Speak & Improve Corpus (Knill et al., 2025; Qian et al., 2025) represents a major milestone, providing large-scale spontaneous L2 speech with detailed proficiency annotations, but its speakers, tasks, and exam parts are not directly comparable to its written counterpart, the Write & Improve Corpus (Nicholls et al., 2024).

In this work, we use the Corpus of English as a Foreign Language (COREFL) (Lozano et al., 2021) (see Section 2). What differentiates this corpus from the above-mentioned ones is that it contains paired written and spoken answers to the same questions from the same L2 learners. This double modality allows us to study how linguistic and speaker-level factors interact in synthetic speech generation, enabling controlled investigation of different speaker–text pairing strategies as discussed below.

In our proposed data augmentation framework, to address the structural differences between modalities, we first introduce a *text speechification* step using a large language model (LLM). This module rewrites written responses into spoken-style transcripts. Next, the resulting texts are used for audio generation with a TTS/voice-cloning model. To select the voice for the generated audio, we investigate different matching strategies for pairing learner voices and written texts based on shared learner attributes (proficiency level, L1, both, or neither). Ensuring consistency between the textual and acoustic dimensions is important because, in real-world scenarios, the downstream grading system is not known a priori and may be text-based, audio-based, or multimodal. Finally, we validate our augmented data by evaluating downstream performance on two automated grading models: an audio-based system using a self-supervised learning (SSL) speech foundation model, i.e., HuBERT (Hsu et al., 2021), and a text-based system using a text foundation model, i.e., ModernBERT (Warner et al., 2025). Both SSL-based (Xu et al., 2021; Peng et al., 2021; Kim et al., 2022; Bannò et al., 2023; McKnight et al., 2023) and BERT-like (Craighead et al., 2020; Raina et al., 2020; Wang et al., 2021) architectures have been extensively used in L2 speaking assessment. The main contributions of this paper are as follows:

- we introduce an LLM-based preprocessing step that converts written responses into spoken-style transcripts, reducing the discrepancy between written and spoken language before TTS generation;
- we study how to construct high-quality synthetic L2 speech for speaking assessment by pairing speakers and texts according to proficiency level and L1, finding that proficiency-level matching provides a principled strategy for synthetic data augmentation that is consistent across downstream assessment models;
- we show that our data augmentation approach consistently improves performance across both text-based and audio-based grading systems.

2 Data

For our experiments, we use the Corpus of English as a Foreign Language (COREFL) (Lozano et al., 2021), a publicly available dataset[1] collected anonymously that contains both written and spoken answers to the same questions by L2 learners of English from 9 different L1 backgrounds, with German and Spanish representing the highest percentage of learners, followed by Greek, Italian, Chinese, Turkish, Czech, Estonian, and French learners. However, only the first 5 L1s are also represented in the spoken section.

A unique feature of this corpus is that some learners were asked to produce both a spoken and written response to the same question with a few weeks of difference between the two tasks. The corpus includes four question prompts. Learners are asked to: describe the picture story *Frog, where are you?* (Mayer, 2003); talk about a famous person; describe a film they have seen recently; describe an excerpt from Charlie Chaplin's *The Kid*.

Although the spoken section of the corpus is relatively small, it contains comparatively long responses: they range from approximately 40 seconds to 8 minutes with an average duration of 3 minutes. All the spoken responses are accompanied by manual transcriptions. The written responses range from roughly 10 to 1,000 words, with an average of 160 words. Table 1 shows the speaker-independent training and test set splits. All learners included in the test set also produced a written response, whereas not all learners included in the training set also have a written counterpart. Additionally, there are 2,820 written responses that are not accompanied by their spoken counterparts. We refer to this set as *Extra*.

The corpus is annotated with proficiency levels based on the Common European Framework of Reference (CEFR) (Council of Europe, 2001). These were given to the learners after they took a placement test, which assigned them a score ranging from 0 to 100 (i.e., A1 to C2). We normalise this score to [0, 1] and use it as the target variable for model training in our language assessment experiments. There is also a control L1 English subcorpus. For our experiments, we use this subcorpus as additional C2 data. Figure 1 shows the placement test score distribution.

[1] corefl.learnercorpora.com/

	Spoken		Written	
Set	**Responses**	**Hours**	**Responses**	**Words**
Train	410	16.7	383	111K
Test	100	4.2	100	27K
Extra	–	–	2,820	765K

Table 1: Composition of COREFL (spoken and written).

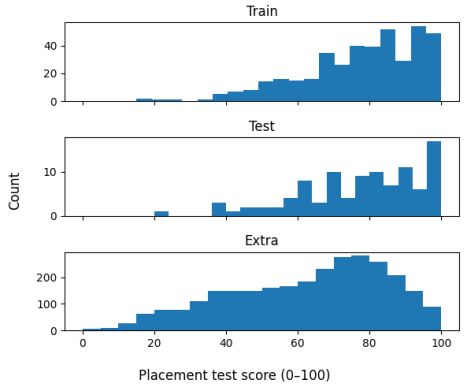


Figure 1: Placement test scores (0-100) distribution across sets.

3 Data Augmentation Framework

We propose a data augmentation framework that leverages L2 learner texts to generate synthetic spoken responses for the automated assessment of holistic speaking proficiency. As outlined in Section 1, it consists of two phases: LLM-based speechification and attribute-conditioned speaker-text matching and synthesis.

3.1 Text "speechification"

To improve the quality of synthetic responses, we introduce a *text speechification* step. Given a learner's written response and CEFR level, we use an LLM to transform the essay-like text into a spoken-style learner transcript. The prompt, reported in Appendix A, instructs the model to preserve the original meaning and all learner errors while rewriting the response as spontaneous speech. Specifically, the generated transcript incorporates characteristics of natural spoken production, including incremental discourse structure, hesitations, repetitions, self-repairs, false starts, and occasional vague references. The frequency and complexity of these phenomena are conditioned on the learner's CEFR level, with lower-proficiency learners exhibiting more disfluencies and simpler discourse structures. The resulting transcript is intended to approximate the characteristics of authentic learner speech while

remaining faithful to the content of the original written response.

We also include a control experiment in which we remove CEFR-conditioning from the prompt. This can be found in Appendix A.

3.2 Speaker-text matching strategies for synthesis

We refer to real spoken data as $Real_{\texttt{Spoken}}$ and to real written data as $Real_{\texttt{Written}}$. In the following, we describe the matching conditions used to generate synthetic data. First, we synthesise a learner's spoken response by using its own manual transcription and the same learner's voice. We refer to this condition as $Synth_{\texttt{Trs}}$. This serves as a best-case reference for evaluating the TTS system, as both the speaker identity and linguistic content from the same spoken response are preserved. We then synthesise a learner's written production using the same learner's voice, referred to as $Matched_{\texttt{Spkr}}$. In this case, although the information is produced by the same learner, the linguistic content differs between the written and spoken productions. Subsequently, we synthesise spoken data from written data using different pairing strategies to investigate how conditioning on learner attributes affects the similarity between synthetic and real speech. For each learner voice, a written response is chosen and assigned to it. A voice may match the selected written response with respect to CEFR level ($Matched_{\texttt{CEFR}}$), L1 ($Matched_{\texttt{L1}}$), both attributes ($Matched_{\texttt{CEFR\&L1}}$), or neither (*Random*). When no eligible text satisfies the primary matching criterion, we progressively relax the constraints. For $Matched_{\texttt{CEFR\&L1}}$, the fallback order is exact CEFR+L1 and adjacent CEFR level with the same L1. For $Matched_{\texttt{CEFR}}$, the order is exact CEFR and adjacent CEFR level. For $Matched_{\texttt{L1}}$, the order is exact L1 followed by an unrestricted global pool. In all cases, except $Synth_{\texttt{Trs}}$ and $Matched_{\texttt{Spkr}}$, the voice's own written responses are excluded from the candidate pool. To promote diversity in the synthetic datasets, we use a balanced selection mechanism which tracks the frequency with which voices and texts are selected and avoids repetition of speakers or responses. To improve robustness, we generate five different draws for each matching condition. Each draw corresponds to a different random assignment of voices to texts under the same matching constraints, such that a voice paired with one text in a given draw may be paired with a different text in another.[2]

4 Experimental Setup

The first objective of this work is to investigate the requirements and criteria for generating synthetic data suitable for L2 spontaneous speaking assessment. To this end, we first synthesise the test set data (see Section 5.1). The underlying idea is that, if an automated grading system trained on real data produces similar results when tested on synthetic and real audio data, then the synthetic data are likely to preserve the characteristics that are relevant for assessment.

Based on this initial analysis, we then revisit the synthetic generation pipeline. In particular, after observing that synthetic speech generated directly from raw written responses leads to a substantial degradation in the text-based grading system, we introduce a text speechification step (see Section 5.2) to transform written responses into spoken-style transcripts before TTS synthesis. We subsequently evaluate whether this modification improves the quality of the generated speech.

Finally, having identified both the most effective speaker-text pairing strategy and the benefit of speechification, we apply the resulting pipeline to the training data (see Section 5.3). The generated synthetic responses are used for data augmentation to investigate whether additional synthetic training examples improve the performance of our graders.

4.1 Models

4.1.1 TTS

As our TTS/voice-cloning system for spoken data augmentation, we use OmniVoice (Zhu et al., 2026),[3] a zero-shot multilingual TTS model supporting over 600 languages, based on a non-autoregressive diffusion language model architecture that directly maps text to multi-codebook acoustic tokens. To generate synthetic spoken responses, the written responses in the test set are first segmented into sentences using the `spaCy`[4] *sentencizer*. Voice cloning is then applied at the sentence-level using the first 10 seconds of the

[2] With respect to the fallback system, the $Matched_{\texttt{CEFR\&L1}}$ test set only has 5 adjacent cases for each draw, while the $Matched_{\texttt{CEFR}}$ test set only has 1 adjacent case for each draw. The $Matched_{\texttt{L1}}$ and *Random* test sets fully satisfy their criteria.

[3] github.com/k2-fsa/OmniVoice

[4] spacy.io

reference audio, and the resulting synthetic utterances are concatenated to reconstruct the full spoken response.

4.1.2 Automatic speech recognition (ASR)

In a real-world deployment, manually transcribing every learner's spoken response is infeasible. Consequently, text-based grading systems must rely on ASR to produce the transcriptions used for assessment. To do this, we use the publicly available CrisperWhisper (Zusag et al., 2024),[5] a variant of OpenAI's Whisper (Radford et al., 2023) designed for verbatim speech recognition. Unlike the original Whisper model, which tends to normalise speech and omit disfluencies in favour of a more fluent transcription style, CrisperWhisper aims to transcribe speech as accurately as possible, preserving all spoken content, including fillers and disfluencies.

4.1.3 Grading systems

For automatic language assessment, i.e., the main task of this paper, we employ an audio-based and a text-based grading system. The former is based on HuBERT (Hsu et al., 2021),[6] with an architecture similar to that described in our previous work (McKnight et al., 2023), where we used wav2vec2 (Baevski et al., 2020). The audio recordings are first segmented into overlapping fixed-length chunks prior to modelling. We use 15-second segments with a 3-second overlap to ensure coverage of long-form responses while preserving local temporal context. At model level, HuBERT is used as a frozen feature encoder, and we introduce a regression head operating on pooled hidden representations. Instead of simple pooling, we apply a hierarchical attention-based aggregation mechanism over frame-level encoder outputs. This mechanism learns to weight temporal frames and segment-level representations before producing a final utterance-level embedding, which is passed through a feed-forward regression head to predict the score.

The text-based grading system is based on ModernBERT (Warner et al., 2025).[7] The model is fine-tuned using a standard sequence regression head.

For both grading systems, we use an ensemble of two graders using two different random seeds, and training minimises mean squared error between predicted and placement scores normalised from 0 to 1 (see Section 2) for a maximum of 20 epochs. The training hyperparameters are reported in Table 2.

Hyperparameter	HuBERT	MBERT
Learning rate	1×10^{-5}	2×10^{-5}
Batch size	8	32
Gradient accumulation steps	4	-
Early stopping patience	3	3

Table 2: Grading systems training hyperparameters.

For each of the two training seeds, we randomly select 10% of the real spoken training data (41 samples) as a development set for optimisation. To avoid data contamination, the candidate IDs selected for the development set in each seed are excluded from that seed's augmented training sets, including both their written texts and speech recordings.

As will be illustrated in Section 5.3, our best-performing system uses a training set composed of real learner data and synthetic data generated by matching learners' texts and voices based on CEFR level. This means that, for the synthetic section of the training data, we have two possible target training labels: one placement score referred to voices and one associated with texts. However, since they are matched based on CEFR level, their differences are tiny, i.e., the mean of the absolute difference between the two sets of labels on the whole dataset is 0.043 and their standard deviation is 0.034. Therefore, we chose the labels associated with voices as training labels.

4.1.4 Text speechifier

To transform the original written responses into spoken-style learner transcripts, we use GPT-4.1 (OpenAI, 2023). The cost of speechifying the 383 written responses in the training set was approximately $2.5. Generating speechified versions of the 2,820 responses in the *Extra* set (see Section 5.3) incurred an additional cost of approximately $15. We used the GPT-4.1 API with data sharing disabled, so that the submitted data were not used for model fine-tuning or other data-sharing purposes.

4.2 Evaluation metrics

4.2.1 Speaker verification

To assess whether speaker identity is preserved after voice cloning, we compute cosine similarity

[5] github.com/nyrahealth/CrisperWhisper
[6] Initialised from facebook/hubert-base-ls960.
[7] Initialised from answerdotai/ModernBERT-base.

between the speaker embeddings of the real and synthetic speech. The speaker embeddings are extracted using `pyannote` (Bredin et al., 2020).

4.2.2 ASR evaluation

We use CrisperWhisper to transcribe both real and synthetic spoken responses. As these transcripts are used in the text-based grading system, transcription quality is critical. We evaluate them using word error rate (WER).

4.2.3 Speech quality

To assess the perceptual quality of the synthetic speech, we use UTMOSv2 (Baba et al., 2024), a non-intrusive speech quality assessment model that predicts human ratings of speech quality from the audio signal. We report the mean UTMOS score across the generated responses ranging from 1 to 5, with higher scores indicating better perceived speech quality. This metric provides a complementary assessment to speaker verification, allowing us to evaluate not only whether the cloned voice preserves speaker identity, but also the overall quality and naturalness of the synthesised speech.

4.2.4 Grading performance

To evaluate grading performance, we report Pearson's correlation coefficient (PCC), Spearman's rank correlation coefficient (SRC), and root mean square error (RMSE) between the reference and predicted scores. For PCC and SRC, the higher value, the better. For RMSE, the lower value the better.

4.2.5 Text analysis

We analyse textual differences between real and synthetic transcriptions using five metrics: average text length (Txt. len.), the average mean dependency distance (MDD) (Liu, 2008) as computed with `textdescriptives`,[8] hesitation ratio (Hes. rt.), Moving Average Type-Token Ratio (MATTR) (Covington and McFall, 2010) with a window size of 50 tokens, and the average number of unique difficult words (Dif. wds.) as computed with `textstat`.[9]

4.2.6 Text similarity

As described in Section 3.1, we speechify written responses. To ensure that this transformation preserves the original semantic content, we measure text similarity between the written responses before and after speechification using cosine similarity between sentence embeddings of original and speechified written responses. These are extracted using `all-mpnet-base-v2` (Song et al., 2020; Reimers and Gurevych, 2019).

[8] `github.com/HLasse/textdescriptives`
[9] `textstat.org/`

5 Experimental results

5.1 Analysis and results on augmented test set without training

As illustrated in Section 4, we first evaluate the quality of the synthetic spoken responses on the test set across all matching conditions, prior to introducing speechification.

First, we evaluate whether speaker identity is preserved. As illustrated in Section 4.2, we report the cosine similarity between the speaker embeddings of the real and synthetic speech for each matching condition in Table 3. Additionally, following our previous work (Karanasou et al., 2025), we report the empirical cumulative distribution function (ECDF) of the cosine distances between speaker embeddings (i.e., real vs. synthetic) in Figure 2. As can be observed, no substantial differences are observed across matching conditions, except for $Synth_{\texttt{Trs}}$ and $Matched_{\texttt{Spkr}}$, which yield the highest cosine similarity. This is expected, as we match learners' voices with their own manual transcriptions for $Synth_{\texttt{Trs}}$ and with their own written production for $Matched_{\texttt{Spkr}}$. The few failures were traced to a recording with prolonged background noise before speech onset, which corrupted the speaker representation used for voice cloning.

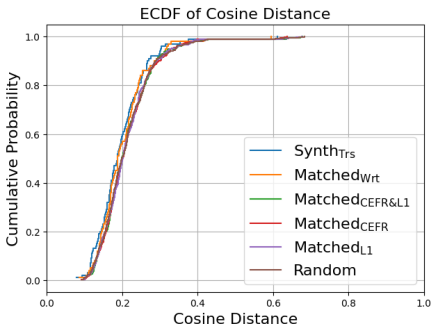


Figure 2: ECDF of cosine distances across matching conditions.

We then compute WER on the CrisperWhisper transcriptions across the different matching conditions. The results are also reported in Table 3.

We do not observe systematic differences in WER across the various conditions.

Linguistic content source	Test Set	WER	Cos. Sim.
Spoken	$\text{Real}_{\texttt{Spoken}}$	11.98	–
	$\text{Synth}_{\texttt{Trs}}$	11.03	0.803
Written	$\text{Matched}_{\texttt{Spkr}}$	12.00	0.797
	$\text{Matched}_{\texttt{CEFR\&L1}}$	$9.60_{\pm 1.45}$	$0.790_{\pm .001}$
	$\text{Matched}_{\texttt{CEFR}}$	$9.87_{\pm 1.16}$	$0.789_{\pm .001}$
	$\text{Matched}_{\texttt{L1}}$	$10.77_{\pm 2.10}$	$0.788_{\pm .002}$
	Random	$10.87_{\pm 1.46}$	$0.788_{\pm .002}$

Table 3: WER and cosine similarity (Cos. sim.) across matching conditions. Mean and standard deviation are computed across the five draws.

Subsequently, UTMOSv2 is used to score the speech quality of the TTS-generated data. Table 4 reports the results across matching conditions. UTMOSv2 is trained primarily on corpora of L1-

Linguistic source content	Test set	UTMOS
Spoken	$\text{Real}_{\texttt{Spoken}}$	2.205
	$\text{Synth}_{\texttt{Trs}}$	2.326
Written	$\text{Matched}_{\texttt{Spkr}}$	2.366
	$\text{Matched}_{\texttt{CEFR\&L1}}$	$2.380_{\pm .04}$
	$\text{Matched}_{\texttt{CEFR}}$	$2.398_{\pm .05}$
	$\text{Matched}_{\texttt{L1}}$	$2.394_{\pm .03}$
	Random	$2.402_{\pm .05}$

Table 4: Mean UTMOS scores results across matching conditions. Mean and standard deviation are computed across the five draws.

speaker and professionally-recorded TTS output, and it is calibrated to treat deviations from fluent, native-like prosody and pronunciation as indicators of lower quality. As our reference recordings consist of authentic L2 English speech, they naturally contain non-native prosodic patterns, disfluencies, and accented pronunciation that are typical and expected features of L2 speech rather than genuine defects. As a result, UTMOS seem to penalise reference recordings for characteristics that are inherent to L2 speech rather than indicative of poor recording or speech quality, as also observed by Onda et al. (2026). This can lead to artificially lower scores for the reference recordings compared with the synthetic systems. Such domain mismatch suggests that UTMOS scores should be interpreted cautiously when applied to L2 speech data. However, if we treat it as a comparative metric, all systems are shown to obtain similar results, with the lowest performance observed in $\textit{Real}_{\texttt{Spoken}}$, suggesting that TTS synthesis may smooth over some of the L2-specific characteristics that would otherwise lower UTMOS scores.

We can then evaluate the performance of the grading systems, as shown in Tables 5 and 6. It is important to note that, at this stage, neither grader has been exposed to synthetic data during training. The HuBERT-based grader is trained solely on the real audio recordings, whereas the ModernBERT-based grader is trained solely on the corresponding CrisperWhisper transcriptions. As a result, across all matching conditions, the former is evaluated on audio recordings, whereas the latter is evaluated on the corresponding CrisperWhisper transcriptions. Due to space constraints, we report only PCC; however, SRC and RMSE exhibit similar trends. Table 5 reports the results on $\textit{Real}_{\texttt{Spoken}}$, $\textit{Synth}_{\texttt{Trs}}$, and $\textit{Matched}_{\texttt{Spkr}}$, for which there is only one reference score label, as both text and voice originate from the same learner. The results in this table serve as best-case references with best results achieved, as expected, on $\textit{Real}_{\texttt{Spoken}}$ test data. $\textit{Synth}_{\texttt{Trs}}$ leads to a drop in performance, indicating that the TTS system already introduces some degradation even without data augmentation with different speakers. For $\textit{Matched}_{\texttt{Spkr}}$, the drop in performance is more pronounced for ModernBERT (0.195 PCC points) than for HuBERT (0.081 PCC points). We discuss this difference in more detail below, after analysing the results reported in Table 6.

Linguistic content source	Test Set	HuBERT	MBERT
Spoken	$\text{Real}_{\texttt{Spoken}}$	0.677	0.684
	$\text{Synth}_{\texttt{Trs}}$	0.570	0.653
Written	$\text{Matched}_{\texttt{Spkr}}$	0.596	0.489

Table 5: Baseline - no synthetic training data - grading results on real and matched transcript/speaker test sets in terms of PCC. One reference score label per test set.

Across all the other pairing strategies, as described in Section 3.2, each synthetic response is generated by combining a text with a voice. Therefore, two reference scores (i.e., one for audio and one for text) can be associated with a generated response. Accordingly, Table 6 reports grading performance with respect to both labels. As expected, performance is generally higher when HuBERT-based predictions are evaluated on audio and ModernBERT-based predictions are evaluated on text, as these represent the most appropriate input modalities for each model (see quadrants highlighted in yellow).

Model	Test Set	Audio ref.	Text ref.
HuBERT	$\text{Matched}_{\text{CEFR\&L1}}$	$0.544_{\pm.02}$	$0.553_{\pm.02}$
	$\text{Matched}_{\text{CEFR}}$	$0.560_{\pm.02}$	$0.534_{\pm.01}$
	$\text{Matched}_{\text{L1}}$	$0.542_{\pm.01}$	$0.399_{\pm.05}$
	Random	$0.554_{\pm.01}$	$0.054_{\pm.11}$
MBERT	$\text{Matched}_{\text{CEFR\&L1}}$	$0.454_{\pm.02}$	$0.484_{\pm.01}$
	$\text{Matched}_{\text{CEFR}}$	$0.468_{\pm.03}$	$0.474_{\pm.02}$
	$\text{Matched}_{\text{L1}}$	$0.331_{\pm.08}$	$0.521_{\pm.02}$
	Random	$0.033_{\pm.10}$	$0.522_{\pm.02}$

Table 6: Baseline - no synthetic training data - PCC grading results on matched CEFR and/or L1 or random paired test sets. Two reference score labels per test set.

Overall, the choice of linguistic content source has relatively little effect on HuBERT, while the choice of synthetic voice has relatively little effect on ModernBERT, as reflected by the comparatively small differences across matching conditions in the yellow quadrants. This suggests that the matching strategy does not have a substantial or consistent effect on either grader when evaluated on its corresponding input modality.

For our data augmentation setting, however, the downstream grading system is not known a priori. We therefore favour matching strategies that provide stable performance across both reference types and both grading modalities. Under this criterion, *Random* fails, while $\textit{Matched}_{\text{CEFR\&L1}}$ and $\textit{Matched}_{\text{CEFR}}$ show relatively consistent performance across the two reference types for both the text- and audio-based graders. $\textit{Matched}_{\text{L1}}$ exhibits less stable performance across evaluation settings. Since there is no substantial difference between $\textit{Matched}_{\text{CEFR\&L1}}$ and $\textit{Matched}_{\text{CEFR}}$, we select the latter for our data augmentation experiments, as it requires less information: only the CEFR level of the source voice and the written response needs to be known.

In Section 5.3, we additionally compare the performance of systems trained using data generated under the *Random* matching strategy. This provides a complementary analysis of the extent to which the choice of matching strategy affects the performance of models trained on the resulting synthetic data, rather than only their evaluation on the held-out test set.

Focusing on model-specific behaviour, a comparison of the results reported in Table 5 and Table 6 shows that the HuBERT-based grader shows only a reasonable drop in PCC when evaluated on the synthetic condition $\textit{Matched}_{\text{Spkr}}$ compared to the $\textit{Real}_{\text{Spoken}}$ setting. In contrast, the ModernBERT-based grader shows a substantially larger degradation (Table 5). This degradation is not recovered by the matching strategies that we applied (Table 6). We hypothesise that this discrepancy arises from the linguistic gap between written and spoken language. Although the TTS system introduces some hesitations and repetitions, written responses still differ fundamentally from spoken production: they lack discourse markers, reformulations, and other indicators of online speech planning such as uncertainty or self-repair. Moreover, written responses tend to exhibit higher syntactic complexity and lexical richness, reflecting greater planning time compared to spontaneous speech.

Our hypothesis is supported by the analysis reported in Table 7, which presents the textual metrics described in Section 4.2 for automatic transcriptions derived from real and synthetic test data. For simplicity, we report results only for one draw of the $\textit{Matched}_{\text{CEFR}}$ condition. Comparing *Real* and $\textit{Matched}_{\text{CEFR}}$ columns, transcriptions of synthetic speech derived from written responses are, on average, shorter than those derived from real learner speech (block Txt. len.). As shown in the MDD block, they also tend to show a slightly higher syntactic complexity than real learner speech. Furthermore, they contain substantially fewer hesitations (block Hes. rt.), reflecting the fact that TTS-generated speech does not fully capture the disfluency patterns characteristic of spontaneous learner production. At the same time, synthetic transcriptions exhibit greater lexical diversity (block MATTR) and contain a larger number of difficult words (block Dif. wds.). This is expected, as the synthetic spoken responses are generated from written productions, which are typically more sophisticated and carefully planned than spontaneous speech. These findings support our hypothesis that the linguistic characteristics of written language are largely preserved in the synthetic responses, despite the fact that the TTS randomly introduces some hesitations and repetitions.

5.2 Effects of speechification

To address the gap between written and spoken language, we revisit the written responses and apply the speechification procedure described in Section 3.1, transforming them into spoken-style learner transcripts using GPT-4.1 prior to speech synthesis. To assess whether semantic content is preserved after speechification, as de-

	Count	Txt. len.			MDD			Hes. rt. (%)			MATTR			Dif. wds.		
		Real	Matched$_{\text{CEFR}}$	+Spf	Real	Matched$_{\text{CEFR}}$	+Spf	Real	Matched$_{\text{CEFR}}$	+Spf	Real	Matched$_{\text{CEFR}}$	+Spf	Real	Matched$_{\text{CEFR}}$	+Spf
A	9	249.8	155.8	263.9	2.38	2.48	2.35	8.2	2.7	15.3	0.56	0.67	0.61	9.0	11.8	11.0
B1	15	234.7	178.8	322.3	2.61	2.68	2.37	4.8	1.0	11.9	0.65	0.72	0.67	14.6	16.9	16.8
B2	25	284.7	190.2	330.4	2.69	2.62	2.32	5.5	0.8	10.8	0.65	0.72	0.68	15.9	15.6	16.3
C1	28	333.9	249.6	434.8	2.64	2.86	2.46	5.7	0.9	7.9	0.66	0.70	0.71	17.9	23.9	25.5
C2	23	328.9	222.4	443.9	2.56	2.68	2.40	3.9	0.7	7.8	0.68	0.73	0.72	19.0	22.3	26.3

Table 7: Text statistics across CEFR levels for CrisperWhisper transcriptions from *Real*$_{\texttt{Spoken}}$, *Matched*$_{\texttt{CEFR}}$, and *Matched*$_{\texttt{CEFR}}$+*Speechify* (+Spf) data. Abbreviations are explained in Section 4.2.

scribed in Section 4.2, we compute the cosine similarity between sentence embeddings extracted from the original (*Real*$_{\texttt{Written}}$) and speechified (*Real*$_{\texttt{Written}}$+*Speechify*) written responses. The value of $0.853_{\pm.052}$ indicates that the speechified responses remain highly similar to the original texts, suggesting that the speechification process largely preserves semantic content. For comparison, the cosine similarity between real written and spoken texts from the same learners is $0.813_{\pm.080}$.

In addition to semantic similarity, we analyse the +*Spf* column in Table 7, which reports the text statistics described in Section 4.2.5 for speechified *Matched*$_{\texttt{CEFR}}$ transcriptions. Overall, speechification shifts several properties of the generated transcripts towards those observed in real learner speech. For MDD, the speechified texts show lower values than the real spoken texts across all CEFR levels, suggesting that the procedure may over-simplify syntactic structure, while still moving away from the higher MDD values observed in the non-speechified transcriptions. Similarly, the hesitation rate increases substantially after speechification and exceeds the values observed in real speech, indicating that the model may overproduce hesitation phenomena. In contrast, MATTR generally moves towards the values observed in real speech. Speechification also substantially increases text length relative to the non-speechified *Matched*$_{\texttt{CEFR}}$ transcripts, bringing it closer to, and at several levels beyond, the length of real spoken responses. Finally, the number of difficult words remains relatively stable across conditions, suggesting that speechification primarily modifies the form of the responses rather than substantially changing their lexical difficulty. Together with the high semantic similarity reported above, these results indicate that speechification introduces several characteristics of spoken language while preserving the core content of the original responses, although some features, particularly hesitations and syntactic complexity, appear to be exaggerated.

Table 8 compares the performance obtained on *Matched*$_{\texttt{CEFR}}$ data before and after speechification. As illustrated in Section 3.1, speechification is conditioned on the CEFR level associated with the learner's text. Importantly, this CEFR information is used here only as a control variable for the generation process and for analysing whether proficiency-specific speech characteristics can be reproduced; it is not provided to the grading models and is not used to determine their predictions. Thus, the use of CEFR conditioning in this analysis should not be interpreted as providing additional proficiency information to the downstream grading systems. We also report the results when running the speechification process without injecting the CEFR-related labels into the prompt given to the LLM. The results indicate that CEFR-conditioned speechification substantially improves the quality of the synthetic data, yielding large gains for the ModernBERT-based grader and more moderate, yet consistent, improvements for the HuBERT-based system. These findings suggest that modelling the discourse- and fluency-related characteristics is at least as important as preserving acoustic characteristics when generating synthetic data for speaking assessment. The fact that, for ModernBERT, the results for *Matched*$_{\texttt{CEFR}}$+*Speechify* are even higher than those for *Real* (see Table 5) should not be surprising, as the two settings are not directly comparable: the former, despite speechification, is generated starting from written productions, whereas the latter consists of authentic spoken responses. Therefore, the linguistic content of the two sets is different.

The analysis in this section is intended to characterise the effect of speechification and its conditioning, rather than to establish its effectiveness as a training data augmentation strategy. We therefore provide a separate ablation experiment in Table 11 at the end of Section 5.3, where HuBERT and ModernBERT are trained with and without speechified synthetic data and evaluated on the real test set. This experiment provides a direct as-

sessment of the contribution of speechification to grading performance under the final training and evaluation setting.

Test Set	CEFR in Spf	HuBERT	MBERT
$\text{Matched}_{\text{CEFR}}$	-	$0.560_{\pm.02}$	$0.474_{\pm.04}$
$\text{Matched}_{\text{CEFR}}$+Spf	×	$0.542_{\pm.02}$	$0.457_{\pm.01}$
$\text{Matched}_{\text{CEFR}}$+Spf	✓	$\mathbf{0.583}_{\pm.01}$	$\mathbf{0.715}_{\pm.01}$

Table 8: Analysis on synthetic test sets in terms of PCC: *Matched*$_{\text{CEFR}}$ vs *Matched*$_{\text{CEFR}}$+*Speechify* (with and without CEFR-conditioning).

Additionally, in Figure 3, we report scatterplots between predictions on *Real*$_{\text{Spoken}}$ and predictions on *Matched*$_{\text{CEFR}}$+*Speechify* for both graders, together with the corresponding PCC, SRC, and RMSE values. The higher performance observed for the HuBERT-based system can be explained as follows: HuBERT operates directly on acoustic signals, which are largely preserved through the TTS process, as shown in Table 3 (see Cosine Similarity column), despite differences in linguistic content. In contrast, ModernBERT operates on text and is therefore more sensitive to differences in linguistic realisation, since *Real*$_{\text{Spoken}}$ derives from spoken responses, while *Matched*$_{\text{CEFR}}$+*Speechify*, despite speechification, still originates from a written source.

5.3 Training and testing with augmented data

At this point, we can use the augmented training data to train our grading systems and evaluate them on the real test set. In Table 9, we compare different training configurations. *Real*$_{\text{Spoken}}$ refers to models trained exclusively on the real training spoken data. *Syn* denotes training on synthetic data generated using the *Matched*$_{\text{CEFR}}$+*Speechify* pipeline by matching the 410 voices with the 383 written responses in the training set (see Table 1). For the audio-based grader, this results in 410 synthetic speech samples, with 27 (i.e., $410 - 383$) written responses reused across different voices. Conversely, for the text-based grader, training is performed on the automatic transcriptions obtained after synthesising the 383 unique written responses without duplicates. Finally, *Syn*$_{\text{Extra}}$ extends *Syn* by additionally incorporating the 2,820 written responses without spoken counterparts (i.e., *Extra* in Table 1).

As can be observed, data augmentation leads to consistent performance improvements across all

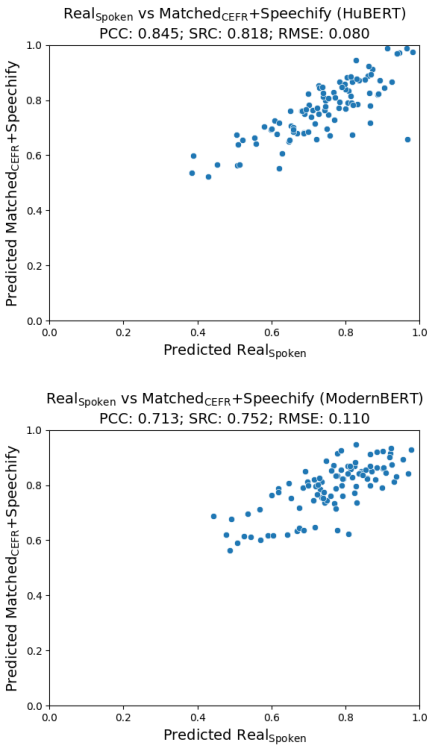


Figure 3: *Real*$_{\text{Spoken}}$ vs *Matched*$_{\text{CEFR}}$+*Speechify* predictions.

evaluation metrics for both models. We hypothesise that this gain is particularly pronounced for the ModernBERT-based system due to the substantially increased amount of training text available. In the *Real*$_{\text{Spoken}}$+*Syn*$_{\text{Extra}}$ setting, the model effectively benefits from a total of 3,613 training automatic transcriptions (i.e., 410 from *Real*$_{\text{Spoken}}$, 383 from *Syn*, and the remaining 2,820 in *Syn*$_{\text{Extra}}$). In contrast, the HuBERT-based system, although exposed to the same augmented synthetic speech pipeline, remains constrained by the diversity of available speaker identities. Specifically, it still relies on the 410 unique voices present in the spoken training set, which limits the effective increase in variability compared to the text-based model. However, for both models, the *Real*$_{\text{Spoken}}$+*Syn*$_{\text{Extra}}$ setting significantly outperforms the *Real*$_{\text{Spoken}}$ baseline.[10]

Interestingly, the combination of HuBERT and ModernBERT[11] using *Real*$_{\text{Spoken}}$+*Syn*$_{\text{Extra}}$ achieves the best performance showing that the two models carry complementary information. As

[10] We used a paired permutation test with 10,000 within-pair swaps comparing *Real*$_{\text{Spoken}}$ to *Real*$_{\text{Spoken}}$+*Syn*$_{\text{Extra}}$. This yielded $p = 0.041$ for HuBERT and $p = 0.011$ for ModernBERT.

[11] I.e., the average of the two grading systems predictions.

Model	Training Set	PCC	SRC	RMSE
HuBERT	$\text{Real}_{\texttt{Spoken}}$	0.677	0.634	0.133
	Syn	0.663	0.629	0.133
	$\text{Syn}_{\texttt{Extra}}$	0.687	0.656	0.134
	$\text{Real}_{\texttt{Spoken}}$+$\text{Syn}_{\texttt{Extra}}$	**0.739**	**0.692**	**0.119**
MBERT	$\text{Real}_{\texttt{Spoken}}$	0.684	0.656	0.132
	Syn	0.643	0.577	0.136
	$\text{Syn}_{\texttt{Extra}}$	0.663	0.607	0.132
	$\text{Real}_{\texttt{Spoken}}$+$\text{Syn}_{\texttt{Extra}}$	**0.769**	**0.710**	**0.113**
Both	$\text{Real}_{\texttt{Spoken}}$	0.730	0.699	0.121
	Syn	0.696	0.652	0.126
	$\text{Syn}_{\texttt{Extra}}$	0.717	0.686	0.123
	$\text{Real}_{\texttt{Spoken}}$+$\text{Syn}_{\texttt{Extra}}$	**0.795**	**0.775**	**0.107**

Table 9: Results on real test data using different training sets. *Syn* and $\textit{Syn}_{\texttt{Extra}}$ use the $\textit{Matched}_{\texttt{CEFR}}$+*Speechify* process.

we observed for the two models taken individually, the $\textit{Real}_{\texttt{Spoken}}$+$\textit{Syn}_{\texttt{Extra}}$ setting is significantly better than the $\textit{Real}_{\texttt{Spoken}}$ baseline.[12]

One of the leitmotifs of this paper is that the downstream grading system is not known a priori. However, if the target system is known to be text-based, TTS is not necessarily required. To examine this, we train a ModernBERT-based grader directly on the original $\textit{Real}_{\texttt{Spoken}}$ CrisperWhisper transcriptions together with the $\textit{Syn}_{\texttt{Extra}}$ data before the synthesis step, i.e., using the speechified written texts directly. This yields PCC = 0.778, SRC = 0.732, and RMSE = 0.110, compared with PCC = 0.769, SRC = 0.710, and RMSE = 0.113 when the same synthetic texts are first converted to speech and subsequently transcribed. The slightly better performance obtained without TTS suggests that speech synthesis and subsequent ASR introduce some noise that is detrimental to text-based grading. Thus, while TTS is necessary when augmenting an acoustic grading system, it appears to be an unnecessary intermediate step for a text-based grader when the latter can operate directly on the speechified transcripts.

Model	Matching	PCC	SRC	RMSE
HuBERT	$\text{Matched}_{\texttt{CEFR}}$	**0.663**	**0.629**	**0.133**
	$\text{Random}_{\texttt{Audio_Label}}$	0.632	0.614	0.139
	$\text{Random}_{\texttt{Text_Label}}$	0.322	0.342	0.175
MBERT	$\text{Matched}_{\texttt{CEFR}}$	**0.643**	0.577	**0.136**
	$\text{Random}_{\texttt{Audio_Label}}$	0.164	0.061	0.185
	$\text{Random}_{\texttt{Text_Label}}$	0.632	**0.586**	0.140

Table 10: Results on real test data comparing $\textit{Matched}_{\texttt{CEFR}}$+*Speechify* to *Random*+*Speechify* training data using the *Syn* training partition.

Furthermore, as mentioned in Section 5.1, we consider a different matching strategy, i.e., *Random*+*Speechify*, to generate synthetic training data. Under this matching strategy, the TTS-generated responses might have two very different text and speaker labels. We therefore train separate graders on the audio ($\textit{Random}_{\texttt{Audio_Label}}$) and the text labels ($\textit{Random}_{\texttt{Text_Label}}$) using the *Syn* training data partition. We compare this matching strategy to $\textit{Matched}_{\texttt{CEFR}}$ in Table 10. These results suggest that matching synthetic voices and texts by CEFR provides a more reliable alignment between the augmented samples and the proficiency labels than random matching. $\textit{Matched}_{\texttt{CEFR}}$ achieves the best PCC and RMSE for both models and the best SRC for HuBERT, with only a small exception for ModernBERT's SRC, where $\textit{Random}_{\texttt{TextLabel}}$ performs marginally better. More importantly, CEFR-based matching avoids a fundamental ambiguity associated with random matching: when the voice and text originate from learners with different proficiency levels, it is unclear whether the proficiency label of the synthetic sample should be determined by the voice or by the text. This is particularly problematic because the appropriate choice depends on the modality used by the downstream grader. By matching the voice and text at the same CEFR level (see Section 4.1.3), $\textit{Matched}_{\texttt{CEFR}}$ ensures that both modalities are associated with the same proficiency level, removing the need to make this decision and making the resulting synthetic data independent of the modality of the downstream grader. This is exactly where *Random* is shown to fail (see row $\textit{Random}_{\texttt{Text_Label}}$ for HuBERT and row $\textit{Random}_{\texttt{Audio_Label}}$ for ModernBERT).

Finally, as mentioned in Section 5.2, we conduct an ablation study in order to provide a direct assessment of the contribution of speechification to grading performance. Specifically, we consider the *Syn* training partition using the $\textit{Matched}_{\texttt{CEFR}}$ setting with and without the speechification process. As can be observed in Table 11, speechification consistently improves performance across all three evaluation metrics for both grading models. The improvement is substantially larger for ModernBERT, suggesting that the spoken-language phenomena introduced through speechification are particularly beneficial for the text-based grading system. This finding supports the analysis presented in Table 8 in Section 5.2, while providing direct evidence that speechification contributes to

[12] The paired permutation test with 10,000 within-pair swaps yielded $p = 0.009$.

downstream grading performance when used as a training data augmentation strategy.

Model	Speechification	PCC	SRC	RMSE
HuBERT	✓	**0.663**	**0.629**	**0.133**
	×	0.640	0.621	0.139
MBERT	✓	**0.643**	**0.577**	**0.136**
	×	0.546	0.522	0.148

Table 11: Results on real test data comparing $Matched_{\text{CEFR}}$ training data with and without speechification using the *Syn* training partition.

6 Conclusions

In this work, we investigated how to generate synthetic L2 speech for proficiency assessment. We proposed a cross-modal augmentation framework that speechifies written L2 responses using an LLM followed by TTS synthesis. Using HuBERT- and ModernBERT-based graders, we showed that pairing speakers and texts by proficiency level is a principled and effective strategy for data augmentation. We further found that raw written text creates a strong mismatch with spoken language, while speechification substantially reduces this gap and improves performance. Overall, synthetic data consistently improve both audio- and text-based grading systems, highlighting the importance of modelling spoken linguistic structure in addition to speaker characteristics.

Future work will explore LLM-generated responses directly, removing reliance on written sources, following Wang et al. (2025) and Voskoboinik et al. (2025). Furthermore, we plan to enhance the speechification step by calibrating it through few-shot prompting or level-specific statistics derived from real learner speech in order to adjust the overestimated hesitation rates and underestimated syntactic complexity observed in Table 7. An additional direction for future work is to investigate whether written texts from a lower CEFR level may better approximate spoken responses at the target level, given that written language tends to exhibit greater linguistic complexity than speech even within the same proficiency level. Finally, we plan to extend our data augmentation framework to other L2 assessment tasks and languages.

Limitations

Our experiments are based on COREFL and therefore involve a relatively limited number of speakers and approximately 16.7 hours of spoken training data. The extent to which our findings generalise to other L2 populations, proficiency distributions, languages, or assessment settings remains to be established. In particular, our experiments use a specific combination of LLM, TTS, ASR, and proficiency models, and different model architectures may respond differently to synthetic data.

A further limitation concerns the calibration of the speechification process. Although the LLM is instructed to modulate disfluency rates and syntactic complexity according to the learner's CEFR level, these instructions are qualitative rather than derived from the empirical distributions observed in real learner speech. As shown in Table 7, the resulting speechified transcriptions overestimate hesitation rates relative to authentic spoken productions at every proficiency level and, conversely, tend to underestimate syntactic complexity as measured by MDD. Explicitly calibrating the speechification prompt through few-shot prompting or against corpus-derived, level-specific statistics (e.g., target hesitation rates and MDD ranges) is a promising direction for future work, but is not explored in this study.

Ethics Statement

The use of voice cloning to generate synthetic L2 speech introduces ethical and privacy risks that require particular consideration. A cloned or synthesised voice may closely resemble that of a real speaker, raising concerns related to consent, speaker identity, impersonation, and unauthorised reuse of voice data. In this work, synthetic speech is generated solely for research and assessment-data augmentation. Voice cloning is not used to provide learner feedback, for example, by generating a synthesised version of a learner's utterance with corrected pronunciation or grammatical errors. Instead, its sole purpose is to increase the amount and diversity of training data available to the grading systems.

Acknowledgments

This paper reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: ChatGPT and Gemini were used only to review grammar and improve the fluency of the text.

References

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2024. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 818–824.

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS 2020*, pages 1–12.

Stefano Bannò, Katherine M Knill, Marco Matassoni, Vyas Raina, and Mark Gales. 2023. Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning. In *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 126–130.

Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt. 2002. Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36(1):9–48.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *Proc. ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain.

Sylvain Coulange, Marie-Hélène Fries, Monica Masperi, and Solange Rossato. 2024. A corpus of spontaneous L2 English speech for real-situation speaking assessment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 293–297, Torino, Italia.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pages 2258–2269, Online.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proc. 31st Second Language Research Forum*, pages 240–254, Somerville.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Lan-fen Huang, Yen-liang Lin, and Tomáš Gráf. 2023. Development of the use of discourse markers across different fluency levels of CEFR: A learner corpus analysis. *Pragmatics*, 33(1):49–77.

S. Ishikawa. 2011. A new horizon in learner corpus studies: The aim of the ICNALE project. In *Corpora and language technologies in teaching, learning and research*, pages 3–11. University of Strathclyde Press.

Penny Karanasou, Mengjie Qian, Stefano Bannò, Mark J.F. Gales, and Kate M. Knill. 2025. Data Augmentation for Spoken Grammatical Error Correction. In *Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 194–198.

Eesung Kim, Jae-Jin Jeon, Hyeji Seo, and Hoon Kim. 2022. Automatic Pronunciation Assessment using Self-Supervised Speech Representation Learning. In *Interspeech 2022*, pages 1411–1415.

Kate Knill, Diane Nicholls, Mark J.F. Gales, Mengjie Qian, and Pawel Stroinski. 2025. The Speak & Improve Corpus 2025: an L2 English Speech Corpus for Language Assessment and Feedback.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Cristóbal Lozano, Ana Díaz-Negrillo, and Marcus Callies. 2021. Designing and compiling a learner corpus of written and spoken narratives: COREFL. In Christiane Bongartz and Jacopo Torregrossa, editors, *What's in a Narrative? Variation in Storytelling at the Interface Between Language and Literacy*, pages 21–46. Peter Lang Verlag, Berlin, Germany.

Mercer Mayer. 2003. *Frog, where are you?* Penguin.

Simon W McKnight, Arda Civelekoglu, Mark Gales, Stefano Bannò, Adian Liusie, and Katherine M Knill. 2023. Automatic Assessment of Conversational Speaking Tests. In *Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 99–103.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve Corpus 2024: Error-annotated and CEFR-labelled essays by learners of English.

Kentaro Onda, Satoru Fukayama, Daisuke Saito, and Nobuaki Minematsu. 2026. How should we evaluate naturalness for synthetic speech with a foreign accent? *arXiv preprint arXiv:2610.13175*.

OpenAI. 2023. GPT-4 Technical Report.

L. Peng, K. Fu, B. Lin, D. Ke, and J. Zhan. 2021. A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis. In *Proc. Interspeech 2021*, pages 4448–4452.

Mengjie Qian, Kate Knill, Stefano Bannò, Siyuan Tang, Penny Karanasou, Mark Gales, and Diane Nicholls. 2025. Speak & Improve Challenge 2025. In *Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 41–45.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. International Conference on Machine Learning (ICML)*, pages 28492–28518. PMLR.

Vyas Raina, Mark J.F. Gales, and Kate M. Knill. 2020. Universal Adversarial Attacks on Spoken Language Assessment Systems. In *Proc. Interspeech 2020*, pages 3855–3859.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Proc. Advances in Neural Information Processing Systems (NeuRIPS)*, volume 33, pages 16857–16867.

Ekaterina Voskoboinik, Anna von Zansen, Nhan Chi Phan, Yaroslav Getman, Tamás Grósz, and Mikko Kurimo. 2025. Enhancing second language speaking assessment: Integrating large language models for Finnish and Finland Swedish proficiency scoring. *Language Testing*, 42(4):508–538.

Chung-Chun Wang, Jhen-Ke Lin, Hao-Chien Lu, Hong-Yun Lin, and Berlin Chen. 2025. A novel data augmentation approach for automatic speaking assessment on opinion expressions. In *Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 199–203.

X. Wang, K. Evanini, Y. Qian, and M. Mulholland. 2021. Automated scoring of spontaneous speech from young learners of english using transformers. In *Proc. 2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma. 2021. Explore wav2vec 2.0 for Mispronunciation Detection. In *Proc. Interspeech 2021*, pages 4428–4432.

Xun Yan, Ping-Lin Chuang, Yulin Pan, Huiying Cai, Shelley Staples, and Mariana Centanin Bertho. 2025. Disfluency doesn't happen in isolation: Exploring how individual disfluency features co-occur in L2 speaking performances. *Studies in Second Language Acquisition*, 47(2):560–591.

Klaus Zechner and Keelan Evanini. 2019. *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762:

An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment. In *Proc. Interspeech 2021*, pages 3710–3714.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-ARCTIC: A non-native English speech corpus. In *Proc. Interspeech 2018*, pages 2783–2787.

Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. 2026. OmniVoice: Towards Omnilingual Zero-Shot Text-to-Speech with Diffusion Language Models.

Mario Zusag, Laurin Wagner, and Bernhad Thallinger. 2024. CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions. In *Proc. Interspeech 2024*, pages 1265–1269.

A Prompt for speechification

Below, we report the prompt used for GPT-4.1 *speechification*, i.e., the transformation of essay-like texts into spoken-style learner transcripts. We highlight in bold the placeholders for the learner's CEFR level and written response.

```
Rewrite the following text
as a spoken response by an
English learner.

Speaker profile:

* Proficiency level
(CEFR): [CEFR LEVEL]

Input text: [WRITTEN
RESPONSE]

Core constraints:

* Keep the original
meaning exactly the same.

* DO NOT correct grammar
mistakes or improve the
language.

* Preserve all learner
errors.

* Do NOT introduce new
factual content.

Goal:

Produce a realistic
transcription of
spontaneous speech, as if
the learner is speaking
and constructing the
response in real time (not
reading or summarising).

Discourse and structure:

* Convert the text into
an incremental spoken
narrative.

* Avoid compact or
essay-like sentences.

* Use step-by-step
development of ideas (as
if recalling or explaining
while speaking).

* Allow partial sentences,
restarts, and local
reformulations.

* Introduce mild
redundancy (rephrasing,
clarifying, or repeating
ideas).

* Expand the text slightly
to resemble spoken length
(do not shorten).

Disfluency modelling
(must be natural and
controlled):

Include a mix of:

* Hesitations (e.g.,
''um'', ''uh'')

* Repetitions (e.g.,
''he-he goes'', ''they
go, they go'')

* False starts and
self-repairs (e.g.,
''go-went'', ''to the-
to the place'')

* Occasional word
fragments (e.g.,
''bi-bit'', ''f-falled'')

Placement rules:

* Insert disfluencies at
natural planning points
(before verbs, noun
phrases, or difficult
words).

* Include some mid-clause
disfluencies, not only at
sentence boundaries.
```

* Do NOT overuse any single type.

* Model disfluencies according to CEFR level.

* Model hesitation frequency according to CEFR level, distributing them naturally across the text (not clustered or periodic).

Cognitive/production effects:

Simulate real speaking behaviour:

* Include signs of uncertainty or memory limits (e.g., ''I think'', ''I don't remember'', ''something like that'') when appropriate.

* Allow vague references (e.g., ''that thing'', ''that kind of things'').

* Do not make the speaker sound perfectly confident or precise.

* Avoid perfectly structured summaries.

Proficiency adaptation:

* Lower CEFR (A1{A2): shorter chunks, more hesitations, more repetitions, simpler structure.

* Mid CEFR (B1{B2): moderate disfluency, some repairs, longer sentences.

* High CEFR (C1{C2): mostly fluent, occasional hesitation for planning.

Formatting:

* Keep punctuation clear and readable.

* Use commas, dashes, and ellipses to reflect pauses and repairs.

* Do NOT produce raw ASR-style text.

* Output should remain compatible with standard written text processing.

Final requirement:

The output must resemble a real learner transcript, capturing both linguistic errors and the cognitive process of speaking, not just adding superficial noise. Only output the transcript without adding any other comments or observations.

Here, we report the prompt used for speechification without CEFR-conditioning:

Rewrite the following text as a spoken response by an English learner.

Input text: **[WRITTEN RESPONSE]**

Core constraints:

* Keep the original meaning exactly the same. * DO NOT correct grammar mistakes or improve the language. * Preserve all learner errors. * Do NOT introduce new factual content.

Goal: Produce a realistic transcription of spontaneous speech, as if the learner is speaking and constructing the response in real time (not reading or summarising).

Discourse and structure:

* Convert the text into an incremental spoken narrative. * Avoid compact or essay-like sentences. * Use step-by-step development of ideas (as if recalling

or explaining while speaking). * Allow partial sentences, restarts, and local reformulations. * Introduce mild redundancy (rephrasing, clarifying, or repeating ideas). * Expand the text slightly to resemble spoken length (do not shorten).

Disfluency modelling (must be natural and controlled): Include a mix of:

* Hesitations (e.g., ''um'', ''uh'') * Repetitions (e.g., ''he-he goes'', ''they go, they go'') * False starts and self-repairs (e.g., ''go-went'', ''to the- to the place'') * Occasional word fragments (e.g., ''bi-bit'', ''f-falled'')

Placement rules:

* Insert disfluencies at natural planning points (before verbs, noun phrases, or difficult words). * Include some mid-clause disfluencies, not only at sentence boundaries. * Do NOT overuse any single type.

Cognitive/production effects: Simulate real speaking behaviour:

* Include signs of uncertainty or memory limits (e.g., ''I think'', ''I don't remember'', ''something like that'') when appropriate. * Allow vague references (e.g., ''that thing'', ''that kind of things''). * Do not make the speaker sound perfectly confident or precise. * Avoid perfectly structured summaries.

Formatting:

* Keep punctuation clear and readable. * Use commas, dashes, and ellipses to reflect pauses and repairs. * Do NOT produce raw ASR-style text. * Output should remain compatible with standard written text processing.

Final requirement: The output must resemble a real learner transcript, capturing both linguistic errors and the cognitive process of speaking, not just adding superficial noise. Only output the transcript without adding any other comments or observations.

Automatic Analysis of Noun Phrase Complexity in Learner spoken Productions

Saint Joannes Monengue Essengo, Bi Yoman Ghislain Aymard Toa, Thomas Gaillat, Christine Evain, Fabienne Moreau
Université Rennes 2, LIDILE
{saint-joannes.monengue-essengo, bi.yoman.ghislain.toa}@etudiant.univ-rennes2.fr
{thomas.gaillat, christine.evain, fabienne.moreau}@univ-rennes2.fr

Abstract

Most tools for measuring syntactic complexity were built for written English, and learner speech, like comparisons across languages, has received far less attention. We present an open pipeline, anchored in Universal Dependencies (UD), that extracts nominal constructions from transcribed speech and applies a single set of metric definitions to every language it processes. We test it on the Corpus InterLangue (CIL), a comparable spoken corpus of learner French and learner English, and on the spoken part of the International Corpus of English – Great Britain (ICE-GB) as a native baseline. As a result ,learner English shows more phrasal complexity than learner French, which leans on appositive and looser structures. Against native speakers the picture changes. Natives pack complexity into the noun phrase, while learners spread it across longer, flatter structures. Against manual gold standards, extraction reaches strict F1 scores between .79 and .95, depending on the corpus, and group differences are tested with Mann–Whitney U tests and Cliff's delta. The tool and its metric definitions are openly available.

1 Introduction

1.1 Motivation and research gap

Complexity research in second language acquisition (SLA) traditionally relied on global measures such as sentence length, subordination ratios and type–token ratios (Wolfe-Quintero et al., 1998; Norris and Ortega, 2009; Lu, 2010). Such measures are computationally undemanding but insufficiently discriminating. Two texts can share a mean clause length and still build their noun phrases in entirely different ways, and that is where learner and native production tend to diverge (Biber et al., 2011, 2020, 2025). The automatic toolkits that compute these indices, and finer-grained ones, were calibrated on written English (Lu, 2010; Kyle, 2016). This leaves two gaps open.

The first is speech. Speech is not writing with the disfluencies taken out; it packages information differently (Biber et al., 2011), and a metric tuned on academic prose will show different values when applied to a semi-guided interview. Even the unit of analysis is unstable, because transcripts lack the sentence boundaries of writing (Foster et al., 2000). One of our own learner transcripts, for instance, contains an unpunctuated unit of 817 words. Parsers add their share, since those trained on L1 English alone are less accurate on spoken learner English than those that have also seen learner data (Kyle et al., 2022). The second gap is comparability across languages. Asking whether learners of French and learners of English behave alike requires the same construct, measured the same way, on both sides; separate tools and post-hoc alignments do not deliver that. Multilingual UD-based profilers exist (Brunato et al., 2020), but they describe texts broadly, not noun phrases in particular.

This paper presents an analysis based on a tool built for that niche. It is an open-source pipeline[1] that extracts noun phrases from learner speech and, because it is anchored in Universal Dependencies (UD), measures their complexity comparably across languages.

1.2 Research questions

Both research questions are questions about the tool, which we test on the Corpus InterLangue (CIL) and on ICE-GB (Section 3).

RQ1. Does the tool accurately capture differences in noun phrase complexity between learners of French and learners of English? To answer this,

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

[1]The tool is openly available on GitLab (https://gitlab.huma-num.fr/lidile/constructions-nominales/-/tree/v0.3.0).

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 214–228, November 19–20, 2026, University of Tartu Library

we compare learner French and learner English from the CIL, focusing on whether the tool can detect structural differences in how complexity is distributed within noun phrases.

RQ2. Can the tool reliably distinguish between learner and native English in terms of noun phrase complexity? Here, we compare learner English from the CIL with native English from ICE-GB, assessing whether the tool can identify developmental differences in how complexity is packed into noun phrases.

2 Theoretical Background

2.1 The shift from global to phrasal complexity

Complexity entered SLA as one of three dimensions of performance, alongside accuracy and fluency (Housen and Kuiken, 2009), and for a long time it was measured globally (Norris and Ortega, 2009; Bulté and Housen, 2012). Length- and subordination-based ratios dominated developmental research (Wolfe-Quintero et al., 1998); tools such as the L2 Syntactic Complexity Analyzer automated fourteen of them (Lu, 2010, 2011). Such ratios trace broad developmental trends well, but they cannot locate the point, inside the sentence, where a learner's grammar departs from a native speaker's (Biber et al., 2020).

Norris and Ortega (2009) defended an organic view of CAF and showed that global measures do capture large-scale developmental variation. Bulté and Housen (2012) then took the construct apart, separating absolute from relative complexity, and syntactic complexity (clausal, phrasal, subclausal) from morphological complexity. More recently, Bulté et al. (2025) drew a sharp line between complexity, a structural property of linguistic components and their relations, and difficulty, the cognitive cost of those structures. We work on the structural side of that line.

An important move was brought forward by Biber and Gray. Global indices, they argued, misfire across registers (Biber et al., 2011; Biber and Gray, 2016), because complexity does not sit in the same place everywhere. It is clausal in conversation, whereas in informational writing it migrates into the noun phrase, where phrasal compression prevails over clausal elaboration (Biber and Gray, 2011). Learner writing follows suit. Less proficient writers of English for Academic Purposes (EAP) lean on attributive adjectives, more advanced ones on nouns as premodifiers and on prepositional phrases (Parkinson and Musgrave, 2014). Hence the current call to describe specific structures and their syntactic functions instead of relying on omnibus indices (Biber et al., 2020, 2025), a call that the present paper takes up.

The Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) turned these ideas into software (Kyle, 2016; Kyle and Crossley, 2018). Its phrasal indices cross seven noun phrase types, defined by function (nominal subject, direct object, and so on), with ten dependent types (determiners, adjectival modifiers, prepositional phrases and relative clauses, among others), which gives 66 indices in all. On TOEFL essays they predicted writing quality better than global and fine-grained clausal indices (Kyle and Crossley, 2018), which made the phrase, and not only the clause, a unit worth measuring. Section 7.1 returns to TAASSC once our own measures have been introduced.

2.2 The noun phrase as the locus of complexity

The noun phrase has become a privileged site for observing complexity in SLA. Global measures average structural differences away; noun phrase measures keep them. Learners, for instance, may pack less into a noun phrase and spread the same content over appositions or separate clauses.

In spoken learner data, this often takes the form of successive reformulations, where each noun phrase remains short and new information is added in a separate, loosely attached phrase instead of being embedded within the head noun phrase. Examples (1a) and (2a) come from CIL interviews, (1a) from a French-speaking learner of English (CIL-EFL) and (2a) from an English-speaking learner of French (CIL-FFL); each is identified by its transcript code in the corresponding sub-corpus. Examples (1b) and (2b) are constructed counterparts.

(1) a. *it's like erh a sausage a big sausage* (CIL_L10_Fr_07_F_15, turn 007)
b. *it's like a big sausage*

(2) a. *une professeur au l'université une professeur de littérature anglophone* (CIL_ENG_JO_86_F_10)
b. *une professeure de littérature anglophone à l'université*

In (1a) and (2a), the speaker produces a first noun

phrase (*a sausage*, *une professeur au l'université*) and then elaborates it through an appositive reformulation (*a big sausage*, *une professeur de littérature anglophone*), so that the information is spread across several short, juxtaposed phrases. In (1b) and (2b), the same content is integrated into a single noun phrase through modification (*big*, *de littérature anglophone*) and embedding (*à l'université*). In Universal Dependencies terms, the first strategy surfaces as `appos` relations between flat phrases, the second as `amod`, `compound` or `nmod` dependents within a deeper phrase. Such reformulations also occur in native speech, so the contrast is one of degree rather than kind; it is precisely this difference in distribution that our metrics are designed to quantify.

Measuring such contrasts automatically requires a syntactic representation that holds across languages, and UD provides one (de Marneffe et al., 2021), although an imperfect one. Several non-canonical nominal phenomena still lack a standard treatment, in English as elsewhere (Schneider and Zeldes, 2021), and every parse-based measure, dependency distance included (Liu, 2008), inherits these gaps. This is why we treat extraction as a step to be evaluated in its own right, not as a by-product of parsing.

3 Data

3.1 The Corpus InterLangue (CIL) as a testbed

To evaluate the tool, we use the CIL,[2] a comparable spoken corpus of learner French and learner English developed by the LIDILE research unit at Université Rennes 2 (Gaillat et al., 2021). We chose it because both of its sub-corpora follow a single protocol, in which learners of French and learners of English performed the same tasks. The corpus includes 292 recordings (3,014 minutes of speech) from 146 learners, mostly at CEFR B1–B2 levels. The protocol comprises three tasks, namely a semi-guided interview, a read-aloud task and a written task in which learners imagine the continuation of a dialogue. The transcripts are time-aligned in ELAN (Wittenburg et al., 2006), with separate Turn Units (TUs) for interviewers and learners, ensuring clean separation of learner productions. One of us took part in curating the 2025 CIL data (character-encoding review, re-annotation of the transcriptions, anonymisation and pseudonymisation of the recordings). Table 1 gives the composition of the corpus and of the subsets analysed for RQ1 and RQ2, which the tool handles in two different modes (Section 4.1).

	CIL-FFL (L2 French)	CIL-EFL (L2 English)	Total
Design			
Speakers' L1	Ar, Cmn, En, Es	French	
Proficiency	B1–B2	A2–B2[a]	
Tasks	Semi-guided interview (∼20 min), read-aloud, written continuation of a dialogue		
Language recorded	French	English	
Size			
Speakers	128	18	146
Recordings	250	36	292
Transcripts	128	18	146
Duration (min)	2,702	316	3,014
Analysed for RQ1 (Mode A)			
Texts	17	17[b]	34
NPs extracted	11,927	5,019	16,946
Analysed for RQ2 (Mode B)			
Texts	—	18	
Words	—	28,679	
NPs extracted	—	6,823	

Table 1: Composition of the Corpus InterLangue (CIL) and of the subsets analysed for each research question. [a]Self-assessed; 15 of the 18 CIL-EFL learners are at B1 or B2. [b]Mode A draws on 17 CIL-EFL transcripts, Mode B on all 18.

3.2 ICE-GB as a native baseline

The native baseline is the spoken component of ICE-GB, the British component of the International Corpus of English, compiled and fully parsed by the Survey of English Usage at University College London (Greenbaum, 1996; Nelson et al., 2002). It holds 637,562 words of spoken English in 300 texts (Table 2), of which we kept only the 99 texts of the private-dialogue category (S1A), unscripted conversation, which is the register closest to an interview; adding broadcast talk or scripted monologue would have mixed part of the speech–writing contrast, along which nominal complexity varies most (Biber and Gray, 2011), into the learner–native one.

4 Design and Implementation of the Tool

4.1 Overview of the pipeline

The tool extracts and measures noun phrase complexity in learner productions. It is a modular pipeline. Raw transcripts are cleaned, parsed with UD-based models and turned into nominal con-

[2] https://corpusinterlangue.nakala.fr/

ICE-GB, spoken component	
Design	
Speakers	Adult L1 British English
Texts dated	1990–1992
Release	2 (2006), ICECUP 3.1
Annotation	Fully parsed treebank
Size	
Texts	300
dialogue	180 (100 private, 80 public)
monologue	120 (70 unscripted, 50 scripted)
Words	637,562
Duration	>70 h
Parse trees	59,640
Analysed here (S1A)	
Texts	99
Words	165,919
NPs extracted	44,025

Table 2: The spoken component of ICE-GB, used as the native baseline.

structions, over which the metrics are computed. Any language with a UD model can in principle go through it.

The pipeline operates in two modes, each tailored to specific comparison tasks. Mode A is designed for cross-linguistic comparisons, such as learner French vs. learner English (Figure 1). It relies on UDPipe 2 (Straka, 2018), queried through its REST service, with models trained on Universal Dependencies 2.17 (Straka, 2025; de Marneffe et al., 2021). To match the spoken nature of our data, we selected models trained on spoken treebanks, namely `french-rhapsodie-ud-2.17`, trained on the Rhapsodie treebank of spoken French (Lacheret et al., 2014; Kahane et al., 2021), and `english-childes-ud-2.17`, trained on UD English-CHILDES, a treebank of spontaneous child and caregiver speech (Liu and Prud'hommeaux, 2023; Yang et al., 2025). Mode B compares learners and native speakers of the same language, here learner and native English. It runs a heavier parser, spaCy's transformer-based `en_core_web_trf` 3.8.0 (Honnibal et al., 2020), built on RoBERTa (Liu et al., 2019), which the denser noun phrases of native spoken English call for.

Both modes share the same six steps, namely input, cleaning, parsing, CoNLL-U structuring, nominal-construction extraction, and mapping to a canonical model. Comparability rests on the last of these. The metric code never reads a parser's raw output, only the canonical records, so that changing

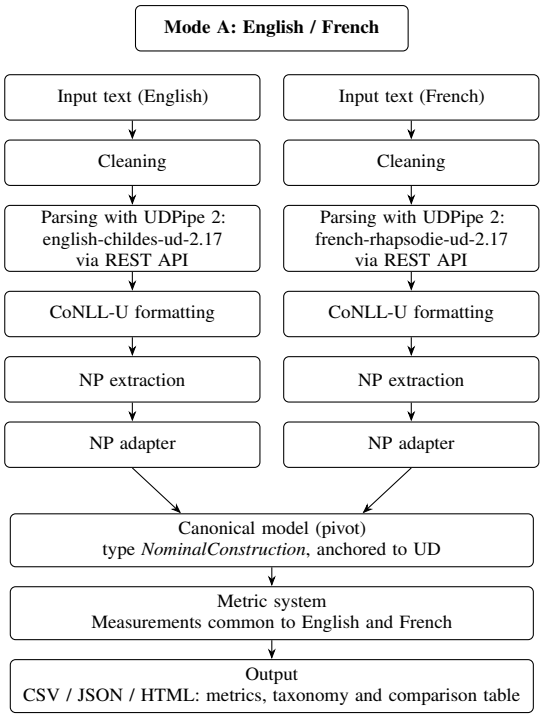


Figure 1: The Mode A pipeline. Each language is cleaned, parsed with UDPipe 2 and converted to CoNLL-U separately; language-specific adapters then map the extracted noun phrases onto a shared canonical model anchored to Universal Dependencies, from which the same metrics are computed for both languages.

the parser or the language leaves the metric definitions untouched. CoNLL-U, the plain-text format of UD treebanks,[3] stores one token per line in ten tab-separated fields (index, form, lemma, universal part of speech, language-specific tag, morphological features, head, dependency relation, enhanced dependencies and a free *MISC* field); every parse reaches the extraction step in that form.

4.2 Nominal-construction extraction in Mode B

With English on both sides, Mode B is free to use measures that have no French counterpart and to rely on spaCy's transformer-based English parser (Section 4.1). That does not make the parse trustworthy on spoken learner data, for the reasons given in Sections 1.1 and 2.2. Mode B therefore never reads noun phrases straight off the tree. It extracts them through a constrained procedure and checks the result against a gold standard (Figure 2; Section 5).

Each input is parsed by spaCy and written as en-

[3] https://universaldependencies.org/format.html

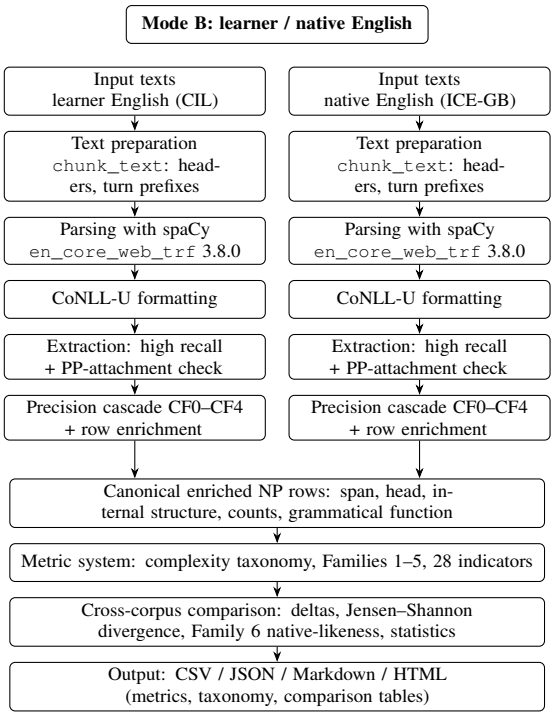


Figure 2: The Mode B pipeline. Learner (CIL) and native (ICE-GB) texts are processed by the same chain of text preparation, parsing with spaCy, CoNLL-U formatting, high-recall extraction and a precision cascade. The resulting noun phrases are stored as canonical records, from which the same metrics and cross-corpus comparisons are computed.

riched CoNLL-U, where spaCy relation labels that have a UD counterpart are mapped onto it in the *MISC* field (*dobj→obj*, *relcl→acl:relcl*). ICE-GB was a special case. We benchmarked spaCy's part-of-speech tags against the hand-corrected ICE-GB tags and found only 85.8% agreement on written and 81.1% on spoken samples, too low to rely on; we therefore kept the ICE-GB tokens and tags and let the parser supply the dependency layer only.

Extraction proper takes three steps. A generator first proposes a span around every nominal head, pronouns and verbless nominal predicates included; at this stage recall comes first. Our error analysis then pointed to one dominant fault, namely right boundaries stretched over a following prepositional phrase. A rule based on the dependency parse now checks whether a prepositional phrase attaches to the head or higher up, and it removed 71% of that error class. Some cases still escape it. In *I'm born in Fougères in France* (CIL, L1), the extractor returns *Fougères in France*, where the annotators marked *Fougères* and *France* as two constructions. Finally, a precision cascade (filters CF0–CF4) drops empty, determiner-only, artefact, idiom and contained spans and collapses duplicate coordinations. Each surviving candidate becomes one canonical record per maximal noun phrase (head, span, modifiers, embedded layers, coordination, grammatical function), and every metric is computed from these records. One rule governed development throughout. A change to the extractor stayed only if it did not lower the strict F1 on the gold standard.

4.3 Metrics for noun phrase complexity

The metrics are organized into eleven families, viz. family A, frequency and density; family B, phrasal modification and compression; family C, clausal modification and elaboration; family D, modification scope and syntactic distribution; family E, recursion, embedding, and depth; family F, nominalization and density of nominalized heads; family G, structural and modifier-type diversity; family H, lexical diversity of heads; family I, head type and referential density; family J, the genitive and possessive system; and family K, coordination. They are shared by the two modes, with 18 metrics in Mode A and 17 in Mode B. Table 3 presents the first four with examples. The full metric table[4] gives, for each of the 35 metrics, its formula, unit, comparability between French and English, and counterpart in the other mode.

Family A counts nominal material, per 1,000 words or per finite clause. On its own it says little; it is the baseline against which the other families are read. In Mode A it also breaks modification down by type (epithets, noun premodifiers, postnominal prepositional phrases, relatives, complement clauses, appositions), because these do not occupy the same positions in French and English.

Families B and C carry the distinction drawn by Biber and Gray (2016). Phrasal modifiers (adjectives, nouns, genitives, prepositional phrases, appositives) compress information inside the noun phrase; clausal ones (relative, complement and non-finite clauses) elaborate it. Family B adds informational compression, the number of modifier-borne propositions per finite clause. Read together, the two families ask whether a speaker packs content into the phrase or unfolds it.

Family D looks at breadth and place, that is, how

[4] https://gitlab.huma-num.fr/lidile/constructions-nominales/-/blob/b522959f914773585d810aeca1e7951e21050932/Metrics_Learners_EN_and_FR_FAMILIES_EN__.ods

Family	What it captures	Example metrics	Examples
A · Frequency and density	Normalised frequencies of nominal material, before any complexity	NP density; NPs per finite clause; complement-clause density	*the fact that there is many things to do* (CIL, L1) and *the fact that she's really good* (ICE-GB S1A-005) contain noun complement clauses, counted per 1,000 words; FR *le fait qu'il est parti*. *the accuracy of my results* (ICE-GB S1A-008) counts as one NP.
B · Phrasal modification and compression	Information packed inside the phrase by adjectives, nouns, genitives, PPs and appositives	Phrasal modifiers per NP; informational compression	Adjective and PP postmodifier in *the best food in the world* (CIL, L13). Noun premodifier and of-phrase in *the response curves of the tape* (ICE-GB S1A-008). FR *une décision sur le budget*.
C · Clausal modification and elaboration	Propositional elaboration of the NP by relative, complement and non-finite clauses	Clausal modifiers per NP	Relative clauses in *someone who had an Airbnb for us* (CIL, L7) and *the one that Philip's just done* (ICE-GB S1A-008). FR *l'homme qui est arrivé*.
D · Modification scope and distribution	How many NPs are modified, and where in the sentence complexity sits	Share of modified NPs; NP complexity by function; subject-NP complexity share	Simple subject *my name* (CIL). Complex NP outside the subject slot in *a small town in the South of Brittany* (CIL, L16). Complex, postmodified NP in *the biggest interest in my life* (ICE-GB S1A-003). Heavy subject in FR *[le rapport sur la crise] nous a surpris*.

Table 3: The first four of the eleven metric families (A–D), with example metrics from both modes. English examples are authentic (CIL learner number or ICE-GB file); French examples are constructed illustrations. Families E–K are described in the text and in the full metric table.

many noun phrases carry any modification at all and where the complex ones end up in the clause. The share of complex subjects is the one we watch most closely.

The remaining families appear in the full table. Family E covers recursion, embedding and depth through recursive depth and two parse-based indicators, mean dependency distance (Liu, 2008) and mean tree depth within the noun phrase. The two do not always agree. Depth grows when modifiers nest inside modifiers, distance when material is strung out linearly, as in coordination, and the disagreement turns out to be informative (Section 6.3). F measures nominalisation; G structural and modifier-type diversity, through a moving-average type–token ratio (Covington and McFall, 2010) over noun phrase patterns and the Shannon entropy (Shannon, 1948) of modifier types; H the lexical diversity of head nouns; I the share of lexical rather than pronominal heads; J the 's- and of-genitives; K coordination inside the noun phrase. Mode B computes 22 scalar indicators and 6 distributional profiles in all; the 17 metrics listed for it in the table are those that pass the source-validation and coverage checks of Section 5.

5 Evaluation Methodology

Before any result can be read, two questions must be settled. Is the extraction right, and are the group differences real?

To answer the first question, we built manual gold standards. For Mode A, a single annotator built a gold standard of 200 nominal constructions per language (French and English), so that the tool's output on the same sample could be compared with a human analysis. For Mode B, the gold standard contains 504 span types, 239 from CIL learner English and 265 from ICE-GB spoken English. Mode B annotation followed a written specification, with one span per maximal construction, embedded layers recorded inside it, and explicit rulings on free relatives, idiomatic prepositional objects and disfluent repairs. The specification was revised once, after an audit of the first annotation rounds showed where two careful annotators could legitimately diverge. Two annotators then worked independently; before adjudication, their span-level F1 agreement was .992 on the learner slice and 1.000 on the native one. The Mode B gold standard was then adjudicated by a human annotator and an AI assistant (Claude Opus 4.7, Anthropic) to ensure consistency; the Mode A gold standard, which comes from one annotator, was not adjudicated.

We report precision, recall and F1 under strict matching, in which an extracted span counts as correct only if it matches a gold span exactly. In Mode B, scoring is type-level within each utterance. A span that finds the right construction but places a boundary differently therefore counts as an error; such boundary errors are analysed separately in Section 6.1.

Group differences are tested with the Mann–Whitney U test (Mann and Whitney, 1947), which suits independent samples of unequal size and does not assume normality. Significance alone is not enough, so each test comes with Cliff's delta (Cliff, 1993), read with the thresholds of Romano et al. (2006), namely .147, .33 and .474 for small, medium and large effects. With 22 simultaneous tests, all p-values are Holm–Bonferroni corrected (Holm, 1979). The unit of analysis is the text. In Mode B, distributional profiles are also compared with the Jensen–Shannon divergence (Lin, 1991). A second-stage profile then synthesises the scalar metrics. Of its 23 candidate metrics (the 22 tested ones plus the share of core-argument noun phrases, which has no test of its own), it keeps 17. It drops four whose per-text source values do not match the tested values, one whose coverage is insufficient and the untested one. It then clusters the retained metrics correlated at $|r| \geq .80$, orients each standardised metric in the direction of the observed contrast, and averages them by family. Because 18 learner texts face 99 native ones, a balanced bootstrap (Efron and Tibshirani, 1993) with 5,000 resamples sets the learner texts against equally sized native subsamples.

6 Results

6.1 Extraction accuracy

For Mode A, the tool's output was compared with the annotator's analysis of the same 200 nominal constructions per language. Strict F1 reaches .950 for learner French and .860 for learner English (Table 4). Figure 3 gives the corresponding confusion matrix and explains why precision, recall and F1 coincide in this balanced design.

In Mode B, strict F1 reaches .954 on learner English and .787 on native English, against 504 gold span types (Table 5). The gap has two sources. Native noun phrases are denser and deeper, and ICE-GB is not segmented like CIL. Most residual errors, however, concern the boundary of a construction that was correctly detected. On the full

Slice	Gold	TP	FP	FN	P	R
Learner EN	200	172	28	28	.860	.860
Learner FR	200	190	10	10	.950	.950
Strict F1 = .860 (learner EN), .950 (learner FR)						

Table 4: Mode A extraction accuracy against the gold standard (strict, type-level matching). Gold = gold nominal constructions; TP, FP, FN = true positives, false positives and missed gold constructions.

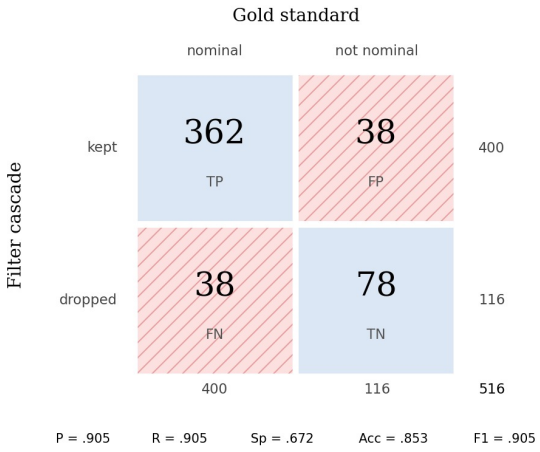


Figure 3: Confusion matrix of the Mode A precision cascade (pooled CIL learner English and learner French gold slices), computed over the 516 candidate span types on which the cascade decides. Because the extractor is constrained to return as many candidates as there are gold constructions (400), false positives and false negatives are equal in number, which yields $P = R = F1$.

gold sample, which also includes an ICE-GB written slice, 117 of the 161 false positives and of the 161 misses form boundary-overlap pairs, leaving 44 residual false positives and 44 residual misses. The extractor also errs on the short side more often than on the long side. In 72 of these pairs the gold span contains the extracted one, in 44 the reverse.

Figure 4 opens up the Mode B scores. Extraction has no natural negative class, so the matrix covers only the 772 span types on which the cascade decides. The cascade keeps 432 genuine noun phrases and 61 spurious ones, and discards 23 genuine and 256 spurious. Its errors lean one way. It rarely throws out a real noun phrase (recall = .949) but lets roughly one spurious candidate in five through (specificity = .808). The 72 end-to-end misses thus come from two places, namely 23 noun phrases the cascade wrongly discarded and 49 that were never proposed at all, a group in which recall can still be

Slice	Gold	TP	FP	FN	P	R
CIL learner	239	227	10	12	.958	.950
ICE-GB spoken	265	205	51	60	.801	.774
Strict F1 = .954 (CIL), .787 (ICE-GB spoken)						

Table 5: Mode B extraction accuracy against the adjudicated gold standard (strict, type-level matching). Gold = gold span types; TP, FP, FN = true positives, false positives and missed gold spans.

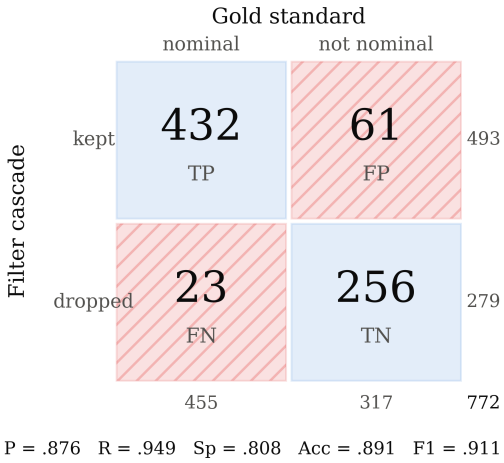


Figure 4: Confusion matrix of the Mode B precision cascade (CIL learner and ICE-GB spoken gold slices), computed over the 772 candidate span types on which the cascade decides. The 49 gold constructions never generated as candidates fall outside this population.

gained.

6.2 Learner French vs. learner English

The main finding for RQ1 is that, across phrasal measures, learner English is consistently more complex than learner French. Phrasal modifiers per noun phrase, for instance, were higher in learner English (0.61) than in learner French (0.41; $\delta = -1.00$, $p < .001$), and the share of modified noun phrases was higher as well (50% vs. 39%; $\delta = -.99$, $p < .001$). Phrasal depth, modifier-type entropy and structural diversity also showed significantly higher values for learner English (Figure 5).

However, the tool also revealed that learner French compensates for its lower phrasal complexity by using more appositive and looser structures. The share of appositive structures, in particular, was higher in learner French ($\delta = +.56$, $p < .01$).

Clause-level density measures, such as the number of noun phrases per finite clause, were largely comparable between the two groups, indicating

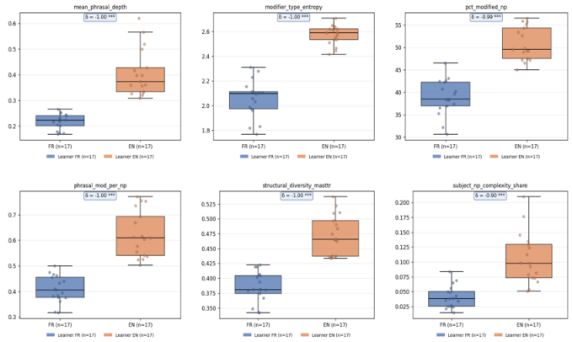


Figure 5: Per-text distributions of the six metrics that separate the two learner corpora most sharply (Learner FR vs. Learner EN, $n = 17$ texts each). Boxes show the median and interquartile range, whiskers the range of the data, and points the individual texts. The top row shows mean phrasal depth, modifier-type entropy, and the percentage of modified NPs; the bottom row shows phrasal modifiers per NP, structural diversity (MATTR), and the share of complexity carried by subject NPs. On all six metrics, the EN texts score higher than the FR texts. Effect sizes are given as Cliff's δ (negative values indicate FR < EN), ranging from -0.90 to -1.00, meaning the two distributions barely overlap or do not overlap at all. All contrasts are Holm-corrected and significant at $p < .001$.

that the differences lie primarily in how complexity is distributed rather than in the overall amount of nominal material produced.

6.3 Learner English vs. native English

Does the tool separate learner from native English? It does, and in a specific way. Natives pack complexity into the noun phrase, whereas learners spread it over longer, flatter structures. Table 6 gives the full panel of per-text means; all 18 CIL-EFL interviews and 99 ICE-GB private dialogues were retained, and the imbalance is handled as described in Section 5.

The native advantage lies in compression. Natives use more phrasal modifiers per noun phrase (0.091 vs. 0.060; $\delta = -.65$, $p < .001$), and their of-genitive share is three times the learners' (1.65% vs. 0.55%; $\delta = -.71$, $p < .001$), the largest single effect in the panel. Informational compression (0.185 vs. 0.105; $\delta = -.67$, $p < .001$), recursive depth and tree depth (both $\delta = -.54$, $p < .01$) point the same way, as do clausal modifiers per NP ($\delta = -.42$, $p = .044$) and NP density ($\delta = -.46$, $p = .028$).

Figure 6 shows the per-text distributions. The corpora overlap on every metric taken alone; they

ID	Metric	L	N	δ	p_H
A.1	NPs per 1,000 words	236.5	249.2	**−.46**	.028
A.2	NPs per finite clause	2.95	2.51	**+.66**	<.001
B.2	Phrasal modifiers/NP	.060	.091	**−.65**	<.001
B.3	Inform. compression	.105	.185	**−.67**	<.001
C.2	Clausal modifiers/NP	.002	.004	**−.42**	.044
D.3	Complex subjects (%)	8.16	7.01	+.08	1
E.3	Recursive depth	.040	.056	**−.54**	.005
E.4	Tree depth	.195	.244	**−.54**	.005
E.5	Dependency distance	.359	.274	**+.62**	<.001
F.2	Nominalisations/NP	.033	.035	+.06	1
G.2	Modifier entropy (bits)	2.45	2.42	+.16	1
G.4	Pattern MASTTR	.165	.164	−.07	1
H.2	Head-noun MATTR	.674	.648	+.39	.082
I.1	Lexical NPs (%)	24.0	24.5	−.12	1
J.1	's-genitive (%)	1.22	1.60	−.29	.362
J.2	of-genitive (%)	0.55	1.65	**−.71**	<.001
K.1	Coordinate NPs (%)	2.80	1.33	**+.54**	.004
Not retained in the metric table					
—	NP length (words)†	1.91	1.70	**+.53**	.005
—	Premodifiers/NP†	.154	.182	**−.42**	.047
—	Postmodifiers/NP†	.045	.058	**−.46**	.028
—	Complex NPs (%)†	15.5	17.7	−.37	.110
—	Phrasal/clausal‡	21.0	24.3	−.20	1

Table 6: Per-text means of the 22 scalar metrics in learner (L, 18 CIL texts) and native (N, 99 ICE-GB texts) English. IDs follow the families of the full metric table (A–K). δ = Cliff's delta (positive values, learner higher); p_H = Holm-corrected p over the 22 tests; bold = significant. †Failed the second-stage source validation. ‡Computable for 7 learner and 77 native texts only. The full metric table, with standard deviations, is available online (https://qrco.de/bh0mtx).

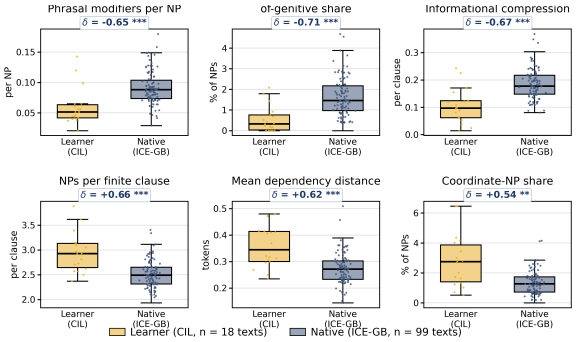


Figure 6: Per-text distributions of the six metrics that separate the two corpora most sharply. Boxes show the median and interquartile range, points the individual texts. The top row holds the three native-higher metrics of nominal compression, the bottom row the three learner-higher metrics of nominal spread. All contrasts are Holm-corrected and significant at $p < .001$, except Coordinate-NP share ($p < .01$).

differ by a shift in central tendency that recurs from one measure to the next, and no single measure provides a clean threshold.

Learners, by contrast, distribute their nominal material over more, and less compact, noun phrases. They produce more noun phrases per finite clause (2.95 vs. 2.51; $\delta = +.66$, $p < .001$), with longer dependency distances inside them (0.359 vs. 0.274; $\delta = +.62$, $p < .001$) and twice as much internal coordination (2.80% vs. 1.33%; $\delta = +.54$, $p < .01$). Their noun phrases are also longer, at 1.91 content words against 1.70 ($\delta = +.53$, $p < .01$). We read this last figure with care, since mean length is one of the four measures flagged by the source validation, although its direction holds across source layers. Length does not equal complexity here. Learners lengthen a noun phrase by adding a conjunct, as in *my mother and my father* (CIL, L2); natives lengthen it by embedding, as in *the response curves of the tape* (ICE-GB S1A-008).

Diversity and external distribution do not separate the groups. No diversity metric survives correction, nor does the share of complex subjects (all $|\delta| \leq .39$, $p \geq .082$), and the distributional profiles are close, with Jensen–Shannon divergences between .001 (coordination, embedding depth) and .024 (head word class). We would not call this equivalence, since with 18 learner texts a null result is inconclusive. On these data, though, learners draw on the same inventory of nominal structures as natives, in similar proportions, and differ in how densely they combine them.

The second-stage profile says the same thing, more strongly. The core profile separates the corpora with $\delta = .87$ ($p < .001$), more than any single metric ($|\delta| = .71$ at best), and does so in every balanced resample (median $\delta = .87$, 95% interval $[.79, .93]$). Grouped by the dimensions the profile uses, the separation sits in the parse-based indicators (.97), frequency (.69) and NP-internal structure (.68), all at $p < .001$, whereas diversity (.16) and external distribution (.08) contribute little. One caveat applies. The profile is oriented by the contrast it measures, so it synthesises the same metrics without validating them independently.

7 Discussion

The two comparisons do not ask the same question, but their answers fit together. Learner English carries more phrasal modification than learner French and less than native English, and in each comparison the group with less modification compensates

elsewhere (Sections 6.2 and 6.3). In both cases the difference lies less in how much nominal material speakers produce than in where they put the structure. The spoken data thus echo, on a small scale, what has been described for writing, where nouns and prepositional phrases as modifiers mark more advanced production (Biber et al., 2011; Parkinson and Musgrave, 2014).

Regarding the study's depth, many other aspects could have been tackled, namely a native English vs. native French comparison, that account for interlinguistic differences that are not learner-specific. Or, a monolingual native French vs. learner French study, parallel to the English substudy that shed light to the interlanguage development process in both languages i.e. English and French. But also, one or more similar substudies on other less closely related languages. However, it is worth mentioning that including all these aspects of study involves time and dedicated corpora building to support it. Given that we already embarked in an analysis tool development, we chose not to engage in compiling all the possible analysis in one work. They will be discussed accordingly in future works.

7.1 Comparison with TAASSC

TAASSC (Kyle, 2016), whose phrasal indices were introduced in Section 2.1, is the closest existing point of reference for our tool, and much of our work builds on the questions it made measurable. Table 7 sets the two tools side by side, to situate them without ranking them.

TAASSC in brief. TAASSC groups its indices into four families (Kyle, 2016). They are the fourteen large-grained indices of the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010), such as mean length of T-unit (MLT), dependent clauses per clause (DC/C) and complex nominals per clause (CN/C); fine-grained clausal indices, which count the dependents of each clause by type; the fine-grained phrasal indices described in Section 2.1; and usage-based indices of syntactic sophistication, which score main verbs, verb argument constructions (VACs) and verb–VAC combinations by their frequency and strength of association in the written sections of the Corpus of Contemporary American English (COCA; Davies, 2009) (see Kyle and Crossley, 2017; Kyle et al., 2021). The original release relied on Stanford CoreNLP and Stanford typed dependencies (de Marneffe and Manning, 2008); Kyle et al. (2021) identify VACs with the Stanford neural dependency parser (Chen and Manning, 2014), and version 2.0 of the tool parses with spaCy, as our Mode B does.

	TAASSC	Mode A	Mode B
Data	Written English essays	Spoken learner FR and EN	Spoken learner and native EN
Parser	Stanford; spaCy (v2.0)	UDPipe 2, spoken UD models	spaCy `trf`; ICE-GB tags
Unit	Dependents of NP heads	Nominal construction, gold-evaluated	Maximal NP span, gold-evaluated
Indices	Means, SDs by NP × dependent type	18 metrics (A–K)	22 scalar, 6 distrib.; 17 validated
Embedding	—	Phrasal depth	Depth, dependency distance
Frame	Dependent types	Phrasal vs. clausal; UD function	Phrasal vs. clausal; *'s* vs. *of*; compression
French	—	Adj. position, *de*/*à*/*en*, relatives, determiner	—
Diversity	SDs of counts	Entropy	MATTR, entropy, JSD
Statistics	Indices only	MWU, Cliff's δ, Holm	Same + JSD, profile
Beyond scope	—	Clausal; sophistication	Clausal; sophistication

Table 7: TAASSC (Kyle, 2016; Kyle and Crossley, 2018) compared with the two modes of our tool.

Common ground. Both tools count the modifiers of nominal heads and relate them to the function of the phrase. Our phrasal modifiers per noun phrase are conceptually close to TAASSC's dependents per nominal, and our share of complex subjects to its indices for nominal subjects. Both also report their values per text. TAASSC returns one row of indices per text, and the text is our unit of analysis as well (Section 5), which makes a text-by-text comparison of the two tools possible. The correspondences between the two sets of measures nevertheless remain conceptual, since the definition of the noun phrase, the parser, the dependency scheme and the normalisation differ. Our tool's contribution lies mainly in these differences, which we take in turn.

(1) French and English on a common basis. TAASSC is designed for English, and its sophistication indices rely on an English reference corpus. Our tool, anchored in Universal Dependencies (de Marneffe et al., 2021), describes a second lan-

guage with the same categories as the first. In Mode A, the French and English parses produced by UDPipe 2 (Straka, 2018) meet in the canonical model of Section 4.1, where modifiers are coded by their UD relation whatever their position. The French postnominal adjective and the English prenominal adjective are thus both counted under *amod*. Mode A also adds measures for properties of the French noun phrase, such as adjective position, prepositional phrases in *de*, *à* and *en*, the type of relative pronoun and the omission of the determiner. The comparison of learner French with learner English (Section 6.2) depends on this design, which is the main extension our tool offers.

(2) The noun phrase as a delimited unit. TAASSC counts the dependents that a parser attaches to a head, without delimiting the noun phrase itself. Our tool delimits each noun phrase as a unit and evaluates that unit, span by span, against a gold standard (Section 5; Tables 4 and 5), so that a reader can check what was counted. Reporting accuracy on the unit of analysis follows a concern that the TAASSC studies themselves raise. Kyle et al. (2021) report that 97.9% of main verbs but 81.25% of verb–VAC combinations were correctly identified in their learner corpus, and note that such errors carry over into any downstream analysis.

(3) Embedding and dependency distance. Treating the noun phrase as a delimited unit also makes its internal structure measurable, through the depth and distance indicators of family E (Section 4.3); Liu (2008) relates the second of these to processing difficulty. A count of dependents does not in itself distinguish dependents attached to the same head from dependents embedded in one another. In our data, depth and dependency distance separate learners and natives more sharply than the other dimensions (Section 6.3), which suggests that this aspect of the noun phrase is worth measuring alongside dependent counts. Because spans depend on segmentation, the dependency-distance half of this result carries the caveat set out in the Limitations section.

(4) A developmental reading. The phrasal indices of TAASSC describe how many dependents of each type occur. Families B and C operationalise instead the distinction between phrasal and clausal modification that Biber et al. (2011) place along a hypothesised developmental sequence, and that Biber and Gray (2016) describe as compression against elaboration. The longitudinal study of Kyle et al. (2021) gives one reason why this reading may be useful. Over two years, their adolescent learners of English grew in DC/C and MLT but not in CN/C, and the relation between time and MLT ($r = .443$) became negligible once DC/C was controlled for ($r = -.046$); the authors interpret this in light of Biber et al. (2011), with dependent clauses developing first and noun phrase features possibly later. A measure that sets phrasal against clausal modification may help to follow that shift, which an overall count of complex nominals does not isolate. Whether it does so over time is a question that longitudinal spoken data would need to answer.

(5) Functional contrasts. TAASSC separates nominal dependents by type. Our tool adds functional contrasts, namely phrasal against clausal modification, the *'s*-genitive against the *of*-genitive (family J), and informational compression (family B). In our data these contrasts produce the strongest effects (Table 6), and they are robust to the segmentation problem discussed in the Limitations section.

(6) Diversity. In TAASSC, variability is expressed through the standard deviation of dependent counts. Our tool adds measures of diversity proper (families G and H, Section 4.3) and, in Mode B, the Jensen–Shannon divergence between distributional profiles. A standard deviation describes how much counts vary across nominals, whereas these measures describe how varied the repertoire of structures is and how close two distributions are.

(7) Transcribed learner speech. TAASSC was developed and validated on written texts, mainly TOEFL essays (Kyle and Crossley, 2017, 2018) and learner essays (Kyle et al., 2021). Our tool was developed for transcribed speech from learners and natives. Mode A relies on the spoken-treebank models presented in Section 4.1, Mode B on the hand-corrected ICE-GB tagging (Section 4.2), and both modes filter disfluencies and interactional forms and handle mass and institutionalised bare nouns. Earlier findings suggest that such an adaptation is worthwhile. With the companion tool TAALES, Kyle and Crossley (2016) found that the lexical sophistication features related to writing scores differed between independent and source-based essays, so that indices do not necessarily behave alike across tasks, and complexity itself is

distributed differently in conversation and in writing (Section 2.1).

(8) Normalisation. TAASSC normalises mainly per clause or per nominal. Our tool reports measures per 1,000 words, per finite clause and per noun phrase, so that the unit can be chosen according to the question, whether it concerns how much nominal material a speaker produces (family A) or how much each phrase or clause carries (families B to D).

(9) Statistical analysis. TAASSC exports indices, and the statistical modelling is carried out separately, as in the mixed-effects analyses of Kyle et al. (2021). Our tool builds the group comparison of Section 5 into the pipeline, in both modes. Since the unit of analysis is the text, values for each speaker remain available next to group results, in line with the caution, recalled by Kyle et al. (2021) after Lowie and Verspoor (2019), that group means may not represent individual trajectories.

What TAASSC covers and our tool does not. TAASSC covers several dimensions that our tool does not attempt. It includes clausal complexity, through the L2SCA indices and its fine-grained clausal counts, whereas we measure only compression per clause. It includes usage-based sophistication, its largest family; in Kyle et al. (2021), main verb frequency and DC/C together predicted the time point of an essay ($R^2_{\text{adj}} = .583$). Such indices presuppose a suitable reference corpus, which a comparison of learner speech in two languages would have to supply on both sides. TAASSC also crosses each nominal function with each dependent type; our family D reports complexity by function and the share of complex subjects, but not by dependent type. The two tools are therefore best seen as complementary. TAASSC offers broad coverage of written English, ours a more detailed description of the noun phrase, comparable across French and English and adapted to learner speech.

7.2 Implications for SLA and CALL research

For SLA and CALL research, the results argue for phrasal measures over global ones when the data are spoken. The noun phrase is sensitive to developmental differences, and it can now be measured automatically. Because it rests on Universal Dependencies, the tool also serves work on cross-linguistic influence between L1s, where comparable measurement across languages has long been a sticking point. The learner results also warn against reading complexity as a single scale. Learners build more noun phrases per clause, and coordinate them more often, whereas natives build fewer but more heavily modified ones, so that a longer noun phrase is not necessarily a more complex one.

For CALL systems, the metrics could be used for two purposes. First, they could be used as features in proficiency prediction models. Second, depending on their statistical significance, they could be exploited to tailor specific feedback messages giving details on what type of nominal complexity (if any) learners employ. Such fine-grained, structure-specific features would contribute to providing actionable language learning analytics. In class, the tool could also support data-driven learning, with learners comparing their own noun phrases to native ones. The tool is open, so these claims can be checked, extended or overturned by others.

8 Conclusion and Future Work

This paper has presented an open, modular tool that measures noun phrase complexity in learner speech through Universal Dependencies, together with an extraction procedure validated against gold standards, with strict F1 between .79 and .95 depending on the corpus. The empirical picture that emerges does not point to missing structures. Learners draw on the native inventory of nominal structures but assemble it into flatter, longer and more coordinated noun phrases, which carry less modifier-borne information per clause than native ones (Table 6, B.3). Because every language-specific step happens before the canonical model, a single metric code serves both the cross-linguistic and the learner–native comparison.

Resegmenting both corpora into AS-units (Analysis of Speech units), that is, units made of an independent clause or sub-clausal unit together with its subordinate clauses (Foster et al., 2000), is a natural next step and, in our view, the least costly improvement available. Adapting the parser to spoken learner English, for instance with the treebank of Kyle et al. (2022), would reduce the parsing errors that the cascade only partly absorbs. Calibration against other measurements would also help, in two ways. Since learner English is measured by both modes, Mode A and Mode B could be run on the same CIL-EFL texts to see which measures agree. The English branch could likewise be compared with TAASSC. Running both tools on

the same transcripts and correlating the measures they share (Table 7), once normalisation is aligned, would provide a test of convergent validity, and any divergence could be traced to the differences discussed in Section 7.1.

Limitations

The samples are small, with 17 speakers per language for RQ1 and 18 learners against 99 native texts for RQ2, and more proficiency levels and more first languages are needed before any finding is generalised. Mode A needs inter-annotator agreement, since its gold standard comes from a single annotator, and both modes need firmer segmentation guidelines. The analysis covers semi-guided interviews and read-aloud tasks in French and English only. Other registers, such as academic speech and spontaneous conversation, and other language pairs, such as Spanish or German, remain untested.

Segmentation is the main weakness of Mode B. Neither corpus carries sentence punctuation, so the units are long and uneven, with 53.8 words on average in the learner transcripts and 19.7 in the native ones, and maxima of 817 and 344. Every extracted noun phrase is bounded by its unit, and a unit that runs across a turn drags the right boundary with it. In ICE-GB S1A-008, a relative clause on the pronoun *one* swallows the following utterance and yields a 97-word "noun phrase". In CIL_L3, a construction that should stop at *my gym* runs through a turn marker into the next turn. Such cases are rare. Nominals of 15 words or more make up 0.16% of learner records and 0.25% of native ones. The measures behind our main finding, namely *of*-genitive share, phrasal modifiers and informational compression, do not depend on span length. Mean dependency distance does, which is why we trust it least, and it qualifies the result on depth and dependency distance discussed in Section 7.1. The four boundary-sensitive metrics that failed the source validation are reported with a flag (Table 6).

Ethics Statement

The CIL transcripts analysed here come from the pseudonymised 2025 release of the corpus (Section 3); learners are referred to by codes only, and no personal information is reported. ICE-GB is used as released by the Survey of English Usage (University College London). The use of an AI assistant in the adjudication of the Mode B gold standard is reported in Section 5.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: [Describe the AI tools used and their purpose, or state: "No AI tools were used in preparing this work."] "'

References

Douglas Biber and Bethany Gray. 2011. Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics*, 15(2):223–250.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge University Press, Cambridge.

Douglas Biber, Bethany Gray, Tove Larsson, and Shelley Staples. 2025. Grammatical analysis is required to describe grammatical (and "syntactic") complexity: A commentary on "complexity and difficulty in second language acquisition: A theoretical and methodological overview". *Language Learning*, 75(2):575–581.

Douglas Biber, Bethany Gray, and Kornwipa Poonpon. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1):5–35.

Douglas Biber, Bethany Gray, Shelley Staples, and Jesse Egbert. 2020. Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, 46:100869.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-UD: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7145–7151, Marseille, France. European Language Resources Association.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2025. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, 75(2):533–574.

Bram Bulté and Alex Housen. 2012. *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar.

Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Pauline Foster, Alan Tonkyn, and Gillian Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3):354–375.

Thomas Gaillat, Leonardo Contreras Roa, and Juvénal Attoumbre. 2021. A data repository for the management of dynamic linguistic datasets. In *CLARIN Annual Conference 2021*, Madrid, Spain (online).

Sidney Greenbaum, editor. 1996. *Comparing English Worldwide: The International Corpus of English*. Clarendon Press, Oxford.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Alex Housen and Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University.

Kristopher Kyle, Scott Crossley, and Marjolijn Verspoor. 2021. Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, 43(4):781–812.

Kristopher Kyle and Scott A. Crossley. 2016. The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, 34:12–24.

Kristopher Kyle and Scott A. Crossley. 2017. Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4):513–535.

Kristopher Kyle and Scott A. Crossley. 2018. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2):333–349.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45. Association for Computational Linguistics.

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken French. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zoey Liu and Emily Prud'hommeaux. 2023. Data-driven parsing evaluation for child-parent interactions. *Transactions of the Association for Computational Linguistics*, 11:1734–1753.

Wander M. Lowie and Marjolijn H. Verspoor. 2019. Individual differences and the ergodicity problem. *Language Learning*, 69(S1):184–206.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1):36–62.

Henry B. Mann and Donald R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Gerald Nelson, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. John Benjamins, Amsterdam.

John M. Norris and Lourdes Ortega. 2009. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics*, 30(4):555–578.

Jean Parkinson and Jill Musgrave. 2014. Development of noun phrase complexity in the writing of English for academic purposes students. *Journal of English for Academic Purposes*, 14:48–59.

Jeanine Romano, Jeffrey D. Kromrey, Jesse Coraggio, and Jeff Skowronek. 2006. Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? In *Annual Meeting of the Florida Association of Institutional Research*, Cocoa Beach, Florida.

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 160–172. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka. 2025. Universal dependencies 2.17 models for UDPipe 2 (2025-11-25). LINDAT/CLARIAH-CZ digital library, Institute of Formal and Applied Linguistics (ÚFAL), Charles University.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Number 17 in Technical Report. University of Hawai'i, Second Language Teaching and Curriculum Center, Honolulu.

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, and Nathan Schneider. 2025. UD-English-CHILDES: A collected resource of gold and silver Universal Dependencies trees for child language interactions. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 52–58, Ljubljana, Slovenia. Association for Computational Linguistics.

A corpus of written Swedish texts and keystroke data from deaf and hard-of-hearing children and adolescents

Moa Gärdenfors
Department of Linguistics
Stockholm University
moa.gardenfors@ling.su.se

Robert Östling
Department of Linguistics
Stockholm University
robert@ling.su.se

Lisa Rudebeck
Department of Swedish Language and Multilingualism
Stockholm University
lisa.rudebeck@su.se

Mats Wirén
Department of Linguistics
Stockholm University
mats.wiren@ling.su.se

Abstract

Deaf and hard-of-hearing (DHH) children constitute a distinct subgroup of language learners, as they are acquiring a written language with reduced access to the corresponding spoken language. Also, many in the group have a sign language as their first language, thus making their visual-domain L1 another unique feature. We present an open access corpus of the written production of such learners, including 74 texts with associated keylogging data and correction annotation, making this the only such resource available for any language. Specifically, it provides a necessary component for CALL applications for DHH children.

1 Introduction

Deaf and hard-of-hearing (henceforth, DHH) children constitute a heterogeneous group with diverse linguistic backgrounds. Their language experiences vary depending on factors such as degree of hearing loss, age of identification, access to hearing technology, family language practices, educational setting, and much more (Gärdenfors, 2023a). DHH children grow up with diverse language experiences: some primarily use a spoken language, some acquire a sign language as their first language alongside a written language, while others are exposed to both signed and spoken languages. Consequently, DHH children differ considerably in their language development and literacy experiences. Despite this heterogeneity, research has consistently shown that, as a group, DHH children face greater challenges in writing than their hearing peers. At the same time, relatively little is known about their writing development, as research has traditionally focused more on reading than on writing (Mayer and Trezek, 2018).

The work reported here has its basis in the dissertation work of Moa Gärdenfors (2023b) on writing in DHH children, and provides the following novel contributions. First, an additional data collection has been made, and secondly, the entire dataset has been annotated, using tools and methodology from SweLL, a learner corpus for Swedish (Volodina et al., 2019, 2025). The resulting resource is called the Signwrite corpus (Signing Children and Adolescents' Writing Development; see Section 3). It is the first corpus for any language that includes both DHH children's writing and keylogging data and that is open access. As such, it is aimed at providing a necessary component for both CALL applications and research on second-language acquisition (SLA) for DHH children. We provide analyses of the data (Section 4) and conclude with ideas for future work (Section 5). The Signwrite corpus is publicly available under a CC-BY 4.0 license (Gärdenfors et al., 2026).

2 Background

2.1 Keylogging data and written text

Studies on writing development, whether in DHH subjects or in general, increasingly rely on data about the writing *process* in addition to the written *product* (text). The products are typically narratives elicited from subjects using a visual stimulus. For example, Wolbers et al. (2014) collected personal-narrative writing samples from DHH students in grade 6–8 who used American Sign Lan-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 229–243, November 19–20, 2026, University of Tartu Library

guage (ASL) as their L1, and analysed features of this in the students' writing of English.

The writing process is analysed by means of keylogging (also known as keystroke logging), which captures the temporal behavior of the user by recording each keystroke. A frequently used keylogging system is ScriptLog (Johansson et al., 2018). This functions as a simple word processor without a spellchecker and records all keystrokes and mouse movements, timestamping each action to generate a temporal log of the writing process. This enables unobtrusive documentation of a subject's writing behavior while allowing researchers to replay the writing session. It is also possible to extract output for further analyses, including descriptive statistics on writing time, the number of inserted and deleted characters, pause duration and location and revisions. Furthermore, ScriptLog output can be exported to the keystroke-logging software Inputlog (Leijten and Van Waes, 2013) to enable more extensive pause and revision analyses.

Writing-process research began in the late 1970s and early 1980s, initially using methods such as think-aloud protocols and video recordings. With the increasing availability of personal computers, keystroke logging emerged in the 1990s as a more automated and fine-grained method for studying the temporal dynamics of writing, and several such systems were developed during this period (Wengelin and Johansson, 2023).

2.2 Keylogging data in CALL

A number of previous resources aimed at supporting CALL applications and containing keylogging data have been published. The Pro-TEXT corpus (Miletic et al., 2022) contains French texts produced by a variety of groups, ranging from young children to academics, annotated with syntactic information. The KUPA-KEYS corpus (Velentzas et al., 2024) consists of essays from English L1 and L2 writers, along with annotations of writer proficiency. The KLiCHe corpus (Tian et al., 2025) is an extensive resource containing nearly 5000 argumentative essays from adult writers in the United States, along with holistic scores. The GENIUS corpus (Schaller et al., 2026) is a similar resource containing short essays from German high school students, which have been annotated with a division into sections, as well as an analysis of the arguments and claims produced. Apart from analyzing the writing process, these resources have been used for CALL tasks such as prediction of essay scores (Tian et al., 2025), of students' overall grades (Schaller et al., 2026), and of general writing proficiency (Babalola et al., 2026).

We note that in none of the studies on CALL applications with keylogging data that we have been able to locate, the DHH group is represented to any significant degree or even mentioned. Among the 4992 participants from Tian et al. (2025), 29 different native languages are reported — none of them a signed language.

2.3 Writing in DHH children

As mentioned above, research on writing in DHH children is less extensive than research on reading, and studies that specifically address bilingualism, including sign language proficiency, are even more limited, see a systematic review by Zhang et al. (2024).

Schönström (2010, 2014) collected and analysed hand-written narratives produced by deaf students in grades 5 and 10, to investigate the development of Swedish as an L2. The results showed an implicational order in the students' development of Swedish that largely followed the grammatical learning order predicted by Processability Theory. The study also found a positive relationship between Swedish and Swedish sign language (STS) proficiency, suggesting that a well-developed first language may support the development of written Swedish.

Vizzi et al. (2023) investigated writing and spelling in 16 deaf young adults educated in bilingual settings. Compared with hearing controls, the deaf participants produced shorter and less informative texts, with fewer adjectives and subordinate clauses, and showed more semantic, morphological, and syntactic errors. In contrast, spelling errors were relatively infrequent, suggesting that their main difficulties concerned broader grammatical and lexical-semantic aspects of written language rather than spelling itself.

Wengelin (2002) investigated written-language production in nine congenitally deaf adults who regarded STS as their first language, alongside adults with reading and writing difficulties and hearing controls. The participants completed five different writing tasks, and their texts and

writing processes were analysed with regard to spelling, grammar, vocabulary, syntactic complexity, pauses, and editing. Wengelin (2002) represents an early study using keystroke logging to investigate written-language production in deaf adults. The results showed that the deaf participants had some spelling difficulties but, more prominently, grammatical difficulties, particularly with the syntactic rules of written Swedish. Their writing processes were also characterised by relatively linear text production, limited initial planning, and a low proportion of content-related editing.

Probably the first study of writing in DHH children using keylogging was conducted by Asker-Árnason et al. (2010). The results showed differences between children with cochlear implants (CI) and their hearing peers in both the writing process and the written product. Children with CI spent more time pausing, produced text more slowly, and showed lower narrative ability and fewer complex clauses. They also made more grammatical errors, although fewer spelling errors, and their texts had higher lexical density. The clearest group differences were found for pause time and lexical density (Asker-Árnason et al., 2012).

Gärdenfors (2023b) studied the writing of DHH children from both bilingual and cognitive writing perspectives, including analyses of the writing process using ScriptLog. The results showed that DHH children's written texts were largely comparable to those of hearing peers, especially at the age of 10–11, although they tended to produce shorter texts and showed differences in lexical density. Their writing processes revealed a slower, more revision-oriented writing style, likely reflecting increased cognitive demands during writing. Age, gender, hearing degree, and sign language proficiency were important predictors, while sign language skills did not hinder written language development and even seemed to support certain aspects of writing. Intragroup analyses further revealed that deaf children, with no functional hearing, and hard-of-hearing children, who still could make use of residual hearing, employed partly different writing strategies, reflecting their differing access to auditory and visual language input (Gärdenfors et al., 2019; Gärdenfors, 2021, 2023a,b; Gärdenfors and Johansson, 2023).

2.4 Corpora of writing in DHH children

In spite of the many studies of writing in DHH children, there are very few open access corpora in this domain. We are only aware of one other such case, namely, the Thai Deaf Corpus (TDC) by Traitruengsakul and Chuangsuwanich (2024), collected among 322 deaf students in grades 7–12. TDC is primarily aimed at CALL applications, and specifically grammatical error correction (GEC). However, it differs from Signwrite in two respects: First, the data collection for TDC involved each child writing isolated *sentences* (on average 85 per participant) rather than connected text. Secondly, TDC does not include keylogging data. It is thus not directly comparable to Signwrite.

2.5 The SweLL infrastructure

For annotation of Signwrite, we used the existing infrastructure of SweLL, a learner corpus of L2 Swedish (Volodina et al., 2019). The end product of the project, referred to as the SweLL-gold corpus, contains 502 essays collected from 321 (presumably hearing) adults (Volodina, Elena et al., 2025). The learners were recruited from schools and universities, largely through placement tests in SFI courses (Swedish for Immigrants) and university admission tests. They represented a variety of L1s, the most frequent being Arabic, Dari, Kurdish, Persian, Tigrinya, Turkish, English, Greek, Russian and Spanish.

A key aim of the SweLL project was to provide annotation of the data that facilitated both development of CALL applications and research in SLA. To this end, the project involved creating a parallel corpus consisting of the original texts along with *normalized* versions of these using a purpose-built tool, Svala (Wirén et al., 2019). A normalized (*target*) version is created by a human annotator by editing a copy of the original according to the norms of written standard Swedish (Bylin and Melander, 2023; Rudebeck and Sundberg, 2024, pp. 138–142) while retaining (the annotator's interpretation of) the meaning of the original. Also, the form of the original should not be unnecessarily altered. Normalizing a text thus involves balancing these conflicting goals (for guidelines, see Rudebeck et al. 2021). The original and target tokens are automatically aligned during editing, with Svala incrementally displaying a graph of this. Links may also be manually adjusted.

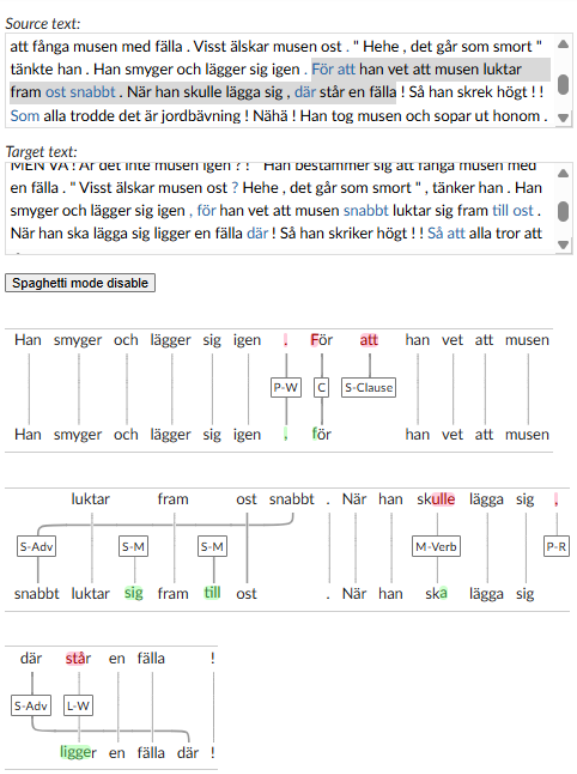


Figure 1: Svala in correction annotation mode (detail)

Once the text is normalized, a *correction annotation* is made, using a different mode in Svala (see Figure 1). (This is also known as error annotation; compare Lüdeling and Hirschmann 2015.) Here, the annotator selects relevant tokens in the source and target texts, or links that connect them, and assigns correction tags using a taxonomy developed in the SweLL project (for guidelines, see Rudebeck and Sundberg 2021). The taxonomy is grouped into orthographic, lexical, morphological, punctuation and syntactic corrections, each of which contains several specific correction tags (see Table 5). In the graph in Figure 1, correction tags are displayed on the links, given that corrections are relations between original and target tokens.[1]

3 The Signwrite corpus

3.1 Data collection

Participants in the Signwrite corpus were recruited from four groups of children and adolescents, defined as follows: deaf, hard-of-hearing, CODA, and hearing.

1. The deaf group refers to individuals who cannot effectively perceive or use spoken Swedish, do not use hearing technology, and primarily communicate through STS as their L1, with written Swedish as their L2.

2. The hard-of-hearing group refers to individuals who are deaf and use cochlear implants or have partial hearing loss and use hearing aids to enhance their auditory perception. This group communicates through spoken and written Swedish alongside STS (with exception for two individuals).

3. The CODA group refers to children of deaf adults, who are normal-hearing, but typically use STS at home with their deaf parents and are therefore often bilingual in a spoken and a signed language.

4. The hearing group refers to hearing children who do not have knowledge in STS.

Although CODA are not part of the DHH group, they increase the value of Signwrite by providing a highly relevant comparative baseline of children proficient in both STS and spoken Swedish. Likewise, hearing children constitute a very useful control group.

Signwrite is based on two data collections. The first was carried out as part of the dissertation work of Gärdenfors (2023b), involving 63 participants. Gärdenfors subsequently decided to make as much as possible of this data open access, and 44 of the participants consented to this. The second data collection was carried out by recruiting another four DHH children and adolescents, and by creating a longitudinal component. The latter was brought about by inviting the participants from the first data collection to take part in an additional session, taking place 5–7 years later. Altogether 26 DHH and CODA participants agreed to this.[2] Thus, the current Signwrite corpus comprises 74 texts and 74 keylogging files produced by 48 individuals.

3.2 Scenario

The two data collections described above were carried out using the exact same scenario. Each

[1] We refer broadly to normalization and correction annotation as "annotation". Normalization can be seen as a kind of annotation in the sense that it adds corrected tokens aligned with the original ones.

[2] The dissertation project was approved by Etikprövningsnämnden in Stockholm (2018/5:6), but for the prolongation project the approval needed to be expanded in order to make the data open access. For this purpose, a supplementary application was submitted to Etikprövningsnämnden, which was also approved (2024-02109-02).

Final text (Swedish original)	Linear input (Swedish original)
En snögig natt, där bor rosa pantern, han älskar sova. Det är bästa som han vet. Han sover jätte djupt. Men plötsligt hörde han någon knackar dörren. Han rusade ut och kikade om någon är där. Men han ser ingen.	<11.7>En<BACKSPACE3><2.1><BACKSPACE8>En snögig <BACKSPACE4>gig natten <2.9><BACKSPACE3>, där bor ett hus<BACKSPACE9>r rosa pantern <BACKSPACE3>n <BACKSPACE2>r n<BACKSPACE2>m <BACKSPACE2>n <BACKSPACE1>, han älskar sova <15.6><LEFT39><RIGHT1><LEFT1><2.2><RIGHT39><2.1>. Han sover djupt och djupt. <BACKSPACE1> Men<2.3>plötsligt hörde han någon knackar <BACKSPACE1>de dörren<15.7>. Han rusar <BACKSPACE2>de u<BACKSPACE1> ut och kikade om någon är där. Men han såg inte att <BACKSPACE11>er ingen.
Final text (English translation)	**Linear input (English translation)**
A snowgy night, there lives the pink panther, he loves sleep. It's best he knows. He sleeps very deeply. But suddenly he heard someone knocking the door. He rushed out and looked to see if someone is there. But he sees no one.	<11.7>A<BACKSPACE3><2.1><BACKSPACE8>A snowgy <BACKSPACE4>ogwy night <2.9> <BACKSPACE3>, there lives a house<BACKSPACE9> r pink panther <BACKSPACE3>n<BACKSPACE2>r n<BACKSPACE2>m <BACKSPACE2>n <BACKSPACE1>, he loves sleeping <15.6><LEFT39><RIGHT1><LEFT1><2.2> <RIGHT39><2.1>. He sleeps deeply and deeply. <BACKSPACE1>But<2.3>suddenly he heard someone knocki <BACKSPACE1>ing the door <15.7>. He rushed out <BACKSPACE2>de u<BACKSPACE1>out and looked to see if someone is there. But he did not see that<BACKSPACE11>ees no one.

Table 1: The upper left cell shows the beginning of a text by a 12-year-old DHH participant of the Signwrite data collection. The upper right cell shows the corresponding input to ScriptLog, that is, a step-by-step representation of all events during editing, including character input, cursor movements and pauses. The threshold for a pause was set at 2 seconds of inactivity, with durations shown inside angle brackets, for example, <11.7> (here rounded from ScriptLog's millisecond resolution). Angle brackets are also used for the backspace key and left and right arrow keys. Notation such as <BACKSPACE3> means that the backspace key was pressed 3 times in succession. The English translation attempts to mimic the participant's behavior in the Swedish original (including lexical, grammatical and spelling errors), and is here provided for illustrative purposes only.

		Gender		Swedish		STS		Word counts	
Hearing background	Age span	Male	Female	Written	Spoken	Yes	No	Total	Mean
Deaf	8–17	2	11	13	0	13	0	3404	261.8
Hard-of-hearing	8–21	7	25	32	32	30	2	11 021	344.4
CODA	10–18	6	14	20	20	20	0	8781	439.1
Hearing	9–11	6	3	9	9	0	9	2341	260.1
Total	8–21	21	53	74	61	63	11	25 547	345.2

Table 2: Participant metadata by hearing background, including age range, gender distribution, and knowledge of written Swedish, spoken Swedish, and STS. The total number of words (tokens excluding punctuation) is given, as well as the mean number of words per text.

session was conducted on a computer using ScriptLog, resulting in a keylogging file and a text file (see Table 1 for an example). The participants were shown a two-page Pink Panther cartoon and were asked to retell the story in written form. They were informed that they could interpret the cartoon freely, choose any name for the Pink Panther, and use their imagination when composing the story.[3] To minimize pressure and reduce the risk of unfinished texts, participants were given unlimited time to complete the task. Also, they were not informed in advance that their keystrokes would be recorded, but told about this only after completing the task. This decision was made to ensure that the writing process remained as natural as possible and to minimize any potential stress or self-consciousness associated with being monitored. Each child completed the writing task individually and without interruption.

3.3 Metadata

Table 2 summarizes personal metadata along with the number of words written in the original texts. Personal metadata includes information about hearing background, age, gender (self-reported at the first recording session) and language proficiency. As expected, the linguistic backgrounds differed across the groups. Nearly all deaf, hard-of-hearing, and CODA participants had knowledge of STS in addition to written and/or spoken Swedish. Everyone outside the deaf group had knowledge of spoken Swedish. In contrast, none of the hearing participants reported any knowledge of STS. As for the gender distribution, it is uneven with girls constituting 72% of the participants (53 girls and 21 boys). Table 3 shows the age distribution of the participants, which varied between 8 and 21 years. Table 4 shows the hearing background of the 27 participants in the longitudinal data collection.

3.4 Annotation

For the purpose of the normalization and correction annotation of Signwrite, we re-used the tools and guidelines from SweLL, as outlined in Section 2.5. This was facilitated by the fact that the annotator in Signwrite (the third author of this paper) also carried out the major part of these tasks in the construction of the SweLL corpus. Table 5 shows the outcome of the correction annotation of Signwrite in terms of tag frequencies.

[3] One advantage of this scenario is that it does not involve personal data in the sense of GDPR (Megyesi et al., 2018), thereby obviating the need for pseudonymization.

Age interval	Participants
8–9	9
10–11	25
12–13	10
14–15	4
16–17	10
18–21	16
Total	74

Table 3: Age distribution among participants in the corpus. Note that individuals in the longitudinal part are here counted twice.

Hearing background	Participants
Deaf	7
Hard-of-hearing	9
CODA	10
Total	26

Table 4: Overview of individuals who contributed longitudinal data. Each individual participated on two occasions with an intervening time of 5–7 years. The resulting 52 texts (twice the number of participants) are included in the 74 texts in Table 2.

We have not checked the inter-annotator agreement for Signwrite, but an analysis of the reliability of the SweLL correction annotation as measured on 50 essays (10% of the corpus) was made by Volodina et al. (2025, p. 272). Here, two annotators got an inter-annotator agreement of 0.88 according to Fleiss's kappa and 0.76 according to Krippendorff's alpha (Artstein and Poesio, 2008; Krippendorff, 2019).

4 Preliminary analyses

4.1 Comparison with SweLL

With Signwrite and SweLL, we are in a position to make comparisons of errors in written Swedish by subjects with different hearing status and language proficiency. Whereas the DHH and CODA subjects in Signwrite are proficient in STS and have Swedish as their strongest written language, the subjects in SweLL-gold are L2 learners of Swedish who most commonly identify another language as their strongest written language. (Hearing status was not documented in SweLL,

Orthography	O	Orthography (spelling)	717
1145	O-Cap	Upper/lower case	270
	O-Comp	Spaces and hyphens within compounds and between words	158
Lexicon	L-Der	Word formation (compounding or derivation)	64
551	L-FL	Non-Swedish word corrected to Swedish word	0
	L-Ref	Choice of anaphoric expression	277
	L-W	Word choice (other)	210
Morphology	M-Adj/adv	Adjective form corrected to adverb form	5
935	M-Case	Nominative vs genitive/accusative	33
	M-Def	Definiteness: articles; forms of nouns, adjectives	212
	M-F	Form corrected, same category (e.g. one plural suffix corrected to another plural suffix)	46
	M-Gend	Gender	50
	M-Num	Number	23
	M-Other	Other inflection-related correction	0
	M-Verb	Verb forms; auxiliaries	566
Syntax	S-Adv	Word order: adverbial placement	67
632	S-Clause	Basic clause structure	91
	S-Comp	Compound versus multi-word expression	3
	S-Ext	Extensive and complex syntactical correction	21
	S-FinV	Word order: placement of finite verb	33
	S-M	Word missing (added in target), other	179
	S-Msubj	Subject missing (added in target)	52
	S-Other	Other syntactical correction	0
	S-R	Word redundant (removed from target)	142
	S-Type	Change of phrase type/part of speech	29
	S-WO	Word order (other)	15
Punctuation	P-M	Punctuation missing (added in target)	988
1306	P-R	Punctuation redundant (removed from target)	56
	P-Sent	Sentence segmentation	31
	P-W	Wrong (i.e. changed) punctuation	231
Other	C	Consistency correction, necessitated by other correction	593
599	X	Unintelligible string	4
	Cit-FL	Non-Swedish word kept (i.e. no correction)	2
	Total		5168

Table 5: The correction taxonomy in Signwrite: main groups, specific correction tags, short descriptions of these, and their frequencies in Signwrite. Note that the taxonomy and annotation guidelines were adopted from SweLL. Further details about the taxonomy are provided by Rudebeck and Sundberg (2021).

but none of the participants reported knowledge of any sign language and they were presumably all hearing.)

For the comparison here, we have used a subset of Signwrite that contains texts written by DHH and CODA subjects who were 16 years or older (26 out of 74 texts), which is the same age cohort as the subjects of SweLL-gold. Furthermore, we have used a subset of SweLL-gold that includes texts solely with the labels ”narrative” and/or ”descriptive” (comprising 135 of 502 texts) to best fit the retellings of a cartoon in Signwrite.

The two groups of writers differ primarily in their language proficiency profiles. The Signwrite subset comprises deaf, hard-of-hearing, and CODA writers, nearly all of whom report proficiency in STS and have Swedish as their strongest written language. In contrast, the SweLL-gold subset comprises learners of Swedish as a second language, most of whom identify another language as their strongest written language. None of the SweLL-gold participants reports proficiency in a sign language. Hearing status is not documented for the SweLL-gold participants.

Table 6 juxtaposes correction annotation data from the SweLL-gold and Signwrite subsets. For the sake of comparison, the number of tag occurrences is related to the number of tokens in the original (not normalized) texts in each corpus. This relative number is presented as a percentage, referring to the number of tags in each category per 100 tokens in the corpus of original texts, in the respective corpus subsets.

The relative frequencies in Table 6 indicate that most correction tags occur more frequently in SweLL-gold than in Signwrite. The exceptions are found among corrections related to punctuation and sentence segmentation. Only three tags have a clearly higher frequency in the Signwrite subset than in the SweLL-gold subset, namely P-W, P-Sent, and C. While the C tag may be used for various corrections necessitated by another correction, their most frequent usage is to indicate that a lower case letter has been changed to upper case, as a consequence of the insertion of a sentence-segmenting punctuation mark in the normalized version of the text. Three tags have about the same frequency in both subsets, namely O-Cap, L-Ref, and P-M.

The higher frequency of corrections relating to punctuation and sentence segmentation may have to do with the material writing conditions. The Signwrite texts were written on computers, contrary to the texts in the SweLL-gold subset, which mainly includes texts which were handwritten and subsequently transcribed (but also some texts which were written on computers).

The correction category that stands out in this context is L-Ref (choice of anaphoric expression), which is slightly more frequent in the Signwrite subset than in the SweLL subset. Among the correction categories with higher or equal frequency in the Signwrite material, this is the only one which is not related to sentence segmentation and punctuation.

A typical L-Ref correction in the Signwrite material is the change of *han* ’he’ to a name or a noun. Such corrections are made when the local textual context supports an anaphoric interpretation of *han* ’he’, which, given the broader context and the cartoon itself, is unlikely to be the one intended by the writer. The higher frequency of corrections of this kind may have to do with the nature of the task. The cartoon may lead the writer to choose a pronoun that depends on a deictic association with an image, whereas the normalization is made without taking the cartoon into consideration. However, it is also possible that the higher frequency of corrections of anaphoric expressions may be attributed to factors related to the specific language experiences and proficiencies of DHH and CODA subjects. This indicates that anaphoric expressions may be an interesting aspect of the Signwrite data to further analyze.

4.2 Longitudinal data

The longitudinal data shows clear writing development across deaf, hard-of-hearing and CODA groups, although the developmental patterns differ. Deaf students show particularly strong improvements in spelling accuracy and syntactic complexity, alongside substantial increases in text length. Hard-of-hearing students display a more stable profile, with development mainly in lexical diversity and syntactic complexity. CODA students show strong growth in both lexical diversity and syntactic complexity, often reaching the highest levels at the second occasion. Overall, the findings suggest that the three groups develop along partly different trajectories, with deaf students showing more pronounced gains in formal accuracy, and the HoH and CODA groups show-

Category	SweLL-gold		Signwrite	
	Count	Percent	Count	Percent
O	1076	4.55	83	0.78
O-Cap	130	0.55	70	0.65
O-Comp	130	0.55	33	0.31
L-Der	88	0.37	25	0.23
L-FL	25	0.11	0	0.00
L-Ref	118	0.50	62	0.58
L-W	639	2.70	63	0.59
M-Adj/adv	20	0.08	1	0.01
M-Case	23	0.10	7	0.07
M-Def	564	2.38	67	0.63
M-F	98	0.41	7	0.07
M-Gend	206	0.87	17	0.16
M-Num	183	0.77	8	0.07
M-Other	25	0.11	0	0.00
M-Verb	550	2.32	158	1.48
S-Adv	114	0.48	21	0.20
S-Clause	203	0.86	21	0.20
S-Comp	17	0.07	2	0.02
S-Ext	67	0.28	0	0.00
S-FinV	158	0.67	5	0.05
S-M	405	1.71	45	0.42
S-Msubj	115	0.49	7	0.07
S-Other	14	0.06	0	0.00
S-R	264	1.12	38	0.35
S-Type	108	0.46	9	0.08
S-WO	38	0.16	1	0.01
P-M	371	1.57	153	1.43
P-R	89	0.38	15	0.14
P-Sent	6	0.03	6	0.06
P-W	70	0.30	110	1.03
C	219	0.93	129	1.20
X	59	0.25	0	0.00
Cit-FL	7	0.03	2	0.02
Total	6199	26.20	1165	10.88

Table 6: Comparison of corrections between comparable subsets (ages 16+, narrative/descriptive texts) of SweLL-gold and Signwrite. Percentages refer to the number of corrections relative to the total number of tokens. SweLL-gold tokens: 23657. Signwrite tokens: 10706.

ing greater refinement in lexical and syntactic aspects of writing.

These preliminary results are supported by correction annotation data. Table 7 presents the frequencies of the correction categories in the entire corpus with longitudinal data (including all students, i.e. deaf, hard-of-hearing, and CODA). The first two columns in Table 7 show the figures for each session. The last column shows the change in frequency between the two sessions, through a ratio calculated as the frequency for each tag in session 2, divided by the frequency for the same tag in session 1. A ratio below 1 indicates a decrease in the correction frequency between the two sessions, while a ratio above 1 indicates an increase. Generally, it may be assumed that the lower this ratio, the higher the change toward standard norms between the two sessions, that is, the greater decrease of "errors" of this particular kind.

In Table 7 orthographic and punctuation corrections have been merged into a common category. Within each general category, the category with the lowest ratio (session 2/session 1), indicating the greatest change toward norm-adherence, is presented first, followed by the second lowest, etc.

Almost all correction categories decrease in frequency between session 1 and session 2, indicating a general development toward adherence to standard norms. Among the general correction categories, the most significant development occurs within orthography and punctuation, followed by syntax. Within lexicon and morphology, the decrease in the correction frequency is somewhat lower. Interestingly, two correction categories are more frequent in session 2 than in session 1, namely M-Gend (gender) and P-W (changes of punctuation marks). (S-Comp and Cit-FL occur only in session 2, but only with 2 instances each.)

Table 8 compares the correction frequencies between the three sub-groups of students: deaf, hard-of-hearing, and CODA.[4] The comparison is made by dividing the frequency of the correction category (instances/100 tokens) in texts written by the specific group of students in one session with the frequency of the same correction category in all texts (by all students from all the three groups) in the same session. For example, the ratio 8.85 for S-FinV (placement of finite verb) among deaf students in session 1 indicates that this correction category is 8.85 times more frequent in texts written by deaf students than in all texts taken together in session 1. This ratio may be compared with 4.29, which is the ratio for the same correction category in texts by the same group (deaf students) in session 2. Although this type of correction (placement of finite verbs) remains much more frequent in texts by deaf students, the difference from texts by the other student groups has decreased between the two sessions, indicating that the deaf students have caught up some with the other students regarding correct placement of finite verbs in their texts.

In Table 8, black cells indicate a higher frequency for the correction category in texts by the specific student group than in the whole sample (ratio > 1.30), and white cells indicate a lower frequency in texts by the specific student group than in the whole sample (ratio < 0.70). Gray cells indicate approximately the same frequency in texts by the specific student group as in the whole sample.

Generally, the figures presented in Table 8 indicate that, with regard to syntax and lexicon, deaf students write texts with a lower degree of norm-adherence than the other student groups, but that this difference decreases between the two sessions. There are interesting differences between specific correction categories that call for a deeper analysis. For example, we find, again, that corrections related to the choice of anaphoric expressions (L-Ref) stand out. Contrary to the other L and S categories, the frequency of L-Ref does not decrease in texts by deaf students between session 1 and session 2.

Among corrections pertaining to morphology, orthography, and punctuation, the distribution patterns are more complex. Some of the M, P, and O categories, presented in Table 8, exhibit the same general pattern as that found within the S and L categories, that is, corrections are more frequent in texts written by deaf students than in texts written by other students, and the difference decreases between session 1 and session 2. This holds for M-Def (definiteness), M-Verb (verb inflection), and P-Sent (sentence segmentation). Two categories, P-M (punctuation added) and O-Cap (upper/lower case), are over-represented in texts by deaf students in session 1, but this over-representation is eliminated in session 2. In the

[4] Correction categories within the general category "other" are not included in Table 8, nor are categories with fewer than 15 instances in the two sessions taken together.

Correction category		session 1	session 2	session 2/session 1
Syntax	S-Ext	0.15	0.00	0.00
	S-FinV	0.24	0.04	0.18
	S-Msubj	0.30	0.09	0.29
	S-Clause	0.48	0.19	0.39
	S-WO	0.10	0.04	0.42
	S-Type	0.18	0.08	0.44
	S-M	0.91	0.48	0.53
	S-R	0.68	0.36	0.53
	S-Adv	0.30	0.22	0.75
	S-Comp	0.00	0.02	
	S total	3.34	1.53	0.46
Lexicon	L-Ref	1.41	0.60	0.43
	L-Der	0.34	0.24	0.71
	L-W	0.87	0.64	0.74
	L total	2.62	1.49	0.57
Morphology	M-Adj/adv	0.03	0.01	0.28
	M-Case	0.13	0.06	0.45
	M-Def	1.23	0.60	0.49
	M-Verb	2.91	1.58	0.54
	M-Num	0.11	0.09	0.76
	M-F	0.18	0.14	0.79
	M-Gend	0.20	0.24	1.24
	M total	4.78	2.72	0.57
Orthography and punctuation	O	3.37	0.79	0.23
	P-M	4.34	1.44	0.33
	O-Comp	0.74	0.27	0.36
	P-Sent	0.11	0.05	0.46
	O-Cap	1.21	0.60	0.50
	P-R	0.22	0.13	0.60
	P-W	0.74	0.94	1.27
	P and O total	10.73	4.22	0.39
Other	X	0.03	0.00	0.00
	C	2.45	1.13	0.46
	Cit-FL	0.00	0.02	
	Other, total	2.48	1.15	0.46
Total		23.95	11.11	0.46

Table 7: Correction frequencies (instances/100 tokens) in all texts in the longitudinal study

Category		Deaf		Hard of hearing		CODA	
		session 1	session 2	session 1	session 2	session 1	session 2
Syntax	S-FinV	8.85	4.29	0.00	0.00	0.09	0.00
	S-Ext	7.40		0.50		0.00	
	S-Type	6.53	1.91	0.44	1.27	0.24	0.27
	S-WO	6.47	4.29	0.00	0.00	0.61	0.00
	S-Msubj	5.10	2.14	0.43	0.57	0.56	0.72
	S-R	4.91	2.45	0.71	0.75	0.37	0.40
	S-Adv	4.78	2.31	0.51	0.55	0.56	0.65
	S-M	3.89	2.68	1.24	0.56	0.16	0.43
	S-Clause	3.35	1.56	0.63	0.65	0.78	0.98
Lexicon	L-Der	4.77	2.60	0.68	0.51	0.43	0.51
	L-W	3.08	2.14	0.92	0.77	0.61	0.55
	L-Ref	2.30	2.33	0.63	0.57	1.02	0.62
Morphology	M-Def	6.76	3.61	0.35	0.12	0.26	0.27
	M-Verb	4.53	2.39	0.63	0.98	0.52	0.24
	M-Gend	3.89	3.83	1.44	0.31	0.00	0.00
	M-Num	0.84	3.43	1.35	0.57	0.74	0.00
	M-Case	0.00	0.61	1.72	0.82	0.63	1.37
	M-F	0.00	2.68	1.02	0.18	1.20	0.75
Orthography and punctuation	P-Sent	8.41	2.86	0.23	0.95	0.00	0.00
	P-M	2.55	0.85	0.79	0.99	0.83	1.09
	O-Cap	2.13	0.55	0.68	1.96	1.01	0.45
	O	1.27	1.18	0.47	0.41	1.38	1.40
	O-Comp	0.77	0.69	0.62	0.55	1.36	1.55
	P-W	0.77	1.30	1.14	0.68	0.94	1.10
	P-R	0.44	0.29	1.89	2.10	0.39	0.48

Table 8: Correction category frequency (instances/100 tokens) within each student group and session, divided by the category frequency in the total sample (all three groups) in the same session. Black cells indicate a higher frequency for the correction category in the specific group than in the whole sample (ratio > 1.3). White cells indicate a lower frequency in the specific group than in the whole sample (ratio < 0.7). Gray cells indicate approximately the same frequency in the specific group as in the whole sample.

case of gender-related corrections (M-Gend), the over-representation in texts by deaf students remains stable across the two sessions. The rest of the M, O, and P correction categories exhibit patterns with even clearer deviations from the general pattern among the S and L corrections, as will be described in the following.

For two M categories, the over-representation in texts by deaf students *increases* between session 1 and session 2. This holds for M-Num (number-related morphology) and M-F, that is, instances when a form associated with a particular grammatical category (e.g. number) has been changed to another form (from another paradigm or declension) associated with the same grammatical category. In other words, most M-F corrections are corrections of errors usually analyzed as over-generalizations.

For a couple of correction categories, the frequency relation is reversed compared to the general pattern, that is, corrections are *less* frequent in texts written by deaf students than in the other texts. This holds particularly for M-Case (see below) and P-R (punctuation mark removed). Moreover, some correction categories are about as frequent in texts by deaf students as in all texts, namely O (spelling), O-comp (see below), and P-W (punctuation mark changed). O and O-comp are relatively more frequent in texts written by CODA students, while P-R is more frequent in texts written by students in the hard-of-hearing group.

Among these correction categories with an equal or lower frequency in texts by deaf students, M-Case and O-comp deserve special commenting. M-Case stands for case-related corrections. The most common M-Case correction is a change between the nominative form *de* 'they' and the accusative form *dem* 'them'. O-comp stands for corrections that have to do with spaces and hyphens in and between words. The most common O-Comp correction is the removal of a space between the two parts of a compound (concatenated in Swedish). Both of these correction types (changes between *dem* 'them' and *de* 'they', and removal of spaces in compounds) correspond to norms of written standard Swedish commonly known to be frequently broken, in many text types and by many types of writer. They are norms which are subject to debate, and which tend to be included in standard descriptions of correctness issues in written standard Swedish (e.g. Bylin and Melander (2023)). It is interesting that the current data seem to indicate that these types of norm deviation are relatively less common in texts written by deaf students than in texts written by other student groups.

5 Conclusions and future work

We have presented Signwrite, a corpus of written Swedish by DHH children and adolescents, complemented with material from CODA and hearing subjects. In addition to the texts, it includes keylogging data recording the writing process and corrected versions of the texts with error annotation. All of this is being made available under a CC-BY license, in formats suitable for use by researchers in writing development, CALL and the broader NLP field. It is the first corpus for any language that includes both DHH children's writing and keylogging data and that is open access. We have here also provided preliminary analyses of the data, suggesting the potential for more targeted contrastive analyses.

Given the relatively small size of the target population and the limited amount of available data, we aim to continue expanding the Signwrite corpus as new data become available. All future data collection includes informed consent for publication in the corpus, allowing for its continued growth. A larger dataset would enable more robust comparisons between learner groups and more detailed investigations of writing development over time.

In light of the significant impact the SweLL corpus has had on Swedish CALL applications, such as grammatical error detection, grammatical error correction, and proficiency level classification, our choice to make the Signwrite annotations follow the SweLL guidelines will enable future research on how to adapt these applications to the DHH group. The addition of keylogging data, which was not present in the SweLL corpus, will allow exploring methods for these CALL applications that depend on data about the writing process.

Acknowledgements

This work has been supported by Språkbanken, a Swedish national research infrastructure consortium jointly funded by its 10 partner institutions and the Swedish Research Council (current phase 2025–28, project id 2023-00161).

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☐ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: Claude Code was used to modify the code generating Table 8 and Table 7 into a different format, and GPT-6 was used for converting bibliography items in other formats into BibTeX. The results have been manually verified by the authors.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Lena Asker-Árnason, Viktoria Åkerlund, Cecilia Skoglund, Ingela Ek-Lagergren, Åsa Wengelin, and Birgitta Sahlén. 2012. Spoken and written narratives in swedish children and adolescents with hearing impairment. *Communication Disorders Quarterly*, 33(3):131–145.

Lena Asker-Árnason, Tina Ibertsson, Malin Wass, Åsa Wengelin, and Birgitta Sahlén. 2010. Picture-elicited written narratives, process and product, in 18 children with cochlear implants. *Communication Disorders Quarterly*, 31(4):195–212.

Damilola Babalola, Naga Buddarapu, Caleb Scott, Piotr Mitros, Paul Deane, and Collin Lynch. 2026. Measuring writing proficiency through keystroke process data: A cross-school generalizability study. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 358–368, Wyndham Grand Pittsburgh Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

Maria Bylin and Björn Melander, editors. 2023. *Språkrådet rekommenderar: Perspektiv, metoder och avvägningar i språkriktighetsfrågor*. Språkrådet, Institutet för språk och folkminnen.

Moa Gärdenfors. 2021. The writing process and the written product in bimodal bilingual deaf and hard of hearing children. *Languages*, 6(2).

Moa Gärdenfors. 2023a. Writing development in dhh students: A bimodal bilingual approach. *Journal of Deaf Studies and Deaf Education*, 28(2):211–225.

Moa Gärdenfors. 2023b. *Writing in deaf and hard-of-hearing children: A bimodal bilingual perspective on their written products and writing processes*. Ph.D. thesis, Stockholm University, Department of Linguistics.

Moa Gärdenfors and Victoria Johansson. 2023. Written products and writing processes in swedish deaf and hard of hearing children: an explorative study on the impact of linguistic background. *Frontiers in Psychology*, 14.

Moa Gärdenfors, Victoria Johansson, and Krister Schönström. 2019. Spelling in deaf, hard of hearing and hearing children with sign language knowledge. *Frontiers in Psychology*, 10.

Moa Gärdenfors, Robert Östling, Lisa Rudebeck, and Mats Wirén. 2026. Signwrite: A corpus of written Swedish texts and keystroke data from deaf and hard-of-hearing children and adolescents [Dataset]. DOI: 10.5281/zenodo.23803517.

Victoria Johansson, Johan Frid, and Åsa Wengelin. 2018. Scriptlog – an experimental keystroke logging tool. In *Proceedings of the 1st Literacy Summit*, page 51, Porto, Portugal. European Literacy Network (ELN).

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications, Los Angeles, CA.

Mariëlle Leijten and Luuk Van Waes. 2013. Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30.

Anke Lüdeling and Hagen Hirschmann. 2015. Error annotation systems. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, Cambridge Handbooks in Language and Linguistics, pages 135–158. Cambridge University Press.

Connie Mayer and Beverly J Trezek. 2018. Literacy outcomes in deaf students with cochlear implants: Current state of the knowledge. *The Journal of Deaf Studies and Deaf Education*, 23(1):1–16.

Beáta Megyesi, Lena Granstedt, Sofia Johansson, Julia Prentice, Dan Rosén, Carl-Johan Schenström, Gunlög Sundberg, Mats Wirén, and Elena Volodina. 2018. Learner corpus anonymization in the age of GDPR: Insights from the creation of a learner corpus of Swedish. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 47–56, Stockholm, Sweden. LiU Electronic Press.

Aleksandra Miletic, Christophe Benzitoun, Georgeta Cislaru, and Santiago Herrera-Yanez. 2022. ProTEXT: an annotated corpus of keystroke logs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1732–1739, Marseille, France. European Language Resources Association.

Lisa Rudebeck and Gunlög Sundberg. 2021. SweLL correction annotation guidelines. Technical report, Göteborgs universitet.

Lisa Rudebeck and Gunlög Sundberg. 2024. On the other side of the error tag: Predefined normalizations as a basis for correction annotation. In *Continuing Learner Corpus Research: Challenges and Opportunities*, pages 123–152. Ackerley, Catherine and Castello, Erik.

Lisa Rudebeck, Gunlög Sundberg, and Mats Wirén. 2021. SweLL normalization guidelines. Technical report, Göteborgs universitet.

Nils-Jonathan Schaller, Thorben Jansen, Lars Höft, Hannah Pünjer, and Andrea Horbach. 2026. GENIUS keylog corpus - a German high school student corpus with keystroke logging data. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 6919–6928, Palma de Mallorca, Spain. ELRA Language Resource Association.

Krister Schönström. 2010. *Tvåspråkighet hos döva skolelever: Processbarhet i svenska och narrativ struktur i svenska och svenskt teckenspråk*. Ph.D. thesis, Institutionen för lingvistik, Stockholms universitet.

Krister Schönström. 2014. Visual acquisition of swedish in deaf children: An l2 processability approach. *Linguistic Approaches to Bilingualism*, 4(1):61–88.

Yu Tian, Scott Crossley, and Luuk Van Waes. 2025. The klicke corpus: Keystroke logging in compositions for knowledge evaluation. *Journal of Writing Research*, 17(1):23–60.

Supachan Traitruengsakul and Ekapol Chuangsuwanich. 2024. Developing a Thai Grammatical Error Correction Tool for Deaf Students. *IEEE Access*, 12:153980–153999.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging keystrokes in writing by English learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

Francesca Vizzi, Paola Angelelli, Marika Iaia, Anthony H. Risser, and Chiara Valeria Marinelli. 2023. Writing composition ability and spelling competence in deaf subjects: A psycholinguistic analysis of source of difficulties. *Reading and Writing*, 36:1201–1226.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Elena Volodina, Arianna Masciolini, Béata Megyesi, Julia Prentice, Lisa Rudebeck, Gunlög Sundberg, and Mats Wirén. 2025. SweLL with pride: How to put a learner corpus to good use. In *Huminfra handbook: Empowering digital and experimental humanities / Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis, Elena Volodina (eds.)*, pages 251–305. University of Tartu Library, Tartu.

Volodina, Elena, Granstedt, Lena, Megyesi, Beáta, Prentice, Julia, Rudebeck, Lisa, Sundberg, Gunlög, and Wirén, Mats. 2025. SweLL-gold [dataset]. DOI: 10.23695/2K47-Y432.

Åsa Wengelin. 2002. *Text Production in Adults with Reading and Writing Difficulties*. Ph.D. thesis, University of Gothenburg.

Åsa Wengelin and Victoria Johansson. 2023. Investigating writing processes with keystroke logging. In Otto Kruse, Christian Rapp, Chris M. Anson, Kalliopi Benetos, Elena Cotos, Ann Devitt, and Antonette Shibani, editors, *Digital Writing Technologies in Higher Education: Theory, Research, and Practice*, pages 405–420. Springer International Publishing, Cham.

Mats Wirén, Arild Matsson, Dan Rosén, and Elena Volodina. 2019. Svala: Annotation of second-language learner text based on mostly automatic alignment of parallel corpora. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, Linköpings universitet. Linköping University Electronic Press, Linköpings universitet.

Kimberly A. Wolbers, Lisa M. Bowers, Hannah M. Dostal, and Shannon C. Graham. 2014. Deaf writers' application of American Sign Language knowledge to English. *International Journal of Bilingual Education and Bilingualism*, 17(4):410–428.

Dongbo Zhang, Sihui Ke, Junhui Yang, and Hannah Anglin-Jaffe. 2024. Sign language in d/deaf students' spoken/written language development: A research synthesis and meta-analysis of cross-linguistic correlation coefficients. *Review of Education*, 12(3):e70016. E70016 RoE-2023-11-1047.R2.

CineLexi: A corpus-based bilingual resource of evaluative vocabulary in film reviews (German–Russian)

Mariia Kopylova
Roma Tre University
Rome, Italy
mashulya.kopylova@list.ru

Abstract

Evaluative vocabulary constitutes the core of the film review genre; however, existing CALL resources primarily focus on general vocabulary, and the domain-specific evaluative vocabulary of specialized genres such as film reviews remains largely unaddressed. Two independent gaps motivate this study: (i) general sentiment lexicons cover only a small subset of the functional-evaluative vocabulary that learners need for productive writing tasks, and (ii) for German, open CEFR-tagged lists are available only up to level B1, whereas the evaluative vocabulary attested in film reviews is predominantly at B2–C1.

The article presents CineLexi[1], a pilot, proof-of-concept corpus-based resource of evaluative vocabulary derived from film reviews for the German–Russian language pair. The corpus comprises 900 user reviews of 15 films selected from the IMDb Top 250 (5 genres × 3 films × 30 reviews × 2 languages), collected from the Kinopoisk (kinopoisk.ru) and Filmstarts (filmstarts.de) platforms. An eight-stage automatic candidate extraction pipeline including a manually curated exclusion list facilitated the extraction of 38 lemmas (18 German, 20 Russian) at proficiency levels B1–C1, each accompanied by normalized frequency data, examples, collocations, and chunks. The resource contains monolingual dictionary entries in both languages and a cross-lingual comparison section organized by evaluative categories rather than by direct translation equivalents.

Cross-validation using SentiWS and RuSentiLex shows that common sentiment resources cover only 23% of the final 38 lemmas (6/18 German; 3/20 Russian). The study identifies corpus-observed differences between the German Filmstarts and Russian Kinopoisk subcorpora: the German subcorpus shows a stronger concentration of craft-related evaluative vocabulary (cf. Bordwell and Thompson, 2008), while the Russian subcorpus shows a stronger concentration of narrative and existential vocabulary (cf. Chatman, 1978). The final 38 lemmas form a curated core; the supplementary candidate list (85 lemmas: 34 German, 51 Russian) provides material for expansion (Section 7). A pedagogical scenario for B2-level writing tasks is proposed but has not yet been empirically evaluated.

1 Introduction

Evaluative vocabulary is central to the film review genre; it is through this vocabulary that reviewers convey their judgments regarding a film's artistic quality, narrative structure, and emotional impact. However, this vocabulary remains largely unexplored within the context of foreign language teaching. Existing corpus-based dictionaries for computer-assisted language learning (CALL) tend to emphasize general language. Kilgarriff et al. (2014) developed frequency lists with CEFR annotations for nine languages "in a corpus-driven, principled way, but with the overriding priority of creating lists which are as useful as possible for language learners." Nevertheless, their source materials consist of general web corpora, and domain-specific evaluative vocabulary from specialized genres is not represented in these lists. Similarly, François et al. (2016) create SVALex as "a lexical resource primarily aimed at learners and teachers of Swedish as a foreign and second language that describes the distribution of 15,681 words and expressions across the Common European Framework of Reference (CEFR)."

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

[1]All resources – including review identifiers, candidate extraction pipeline, dictionaries in CSV and XML formats, RelaxNG schema, HTML prototype of a dictionary entry, supplementary candidate list (85 lemmas), and Jupyter notebook – are publicly available at https://github.com/mariiakopylova/CineLexi under a CC BY 4.0 license (CEFR labels derived from GermanFluent: CC BY-SA 4.0).

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 244–255, November 19–20, 2026, University of Tartu Library

However, SVALex is derived from a corpus of educational texts (textbooks and teaching materials) rather than from authentic genre-specific discourse; film-review texts are not part of its source material.

The use of film texts in language learning is well supported by research literature. Webb (2010) analysed the scripts of 143 films (over one million word occurrences) and showed that repeated viewing provides multiple encounters with low-frequency vocabulary. Specifically, as the number of films analysed increased, so did the number of words encountered ten or more times – i.e., the relevant point is not raw frequency but repeated exposure across films. Sánchez-Auñón et al. (2023), in a systematic review, highlight the pedagogical potential of film as an authentic context for language learning. However, most of these studies focus on subtitles or dialogues, while user reviews as a source of evaluative vocabulary for computer-assisted language learning (CALL) remain virtually unexplored.

An additional challenge arises with the German–Russian language pair: unlike English–other language pairs, it is virtually absent from existing computer-assisted language learning (CALL) resources. For German, official open lists with CEFR tagging are available only up to level B1 (Goethe-Institut, 2016), whereas the evaluative vocabulary found in film reviews, as this study demonstrates, tends to be at levels B2–C1. Cross-validation against general sentiment lexicons – SentiWS (Remus et al., 2010) and RuSentiLex (Loukachevitch and Levchik, 2016) – reveals that only 23% of the candidate lemmas extracted in this study (6/18 DE; 3/20 RU) appear in these general resources. The functional-evaluative vocabulary of film reviews – words that assess a film through its aspects rather than direct polarity – systematically falls outside general sentiment resources, highlighting the need for a domain-specific lexicon.

Evaluative vocabulary is defined as lexical items that recurrently participate in linguistic constructions used to assess a film, its narrative, technical execution, emotional effect, or thematic significance, including both overtly evaluative expressions and aspect-denoting nouns that acquire evaluative function in genre-specific contexts. This article introduces CineLexi, a corpus-based vocabulary resource developed from film reviews for the German–Russian language pair. The corpus consists of 900 user reviews of 15 films from the IMDb Top 250 list, balanced across five genres: drama, sci-fi, crime/thriller, action, and comedy/romance. These reviews were collected from the Kinopoisk (kinopoisk.ru) and Filmstarts (filmstarts.de) platforms, with 450 reviews in each language. Using an automated pipeline with multi-stage filtering, 38 lemmas (18 German, 20 Russian) at proficiency levels B1–C1 were extracted and presented as monolingual dictionary entries. Each entry includes normalized frequency data, corpus examples, collocations, and productive multi-word chunks. Additionally, the resource features a cross-lingual comparison section organized by evaluative categories rather than by direct translation equivalents.

The central research question of this study is: What lexical profile characterizes the film review genre in German and Russian, and what cross-lingual asymmetries exist within this profile?

The cross-lingual differences observed in the corpus are reported in Sections 4.3 and 5.2.

Beyond the resource itself, the contribution to NLP4CALL is methodological: the pipeline combines spaCy lemmatisation and POS filtering, TF-IDF-based detection of title artefacts, regex-based frequency normalisation, and automatic CEFR tagging against GermanFluent, Goethe, and Kelly lists. The same pipeline can be transferred to other genres and language pairs, and the resource is designed to support keystroke-logging studies of L2 writing, in which learners' consultations of CineLexi entries can be logged and analysed at the level of evaluative category.

The article is organised as follows: Section 2 reviews the related literature; Section 3 describes the corpus and the candidate extraction pipeline; Section 4 presents the resource and key findings; Section 5 discusses the limitations and implications for CALL; Section 6 proposes a pedagogical scenario for using the resource; and Section 7 concludes the article.

2 Related Work

Corpus-based lexical resources for Computer-Assisted Language Learning (CALL) are developing along two primary lines: frequency lists with CEFR tagging and domain-specific dictionaries of evaluative vocabulary. These two lines are reviewed in Sections 2.1 and 2.2 respectively; Sec-

tion 2.3 then turns to film texts in language learning.

2.1 CEFR-graded lexical resources

The first approach is exemplified by the KELLY project (Kilgarriff et al., 2014) and the CEFR-Lex family – SVALex for Swedish (François et al., 2016) and EFLLex for English (Dürlich and François, 2018). SVALex describes the distribution of 15,681 Swedish words and expressions across the six CEFR levels (François et al., 2016), while EFLLex covers 15,280 English words in pedagogical materials (Dürlich and François, 2018). All three resources rely on educational or general web corpora; however, domain-specific vocabulary from specialized genres is not represented in them. The reliability of automatic CEFR tagging remains an open question: Alfter (2024) demonstrates that generative models grade vocabulary "with moderate but stable success", although their reliability for domain-specific units remains untested. Earlier, Alfter et al. (2016) showed that CEFR labels derived from different source corpora (general web vs. learner essays) yield different distributions, which suggests that frequency-based proxies derived from general corpora may overestimate the level of domain-specific vocabulary.

2.2 Sentiment lexicons

The second approach involves the use of sentiment lexicons. SentiWS (Remus et al., 2010) contains German words "weighted within the interval of $[-1; 1]$" and covers "not only adjectives and adverbs explicitly expressing a sentiment, but also nouns and verbs implicitly containing one"; RuSentiLex (Loukachevitch and Levchik, 2016) is compiled from "opinionated words from domain-oriented Russian sentiment vocabularies" and news collections. Both resources focus on direct polarity; however, functional-evaluative vocabulary – which assesses an object through its various aspects – is systematically absent, as suggested by the cross-validation reported in Section 5.

2.3 Film texts in language learning

Film texts in language learning are primarily examined through subtitles and dialogues (Webb, 2010; Sánchez-Auñón et al., 2023). Singh (2024) emphasizes the importance of corpus analysis of film texts for enhancing students' critical language awareness. However, user reviews – as a genre and as a source of evaluative vocabulary for computer-assisted language learning (CALL) – are not considered in these studies.

Corpus-based lexicography for computer-assisted language learning (CALL) is grounded in the principle of authenticity, as articulated by McEnery and Xiao (2011): authentic texts provide lexical contexts that are unavailable in traditional teaching materials. This study applies this principle to the genre of film reviews for the German–Russian language pair – a pair not represented in any of the referenced resources.

3 Corpus and Candidate Extraction Pipeline

3.1 Corpus

CineLexi is built on a corpus of user reviews in two languages. Russian reviews were collected from the Kinopoisk platform (kinopoisk.ru) via an unofficial API (poiskkino.dev), following the platform's terms of use for research purposes. German reviews were gathered from Filmstarts (filmstarts.de) using Selenium 4.52.0 (Selenium, 2026) with a headless Chrome browser, as the page content is dynamically loaded via JavaScript.

Fifteen films from the IMDb Top 250 list that appear on both platforms were selected, balanced across five genres (manually assigned based on platform metadata): drama ("The Shawshank Redemption," "The Green Mile," "Fight Club"), sci-fi ("Interstellar," "Inception," "The Matrix"), crime/thriller ("Pulp Fiction," "The Silence of the Lambs," "No Country for Old Men"), action ("Gladiator," "Léon," "The Dark Knight"), and comedy/romance ("Amélie," "The Grand Budapest Hotel," "Forrest Gump"). For each film, 30 reviews were collected in each language, totaling 900 texts (450 German and 450 Russian; 56,098 German tokens, 182,366 Russian tokens). Russian reviews are on average 3.4 times longer than German reviews (2,769 vs. 824 characters), necessitating frequency normalization per 1,000 tokens throughout the analysis. Table 1 summarizes the corpus statistics.

The corpus is a purposive sample: the 15 films were manually selected from the IMDb Top 250 and balanced across five genres (three films per genre). The repository does not redistribute the full review texts; it publishes review identifiers, extracted lexical data, and short illustrative frag-

Metric	DE	RU
Reviews	450	450
Tokens	56,098	182,366
Avg. length (chars)	824	2,769
Films covered	15	15
Genres	5	5
Reviews per film	30	30

Table 1: CineLexi corpus statistics.

ments. CEFR labels derived from GermanFluent are distributed under CC BY-SA 4.0; all other components are distributed under CC BY 4.0.

3.2 CEFR Reference Lists

CEFR annotation for German was conducted using three sources, prioritized as follows. The primary source is GermanFluent (germanfluent.com, CC BY-SA 4.0), a frequency-graded list of 14,931 lemmas covering all six CEFR levels from A1 to C2. This resource addresses the gap left by the official Goethe-Institut lists, which are publicly available only up to level B1. As a secondary cross-check, the Goethe-Institut Wortlisten (A1–B1) were used. For items absent from both sources, word frequency Zipf scores served as a fallback. For Russian, the Kelly word list (Kilgarriff et al., 2014; `ru_m3.xls` version) was the primary source, covering nouns and adjectives across CEFR levels A1–C2, with 5,925 entries retained after filtering to nouns and adjectives. Final CEFR coverage includes GermanFluent accounting for 88% of German entries in the core dictionary, while Kelly covers 100% of Russian entries.

3.3 Candidate Extraction Pipeline

The candidate extraction process involves eight successive stages, each guided by a specific methodological consideration. Table 2 shows the reduction of the candidate pool at each stage – separately for German and Russian. All thresholds listed below are heuristic and reflect the author's decisions; ablation studies are reserved for future work.

The first stage involves a part-of-speech (POS) filter. For lemmatizing all reviews, spaCy models of the medium (md) class (`de_core_news_md`, `ru_core_news_md`) were used. Subsequently, only nouns (NOUN) and adjectives (ADJ) were retained – two parts of speech that carry the primary evaluative content in film reviews. At this stage, stop words and tokens shorter than three characters were excluded. As a result, 3,895 German and 8,884 Russian lemmatized types were obtained.

Stage	DE	RU
POS filter (NOUN + ADJ)	3,895	8,884
Frequency ≥ 5	456	2,037
Domain stopword removal	430	2,012
CEFR B1–C1 + anglicism filter	233	1,201
Film coverage $\geq 6/15$	107	724
Filter A + Filter B + frequency norm + REMOVE	18	34
Top-20 selection – final dictionary	18	20

Table 2: Candidate extraction funnel.

The second stage involves a frequency filter. Lemmas with a corpus frequency below five were discarded as statistically marginal; a word occurring fewer than five times in 900 reviews cannot be considered a stable candidate for a pedagogical dictionary. This threshold is consistent with the practice in corpus-based lexicography of discarding low-frequency items before applying association measures, since such measures are unstable on small counts (Church and Hanks, 1990; Evert, 2008). After applying this filter, 456 German and 2,037 Russian lemmas remained.

The third stage involves the removal of domain-specific stop words. A short list of generic nouns was manually compiled – words that are common in any film review but do not carry evaluative content (for example, *Film*, *Szene*, *Regisseur* in German; *фильм*, *история*, *режиссёр* in Russian). The list is deliberately minimal: only unambiguously generic domain words are excluded to avoid accidentally removing potentially evaluative vocabulary. After this stage, 430 German and 2,012 Russian candidates remained.

The fourth stage involves applying the CEFR filter (B1–C1) and removing Anglicisms. Candidates outside the B1–C1 range were excluded: items at levels A1/A2 are too simple for the target audience (students at level B1+), while items at level C2 are outside the target pedagogical band for this pilot resource. The B1–C1 range was selected as the target pedagogical band; CEFR levels are inherited labels, not independently validated proficiency levels. CEFR levels were assigned au-

tomatically based on reference lists, prioritizing GermanFluent → Goethe → wordfreq for German and Kelly → wordfreq for Russian. Additionally, Anglicisms were removed from the pool of German candidates. An Anglicism was identified using a frequency-based criterion: the difference between the English Zipf score and the German Zipf score of the same string exceeded 0.8. A stricter filter for transparent English cognates (Filter B) is described separately in Stage 7. This filter rests on a transfer assumption: a German lemma that is orthographically and semantically transparent through English is likely to be already known to learners with English proficiency, and is therefore not a productive learning target. This assumption is plausible for the German–Russian pair, where English is the dominant L2, but it has not been validated empirically; a learner study testing whether such items are actually known is part of future work. This stage yielded 233 German and 1,201 Russian candidates.

The fifth stage is the film coverage filter. The lemma must appear in reviews of at least 6 out of 15 films (40% of the corpus). Words concentrated in reviews of a single film tend to reflect thematic rather than genre-specific vocabulary – for example, *prison* appears almost exclusively in reviews of "The Shawshank Redemption." This filter reduced the pool to 107 German and 724 Russian candidates.

The sixth stage is Filter A, which detects title artifacts. TF-IDF was calculated based on film subcorpora, with each film treated as a single document. A word was identified as a title artifact if its maximum TF-IDF value across films exceeded the average by more than 8.0 times. This process removed 8 German and 80 Russian candidates – words whose frequency in the corpus is attributable to the subject matter of a single film rather than to genre-based evaluative usage.

Stage seven – filter B: detection of transparent internationalisms (German Only). This filter is applied exclusively to German candidates. A word was flagged if its umlaut-normalised form is attested in English with a Zipf frequency above 3.0; the intended orthographic-similarity check against a distinct English form was not effective in the current implementation (see Section 5.3). At this stage, 19 German candidates were removed. The complete list of identified internationalisms, along with their similarity coefficients and English Zipf frequencies, is available in the repository. It should be noted that Filter B focuses on English cognates and therefore does not cover Russian–German cognates. For example, the German word *genial* is transparent to a Russian-speaking learner through the Russian word *гениальный*, but it does not meet the criterion of English frequency. This limitation is discussed in Section 5.3.

The eighth stage involves frequency normalization and recalculation using regular expressions (regex). Since spaCy lemmatization and regex matching based on surface forms differ for inflectional variants, all frequencies were recalculated using language-specific regex patterns. For Russian, the patterns match word roots followed by the \w* extension (e.g., атмосфер\w* for *атмосфера*, \bфинал\w* for *финал*). For German, the patterns match the lemma in its capitalized form along with possible inflectional variations. A minimum normalized frequency threshold of 0.35 occurrences per 1,000 tokens was applied to filter out statistically marginal units, and a manually curated REMOVE list excluded regex false positives, overly general items, and intensifiers. Together with Filter A and Filter B, this reduced the candidate pool to 18 German and 34 Russian candidates.

The final stage is selection (top–20 per language). Of the 18 German and 34 Russian candidates remaining after Stage 8, the top 20 per language by regex frequency were retained (all 18 German items). The manually curated REMOVE list applied in Stage 8 comprised three categories: regex false positives, i.e., lemmas whose surface form matched a different, unrelated word (*быт* "daily life" matched forms of *быть* "to be"; *псих* "psych" matched *психологический* "psychological"; *благо* "good" matched *благодаря* "thanks to"; *поп* "priest" matched *попытка* "attempt"); items too general to be genre-specific, i.e., words that occur in any review but do not carry evaluative content (*момент* "moment", *ситуация* "situation", *уровень* "level"); and ordinal numerals misclassified as adjectives by spaCy (*первый* "first", *второй* "second", *третий* "third"). After selection, the final resource comprises 18 German and 20 Russian lemmas at CEFR levels B1–C1.

The candidate pool after Filters A and B, before the normalised-frequency threshold and the REMOVE list, comprising 34 German and 51 Russian lemmas (85 in total), is available at `https:`

//github.com/mariiakopylova/CineLexi.git. Units with known false positives from the regex require further validation before being used in pedagogical materials.

4 Description of the Resource

CineLexi is a corpus-based bilingual resource of evaluative vocabulary in the genre of film reviews for the German–Russian language pair, designed for use in computer-assisted language learning (CALL). The resource includes 38 dictionary entries – 18 in German and 20 in Russian – and a cross-lingual section organized not by direct translation equivalents but by evaluative categories. This structure reflects the differences in the evaluative strategies of the two languages and enhances their pedagogical accessibility.

4.1 Structure of the Dictionary Entry

Each CineLexi dictionary entry follows a standardized format and includes eight fields. The headword is presented as a lemma in its base form. The part of speech is determined automatically using spaCy; the final dictionary contains only nouns and adjectives, as these are the primary carriers of evaluative meaning in the review genre. The CEFR level is assigned automatically, with an explicit indication of the source: GermanFluent (CC BY-SA 4.0) for 88% of German entries, and Kelly (Kilgarriff et al., 2014) for 100% of Russian entries. All CEFR labels are indicative and have not been validated using student data (see Section 5.3).

The normalized frequency is presented as the number of occurrences per 1,000 tokens, separately for the German and Russian subcorpora. This normalization is necessary because the average length of a Russian review in the corpus is 3.4 times greater than that of a German review. The film coverage indicates how many of the 15 films in the corpus contain the given lemma; a minimum threshold of at least six films ensures that the word belongs to genre-specific vocabulary rather than the thematic vocabulary of a single film. Each article is accompanied by two corpus examples from reviews of different films (preference was given to short sentences no longer than 250 characters), up to three collocations – the most frequent content words immediately preceding the target lemma – and up to five chunks: frequent two- and three-word combinations that include the target word.

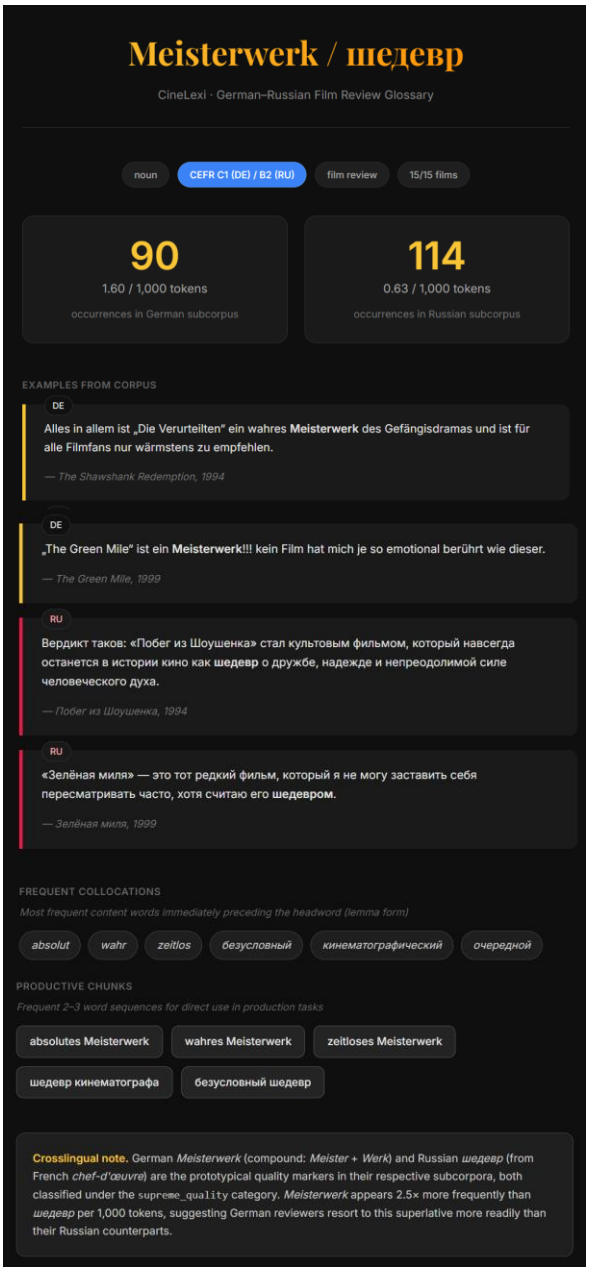


Figure 1: HTML renderer of the *Meisterwerk / masterpiece* article: eight CineLexi fields – headword, part of speech, CEFR level, frequency, film coverage, examples, collocations, chunks, plus a cross-lingual note.

Following Sinclair (1991), a collocation is treated as "the occurrence of two or more words within a short space of each other in a text", with the usual measure of proximity being "a maximum of four words intervening". A chunk, by contrast, is a frequent two- or three-word sequence containing the target lemma that functions as a ready-made production unit. The distinction is thus between a lexical association (collocation) and a prefabricated production unit (chunk). In the current implementation, collocations are oper-

ationalised more narrowly than in Sinclair's definition, as the most frequent content word immediately preceding the lemma; as a result, the two lists may overlap for adjective + noun patterns (e.g., *absolutes Meisterwerk*, *смысл жизни*), and a stricter separation is left for future work. Collocations and chunks are automatically extracted candidates and have not been manually validated; they are included in the resource as candidates rather than as validated pedagogical units. Known error types (over-generalising regex patterns, noisy preceding-word heuristics for adjectives) are documented in Section 5.3. Chunks are illustrated by such examples as *absolutes Meisterwerk* "absolute masterpiece", *düstere Atmosphäre* "gloomy atmosphere", *смысл жизни* "the meaning of life", and *человеческая природа* "human nature", intended for direct use in production tasks.

Figure 1 shows the HTML renderer of the dictionary entry *Meisterwerk* / *masterpiece* – an example of how eight fields are presented to the learner.

4.2 Final List of Lemmas

The complete list of 38 lemmas – with indications of the CEFR level, source of the label, corpus frequency, normalized frequency, and film coverage – is presented in the repository. In the main text, only a qualitative description of the distribution is provided. In the German sub-dictionary, lemmas at levels B2 (9 entries) and C1 (7 entries) predominate, which corresponds to the observation that evaluative lexis in German film reviews is predominantly at advanced levels, not covered by the official open lists of the Goethe-Institut. In the Russian sub-dictionary, levels B2 (11 entries) and B1 (7 entries) dominate – the distribution is lower, reflecting both the characteristics of the Russian subcorpus (longer analytical reviews with commonly used abstract vocabulary) and those of the Kelly list, which also covers more basic lexical units.

The most frequent lemmas in each language are indicative of the genre profile. In the German sub-dictionary, the core is formed by *Meisterwerk* "masterpiece" (1.60 per 1,000 tokens), *Darsteller* "actor" (0.94), *Spannung* "tension" (0.94), *Handlung* "plot" (0.89), and *Drehbuch* "screenplay" (0.66). The most frequent German items overall are the adjectives *perfekt* "perfect" (1.85) and *genial* "brilliant" (1.62). *Regie* "direction" (0.36) is also present in the final list as a craft-related item. The high-frequency discourse marker *Fazit* "conclusion" (1.19) is discussed separately in Section 4.3 as a genre-structuring item rather than a craft-evaluative noun. In the Russian sub-dictionary, the core is different: *сюжет* "plot" (2.65/1000t), *смысл* "meaning" (1.29), *драма* "drama" (0.97), *атмосфера* "atmosphere" (0.91), *идея* "idea" (0.89), *кадр* "shot/frame" (0.87), and *сценарий* "screenplay" (0.78) – predominantly nouns that evaluate narrative structure and thematic depth.

4.3 Cross-Lingual Mapping

A key feature of CineLexi is the <crosslingual> block – a structure that organizes lemmas in both languages not based on direct translation but according to genre assessment dimensions. Six such dimensions were identified, each described below. A complete representation of the categories, including lemmas and analytical notes, is provided in the resource's XML file.

Before the six dimensions are described, the two key analytical categories are defined. Craft is operationalised as the production-side dimensions of acting, directing, and screenwriting, following Bordwell and Thompson's (2008) account of film production. Narrative and existential vocabulary, by contrast, targets the story-side dimensions: plot structure, narration, thematic meaning, and fate (cf. Chatman, 1978).

The first dimension represents the highest quality rating. Both languages feature a frequently used noun denoting the pinnacle of artistic achievement: the German term *Meisterwerk* "masterpiece" (a compound of *Meister* + *Werk*) and the Russian term *шедевр* "masterpiece" (borrowed from the French *chef-d'œuvre*) are the most common markers of the highest quality in their respective subcorpora.

The second dimension is narrative structure. The German subcorpus evaluates narrative using two distinct nouns: *Handlung* "plot" (plot as experienced) and *Drehbuch* "screenplay" (script as a crafted object). In contrast, Russian distributes this evaluation across three units: *сюжет* "plot" (encompassing both German senses of *Handlung* "plot as experienced" and *Drehbuch* "script as a crafted object"), *повествование* "narration" (focusing on the mode of narration), and *сценарий* "screenplay" (the text as an object of evaluation). This threefold Russian distinction has no direct

counterpart in German.

The third dimension – emotional and atmospheric impact. The German subcorpus exhibits a more nuanced emotional vocabulary, employing three nouns with distinct connotations: *Atmosphäre* "atmosphere" (atmosphere as mood), *Gefühl* "feeling" (feeling as a subjective experience), and *Spannung* "tension" (tension as a structural arc). In contrast, the Russian subcorpus primarily conveys this evaluative dimension through a single term, *атмосфера* "atmosphere", which covers the semantic range distributed across the three German nouns.

The fourth dimension – acting and craft – is explicitly evaluated by German reviewers, who use specialized terms such as *Darsteller* "actor", *Hauptdarsteller* "lead actor", and *Drehbuch* "screenplay" to assess acting and script quality. The Russian final list also contains craft-related items (*состав* "cast", predominantly in *актёрский состав* "cast"; *кадр* "shot/frame"; *сценарий* "screenplay"; *визуальный* "visual"), but no dedicated actor noun, partly because *актёр* "actor" was treated as a generic stopword.

The fifth dimension – meaning and thematic depth. Russian reviewers assess films using existential and philosophical categories such as *смысл* "meaning", *идея* "idea", and *судьба* "fate". Among the German final lemmas, only *Tiefe* "depth" is directly comparable; *Sinn* "sense/meaning", the closest counterpart of *смысл*, was removed by Filter B (Section 5.3).

The sixth dimension involves reviewing discourse markers. The German term *Fazit* "conclusion" and the Russian *итог* "conclusion" are functionally equivalent, as both serve to frame the reviewer's final evaluation. These units function as genre-structuring markers rather than content-evaluative elements; however, their high frequency underscores their pedagogical importance for learners producing film reviews.

4.4 Resource Storage Format

The resource is stored in three formats: CSV, XML, and HTML. The CSV file (`cinelexi_entries.csv`) contains a tabular representation of all 38 dictionary entries with the following fields: lemma, language, CEFR level, CEFR source, corpus frequency, normalized frequency, film coverage, two examples, up to three collocations, and up to five chunks. The XML file (`cinelexi_dictionary.xml`) has a two-block structure. The <entries> block contains monolingual entries, each with the fields <headword>, <pos>, <cefr>, <domain>, <frequency>, <examples>, <collocations>, and <chunks>. The <crosslingual> block contains six categories described above; each category includes links to the corresponding dictionary entries and an analytical note.

The XML structure is formally defined using the RelaxNG Compact schema (`cinelexi_schema.rnc`). All 38 entries and six cross-lingual categories are validated against this schema. The HTML renderer for a single dictionary entry (`cinelexi_entry_meisterwerk.html`) is published as a prototype of the end-user interface, as illustrated in Figure 1.

This structure reflects a fundamental choice to create a cross-lingually annotated monolingual resource rather than a translation dictionary. CineLexi does not aim to provide direct translation equivalents for evaluative vocabulary; instead, it highlights the differences in the evaluative strategies of the two languages and offers students a tool to develop interlingual pragmatic awareness.

5 Discussion and Limitations

5.1 Cross-Validation Using Sentiment Lexicons

Cross-validation against SentiWS and RuSentiLex is consistent with the expectation that aspect-denoting evaluative vocabulary is underrepresented in general sentiment lexicons. However, no baseline is available for comparison: it is not known what coverage random nouns and adjectives of comparable frequency would receive. The observed overlap should therefore be interpreted as suggestive rather than conclusive. Of the 38 final lemmas, only 9 (23.7%) occur in SentiWS or RuSentiLex; a lemma is considered covered if it appears in at least one of the two resources — 6 out of 18 German lemmas (33%) and 3 out of 20 Russian lemmas (15%). Moreover, two of the three RuSentiLex matches are tagged as neutral (*реальность*, *человеческий*), which is methodologically uninformative for the evaluation task. The only genuinely useful match is *шедевр*, which is tagged as positive.

The low overlap suggests that general sentiment lexicons capture a substantially different subset of evaluative vocabulary from the one targeted by

CineLexi. The absence of items such as *Handlung*, *Drehbuch*, *сюжет*, *сценарий*, *итог*, and *Fazit* from both resources can be attributed to the nature of these words: they belong to functional-evaluative vocabulary, which assesses an object through its various aspects rather than through direct polarity. This particular class of words forms the core of CineLexi and is systematically absent from resources derived from news and web corpora. The observed coverage of 23% is therefore best interpreted as an indication that general sentiment lexicons and domain-specific evaluative resources capture different types of evaluativity, rather than as a measure of resource quality.

5.2 Asymmetry in Evaluative Strategies

The distribution of lemmas across evaluative categories differs between the two subcorpora. In this corpus, the German Filmstarts subcorpus shows a stronger concentration of craft-related vocabulary (*Darsteller*, *Drehbuch*, *Regie*, *Spannung*), whereas the Russian Kinopoisk subcorpus shows a stronger concentration of narrative and existential vocabulary (*смысл*, *судьба*, *идея*, *повествование*). This difference should be interpreted as a distributional tendency within this specific corpus rather than as a categorical opposition: several individual lemmas (e.g., *сценарий*, *визуальный*, *кадр*) pattern differently from the aggregate trend. This pattern may reflect differences in linguistic and/or platform-specific reviewing conventions. Kinopoisk encourages extended analytical reviews (average length: 2,769 characters), whereas Filmstarts favors shorter ones (824 characters).

The pedagogical significance of this asymmetry is considerable: a hypothesis is that a student who applies Russian-language assessment patterns to German may produce pragmatically marked texts, even at an advanced level of language proficiency. CineLexi makes this asymmetry explicit and pedagogically accessible through the <crosslingual> block, as well as through the instructional scenario proposed in Section 6.

5.3 Limitations

The first limitation is the size of the corpus. While 900 reviews for 15 films provide a sufficient volume for a pilot study, this sample is not large enough to support generalizations at the genre level. Expanding the corpus to include additional films, genres, and time periods is a priority for future iterations of the resource. Additionally, the 15 films represent a purposive sample, manually selected from the IMDb Top 250. This selection limits the external validity of conclusions regarding genre asymmetry, which should be interpreted as an observed trend within this specific corpus rather than a universal statement about the two languages.

The second limitation concerns the preliminary status of the CEFR labels. All CEFR levels in CineLexi are assigned automatically based on frequency lists and have not been validated using student data. This approach aligns with the practice of the CEFRLex family (François et al., 2016), but it necessitates a caveat: the levels should be considered indicative rather than normative. As Alfter (2024) demonstrates, automatic level assignment grades vocabulary "with moderate but stable success," even when employing generative models; however, reliability is lower for domain-specific vocabulary. Therefore, the CEFR labels in CineLexi should be regarded as a starting point for developing learning materials rather than as validated pedagogical recommendations.

The third limitation pertains to regex patterns and false positives. Several candidates in the supplementary list (85 lemmas) contain regex false positives, i.e., cases where a lemma's surface form matched an unrelated word: *быт* "daily life" matched forms of *быть* "to be"; *псих* "psych" matched *психологический* "psychological"; *благо* "good" matched *благодаря* "thanks to"; and *поп* "priest" matched *попытка* "attempt". In each case, the regex pattern for the intended lemma over-generated matches, inflating its frequency. These units are included in the supplementary data and require manual verification before use.

The fourth limitation concerns Filter B. Filter B (transparent internationalisms) targets English cognates and therefore does not cover Russian–German cognates. For example, the German word *genial* "brilliant" is transparent to a Russian-speaking learner through the Russian word *гениальный* "brilliant", yet it does not meet the English frequency criterion. In addition, the current implementation of Filter B has a methodological limitation: the orthographic similarity between a German lemma and its English counterpart was approximated by comparing the lemma with its own umlaut-normalised form, which yields similarity scores close to 1.0 for most items. As a

result, Filter B effectively reduced to a frequency check against English, and some items were removed that are not transparent English cognates – for example, *Sinn* "sense/meaning" (attested in 10 films), a direct counterpart of Russian *смысл* "meaning". This means that the German half of the meaning-depth category may be underrepresented in the final resource. A corrected implementation comparing German lemmas against English forms, and a symmetric filter using Russian as the bridge language, remain tasks for future work.

The fifth limitation is platform-specific. Reviews on Kinopoisk and Filmstarts reflect the conventions of their respective platforms and audiences, rather than the German and Russian languages per se. The observed asymmetries may be partly attributed to platform-related factors (such as review length and user demographics) rather than solely to linguistic differences. Testing this hypothesis requires a parallel corpus from other platforms – for example, IMDb for both languages.

The sixth limitation concerns the automatically generated lexicographic content. Several collocations and chunks were extracted using a heuristic that takes the word immediately preceding the target lemma; for adjectives, this heuristic produces noisy results (e.g., *Sache* "thing" for *hervorragend* "outstanding", *gesund* "healthy" for *Spannung* "tension"). The prefix-based regex patterns also over-generalise in some cases: for example, the pattern for *роман* "novel" matches *романтики* "romance", and the pattern for *состав* "cast" matches *составить конкуренцию* "compete", inflating the counts for these lemmas. The corpus examples were selected from the first two films in which the lemma appeared, which in practice means that most examples come from "The Shawshank Redemption" and "The Green Mile." The POS field in the XML is currently set to "noun" for all entries, including adjectives. These issues affect the presentation of individual entries rather than the aggregate lexical profile and will be addressed in a future iteration of the resource.

6 Pedagogical Scenario

Here is an example of using CineLexi in the context of Computer-Assisted Language Learning (CALL) within the framework of the PPP (Presentation–Practice–Production) model, one of the fundamental models for organizing educational materials in foreign language teaching.

Audience: B2-level students studying German or Russian as foreign languages; the task is to write a review of a film they have watched in the foreign language.

Stage 1: Presentation. The teacher presents the students with a cross-lingual CineLexi unit (Figure 1 shows a rendering of one article). The students analyse which evaluative categories are present in both languages and which appear in only one. Key questions include: "How does the German reviewer evaluate the actors' performances? What word does he use for this?" (*Darsteller*, *Hauptdarsteller*). "How does the Russian reviewer discuss the meaning of the film?" (*смысл*, *идея*).

Stage 2: Practice. Students receive flashcards containing case examples from CineLexi with missing target lemmas and reconstruct them using collocations and chunks. For example, for *Handlung*, the context "baut die ___ Stück für Stück geschickt auf" and the collocations (*emotionale*, *verständliche*) provide sufficient clues for recovery. For *сюжет*, the context "___ фильм основан на" and chunks (*сюжетная линия*, *сюжетные повороты*) guide the reconstruction.

Stage 3: Production. Students write a film review in the language they are learning, incorporating at least three lemmas from CineLexi within the appropriate language assessment categories. For native Russian speakers learning German, it is mandatory to include at least one item from the "Craft" category (*Darsteller* / *Drehbuch* / *Regie*), which is less frequent in the Russian subcorpus.

Beyond the product-oriented PPP scenario described above, CineLexi can also support process-oriented research on L2 writing. In a keystroke-logging setup, learners write a film review in a controlled environment (e.g., using Inputlog or a browser-based logger) while CineLexi entries are available in a side panel. The logger captures the entire writing process, including pauses, revisions, deletions, and insertions. This enables investigation into when and how learners consult CineLexi: whether they retrieve a lemma during the planning stage, formulation, or only during revision; whether they copy a chunk verbatim or adapt it; and whether consultations correlate with longer or shorter pauses. Because CineLexi entries are tagged by evaluative category

(craft, narrative, meaning, atmosphere, discourse markers), the log can be annotated at the category level, providing a detailed picture of which evaluative dimensions learners struggle with during production. For the German–Russian language pair, this is particularly informative: a Russian-speaking learner writing in German may consult craft-related entries (*Darsteller*, *Drehbuch*, *Regie*) more frequently than a German-speaking learner writing in Russian, or may avoid them altogether, reproducing the asymmetry observed in the corpus. Such data would allow to test whether the cross-lingual asymmetry documented in this study also manifests in learner production, and whether targeted CineLexi prompts can mitigate it.

This scenario illustrates the unique pedagogical application of CineLexi's cross-lingual structure: the resource serves not only as a dictionary but also as a tool for developing interlingual pragmatic awareness.

From an NLP4CALL perspective, the scenario illustrates how a domain-specific lexical resource can be integrated into process-oriented writing research: the spaCy-based pipeline, TF-IDF filtering, regex normalisation, and CEFR tagging that produced the resource also make it possible to align logged consultations with evaluative categories, and thus to connect NLP-derived lexical data with real-time learner behaviour.

The hypothesis to be tested is as follows: students who receive collocation exercises based on CineLexi will use more domain-appropriate evaluative vocabulary in their reviews than those in the control group. This scenario represents a proposal rather than an evaluated intervention; a pilot study involving B2-level students is planned as the next step.

7 Conclusion

This article presents CineLexi, a bilingual corpus-based resource of evaluative vocabulary in the film review genre for the German–Russian language pair. The corpus comprises 900 user reviews of 15 films (5 genres × 3 films × 30 reviews × 2 languages), collected from the Kinopoisk and Filmstarts platforms. An automatic pipeline consisting of eight automatic stages including a manually curated exclusion list enabled the extraction of 38 final lemmas (18 German and 20 Russian) at levels B1–C1. Each lemma is accompanied by normalized frequency data, corpus examples, collocations, and productive chunks.

Cross-validation with SentiWS and RuSentiLex suggests that general sentiment resources cover only 23% of the final candidates, motivating the need for specialized, domain-specific resources. In this corpus, reviews in the German Filmstarts subcorpus show a stronger concentration of craft-related vocabulary (*Darsteller*, *Drehbuch*, *Regie*), whereas reviews in the Russian Kinopoisk subcorpus show a stronger concentration of narrative and existential vocabulary (*смысл*, *судьба*, *повествование* – meaning, fate, narrative). This asymmetry has direct pedagogical implications for students learning German or Russian as foreign languages.

The study's limitations include the small corpus size (900 reviews for 15 films), the preliminary status of the CEFR labels, and the lack of validation using student data. Future plans include:

- expanding the corpus to cover more than 150 films, encompassing a wider range of genres and time periods
- manual verification of an additional list of candidate lemmas (85 in total) with the involvement of native speakers
- validation of the CEFR labels on student data
- conducting a pilot study of the pedagogical scenario described in Section 6.

The extraction pipeline, review identifiers, final dictionary (CSV, XML), RelaxNG schema, HTML prototype, and the supplementary candidate list are available in the repository. CEFR labels derived from GermanFluent are released under CC BY-SA 4.0; all other components under CC BY 4.0. The Russian CEFR reference list is based on Kelly (Kilgarriff et al., 2014).

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: AI tools were used for assistive purposes (idea generation, argument development), editing

(grammar, spelling, flow improvement), and compression/expansion (shortening or expanding existing human-generated input to meet page limits). All AI-assisted output was reviewed and edited by the author, who takes full responsibility for the final content.

References

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*, pages 1–19, Rennes, France. LiU Electronic Press.

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC*, Umeå, Sweden. LiU Electronic Press.

David Bordwell and Kristin Thompson. 2008. *Film Art: An Introduction*, 8th edition. McGraw-Hill, New York.

Seymour Chatman. 1978. *Story and Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stefan Evert. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics: An International Handbook*, volume 2, pages 1212–1248. Mouton de Gruyter, Berlin.

Filmstarts. 2026. Filmstarts user reviews. filmstarts.de. Accessed September 2026.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: A CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 213–219, Portorož, Slovenia. European Language Resources Association (ELRA).

GermanFluent. 2026. GermanFluent: Frequency-graded German vocabulary lists (A1–C2). germanfluent.com. CC BY-SA 4.0. Accessed September 2026.

Goethe-Institut. 2016. *Goethe-Zertifikat B1: Wortliste*. Goethe-Institut, München.

ilkermeliksitki. 2023. Goethe Institute word list (GitHub mirror). github.com/ilkermeliksitki/goethe-institute-wordlist. Accessed September 2026.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation*, 48(1):121–163.

Kinopoisk. 2026. Kinopoisk user reviews. kinopoisk.ru. Accessed September 2026. Reviews collected via unofficial API: poiskkino.dev.

Natalia Loukachevitch and Anatolii Levchik. 2016. Creating a general Russian sentiment lexicon. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1171–1176, Portorož, Slovenia. European Language Resources Association (ELRA).

Tony McEnery and Richard Xiao. 2011. What corpora can offer in language teaching and learning. In E. Hinkel, editor, *Handbook of Research in Second Language Teaching and Learning*, volume 2, pages 364–380. Routledge, London and New York.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS – a publicly available German-language resource for sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1168–1171, Valletta, Malta. European Language Resources Association (ELRA).

Estefania Sánchez-Auñón, Pedro Antonio Férez-Mora, and Fuensanta Monroy-Hernández. 2023. The use of films in the teaching of English as a foreign language: A systematic literature review. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1):10.

Selenium. 2026. Selenium WebDriver (version 4.52.0). selenium.dev. Accessed September 2026.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Manvender Kaur Miriam Singh. 2024. Corpus-based cultivation of critical language awareness: A case study on English film subtitles. *Journal of Interdisciplinary Studies in Education*, 13(S1).

Stuart Webb. 2010. A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics*, 15(4):497–519.

Translation Surprisal Predicts Learner Error in L2 French Production: A Language-Pair-Specific Difficulty Feature for ICALL

Matteo Lanza
King's College School
Southside, Wimbledon Common, London SW19 4TT, United Kingdom
matteo.lanza@icloud.com

Abstract

Word-difficulty measures in Computer-Assisted Language Learning (CALL) typically describe the target word itself, regardless of the learner's L1. Psycholinguistic research shows that translation probability predicts production, but evidence on learners comes largely from small laboratory studies using elicited norms, which are unavailable for English-French and do not distinguish the inflected form a learner must produce. From word alignments in a subtitle parallel corpus, we estimate the probability that each English word is rendered as each French form and score each learner production by the surprisal of the required form (how unexpected it is, given the English word). A mixed logistic model fitted to 384,047 typed productions from 1,213 learners in Duolingo's SLAM dataset shows that surprisal predicts error: a one-SD increase raises error probability from 21% to 30% and explains 24% of residual between-word variance. The effect lies in the probability of the specific form rather than the word's average uncertainty (entropy) or number of alternatives. Both lexical and inflectional choice contribute, and reverse-direction uncertainty does not increase error. Adding surprisal also doubles the correlation with held-out word-form difficulty. The measure is an interpretable, language-pair-specific difficulty feature for intelligent CALL (ICALL) sequencing and feedback, computable without learner data.

1 Introduction

Systems that predict learner error from platform logs, such as those developed for the Duolingo SLAM shared task, primarily use learner history and large feature sets (Settles et al., 2018); the few language-pair-specific features were form-similarity proxies for cognate status, such as edit distance between a word and its machine translation (Rich et al., 2018; Tomoschuk and Lovelett, 2018). Word-level difficulty in CALL has instead typically been predicted from properties of the target word, including frequency, length, age of acquisition, and CEFR level (Alfter and Volodina, 2018; Strohmaier and Buttery, 2024). Where language-pair-specific features have been used, they measure orthographic or semantic similarity between a word and its L1 translation (Hopman et al., 2018; Felice and Skidmore, 2026; Mayer Martins et al., 2026), or derive item difficulty from alignment-based translation distributions without testing it against learner error (Graën and Schneider, 2017; Chaudhary et al., 2021). To our knowledge, the probability of the translation a learner must produce has not been tested as a difficulty feature against learner production error.

Psycholinguistic research provides a potential language-pair-specific measure. Words with multiple translations are harder to learn and produce (Tokowicz and Kroll, 2007; Degani and Tokowicz, 2010) and slower to recognise, including for French-English bilinguals (Laxén and Lavaur, 2010), and the probability of the specific translation is a stronger predictor of performance than the number of alternatives (Prior et al., 2013). However, this evidence comes from laboratory studies with relatively small numbers of participants and items, using translation probabilities elicited for isolated words. Such norms are available for several language pairs, including English-Spanish (Prior et al., 2007), English-Dutch (Tokowicz et al., 2002), English-Chinese (Wen and van Heuven, 2017), and English-German and English-Hebrew (Degani et al., 2016), but not English-French. Moreover, isolated-word norms cannot capture which inflected form a learner must produce in context. Prior et al. (2011) found moderate convergence between elicited norms and corpus-derived translation probabilities, and alignment-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 256–268, November 19–20, 2026, University of Tartu Library

based entropy correlates with elicited norms across language pairs (Lim et al., 2023), suggesting that corpus distributions can capture part of the same information. For translators, surprisal of the chosen rendering (how unexpected it is, given the source sentence) predicts production time (Lim et al., 2024); learners have not been tested.

This paper tests whether a corpus-derived translation probability predicts learner production at scale. We estimate, for each English word, the probability that it is rendered as each French form using word alignments in an English-French subtitle corpus (Lison and Tiedemann, 2016), and score each learner production by the surprisal of the required French form. For example, *know* is usually rendered by *savoir*: *sais* has a surprisal of 0.98 bits, whereas *connais* (from *connaître*) has 4.00 (Figure 2b). We test the measure on 384,047 typed productions from 1,213 learners in Duolingo's SLAM dataset (Settles et al., 2018), using a mixed logistic model with baselines for learner, word form, and exercise.

We ask:

1. Does the probability of the French form an English speaker must produce predict production error?
2. Is the effect driven by the probability of the specific form, the word's average uncertainty, or simply the number of alternatives?
3. Do lexical and inflectional surprisal contribute independently?
4. Does surprisal remain predictive after accounting for alternative lexical properties, practice, and direction of translation?
5. Can surprisal predict the difficulty of word forms that no learner in the training data has previously attempted?

2 Methods

2.1 Overview

Testing whether the probability of a required French form predicts learner error requires an independent estimate of that probability, a method for linking each French token to its English prompt word, and a model that separates this effect from learner ability, word difficulty, and exercise difficulty (Figure 1).

The outcome data are per-token correctness labels from Duolingo's SLAM 2018 dataset (Settles et al., 2018). The probability estimates come from an unrelated English-French subtitle corpus. Word alignments produce a distribution of French forms for each English word. Each Duolingo token is then linked to its English prompt word and scored by the surprisal of the required French form. Logistic mixed models with crossed random intercepts for learner, word form, and exercise test the resulting predictors.

2.2 Datasets

The learner corpus is the SLAM 2018 dataset (Settles et al., 2018), released by Duolingo for a second-language acquisition modeling shared task. It contains learners' first 30 days across English-French, English-Spanish, and Spanish-English tracks. We use the French-for-English-speakers training split only, with no learner-level filtering. Each exercise records the English prompt, learner ID, exercise format, and correctness label for each reference token. For example, if the prompt is *I know the girl* and a learner types *Je sais la fille*, the closest accepted answer is *Je connais la fille*, and only its token *connais* is labelled an error. The SLAM exercises consist of translating or transcribing short sentences or phrases, and each learner's exercises are recorded by the SLAM dataset, tagged by session type (lessons introducing new material, practice and tests), so the same words recur across sessions.

There are three exercise formats: (i) `reverse_translate` where a learner is given an English sentence and types its French translation; (ii) `reverse_tap` where a learner is shown an English sentence and constructs the French translation by tapping word tiles from a bank that contains the correct words but also distractors; and (iii) `listen` where a learner hears a French sentence and types what they hear.

The primary analysis uses `reverse_translate`, as it requires the learner to produce the form from memory, so the forms competing for production are those in the learner's own lexicon, which the Course Set (Section 2.5) approximates. `reverse_tap` is used only as a secondary replication because its available tile set is not recorded; `listen` is excluded because it provides no English prompt. Labels are based on the full accepted-answer set rather than a single reference answer.

The primary frame (analysis set) contains 384,047 tokens from 1,213 learners, 1,685 French word forms, 1,467 English prompt words, and 117,745 exercises, with a global error rate of 0.206.

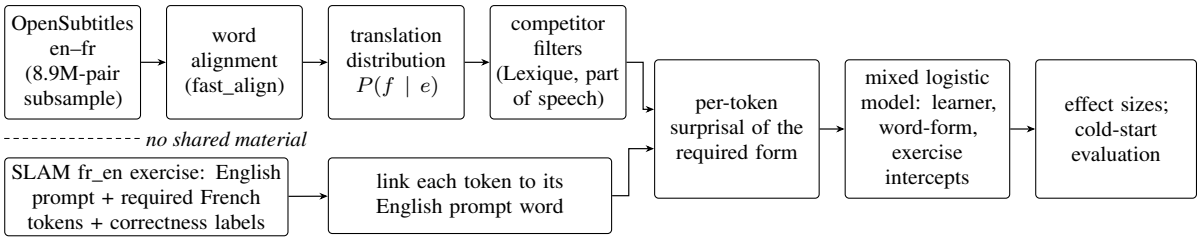


Figure 1: Method overview.

The parallel corpus is OpenSubtitles v2018 en-fr (Lison and Tiedemann, 2016; Lison et al., 2018), containing 41.8M aligned sentence pairs. The corpus is unrelated to the Duolingo material, so no sentence, learner or correctness label contributes to both the predictor and the outcome.

2.3 Translation distribution

For an English word e, the translation distribution is $P(f \mid e)$, the probability that e is rendered as French form f. We estimate this distribution from word alignments in the parallel corpus. Words are aligned with fast_align (Dyer et al., 2013) on a random 8.9M-pair subsample (Appendix A). Counts of English-French links are normalised by the English word's total links to obtain $P(f \mid e)$. The reverse distribution $P(e \mid f)$ is also calculated for token linking and reverse entropy.

2.4 Linking tokens to prompt words

The SLAM data provide an English prompt sentence but no direct word-level links. For each French token, every English word in the prompt is therefore scored using the product of forward and reverse translation probabilities and a positional distance penalty, $\exp(-4|\Delta \text{ normalised position}|)$. The highest-scoring English word is selected as the counterpart. Tokens with no scorable candidate (0.69% of translation-exercise tokens) are dropped.

Possessives and idioms can produce incorrect links and artificially high surprisal. A total of 372 tokens (0.1%) exceed 12 bits, mostly for this reason, and are excluded in a sensitivity analysis.

2.5 Competitor sets

An English word's competitors are the French forms it is aligned to. The distribution ignores sentence context: *le*, *la* and *les* are not free choices for *the* in a given sentence, so surprisal measures how dominant the required form is out of context. The raw alignment table contains substantial noise: the English word *the*, for example, has 12,275 competitors (the majority of which are a product of alignment noise across 9 million sentence pairs), compared with 586 for the average word. We therefore apply two filters. A French candidate must occur in Lexique (New et al., 2004), and its part of speech must match that of the English word (Appendix A). Elided forms are restored before dictionary lookup, and contracted articles (*au*, *du*, *aux*, *des*) are mapped to the lemma *le*, as for UD determiners (including indefinite *des*).

These filters reduce the average competitor set from 586 to 176 forms and that for *the* from 12,275 to 64, and translation probabilities are renormalised over the cleaned set. Most of these 64 forms are very improbable and carry little probability mass, so these low-probability tails have little effect on entropy or surprisal. Perplexity (2^H, the equivalent number of equally likely forms) is likewise barely affected by these tails: for *the* it is 6.4 (Figure 2a), about as uncertain as choosing among six equally likely forms.

The required French form is always retained; without this exemption, 24,351 tokens (6.3%) would receive infinite surprisal values.

The default competitor set contains all forms surviving these filters. A second "Course Set" restricts competitors to words in Duolingo's French vocabulary. Median coverage of the Course Set is 0.83 of each English word's translation probability (IQR 0.66–0.93). For competitor count, a French form must be linked at least five times, a floor fixed before any label was examined (Appendix A).

2.6 Entropy and surprisal

Entropy (Shannon, 1948) measures how evenly an English word's translation probability is distributed across French forms:

$$H(F \mid e) = -\sum_{f} P(f \mid e) \log_2 P(f \mid e). \quad (1)$$

A Miller-Madow small-sample bias correction is applied (Miller, 1955). Because French production

requires choosing both a lemma and an inflected form, entropy is decomposed as:

$$H(F \mid e) = H(L \mid e) + H(F \mid L, e), \quad (2)$$

where L is the lemma.

Surprisal instead describes the required token. If f^* is the required French form and l^* its lemma,

$$S = -\log_2 P(f^* \mid e). \quad (3)$$

Lexical and inflectional surprisal are defined as:

$$S_{\text{lex}} = -\log_2 P(l^* \mid e) \quad (4)$$

and

$$S_{\text{infl}} = -\log_2 P(f^* \mid l^*, e), \quad (5)$$

so that $S_{\text{lex}} + S_{\text{infl}} = S$ (Figure 2). Entropy captures the English word's average uncertainty, whereas surprisal captures how probable the particular form a learner must produce is. This parallels the translation-probability measure used in psycholinguistic work (Prior et al., 2013), estimated from parallel text as in Lim et al. (2023) rather than from elicited norms.

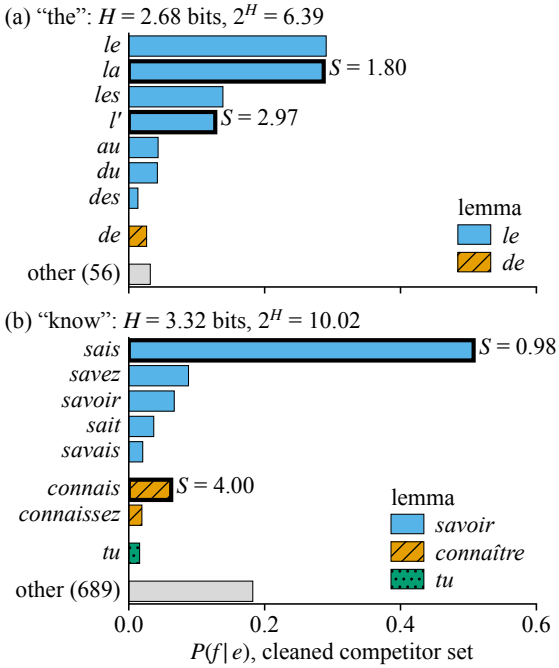
2.7 Comparison measures

We test whether surprisal is simply capturing other properties of the English word. Competitor count measures the number of distinct French forms in the cleaned competitor set (≥5 links). Contextual diversity (Adelman et al., 2006) is the number of distinct OpenSubtitles sentences containing the English word. English sense count is the number of WordNet synsets (Fellbaum, 1998). Because translation ambiguity can correlate with within-language polysemy (Degani et al., 2016), sense count tests whether surprisal is merely a proxy for English polysemy.

Finally, reverse entropy measures uncertainty over English words given the required French form, $H(E \mid f^*)$. If reverse entropy predicted error, the effect would be consistent with general translation uncertainty rather than the demands of producing French. All count measures are log-transformed.

2.8 Exposure variables

Four exposure measures are calculated from each learner's chronological record, using only events before the current token: prior productions of the required form, prior errors, days since last production, and a never-seen indicator. Prior productions are entered as $\log(1 + \text{count})$; days since last production is set to zero for unseen forms.



Prompt	Required form	surprisal (bits) Lexical	Inflectional	Total S
"the"	*la*	0.07	1.73	1.80
"the"	*l'*	0.07	2.90	2.97
"know"	*sais*	0.35	0.63	0.98
"know"	*connais*	3.16	0.85	4.00

Figure 2: Translation distributions and surprisal for two English prompt words. Each bar represents a French form coloured by lemma; bar length is $P(f \mid e)$ on the cleaned competitor set. Remaining forms are pooled as "other". *Tu* is alignment noise from *tu sais* ('you know'); it passes the part-of-speech filter as the past participle of *taire*. Table values are rounded independently, so Lexical + Inflectional may differ from Total S by 0.01.

2.9 Covariates

All models include log frequency of the French form (Lexique), log alignment count of the English word, token length, part of speech, sentence position, sentence length, days since course start, and the four exposure variables. Form frequency and English-word alignment count correlate at $r = .90$; neither coefficient is interpreted. Continuous predictors are z-scored once on the primary frame.

2.10 Model

Binary error is modelled with logistic mixed-effects regression. Crossed random intercepts for learner, word form, and exercise (Baayen et al., 2008) account for learner ability, word-form difficulty, and sentence-level difficulty, respectively. The word-form intercept is defined on the surface form rather than the lemma.

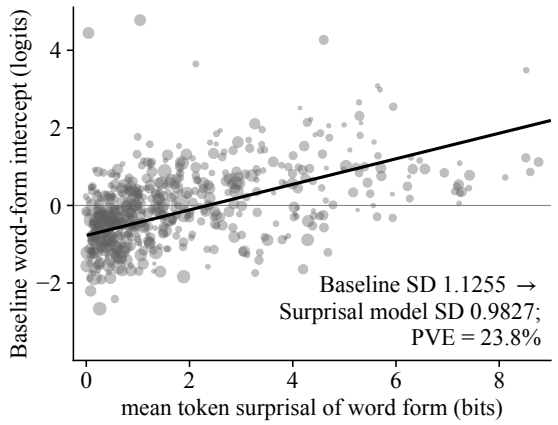


Figure 3: Representation of what PVE measures. Each point is one of the 593 word forms with at least 30 tokens (marker area increases with log token count); the line is a token-weighted linear fit. The y-axis represents difficulty in the Baseline model, after covariates, and the x-axis represents mean token surprisal. Harder words tend to be higher-surprisal words. Three word forms with mean surprisal above 9 bits lie beyond the axis and are included in the fit.

Model	Adds to Baseline
Baseline	Covariates + random intercepts for learner, word form, exercise
Count	Log competitor count
Entropy	Entropy of the translation distribution
Surprisal	Surprisal of the required form
Surprisal split	Lexical + inflectional surprisal
Entropy split	Lexical + inflectional entropy
Alternatives	Surprisal + competitor count + contextual diversity + English sense count
Practice	Surprisal + surprisal × prior productions + surprisal × days since last production
Course set	Entropy or surprisal (separate models) on the Course Set
Direction	Surprisal + reverse entropy

Table 1: Model specifications; each is the Baseline plus the listed terms.

The model containing only covariates and random intercepts is the Baseline. Named models add the predictors listed in Table 1. Because the word-form intercept captures residual between-word difficulty, adding a word-level predictor moves explained variation from the intercept into the predictor coefficient. We quantify this as the proportion of Baseline word-form intercept variance explained (PVE; Figure 3):

$$\mathrm{PVE} = \frac{\sigma^2_{w,0} - \sigma^2_{w,m}}{\sigma^2_{w,0}}, \tag{6}$$

where $\sigma^2_{w,0}$ and $\sigma^2_{w,m}$ are the word-form intercept variances in the Baseline and in model m.

We also report odds ratios (ORs). An OR of 1.62 means that a one-standard-deviation increase in the predictor multiplies the odds of error by 1.62, holding other model terms constant. Confidence intervals are 99% Wald intervals. Likelihood-ratio tests compare each model with the model it extends (Table 1).

The Baseline corresponds to a Rasch model, with learner intercepts representing ability and word-form intercepts representing difficulty; adding word-level predictors produces an explanatory item-response model (De Boeck and Wilson, 2004).

Models were fitted in R 4.6.1 with `lme4` 2.0.6 (Bates et al., 2015), using `glmer`, `nAGQ` = 0, and the nloptwrap BOBYQA optimiser. The Surprisal and Surprisal-split models were additionally refitted with the Laplace approximation (`nAGQ` = 1).

2.11 Cold-start evaluation

To test whether surprisal transfers to unseen vocabulary (a cold-start setting; McCarthy et al., 2021), word forms are split 80/20 five times using fixed seeds, with 337 held-out forms per split. The Baseline, Count, Entropy, and Surprisal models are trained on the 80%; held-out tokens are predicted using fixed effects and learner intercepts, without word-form intercepts.

We report token-level AUC, log loss, and the Spearman correlation between predicted and observed error rates for held-out word forms. We also estimate split-half reliability of observed error rates across learner halves; after Spearman-Brown correction, its square root provides an upper bound on the correlation attainable by a word-level predictor. Stratified splits by part of speech and frequency tercile are reported as a replication.

2.12 Robustness

Eight robustness checks test whether the surprisal effect depends on particular modelling choices. The lexical and inflectional components are permuted across word forms within frequency × length strata (200 replicates; word-level binomial model, deviance-drop statistic). The Surprisal-split model is refitted on random learner halves, within major parts of speech, and excluding determiners and pronouns. The Surprisal model is refitted excluding the 372 potentially mislinked tokens, with surprisal winsorised at the 99th percentile, using the Laplace approximation, and on the tile frame. Highly correlated predictor pairs are reported in

Appendix A.

3 Results

3.1 Descriptives and variance components

Baseline standard deviations are 0.73 for learner, 1.13 for word form, and 0.91 for exercise, indicating greater between-word than between-learner variation in error propensity.

3.2 Surprisal predicts error

Adding surprisal to the Baseline (Table 2) yields OR = 1.62 per standard deviation (99% CI [1.58, 1.66], $p < .001$; PVE = 23.8%). At the base error rate of 0.206, this corresponds to an increase in predicted error probability from 0.206 to 0.296.

3.3 Surprisal versus entropy and count

Competitor count (OR = 1.18, PVE = 3.4%) and entropy (OR = 1.17, PVE = 2.7%) both predict error ($p < .001$), but each explains substantially less between-word variance than surprisal. The distinction is particularly clear for inflection: inflectional entropy is weak or null (OR = 1.00; 1.10 with Lexique rather than UD determiner lemmas), whereas inflectional surprisal predicts error (OR = 1.31; Table 2). Thus, the probability of the required form is more informative than the word's average uncertainty.

3.4 Lexical and inflectional surprisal

Both components predict error: lexical surprisal OR = 1.47 [1.44, 1.50], and inflectional surprisal OR = 1.31 [1.26, 1.36], both $p < .001$. Lexical surprisal has the larger coefficient and together the two components account for 21.2% of between-word variance.

3.5 Alternative explanations

Adding competitor count, contextual diversity, and English sense count leaves the surprisal effect essentially unchanged (OR = 1.61 [1.57, 1.65]). Competitor count has no additional effect; contextual diversity is collinear with the alignment-count covariate ($r = 1.00$) and is not interpreted; English sense count retains a small effect (OR = 1.09 [1.05, 1.13]).

3.6 Direction

Reverse entropy does not increase error (OR = 0.91 [0.82, 1.01]), while the surprisal effect remains unchanged (OR = 1.62). This pattern is consistent with the effect being specific to the demands of producing French rather than reflecting general translation uncertainty.

3.7 Course Set

Course-set surprisal performs similarly to full-set surprisal (OR = 1.61 [1.57, 1.66], PVE = 26.5%; full set OR = 1.62, PVE = 23.8%). Course-set and full-set surprisal correlate at 0.98 across tokens, consistent with the median English word having 83% of its translation probability on forms in the course vocabulary.

3.8 Practice

Prior productions reduce error overall (Baseline OR = 0.65 [0.62, 0.67]) but do not reduce the surprisal gap. The surprisal $\times$ prior-productions interaction is 1.065 [1.046, 1.084], indicating that the surprisal effect increases slightly with practice. Days since last production does not moderate the effect (OR = 1.00 [0.99, 1.01]). Among tokens for which the learner had never previously failed the form, the interaction remains positive (OR = 1.046), suggesting that error-driven scheduling is not the sole explanation.

3.9 Cold start

For held-out word forms, adding surprisal increases AUC from 0.673 to 0.714, reduces log loss from 0.456 to 0.440, and increases the Spearman correlation between predicted and observed word difficulty from 0.18 to 0.37 (Table 6). Count and entropy each reach only 0.21. The result is stable across five splits (SD = 0.03) and stratified splits (0.35 versus 0.19 for the Baseline).

Observed held-out error rates have split-half reliability of 0.45, or 0.62 after Spearman-Brown correction, giving an approximate ceiling of $\sqrt{0.62} \approx 0.79$. On word forms observed in both learner halves, the Surprisal model reaches 0.41, or roughly half of this ceiling.

3.10 Sensitivity and tile exercises

The surprisal effect is stable across sensitivity analyses (Table 7). Its OR is 1.68 after excluding the 372 potentially mislinked tokens, 1.82 after winsorisation, and 1.71 with the Laplace approximation. Lexical and inflectional surprisal exceed the permutation null in 198 and 200 of 200 replicates, and remain positive under learner split-half, exclusion of determiners and pronouns (lexical

Model	Term	OR per SD	99% CI	PVE_{word}
Count	log competitor count	1.18	[1.13, 1.23]	3.4%
Entropy	entropy	1.17	[1.13, 1.22]	2.7%
Surprisal	surprisal	1.62	[1.58, 1.66]	23.8%
Surprisal split	lexical surprisal	1.47	[1.44, 1.50]	21.2%
Surprisal split	inflectional surprisal	1.31	[1.26, 1.36]	
Entropy split	lexical entropy	1.23	[1.19, 1.28]	−0.3%
Entropy split	inflectional entropy	1.00	[0.96, 1.04]	
Alternatives	surprisal	1.61	[1.57, 1.65]	24.3%
Alternatives	log competitor count	0.97	[0.92, 1.02]	
Alternatives	contextual diversity	0.86	[0.43, 1.72]	
Alternatives	English sense count	1.09	[1.05, 1.13]	
Practice	surprisal × log prior productions	1.065	[1.046, 1.084]	
Practice	surprisal × days since last	1.00	[0.99, 1.01]	
Course set	surprisal (Course Set)	1.61	[1.57, 1.66]	26.5%
Course set	entropy (Course Set)	1.14	[1.10, 1.18]	3.7%
Direction	reverse entropy	0.91	[0.82, 1.01]	

Table 2: Main models, primary frame (typed translation, 384,047 tokens). OR per SD of the standardised predictor with 99% Wald CI; PVE_{word} is the proportion of between-word variance explained relative to the Baseline. PVE is slightly negative when the word-form variance estimate does not shrink.

1.42, inflectional 1.31), and part-of-speech analyses. On tile exercises, estimates are similar to the primary frame: overall surprisal OR = 1.60 [1.54, 1.66], lexical surprisal = 1.45, and inflectional surprisal = 1.29.

4 Discussion

The results extend laboratory findings on translation probability to 384,000 learner productions. A corpus-derived measure of the probability of the required French form predicts learner error, despite being estimated from an unrelated parallel corpus. This supports the use of corpus-derived translation distributions where elicited translation norms are unavailable and allows the analysis to distinguish lexical from inflectional surprisal and to test production direction.

Several results narrow the interpretation of the effect. First, surprisal substantially outperforms competitor count and entropy. In particular, inflectional surprisal predicts error even though inflectional entropy barely does, indicating that the probability of the specific required form matters more than the English word's average uncertainty. Second, competitor count contributes no additional information once surprisal is included. Third, reverse entropy does not increase error, while production-direction surprisal does. Together, these findings are consistent with difficulty arising from competition among candidate translations weighted by their probability, although the correlational design cannot directly establish this mechanism.

Both lexical and inflectional surprisal contribute, with a larger lexical effect. The lexical result is consistent with accounts in which candidate translations compete according to their probability (Tokowicz and Kroll, 2007; Prior et al., 2013). The inflectional result indicates that probabilistic difficulty also extends to grammatical distinctions in the required form. The difference between the two results is also consistent with staged models of production where lexical selection is considered a competition between candidate lemmas, followed by the inflected form being built from the selected lemma and the sentence's grammatical features (Levelt et al., 1999): probability should be most relevant at the competitive first stage, and should matter less at the second stage, which relies on rules to determine forms. Some residual inflectional effect may reflect forms that beginners had memorised rather than derived.

Locating inflectional difficulty at the form level supports the use of a difficulty feature with separate lexical and inflectional components in CALL tasks, since a difficulty assigned per lemma cannot separate the probable and improbable realisations of the same lemma. Examples are learner modelling, where a learner producing one inflectional form does not imply the ability to produce others, and preparing assessments, where a lemma-level

difficulty measure will misjudge the difficulty of a specific required form.

Practice reduces errors overall but does not eliminate the difference between more and less probable forms. The surprisal effect increases slightly with prior productions and is unchanged by time since last exposure. Because Duolingo's scheduling is error-sensitive (Settles and Meeder, 2016), this interaction should not be interpreted causally. Separating learning from adaptive scheduling would require data from a system with non-adaptive exposure.

For ICALL, the cold-start result is particularly relevant. Surprisal is defined by a language pair rather than by the target word alone and can therefore provide a difficulty estimate before any learner has attempted the word. On held-out word forms, it substantially improves prediction over the Baseline. Unlike a purely predictive learner-history feature, it can be computed independently of prior learner performance and remains interpretable in terms of the translation distribution.

One application of the measure is to provide initial difficulty estimates for new items a course introduces. This set of initial difficulty scores then allows for ordering of items in the course such that probable forms are introduced earlier than improbable forms, while improbable forms are allotted more repetitions in the spacing schedule from the outset, rather than waiting for failures to accumulate. For lemmas already included in the course, the inflectional component can flag forms likely to be produced incorrectly before a learner has attempted them, so practice can focus on them early on.

A difficulty measure that separates lexical from inflectional difficulty is also useful in providing a system with an estimate of what type of error is more likely to arise from a token. This would allow a system to prepare feedback, whether it be a contrast between the meanings of competing words where lexical surprisal is high, or an agreement or conjugation rule where inflectional surprisal is high. Error-annotated learner data are necessary to check this against what types of errors learners actually make.

5 Limitations

The SLAM data consist largely of beginner vocabulary and cover only learners' first 30 days, so whether the effect generalises to higher proficiency is unknown. OpenSubtitles distributions also differ from language-learning materials; the Course Set analysis partly addresses this mismatch but does not remove it. Competitor sets remain sensitive to the mapping between UD and Lexique categories, although the required-form exemption protects the primary surprisal measure.

Exposure measures partly reflect Duolingo's error-sensitive scheduling, so practice interactions are descriptive rather than causal. The study also tests only English-French; although the measure can be computed for other language pairs with suitable parallel corpora, cross-linguistic generalisation remains untested.

6 Conclusion

We tested whether the probability of the French form an English speaker must produce predicts learner error, estimating this probability from parallel text without learner data. Across 384,000 Duolingo productions, surprisal predicts production error, with the effect concentrated on the probability of the required form rather than average translation uncertainty or the number of alternatives. Both lexical and inflectional surprisal contribute, and the effect is asymmetric: uncertainty in the reverse direction does not increase error.

The result identifies a language-pair-specific, interpretable difficulty feature that can be computed before a learner attempts a word. Surprisal also predicts unseen word forms substantially better than the Baseline, making it a potential feature for ICALL systems that estimate item difficulty for new vocabulary. The evidence is limited to beginner English-French learners and is correlational, so the proposed mechanism remains indirect. Future work should test additional language pairs, including the SLAM Spanish track, and use learner data annotated by error type to examine the mechanisms underlying surprisal-based difficulty more directly.

AI Disclosure

An AI assistant was used for parts of the analysis code and for language editing/polishing of our own text for clarity.

References

James S. Adelman, Gordon D.A. Brown, and José F. Quesada. 2006. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9):814–823.

David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88, New Orleans, Louisiana. Association for Computational Linguistics.

R.H. Baayen, D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Aditi Chaudhary, Kayo Yin, Antonios Anastasopoulos, and Graham Neubig. 2021. When is Wall a Pared and when a Muro?: Extracting rules governing lexical selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6911–6929, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul De Boeck and Mark Wilson, editors. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York.

Tamar Degani, Anat Prior, Chelsea M. Eddington, Ana B. Arêas da Luz Fontes, and Natasha Tokowicz. 2016. Determinants of translation ambiguity: A within and cross-language comparison. *Linguistic Approaches to Bilingualism*, 6(3):290–307.

Tamar Degani and Natasha Tokowicz. 2010. Ambiguous words are harder to learn. *Bilingualism: Language and Cognition*, 13(3):299–314.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Mariano Felice and Lucy Skidmore. 2026. Findings of the BEA 2026 shared task on vocabulary difficulty prediction for English learners. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 964–984, San Diego, California, USA. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Johannes Graën and Gerold Schneider. 2017. Crossing the border twice: Reimporting prepositions to alleviate L1-specific transfer errors. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 18–26, Gothenburg, Sweden. LiU Electronic Press.

Elise Hopman, Bill Thompson, Joseph L. Austerweil, and Gary Lupyan. 2018. Predictors of L2 word learning accuracy: A big data investigation. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pages 511–516.

Jannika Laxén and Jean-Marc Lavaur. 2010. The role of semantics in translation recognition: effects of number of translations, dominance of translations and semantic relatedness of multiple translations. *Bilingualism: Language and Cognition*, 13(2):157–183.

Willem J. M. Levelt, Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–38.

Zheng Wei Lim, Trevor Cohn, Charles Kemp, and Ekaterina Vylomova. 2023. Predicting human translation difficulty using automatic word alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11590–11601, Toronto, Canada. Association for Computational Linguistics.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1742–1748, Miyazaki, Japan. European Language Resources Association (ELRA).

Jonas Mayer Martins, Zhuojing Huang, Aaricia Herygers, and Lisa Beinborn. 2026. BoostedCats at BEA 2026 shared task 1: What makes a word hard to learn? Modeling L1 influence on English vocabulary difficulty. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 1047–1064, San Diego, California, USA. Association for Computational Linguistics.

Arya D. McCarthy, Kevin P. Yancey, Geoffrey T. LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George A. Miller. 1955. Note on the bias of information estimates. In Henry Quastler, editor, *Information Theory in Psychology: Problems and Methods*, pages 95–100. Free Press, Glencoe, IL.

Boris New, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Anat Prior, Judith F. Kroll, and Brian MacWhinney. 2013. Translation ambiguity but not word class predicts translation performance. *Bilingualism: Language and Cognition*, 16(2):458–474.

Anat Prior, Brian MacWhinney, and Judith F. Kroll. 2007. Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, 39(4):1029–1038.

Anat Prior, Shuly Wintner, Brian MacWhinney, and Alon Lavie. 2011. Translation ambiguity in and out of context. *Applied Psycholinguistics*, 32(1):93–111.

Alexander Rich, Pamela Osborn Popp, David Halpern, Anselm Rothe, and Todd Gureckis. 2018. Modeling second-language learning from a psychological perspective. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 223–230, New Orleans, Louisiana. Association for Computational Linguistics.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1848–1858, Berlin, Germany. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

David Strohmaier and Paula Buttery. 2024. Semantic error prediction: Estimating word production complexity. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 209–225, Rennes, France. LiU Electronic Press.

Natasha Tokowicz and Judith F. Kroll. 2007. Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language and Cognitive Processes*, 22(5):727–779.

Natasha Tokowicz, Judith F. Kroll, Annette M. B. de Groot, and Janet G. van Hell. 2002. Number-of-translation norms for Dutch–English translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments, & Computers*, 34(3):435–451.

Brendan Tomoschuk and Jarrett Lovelett. 2018. A memory-sensitive classification model of errors in early second language learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 231–239, New Orleans, Louisiana. Association for Computational Linguistics.

Yun Wen and Walter J. B. van Heuven. 2017. Chinese translation norms for 1,429 English words. *Behavior Research Methods*, 49(3):1006–1019.

A Alignment, token linking and correlated predictors

Alignment. Word alignments were trained with fast_align (Dyer et al., 2013) in both directions on a uniform random subsample of 8,930,148 sentence pairs of OpenSubtitles v2018 en–fr (Lison and Tiedemann, 2016; Lison et al., 2018) (5 EM iterations, diagonal prior with optimised tension, Dirichlet prior on lexical distributions) and symmetrised with grow-diag-final-and, giving 62,057,793 links. $C(e, f)$ is the symmetrised link count; $P(f \mid e) = C(e, f)/C(e, \cdot)$ and $P(e \mid f) = C(e, f)/C(\cdot, f)$.

Token linking. Each SLAM French token f at normalised position pos_f is linked to the prompt word e at normalised position pos_e maximising

$$\begin{aligned}\mathrm{score}(e, f) = P(f \mid e)\, P(e \mid f) \\ \times \exp\big(-4\,|\mathrm{pos}_f - \mathrm{pos}_e|\big);\end{aligned} \tag{7}$$

tokens with all scores zero (5,136; 0.69%) are excluded.

Part-of-speech compatibility. Competitor forms are kept only if some Lexique category of the form is compatible with the prompt word's class, the majority UD part of speech of its linked SLAM tokens (Table 3).

SLAM UD POS class	Compatible Lexique cgram prefixes
NOUN	NOM
PROPN	NOM
VERB	VER, AUX
AUX	VER, AUX
ADJ	ADJ
ADV	ADV
DET	ART, ADJ:dem, ADJ:pos, ADJ:ind, ADJ:int, ADJ:num
PRON	PRO
ADP	PRE
CONJ	CON
SCONJ	CON
NUM	ADJ:num, NOM
INTJ	ONO
other	no POS filter

Table 3: Lexique categories (cgram prefixes) accepted for each SLAM UD part-of-speech class.

Count floor. The pre-specified rule (the smallest floor k at which $r(\log \mathrm{COUNT}_k, \log N) < 0.5$, otherwise $k = 5$) selected $k = 5$ by its otherwise-clause: the correlation rises with the floor (Table 4).

Correlated predictor pairs. Table 5 lists every pair of model terms with $|r| > 0.7$ on the primary frame.

floor k	words	$r(\log \mathrm{COUNT}_k, \log N)$	$r(\log \mathrm{COUNT}_k, \log \mathrm{CONTEXT_DIV})$	median COUNT_k
1	1558	0.521	0.537	90.000
3	1555	0.709	0.722	29.000
5	1552	0.752	0.764	19.000
10	1549	0.791	0.802	11.000
25	1538	0.825	0.833	6.000

Table 4: Count-floor diagnostic (English words with $N \geq 50$ alignment links).

Term	**Term**	r
form frequency (log)	English-word alignment count (log)	0.90
form frequency (log)	token length	−0.81
form frequency (log)	contextual diversity (log)	0.89
English-word alignment count (log)	token length	−0.72
English-word alignment count (log)	contextual diversity (log)	1.00
token length	contextual diversity (log)	−0.71
surprisal	surprisal (Course Set)	0.98
inflectional surprisal	inflectional entropy	0.74
entropy	lexical entropy	0.75
entropy	lexical entropy (Lexique lemmas)	0.86
entropy	entropy (Course Set)	0.86
lexical entropy	lexical entropy (Lexique lemmas)	0.76
lexical entropy (Lexique lemmas)	entropy (Course Set)	0.81
competitor count (log)	competitor count (Course Set, log)	0.92
contextual diversity (log)	competitor count (Course Set, log)	0.71

Table 5: Predictor pairs with $|r| > 0.7$ (primary frame, token level).

B Cold-start and robustness detail

Model	Split	AUC	log loss	Spearman
Baseline	random	0.673 ± 0.006	0.456 ± 0.019	0.177 ± 0.034
Count	random	0.676 ± 0.004	0.452 ± 0.019	0.210 ± 0.032
Entropy	random	0.685 ± 0.004	0.449 ± 0.019	0.208 ± 0.038
Surprisal	random	0.714 ± 0.006	0.440 ± 0.017	0.365 ± 0.029
Baseline	stratified	0.662 ± 0.028	0.505 ± 0.041	0.190 ± 0.055
Surprisal	stratified	0.702 ± 0.026	0.487 ± 0.043	0.353 ± 0.062
Split-half reliability	learner halves	–	–	0.45 (Spearman-Brown 0.62; ceiling 0.79)

Table 6: Cold-start evaluation on held-out French word forms (mean $\pm$ SD over five splits).

Check	Item	Value
permutation (200 replicates)	lexical surprisal	exceeds the null in 198 of 200 replicates
permutation (200 replicates)	inflectional surprisal	exceeds the null in 200 of 200 replicates
learner half A	lexical; inflectional surprisal	OR 1.488; 1.346 (n = 195,427)
learner half B	lexical; inflectional surprisal	OR 1.434; 1.325 (n = 188,620)
excluding determiners and pronouns	lexical; inflectional surprisal	OR 1.423; 1.307 (n = 243,998)
within part of speech (within-class SD)	NOUN: lexical; inflectional	OR 1.275; 1.372 (n = 90,506)
within part of speech (within-class SD)	VERB: lexical; inflectional	OR 1.553; 1.338 (n = 80,388)
within part of speech (within-class SD)	DET: lexical; inflectional	OR 1.479; 1.393 (n = 80,070)
within part of speech (within-class SD)	PRON: lexical; inflectional	OR 1.573; 1.037 (n = 59,979)
within part of speech (within-class SD)	ADJ: lexical; inflectional	OR 1.217; 1.280 (n = 32,279)
within part of speech (within-class SD)	ADV: lexical; inflectional	OR 1.322; 1.090 (n = 9,698)
winsorised at the 99th percentile	surprisal	OR 1.817 [1.762, 1.873]
winsorised at the 99th percentile	lexical; inflectional surprisal	OR 1.697 [1.653, 1.743]; 1.417 [1.353, 1.484]
excluding tokens above 12 bits	surprisal	OR 1.680 [1.632, 1.730]
excluding tokens above 12 bits	lexical; inflectional surprisal	OR 1.522 [1.487, 1.558]; 1.319 [1.267, 1.374]
Laplace approximation (`nAGQ` = 1)	surprisal	OR 1.71
tile frame (`reverse_tap`)	surprisal	OR 1.597 [1.537, 1.660]
tile frame (`reverse_tap`)	lexical; inflectional surprisal	OR 1.448 [1.405, 1.492]; 1.293 [1.207, 1.384]

Table 7: Robustness checks listed in Section 2.12: odds ratios per SD, with 99% CI where computed.

Behind the Smokescreen: Uncovering Cross-LLM Bias and Reasoning Contradictions of LLM-as-a-judge for Generated Feedback

Céline Leuzinger
University of Gothenburg, Sweden
celine.leuzinger@gu.se

Abstract

The field of Feedback Generation for language learners has received growing attention over the last few years, bringing exciting perspectives for students and teachers alike. However, the lack of reliable and efficient evaluation method remains a major challenge. Concurrently, LLM-as-a-judge has appeared as a possible solution to evaluate natural language. Concerns have yet been raised regarding their capabilities in certain tasks, and their tendency to prefer their own output, known as self-bias. The present work tests whether LLMs are capable of assessing feedback similarly to human through G-Evaluation, an LLM-as-a-judge pipeline enhanced with automatic Chain-of-Thought. We also test whether models prefer GPT-generated feedback to human feedback, and control if the reasons provided by the models match the scores they attribute. Our results show that the (dis)agreement patterns of LLMs mirror the (dis)agreement patterns of human raters. Further, there exists a clear preference for GPT-generated texts compared to human texts, even when equal in quality. Finally, LLMs tend to give high scores even when their reasonings are unfavourable, raising further concerns regarding the reliability of LLM-as-a-judge.

1 Introduction

The field of Feedback Generation (FG) for language learners has gained growing attention over the last few years. FG was the object of a shared task in 2022, and numerous studies have since continued to advance the field, making use of powerful Large Language Models (LLMs) to generate explanatory notes for learner errors (Song et al. 2024, Coyne et al. 2025, Ye et al. 2025, Vainikko et al. 2025).

A major challenge in Feedback Generation is the lack of efficient and reliable evaluation method to determine the quality of the generated feedback. Studies have had to turn to either unreliable automatic metric (Nagata et al. 2023, Ye et al. 2025), or costly and time-consuming manual evaluation (Song et al. 2024, Coyne et al. 2025).

However, the advent of a technique commonly dubbed "LLM-as-a-judge" (Gu et al., 2026) has brought hope to break free from the trade-off between quality and efficiency. This technique consists of prompting an LLM to evaluate a piece of text in the form of a score, according to one or several criteria, and has been adopted in various studies (Chen et al. 2023, Chiang and Lee 2023, Han et al. 2025). LLM-as-a-judge has evolved throughout the years, giving rise to more robust and efficient systems (Liu et al., 2023). Studies have also concentrated on explainability, either by dividing the evaluation task into small, explainable steps (Lee et al., 2025) or prompting the LLMs to explain the reason for their scores (Xu et al. 2023, Liu et al. 2023).

While this brings exciting perspectives for the field of Feedback Generation, a number of concerns have also been raised regarding the use of LLMs as evaluators. The performance of LLMs depends heavily on the task at hand (Bavaresco et al., 2025), and they suffer from a variety of biases (Gu et al., 2026), among which self-bias, a tendency through which they prefer their own outputs to those of other models (Spiliopoulou et al., 2025). LLM-as-a-judge pipelines must therefore be carefully tested before they can be applied in Feedback Generation, or any other NLG tasks.

The present paper tests the performance of G-Evaluation in Feedback Generation. We also check whether LLMs prefer GPT-generated texts to human texts, even when not self-generated. We dub this phenomenon "cross-LLM bias". This is to be differentiated from "self-bias", which is only concerned with the LLMs' own outputs. Finally, we also control whether the reasons provided by

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 269–287, November 19–20, 2026, University of Tartu Library

the LLM, dubbed "reasonings", match the scores attributed.

Our contributions are the following:

1. We show that certain LLMs show high agreement with human raters, but only for majority labels describing high-quality feedback. For minority classes describing low-quality feedback, the agreement is low both among human raters and between human raters and LLMs;

2. We provide evidence that LLMs prefer GPT-generated texts even when not self-generated, a phenomenon we dub "cross-LLM bias";

3. We reveal that reasonings and scores do not always align, with LLMs giving overly high scores even when the reasoning is unfavourable.

The code for our G-Evaluation pipeline is available in our Github repository.[1]

2 Background

In 2022, Nagata et al. (2023) organised the very first shared task on Feedback Generation for language learners. The goal was to create systems capable of generating a short explanatory note given a span containing a preposition error in English. The field of Feedback Generation has continued to gain attention in the following years (Fei et al. 2023, Song et al. 2024, Maity and Deroy 2025, Vainikko et al. 2025), opening up opportunities for both language learners and teachers. However, the field is facing a major challenge: the lack of efficient and reliable evaluation method.

Studies in Feedback Generation have adopted one of two techniques to evaluate the generated feedback: automatic metrics (Nagata et al. 2023, Ye et al. 2025) and manual evaluation (Song et al. 2024, Coyne et al. 2025), both of which have their advantages and drawbacks. Automatic metrics include BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005), and measure textual quality based on n-gram overlap with a reference text. While automatic metrics are convenient and resource-efficient, important concerns have been raised regarding their misalignment with certain tasks and poor correlation with human judgments (Reiter, 2018; Schmidtová et al., 2024). Further, automatic metrics require the creation of a reference dataset, a long and costly process in the case of Feedback Generation.

Manual evaluation entails the recruitment of human experts to assess the quality of the generated feedback based on evaluation criteria. While more reliable than automatic metrics, human evaluation also has its drawbacks: it is a costly and timely procedure that requires extensive expertise, both for the evaluation itself and the definition of evaluation criteria. Coyne et al. (2025) proposed an evaluation grid for generated feedback, based on a set of seven criteria, taking into account the quality of the feedback comment as well as its impact on the learner. While this grid offers a comprehensive framework to evaluate feedback, the problems of time, cost and expertise remain.

These issues, however, could potentially be tackled by a new reference-free, automated evaluation method commonly dubbed "LLM-as-a-judge" (Gu et al., 2026). This technique consists of providing an input and evaluation criteria to an LLM, and prompt it to evaluate the output based on the criteria provided. A large number of studies have shown that LLMs align with human judgments in a range of NLP tasks, such as story generation (Chen et al., 2023; Chiang and Lee, 2023), translation (Xu et al., 2023), as well as dialogue generation and text summarisation (Liu et al., 2023). Several enhanced versions of LLM-as-a-judge have been developed to further improve LLM-based evaluation, among which G-Evaluation (Liu et al., 2023), a pipeline that divides the rating task into evaluation steps using Chain-of-Thought (CoT), and shows higher correlation with human judgments than traditional LLM-as-a-judge. There has also been a push towards explainability, through both task division into small, explainable steps (Lee et al., 2025) and provision of a reasoning to justify the scores provided (Liu et al. 2023, Xu et al. 2023).

Several issues have also been raised regarding the use of LLM-as-a-judge. Performance varies greatly depending on the task at hand and model used (Bavaresco et al., 2025). Further, LLMs suffer from self-bias, also called self-preference or self-enhancement bias, a tendency through which they prefer their own output to those of other models (Gu et al., 2026). Previous studies on the matter have sometimes failed to distinguish self-bias from genuine differences in quality (Spiliopoulou et al., 2025), as self-bias can only be evidenced if the out-

[1] https://github.com/CelineNausicaa/g_eval_generated_feedback

Criterion	Scoring	Defining question
Relevance	Binary	Is the feedback comment relevant to the error?
Factuality	Binary	Is the information in the feedback comment factually correct?
Causality	Binary	Does the feedback comment explain what is wrong and why?
Action.*	Binary	Does the feedback comment explain what to do to fix the error?
Comp.*	Binary	Is the feedback comprehensible to the assumed learner (B1-B2 level)?
Scope	Binary	Does the feedback comment contain unnecessary or out-of-scope content?
Quality	Likert (1-5)	Overall, how good is this feedback comment?

Table 1: Criteria of "good" feedback, along with their scoring type and defining question. A detailed description of each is provided in Appendix B. These criteria are directly taken from Coyne et al. (2025). *"Comp" and "Action." stand for "Comprehensibility" and "Actionability" respectively.

put is compared to a gold standard reference (Chen et al., 2025). Spiliopoulou et al. (2025) provided evidence that LLMs tend to favour both their own outputs and the outputs of models coming from the same family, dubbing this phenomenon "family bias". Wataoka et al. (2025) suggested that the cause of self-bias laid in perplexity, with LLMs favouring texts that display lower perplexity, regardless of the source. Lastly, LLM-as-a-judge pipelines often use large, proprietary models (Chen et al. 2023, Liu et al. 2023, Chen et al. 2025, Gu et al. 2026). This poses several issues, as such models require a large amount of computational resources, lack transparency, and can be taken out of the market without prior notice.

3 Methodology

To test whether G-Evaluation can be used in FG, we create a complete pipeline that grades feedback comments based on the seven criteria of good feedback (Coyne et al., 2025). To check for cross-LLM bias, we run our system on both human and AI-generated feedback. Finally, we annotate a subset of the reasonings provided by the models to control whether they match the scores attributed.

3.1 Dataset

For our experiments, we use W&I+WCF[2] (Coyne et al., 2025), a dataset containing 1156 feedback comments on learner language errors.

Learner data: the W&I+WCF comprises 197 sentences written by learners of English, as well as their corrections. Each sentence contains one grammatical error, highlighted in an error span. Each error is annotated with three error tags, corresponding to three error typologies: EXPECT (Fei et al., 2023), ERRANT (Bryant et al., 2017) and Coyne et al. (2025)'s own typology, which we label "WCF".

Feedback data: each erroneous sentence is accompanied by six pieces of feedback, obtained through six different experimental settings[3]:

- Human setting (*Human feedback*): human teachers manually provided feedback based on the learner sentence and highlighted error span.
- Template setting (*Human-GPT feedback*): a manually crafted feedback template was selected based on the error type. GPT-4o was then used to fill in slots in the template.
- Tagless setting (*GPT feedback*): GPT-4o was used to provide feedback based solely on the learner sentence and highlighted error span.
- EXPECT setting (*GPT feedback*): GPT-4o was used to provide feedback based on the learner sentence, the highlighted error span, and the EXPECT error tag.
- ERRANT setting (*GPT feedback*): GPT-4o was used to provide feedback based on the learner sentence, the highlighted error span, and the ERRANT error tag.
- WCF setting (*GPT feedback*): GPT-4o was used to provide feedback based on the learner sentence, the highlighted error span, and the WCF error tag.

[2]The creators of the dataset did not name it in the original paper. The name W&I+WCF was suggested by the first author of the original paper through personal communication. It stands for "Write & Improve + Written Corrective Feedback".

[3]See Coyne et al. (2025) for the full description of the dataset creation process.

Each setting, except for the Template system[4], generated one piece of feedback per learner sentence, yielding a total of 1156 feedback comments.

Feedback evaluation data: each of the 1156 feedback comments was rated by two human evaluators out of a pool of four, yielding a total of 2312 rated feedback comments. The annotators reported no significant difference in feedback quality between the four GPT experimental settings (Tagless, EXPECT, ERRANT, WCF). We therefore merge these four settings into one "GPT feedback" setting in the rest of this paper. Our dataset therefore contains a mix of human, human-GPT and GPT feedback.

The evaluators annotated the feedback comments based on the seven criteria of good feedback (Coyne et al., 2025), defined in Table 1. See Appendix A for a complete example datapoint from the W&I+WCF dataset, and Appendix B for a detailed description of each criterion.

3.2 Models

We select a total of six open-weights and open-source models, of different sizes and coming from different families:

LLM	N. of parameters
Gemma 3	4 billions
Gemma 3	12 billions
Llama 3.2	3 billions
Llama 3.1	8 billions
Qwen 2.5	32 billions
Deepseek-r1	32 billions

Table 2: Open-weights and open-source models selected for the experiments.

We do not use proprietary models, as we place emphasis on transparency and reproducibility for our experiments, both of which can be problematic with proprietary LLMs (Section 2). We therefore focus on testing small-to-medium open models, which are easily available to most.

3.3 G-Evaluation pipeline

To build our G-Evaluation pipeline, we use the deepeval package,[5] which allows for the creation of custom metrics based on the user's chosen evaluation criteria. We therefore create seven custom metrics based on the criteria of good feedback defined in Table 1. Our metric definitions also include a detailed bullet point list of when to choose each option for binary metrics ("Use Yes if [...]", "Use No if [...]"), and a rubric score range indication for the Likert scale criterion *Quality*. We use the exact criteria definition as provided by Coyne et al. (2025), so as to preserve equity between human evaluators and LLM evaluators. Only a minor change was made for the criteria *Scope*: the human evaluators of the W&I+WCF used 0 to indicate that the feedback did not contain out-of-scope content and 1 if the feedback did contain out-of-scope content. We reversed this logic in our G-Evaluation pipeline, after noticing that the LLMs struggled to assign a score of 0 if the outcome was positive during pre-experiments. The complete metric definitions are included in Appendix B.

The deepeval package includes a default prompt, which asks the LLM to provide a score and a reasoning for a given criterion, based on an input (the erroneous learner sentence) and actual output (the feedback). The pipeline first automatically generates evaluation steps through CoT based on the descriptions of the criteria. Following the steps generated, it attributes a score between 1 and 10 to the actual output. Rather than directly outputting this score, the G-Evaluation pipeline computes the weighted summation of the probability of each scoring outcome (1-10) to obtain a final discrete score between 0 and 1, in increments of 0.1.

For the binary criteria, we then convert the discrete scores to a binary outcome (True/False), and keep the default threshold of 0.5, meaning that any score at 0.6 or higher is considered a success and returns True. For the *Quality* criterion, scores are scaled to the 1-5 Likert scale.

Temperature is set to 0 and we collect our results after one run of our G-Evaluation pipeline for each model. We do not fine-tune our models nor provide any few-shot annotated examples.

3.4 Evaluation

Measuring agreement: we evaluate the models' performance by measuring the alignment of the G-Evaluation outputs with human judgments. To do so, we use both raw agreement and Krippendorff's alpha. Raw agreement is calculated by dividing the number of times the evaluators agree by the total

[4]Only 171 instances of rated feedback were produced with the Template system, against 197 for the other settings. Manual checks of the data revealed that the template system occasionally generated empty feedback, or that no human rating was provided. These datapoints were therefore removed from the present experiments.

[5]https://deepeval.com/docs/metrics-llm-evals

number of evaluations, while Krippendorff's alpha (Krippendorff, 2011) takes into account agreement by chance.

We compute the raw agreement between the LLM and each of the four human raters, for all seven criteria. We also compute the raw agreement between each pair of human raters, in order to compare the LLM-human agreement to the human-human agreement for each piece of feedback. We then average the agreement across pairs of LLM-human and human-human. The average agreement is presented in Table 3. The non-averaged agreement results are presented in Appendix C. We follow the same methodology for Krippendorff's alpha: we calculate alpha between the models and each of the four raters for all criteria, as well as the alpha between each rater pair. The average alpha is presented in Table 4. The non-averaged alpha is presented in Appendix D.

To deepen our analysis, we check for class prevalence for each binary criterion. Given that our dataset is indeed imbalanced (see section 4.1.1), we carry out further analyses on the minority classes. We calculate the proportion of agreement for minority classes (1) between human raters and (2) between human raters and our G-Eval pipeline, using the formula below. With each pair of brackets containing the labels from a rater pair, Min the minority class label and Maj the majority class label:

$$Agr = \frac{N(Min,Min)}{N(Min,Min)+N(Min,Maj)+N(Maj,Min)}$$

Measuring cross-LLM bias: we base our cross-LLM bias test on the *Quality* criterion, as it illustrates the overall quality of the feedback.

We run a linear mixed-effects model to test for the effect of evaluator type (LLM / human) on the scoring of the criterion *Quality*, for both GPT-generated feedback and human feedback. The results are shown in Table 8. We also visualise the score distribution of each LLM as well as human evaluators in Figure 1 and 2.

Measuring alignment between scores and reasonings: to test whether the scores and reasonings provided by the LLMs align, we manually annotate the same random sample of 15 feedback comments for each LLM. Given that each feedback comment is evaluated across 7 criteria, this yields a total of 105 score-to-reasoning pairs per LLM.

We use sentiment analysis as the basis of our annotations: each reasoning is annotated as either "Positive", "Neutral" or "Negative" based on the set of rules below[6]:

- Positive: the reasoning is positive or mostly positive.
  Example: *"The feedback correctly identifies the error by changing 'for' to 'to', addressing the highlighted span. It explains that a to-infinitive is needed for purpose clauses, which is accurate."*

- Neutral: the reasoning contains both positive and negative comments, in equal quantity.
  Example: *"The feedback provides a direct edit suggestion by changing 'for' to 'to', which aligns with evaluation step 1. It also offers a viable candidate edit, fulfilling step 2. However, it doesn't provide multiple options or address broader issues beyond the specific grammatical change, so steps 3 and 4 are not fully addressed."*

- Negative: the reasoning is negative or mostly negative.
  Example: *"The feedback incorrectly states that 'waiting' is a verb or gerund when it's actually a noun in this context. It also suggests changing to 'wait', which isn't the correct noun form for duration."*

We visualise the distribution of scores for each sentiment in Figure 3.

To obtain a more fine-grained analysis, we also convert the corresponding discrete scores into sentiment, following the set of rules below:

- Score $> 0.6 \rightarrow$ Positive
- $0.4 \leq$ Score $\leq 0.6 \rightarrow$ Neutral
- Score $< 0.4 \rightarrow$ Negative

The reasoning is considered to align with the score if both receive the same sentiment label. We compute the accuracy of alignment and extract the mean and mode score per sentiment. These results are in Appendix E.

4 Results and discussion

The results show the complexity of assessing the performance of G-Evaluation. While the high raw

[6]The examples are retrieved from deepseek-r1:32b outputs.

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	96.38	79.25	98.16	98.64	43.91	95.99	28.92
Gemma3:12b	**99.31**	**92.22**	97.38	**99.92**	95.53	**99.10**	**60.51**
Llama3.2:3b	16.01	46.90	32.72	73.23	73.00	93.03	15.08
Llama3.1:8b	96.23	59.17	**98.08**	93.62	58.12	90.15	32.00
Deepseek-r1:32b	86.89	69.80	71.58	99.31	**97.10**	98.05	52.95
Qwen2.5:32b	89.96	73.88	38.12	94.72	93.47	98.32	24.00
Human raters	98.74	93.46	97.30	99.75	95.36	98.10	45.49

Table 3: Average raw agreement between the LLM and the human raters as well as between human raters only. The detail of agreement between each rater-rater pair and each llm-rater pair is provided in Appendix C. All scores are in percentage.

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	**0.024**	-0.054	-0.008	-0.006	-0.338	-0.020	-0.154
Gemma3:12b	-0.002	**0.057**	**0.019**	N/A	**0.165**	-0.004	-0.114
Llama3.2:3b	-0.717	-0.299	-0.483	-0.150	-0.111	-0.035	-0.427
Llama3.1:8b	0.014	-0.186	-0.009	-0.032	-0.201	-0.017	-0.110
Deepseek-r1:32b	-0.044	-0.0865	-0.133	**-0.002**	0.097	0.027	**-0.021**
Qwen2.5:32b	-0.018	-0.056	-0.420	-0.011	0.117	**0.134**	-0.307
Human raters	-0.004	0.159	N/A	N/A	0.251	-0.008	0.015

Table 4: Average alpha score between the LLM and the human raters as well as between human raters only. The detail of agreement between each rater-rater pair and each llm-rater pair is provided in Appendix D. N/A values suggest little to no variability in the data.

agreement suggests that LLMs evaluate similarly to humans, the low alpha scores show that the pipeline struggles with lower-quality feedback. The analyses of cross-LLM bias and alignment between score and reasoning also depict a very contrasted picture.

4.1 Agreement analysis

The results presented in Table 3 show that Gemma3:12b is seemingly the best performing model, with a raw agreement above 90% across all binary metrics. For the Likert scale criterion *Quality*, Gemma3:12b even obtains a higher average agreement with humans than human raters do with each other.

Other models exhibit important differences depending on the criterion, with higher performances recorded for the binary criteria than for the Likert scale criterion. All models obtain a high raw agreement in *Actionability* and *Scope*, while all other criteria show important variation between models. For example, Gemma3:4b underperforms in *Comprehensibility*, and Qwen2.5:32b in *Causality*, but both show high agreement for all other binary criteria. *Quality* is the criterion that shows the lowest agreement, both with LLMs and human raters.

Interestingly, the agreement patterns of most LLMs seem to mirror the agreement patterns of humans: when the Inter-Annotator Agreement (IAA) is high between humans, it is also high between humans and LLMs. Conversely, IAA drops for the criterion *Quality* between humans, as well as between LLMs and humans. Llama3.2:3b is an exception, showing low performance even when human agreement is high, similarly to Qwen2.5:32b for the criterion *Causality*.

The results for Krippendorff's alpha, presented in Table 4, tell a drastically different story: while the high raw agreement suggests that our G-Eval pipeline performs similarly to human raters for binary criteria, the alpha scores show that such results could be due to pure chance. Given that the raw agreement is high, the low alpha scores does not signify that the raters disagree. Rather, it could be due to important class imbalance, rendering chance-corrected agreement obsolete.

Criteria	Relevance	Factuality	Causality	Action.	Comp.	Scope
Proportion of True	> 95%	> 95%	> 95%	> 95%	> 95%	< 5%
Proportion of False	< 5%	< 5%	< 5%	< 5%	< 5%	> 95%

Table 5: Prevalence of each class across the binary criteria, combined for all four human raters.

Criteria	Relevance	Factuality	Causality	Action.	Comp.	Scope
Proportion (%)	0	11.84	6.061	0	12.72	0
Minority class label	False	False	False	False	False	True

Table 6: Proportion of agreement cases, i.e. cases where human annotators agree on the minority class label. For example, a score of 0 for *Relevance* means that the pair of raters never agreed on the minority class label: whenever a rater labeled a feedback as irrelevant (False), the corresponding rater always labeled it as relevant (True). On the other hand, for the criterion *Comprehensibility*, if a rater labeled a feedback comment as incomprehensible (False), the paired rater agreed (also False) in 12.72% of the cases.

4.1.1 Class imbalance

To investigate why the alpha scores are low, we compute the prevalence of each class ("True" / "False") across binary criteria. Table 5 shows a simplified overview of the proportion of each class per criteria across all four raters.

These results show a clear data imbalance, with the label "True" occurring for the vast majority of the data across all criteria. The logic is reversed for the criteria *Scope*, with the label "False" occurring in more than 95% of the cases. This shows that our base dataset is of high quality: human evaluators chose the positive outcome in the vast majority of cases. We therefore carry out further analyses on the rare yet pertinent low-quality feedback (< 5%), to analyse how human evaluators as well as our G-Evaluation pipeline handle such cases.

4.1.2 Minority class agreement

The performance of our G-Evaluation pipeline is evaluated based on whether it mirrors human evaluators. It is therefore of primary importance to first study how human raters handle minority class labels by analysing whether humans agree on such minority cases. High agreement between humans on minority classes would suggest that such cases are rather objective, in the sense that a "ground truth" can be agreed upon. Conversely, low agreement suggests a higher level of subjectivity, where no label is universally accepted. Without such "ground truth" to rely on, evaluating the performance of G-Evaluation based on human labels might reveal itself unfeasible. We calculate the agreement for minority classes between human raters using the formula presented in Section 3.4. The results are presented in Table 6.

Interestingly, the results are low across all criteria. The criteria *Relevance*, *Actionability* and *Scope* even show 0% agreement for their minority classes, meaning that the rater pair never agreed. The numbers are slightly higher for the criteria *Factuality* and *Comprehensibility*, but remain low overall. Together with the raw agreement results from Section 4.1, these scores reveal an interesting insight about the task of feedback evaluation: humans seem to agree on what constitutes "good" feedback, but disagree on what constitutes "bad" feedback. The minority categories (which are always the ones describing lower quality feedback) could therefore be more subjective, with annotators showing different attitudes towards what should be considered flawed or unacceptable feedback.

We also compute the proportions of times when the G-Evaluation pipeline agreed with *at least* one human rater on the minority class label. Results are reported in Table 7, and show that agreement for minority clases is equally low between our G-Evaluation pipeline and human raters. The criteria *Factuality* and *Comprehensibility* show slightly higher agreement than all other criteria, echoing the human agreement results of Table 6.

Once more, the agreement pattern of our G-Evaluation pipeline mirrors the agreement pattern of human raters: there is a lack of agreement on minority classes, which could indicate that evaluating low quality feedback is highly subjective. Given such low agreement, it is impractical to evaluate the performance of our G-Evaluation pipeline based on human raters' answers for the task of feedback

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope
Gemma3:4b	**2.325**	3.414	0	0	3.351	0
Gemma3:12b	0	4.737	1.613	0	**10.34**	0
Llama3.2:3b	0.614	3.840	1.642	0.162	2.5	0
Llama3.1:8b	1.124	4.069	0	0	3.682	1.716
Deepseek-r1:32b	1.299	5.156	**1.788**	0	8.108	2.174
Qwen2.5:32b	1.680	**5.321**	1.784	**0.813**	8.982	**7.142**
Minority class label	False	False	False	False	False	True

Table 7: Proportion of agreement cases, i.e. cases where our G-Evaluation pipeline agrees with at least one human rater on the minority class label. All results are in percentage.

evaluation. A venue for further research is to test whether raters could come to a consensus with further discussion or clarification of the evaluation guidelines. If disagreements can be solved, then the task of evaluating low-quality feedback could reveal itself more objective than the low minority agreement scores imply.

4.2 Cross-LLM bias

The coefficients in Table 8 show that our LLM evaluators systematically rate human feedback lower than GPT-generated feedback. The p-values are significant for all models except Gemma3:12b. The average score difference between GPT-generated texts and human texts is highest for Gemma3:4b and Llama3.1:8b.

LLM	Coeff.	p-value
Gemma3:4b	-0.939	<0.001
Gemma3:12b	-0.157	0.09
Llama3.1:8b	-0.929	<0.001
Llama3.2:3b	-0.507	<0.001
Qwen2.5:32b	-0.381	<0.001
Deepseek-r1:32b	-0.482	<0.001

Table 8: Results of the linear mixed-effects model for the *Quality* criterion. Coefficients describe the average score difference of *Quality* between the GPT-generated texts and the human texts.

On the other hand, the scores given by human evaluators show little to no difference of quality between GPT-generated feedback and human feedback, as shown in Figure 2.

Figure 1 shows that the score distributions are also strikingly different, both between human and LLM evaluators, and between LLM evaluators themselves. Gemma3:12b, Qwen2.5:32b and Deepseek-r1:32b give mostly high scores of 4 or 5, while the distribution is flatter for all other models.

Previous studies have reported that LLMs tend to prefer their own output (Chen et al., 2025) or output of models from their own family (Spiliopoulou et al., 2025). Our experiments ratify that LLMs favour AI-generated texts even when not self-generated, justifying the name "cross-LLM bias".

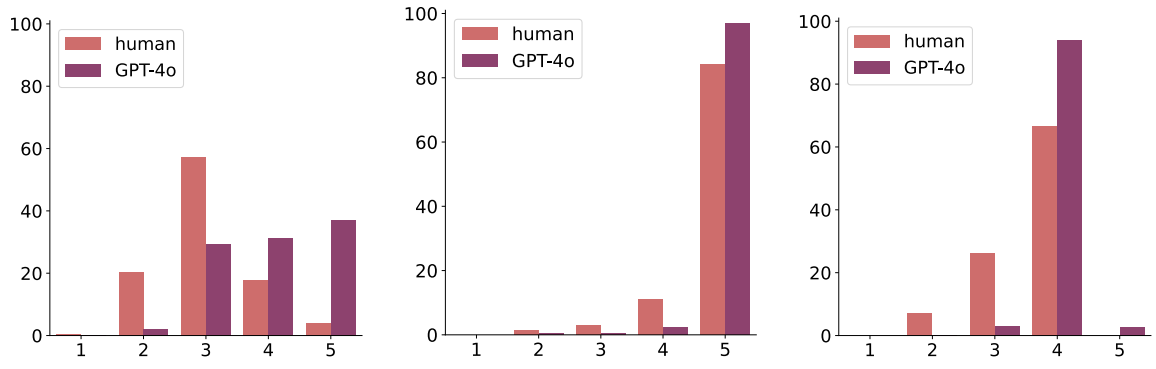
4.3 Alignment between score and reasoning

Figure 3 shows that while models give lower scores when the reasoning provided expresses negative sentiment, the score range remains largely spread out, ranging from 0.0 to 0.7. Only Deepseek-r1-32b displays a slightly more constrained score range. However, it still shows inconsistent scores of up to 0.7. Mode scores for "Negative" reasonings are also high (0.5, 0.6), except for Llama3.1:8b and Qwen2.5:32 (0.3, 0.4).

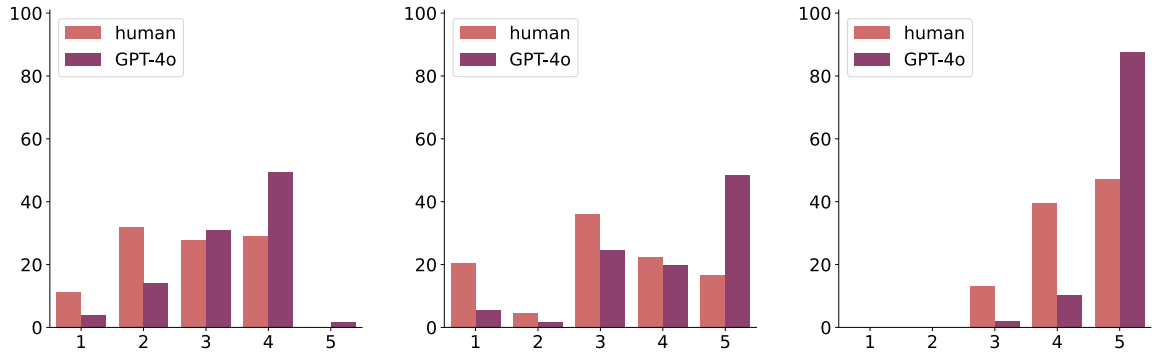
Score range are more constrained for "Neutral" reasonings. However, the scores remain high for most models, spreading from 0.6 to 0.8, with 0.7 and 0.8 being the most common scores. Llama3.2:3b breaks this pattern by displaying an even wider score span, from 0.4 to 0.8. Gemma3:12b also represents a special case for the "Neutral" sentiment, as only one datapoint was annotated as such. We therefore removed it from the results table and plots.

All models show better alignment when the reasoning is "Positive", with mode scores between 0.8 and 1.0, and a score range always higher than the passing threshold (>0.5).

The accuracy of alignment between scores and reasonings varies greatly across models. Gemma3:12b obtains the highest result (98%), despite the mode score for "Negative" sentiment being overly high (0.6). Llama3.2:3b is the lowest



(a) Gemma3:4b (b) Gemma3:12b (c) Qwen2.5:32b



(d) Llama3.2:3b (e) Llama3.1:8b (f) Deepseek-r1:32b

Figure 1: Score distribution for the criterion *Quality* in percentage for each LLM evaluator.

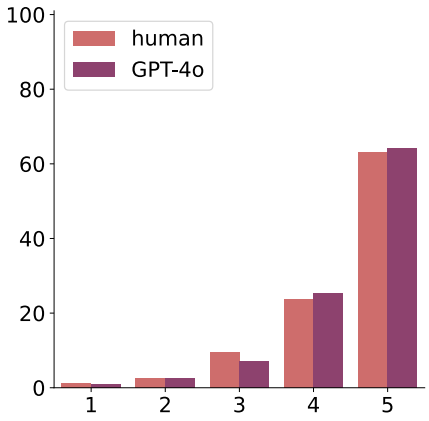


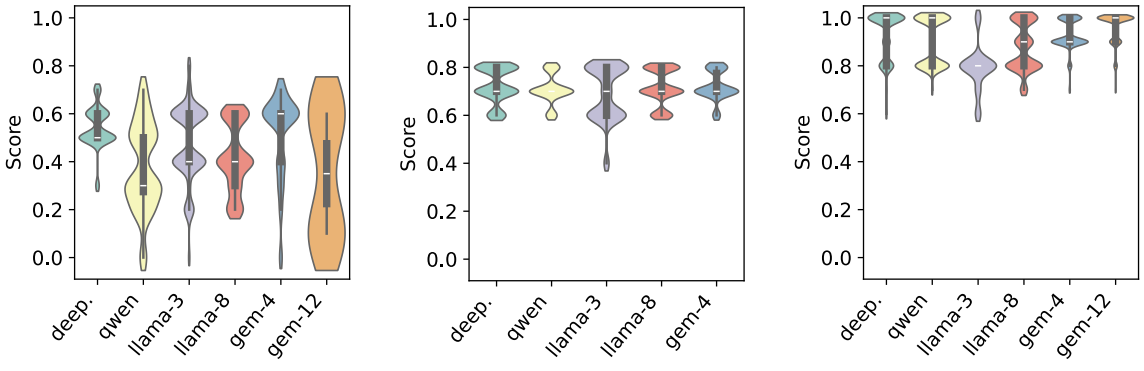
Figure 2: Score distribution for the criterion *Quality* in percentage for human evaluators.

performing model with 53% accuracy, while other models display mixed results (61-83%).

These results suggest that the tested LLMs are capable of attributing high scores to what they consider high quality feedback, but struggle to align their score and reasonings when the feedback is considered lower quality. This is particularly problematic when scores are converted into a binary "True"/"False" label, as even mixed or "Negative" reasonings will give a positive outcome if the threshold is set to the default middle value 0.5. Given that the W&I+WCF contains mostly high quality feedback, the raw agreement results might be inflated by the tendency of LLMs to assign overly high scores. We therefore recommend aligning the threshold based on the scoring pattern of the chosen LLMs. Interestingly, we are facing the same conclusion as Section 4.1: low-quality feedback is more problematic to evaluate than high-quality feedback.

5 Conclusion

In this paper, we show that certain open-source and open-weights LLMs yield results that align with humans in feedback evaluation, but only for majority classes. For minority classes, the agreement between human raters and LLMs (human-LLM) mirrors the agreement among human raters (human-human): evaluators agree on what high-quality feedback is, but disagree on what should be considered low-quality feedback. We also provide evidence that five out of the six LLMs tested suffer from cross-LLM bias and prefer GPT-generated feedback, even when not self-generated. Finally, we expose the tendency of LLMs to attribute high scores even when the reasoning provided is "Neg-



(a) Negative (b) Neutral* (c) Positive

Figure 3: Score distribution per sentiment. For the sake of readability, model names have been abbreviated. deep = Deepseek-r1:32b, qwen = Qwen2.5:32b, llama-3 = Llama3.2:3b, llama-8 = Llama3.1:8b, gem-4 = Gemma3:4b, gem-12 = Gemma3:12b. *Only one datapoint was annotated as "Neutral" for Gemma3:12b over a total of 105. It was therefore removed from the figure.

ative" or "Neutral", providing evidence that there are misalignments between scores and reasonings.

These findings show the complexity of assessing LLM-as-a-judge pipelines as well as the numerous risks of using them. We therefore recommend using G-Evaluation pipelines with great care: pre-experiments should be run to select a model that displays high agreement with human raters across both majority and minority labels. Researchers should also carefully consider the nature of their task, and refrain from using G-Evaluation if the task is highly subjective, as the evaluation of the LLM output becomes impractical without a ground truth. Further, the selected LLMs must show consistency between scores and reasonings. The scoring behaviour of the models must also be tested beforehand, so as to select an appropriate threshold for success. The problem of cross-LLM-bias should also be addressed before LLMs can reliably be used as evaluators, particularly when comparing both human and AI-generated feedback.

In future work, we plan on testing whether some of the disagreements between human raters can be resolved for minority classes, and reflect on whether feedback evaluation can really be "objective". We will also test whether training the models on a subset of annotated feedback comments reduces cross-LLM bias. Finally, we plan on testing how consistent the models are across several rounds of evaluation.

Limitations

While our agreement analysis did account for minority classes, the imbalance of data can still be an issue, as some of the proportions calculated are based on a small number of datapoints. Another limitation of the present paper is that the data for the sentiment analysis was annotated by only one person. Given that sentiment is subjective, the results might vary with a different annotator. Further, the dataset for the sentiment analysis is small, 105 datapoints per LLM.

Ethical considerations

The use of LLM-as-a-judge raises ethical questions that must be discussed. While LLMs can show high raw agreement with human evaluators, their rating behaviour differs widely from humans, due among others to cross-LLM bias and misalignment between scores and reasonings. This could cause important issues if LLMs were to be used as judges in real-world settings, for example to evaluate human teachers or determine the outcome of a language exam for learners. Our G-Evaluation pipeline is built solely for the sake of research and is not meant to be used in any real-world setting.

Acknowledgments

The author of the present paper would like to thank Steven Coyne for his valuable insights and Elena Volodina for her constructive feedback. This work has been supported by Språkbanken - jointly funded by its 10 partner institutions and the Swedish Research Council (2025-2028; project id 2023-00161).

AI Disclosure

☒ Not applicable (no AI tools used)
☐ Assistive use
☐ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: No AI tools were used in preparing this work.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805. Association for Computational Linguistics.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374. Association for Computational Linguistics.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631. Association for Computational Linguistics.

Steven Coyne, Diana Galvan-Sosa, Ryan Spring, Camélia Guerraoui, Michael Zock, Keisuke Sakaguchi, and Kentaro Inui. 2025. Annotating errors in english learners' written language production: Advancing automated written feedback systems. In Alexandra I. Cristea, Erin Walker, Yu Lu, Olga C. Santos, and Seiji Isotani, editors, *Artificial Intelligence in Education*, volume 15880, pages 292–306. Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on LLM-as-a-judge. *The Innovation*, 7(6):101253.

Steve Han, Gilberto Titericz Junior, Tom Balough, and Wenfei Zhou. 2025. Judge's verdict: A comprehensive analysis of LLM judge capability through human agreement. (arXiv:2510.09738).

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Technical report, University of Pennsylvania.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Subhankar Maity and Aniket Deroy. 2025. Leveraging prompt-tuning for bengali grammatical error explanation using large language models. (arXiv:2504.05642).

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, and Masato Mita. 2023. A report on FCG GenChal 2022: Shared task on feedback comment generation for language learners. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 45–52. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondřej Plátek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. GEE! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781. Association for Computational Linguistics.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. 2025. Play favorites: A statistical method to measure self-bias in LLM-as-a-judge. (arXiv:2508.06709).

Martin Vainikko, Taavi Kamarik, Karina Kert, Krista Liin, Silvia Maine, Kais Allkivi, Annekatrin Kaivapalu, and Mark Fishel. 2025. Paragraph-level error correction and explanation generation: Case study for estonian. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 953–967. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-preference bias in LLM-as-a-judge.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994. Association for Computational Linguistics.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Ying Shen, Peng Xing, Zishan Xu, Guo Cheng, and Wenhao Jiang. 2025. EXCGEC: A benchmark for edit-wise explainable chinese grammatical error correction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25678–25686.

A Example datapoint from W&I+WCF

Example datapoint	
Learner sentence	It *became* a big problem now.
Correction	It has become a big problem now.
EXPECT tag	Verb tense
ERRANT tag	R:VERB:TENSE
WCF tag	Tense: Perfect
Feedback setting	Tagless (*GPT-feedback*)
Feedback	The verb 'became' is in the past tense, but the word 'now' indicates the present. To match the present time, the present perfect tense 'has become' is needed. Change 'became' to 'has become' to correctly match the present time indicated by 'now'.
Rater ID	1
Relevance	True
Factuality	True
Causality	True
Actionability	True
Comprehensibility	True
Scope	False
Quality	4

Table 9: Example datapoint from the W&I+WCF dataset.

B Criteria description

Criterion	Definition
Relevance	Is the feedback relevant to the error? A feedback comment is relevant to the error if it attempts to address the issue represented by the highlighted spans in the Source and Corrected sentences. A feedback comment can refer to words outside the highlight, such as to establish context or a reason to make a correction. However, the "error" that is focused on should be within the highlight. Note that a feedback comment can be "relevant" but incorrect or otherwise flawed. Use Yes If: The feedback identifies an error in the highlighted span. The feedback makes some attempt to describe or address the error. Use No If: The feedback comment seems to be addressing a different error outside of the highlight. The feedback comment is nonsensical or does not seem to apply to the sentence. The feedback comment does not talk about a language error, but is focused on something else like content or praising the learner. The feedback seems like a response to the text instead of an instructor's constructive writing feedback (this would also impact the "Unnecessary Content" annotation).
Factuality	Is the information in the feedback comment factually correct? A feedback comment is factual if there are no inaccurate statements about the error, the possible correction(s), or any rule of English mentioned in the feedback text. Use Yes If: Statements referring to the source sentence, such as "you are using A with B," are accurate. The error described by the feedback is indeed an error. The English language rules mentioned in the feedback are valid. Statements like 'must' or "usually" have the correct scope. Use No If: The feedback asserts a rule that is not real or valid. The feedback overgeneralizes a rule of English.
Causality	Does the feedback comment explain what is wrong and why? A feedback comment "explains what is wrong and why" if it successfully identifies the error and provides at least a simple explanation for it. This may or may not involve a grammatical keyword like 'past tense.' A simple example of feedback that does not explain the error is a grammar-check system that provides only underlines and click-to-fix replacements for complex changes. Learners can observe the suggestions and learn, but there is no explicit explanation.The necessary complexity of an explanation varies with error type. Simple errors like spelling can be "explained" with a basic line like "This word may be misspelled". Use Yes If: The feedback identifies a specific grammar point that was attempted unsuccessfully. The feedback provides a pattern that describes the right way to form a construction. The feedback establishes that a certain combination of words is not natural, ideally with a natural alternative or reference to a certain category of words that do or do not fit. Use No If: The feedback consists only of an edit suggestion like "replace A with B," "remove this word," or "add a comma here." The feedback is very vague or open-ended, such as "This is wrong."

Table 10: Part 1: binary criteria of good feedback, along with their defining question and detailed description. Based on Coyne et al. (2025).

Criterion	Definition
Action.	Does the feedback comment explain what to do to fix the error? A feedback comment "explains what to do" if it provides either a concrete edit suggestion or a hint related to the error. Note that a later field will ask which approach was used. Use Yes If: The feedback provides a direct word replacement, deletion, or insertion, such as "add "to" to make this a to-infinitive" or "Remove the extra comma." The feedback provides one viable candidate edit of many, e.g., "replace "bigly" with another adverb such as greatly." The feedback makes a general suggestion that fixes a broad problem, such as "add a verb to complete this sentence." The feedback outlines a grammatical change but does not go into specifics, e.g., "change this verb to the past progressive." Use No If: The feedback describes an error but does not communicate how to fix it. The feedback gives a suggestion that is very vague, such as 'try again.' Note that a broad error may require a broad suggestion, but it should at least be based on the type of error in the sentence. The feedback tells the learner to consult external resources like a dictionary. This is one form of "what to do," but does not explain how to fix the error.
Comp.	Is this feedback comprehensible to the assumed learner? This annotation should reflect your judgement on whether the "assumed learner" - a CEFR B1 or B2 learner who has used academic resources, is likely to understand the comment if they try to read it. These learners can be assumed to know a certain amount of metalinguistic words, and may even be interested in keywords related to an error they made, but there is a limit to how complex a feedback comment can be before it becomes too difficult to be effective. Learners may be intimidated by a long and seemingly difficult comment and not even read it. The best way to measure this is to ask students about feedback they received on their own writing. However, it is not practical in our current research. We are thus asking teachers whether the comment seems reasonably comprehensible for this kind of learner. Use Yes If: The feedback uses a relatively simple way to explain the error and/or correction, given the options. The feedback is not overwhelmingly long or complex. Use No If: The feedback uses linguistic or grammatical terms that even an academic language learner would not know, or uses a large number of terms that would be overwhelming. The feedback has unnecessarily high length or syntactic complexity.
Scope	Does the feedback comment contain unnecessary or out-of-scope content? A good feedback comment is relatively minimal, providing guidance without extraneous information. Use this annotation to flag comments that have unnecessary, questionable, or out-of-scope content. Use Yes If: All the content in the feedback pertains to the error, the way to resolve it, or the concept the learner should learn or review. Use No If: The feedback has something that only makes sense in an AI chat, like 'Let me know if I can help you with anything else!' The feedback contains an assumption about the learner or task, such as nationality, first language, real-world contexts like a class, etc. The feedback is about something other than language correctness, such as the quality of a story or the strength of an argument. The feedback is somehow rude or offensive. The feedback contains praise for the learner for trying hard or for correct text elsewhere. Praise makes sense in many feedback contexts, but not the setting we are building. Please cooperate and flag praise in feedback text.

Table 11: Part 2: binary criteria of good feedback, along with their defining question and detailed description. Based on Coyne et al. (2025). "Comp." and "Action." stand for "Comprehensibility" and "Actionability" respectively.

Criterion	Definition
Quality	Overall, how good is this feedback comment? Overall Quality is a 1-5 measure of a rater's impression of the feedback comment. We place it last so that all prior questions should be answered first to help inform this judgement. Nevertheless, this field is not a direct reflection of prior questions, and there is no formula for deducting points for each "undesirable" answer above it. The score should be attributed following the below guidelines: 1/5: Irrelevant or useless comment. 2-3/5: Potentially valid but flawed comment. 4-5/5: The comment is likely to help the learner if interacted with.

Table 12: Criterion of good feedback *Quality*, along with its defining question and detailed description. Based on Coyne et al. (2025).

C Raw agreement scores per rater

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	0.942	0.787	0.996	0.978	0.458	0.947	0.236
Gemma3:12b	0.985	0.885	0.988	1.000	0.928	0.982	0.981
Llama3.2:3b	0.160	0.480	0.293	0.740	0.723	0.918	0.115
Llama3.1:8b	0.946	0.599	0.995	0.937	0.589	0.885	0.335
Deepseek-r1:32b	0.845	0.682	0.710	0.993	0.938	0.970	0.578
Qwen2.5:32b	0.867	0.715	0.350	0.945	0.907	0.970	0.170

Table 13: Raw agreement per criterion between the LLMs and rater 1.

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	0.971	0.780	0.990	0.995	0.440	0.962	0.278
Gemma3:12b	0.994	0.948	0.984	1.000	0.971	0.998	0.691
Llama3.2:3b	0.156	0.469	0.335	0.721	0.732	0.932	0.112
Llama3.1:8b	0.977	0.568	0.995	0.946	0.577	0.923	0.335
Deepseek-r1:32b	0.872	0.721	0.745	0.991	0.984	0.982	0.585
Qwen2.5:32b	0.910	0.759	0.363	0.946	0.946	0.991	0.174

Table 14: Raw agreement per criterion between the LLMs and rater 2.

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	0.949	0.796	0.949	0.981	0.454	0.972	0.301
Gemma3:12b	0.998	0.935	0.948	0.997	0.963	0.998	0.853
Llama3.2:3b	0.156	0.470	0.325	0.743	0.742	0.943	0.057
Llama3.1:8b	0.957	0.617	0.953	0.925	0.590	0.892	0.378
Deepseek-r1:32b	0.868	0.695	0.677	0.993	0.982	0.992	0.702
Qwen2.5:32b	0.891	0.735	0.403	0.948	0.942	0.988	0.083

Table 15: Raw agreement per criterion between the LLMs and rater 3.

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	0.970	0.970	0.980	1.000	0.475	0.950	0.340
Gemma3:12b	0.994	0.921	0.975	1.000	0.959	0.986	0.182
Llama3.2:3b	0.167	0.457	0.356	0.725	0.723	0.928	0.320
Llama3.1:8b	0.969	0.592	0.980	0.937	0.577	0.906	0.232
Deepseek-r1:32b	0.890	0.694	0.732	0.995	0.982	0.978	0.254
Qwen2.5:32b	0.930	0.746	0.408	0.950	0.944	0.984	0.532

Table 16: Raw agreement per criterion between the LLMs and rater 4.

D Krippendorff's alpha per rater

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	0.054	-0.029	-0.001	-0.008	-0.280	-0.028	-0.204
Gemma3:12b	-0.007	-0.006	-0.005	N/A	0.152	-0.008	0.005
Llama3.2:3b	-0.708	-0.237	-0.545	-0.148	-0.046	-0.042	-0.465
Llama3.1:8b	-0.027	-0.133	-0.002	-0.032	-0.134	-0.005	-0.103
Deepseek-r1:32b	-0.042	-0.030	-0.169	-0.003	0.146	-0.014	0.074
Qwen2.5:32b	-0.023	-0.005	-0.480	-0.027	0.173	0.086	-0.337

Table 17: Krippendorff's alpha per criterion between the LLMs and rater 1.

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	-0.014	-0.050	-0.002	-0.002	-0.360	-0.015	-0.093
Gemma3:12b	-0.001	0.095	-0.007	N/A	0.186	0	-0.026
Llama3.2:3b	-0.724	-0.321	-0.497	-0.161	-0.141	-0.034	-0.452
Llama3.1:8b	0.122	-0.230	-0.002	-0.027	-0.236	0.005	-0.064
Deepseek-r1:32b	-0.040	-0.081	-0.145	-0.004	-0.007	-0.008	0.055
Qwen2.5:32b	-0.007	-0.068	-0.466	-0.027	0.091	0.282	-0.298

Table 18: Krippendorff's alpha per criterion between the LLMs and rater 2.

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	-0.019	-0.095	-0.022	-0.010	-0.384	-0.017	-0.360
Gemma3:12b	0	0.060	-0.026	-0.001	0.186	0	0.132
Llama3.2:3b	-0.727	-0.328	-0.448	-0.134	-0.111	-0.028	-0.682
Llama3.1:8b	-0.021	-0.212	-0.023	-0.038	-0.200	-0.056	-0.247
Deepseek-r1:32b	-0.069	0.120	-0.127	-0.003	0.258	-0.003	0.056
Qwen2.5:32b	-0.056	-0.094	-0.348	0.035	0.117	-0.005	-0.625

Table 19: Krippendorff's alpha per criterion between the LLMs and rater 3.

LLM	Relevance	Factuality	Causality	Action.	Comp.	Scope	Quality
Gemma3:4b	0.078	-0.042	-0.009	-0.004	-0.330	-0.018	0.040
Gemma3:12b	-0.002	0.079	0.113	N/A	0.128	-0.006	-0.569
Llama3.2:3b	-0.708	-0.309	-0.445	-0.158	-0.147	-0.036	-0.108
Llama3.1:8b	-0.015	-0.170	-0.009	-0.032	-0.236	-0.011	-0.023
Deepseek-r1:32b	-0.025	-0.116	-0.092	-0.018	-0.009	0.133	-0.270
Qwen2.5:32b	0.013	-0.056	-0.386	-0.025	0.086	0.174	0.033

Table 20: Krippendorff's alpha per criterion between the LLMs and rater 4.

E Alignment scores

LLM	Accuracy	Average score			Mode score		
		Negative	Neutral	Positive	Negative	Neutral	Positive
Gemma3:4b	0.75	*0.521*	*0.720*	0.930	*0.6*	*0.7*	0.9
Gemma3:12b	**0.98**	0.350	N/A*	0.963	*0.6*	N/A*	1.0
Llama3.2:3b	0.53	0.477	*0.690*	0.790	*0.6*	*0.8*	0.9
Llama3.1:8b	0.61	0.417	*0.714*	0.874	0.4	*0.7*	0.8
Deepseek-r1:32b	0.67	*0.543*	*0.724*	0.930	*0.5*	*0.7*	1.0
Qwen2.5:32b	0.83	0.375	0.*708*	0.901	0.3	*0.7*	1.0

Table 21: Accuracy of the reasoning alignment, as well as the average and mode score per sentiment. Scores that misalign with the reasonings are marked in italics. *Only one datapoint was annotated as "Neutral" out of 105 for Gemma3:12b. It was therefore removed from the table above.

Reading Ahead: Pace-Based Prediction for Real-Time Tracking of Children's Oral Reading in the Browser

Bohdan Podziubanchuk and **Einar Meister** and **Tanel Alumäe**
Department of Software Science
Tallinn University of Technology
Estonia
{bohdan.podziubanchuk,einar.meister,tanel.alumae}@taltech.ee

Abstract

Interactive reading tutors can support children by highlighting the word currently being read. However, when reading position is determined with the help of a streaming automatic speech recognition (ASR) system, the highlight necessarily lags behind the reader because a word can only be confirmed after it has been emitted by the recogniser. We present a browser-based Estonian reading tutor that reduces this delay by predicting the reader's current position between ASR updates. The proposed prediction layer estimates the reader's pace in voiced seconds per letter and uses the voiced time elapsed since the last confirmed word to extrapolate over the expected durations of subsequent words. Voice activity detection prevents pauses from advancing the predicted position. Compared to ASR-based alignment alone, this approach increases exact word-level tracking accuracy from 49.7% to 57.7% and reduces the median word-transition delay from 264 ms to 87 ms. Both ASR and voice activity detection run locally in the browser using WebAssembly, ensuring that no audio data leave the user's device.

1 Introduction

Systems that listen to a child read aloud need to continuously track which word the child is currently reading. This position determines the visual feedback shown by the system, typically by highlighting the current word to help the child keep track of their place in the text. For this feedback to be useful, the highlighted position should closely match the child's actual reading position. Highlighting the wrong word may be confusing or distracting and can even disrupt reading by causing the child to lose their place. Similarly, if the display does not advance after a word has been read, the child may interpret this as the system failing to recognize or accept the word.

Tracking reading position from streaming automatic speech recognition (ASR) is subject to inherent delay. A recogniser cannot identify a word before sufficient acoustic evidence has been observed, while architectural factors such as audio buffering and model lookahead introduce additional latency. In our system, the ASR model begins emitting the tokens of a word approximately 640 ms after the child starts pronouncing it. Delays of this size are perceptible: response times above about 100 ms are generally noticed by users of interactive systems (Miller, 1968; Card et al., 1983). Any matcher that waits for these tokens is therefore delayed by design, causing the highlighted position to trail the child's actual reading position.

Li and Mostow (2012) measured this effect for an English reading tutor and showed that tracking accuracy improves substantially if the tutor is allowed to lag behind the recogniser's frontier: exact accuracy rose from 58% at zero lag to 70% at 200 ms and 76% when the final hypothesis was used. Accuracy and timeliness were traded against each other.

In this paper, we aim to reduce this delay rather than trade it for accuracy. Instead of waiting for further recogniser evidence, we predict the reader's current position. Oral reading proceeds at a pace that can be estimated from the words the recogniser has already confirmed. Our system estimates this pace in voiced seconds per letter and uses it to carry the position forward from the last confirmed word, by at most two words. Two supporting components are required: voice activity detection, which prevents silence from advancing the position, and a correction for the latency of the recogniser's token timestamps. Throughout the paper, we use *tracking* to refer to estimating the reader's current position, which in our system

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 288–294, November 19–20, 2026, University of Tartu Library

combines recogniser evidence with a pace-based prediction.

We demonstrate the approach in a web application for Estonian, where both the speech recogniser and the voice activity detector run locally in the browser using WebAssembly. Since no audio is transmitted to a server, the system offers stronger privacy guarantees. On 134 recordings from nine children, manually annotated at the word level, the prediction layer improves exact tracking accuracy by 8.0 percentage points and reduces the median transition delay by approximately a factor of three, with improvements observed for every child.

2 Related work

Reading tutors that listen have a long history. Project LISTEN (Mostow and Aist, 1999) used ASR to follow children reading aloud, detect miscues and give feedback, and established most of the evaluation practice in this area (Mostow, 2006). Position tracking was recognised early as a distinct problem from recognition accuracy: word error rate does not measure whether the tutor knows where the reader is (Mostow, 2006).

The work closest to ours is the evaluation of real-time tracking by Rasmussen et al. (2011) and Li and Mostow (2012). The latter distinguishes tracking computed from the recogniser's successive partial hypotheses, which is what a live tutor must use, from tracking computed after the fact from the final hypothesis, which can overestimate real-time performance. It defines exact accuracy as the share of time the displayed position matches the true position and near accuracy as the share within one word. We adopt both metrics. The main difference is that Li and Mostow (2012) improve tracking accuracy by introducing additional lag, whereas our goal is to reduce this lag by predicting the reader's position during the recognition delay.

A recent alternative avoids the cascade entirely: Sunder et al. (2023) train a pointer network that predicts positions in the reference text directly from streaming audio, reporting accuracies of 77.1% on CMU Kids and 65.3% on Reading Races. That approach needs aligned training data in the target language and a trained model at inference time. Our setting is a cascaded one, built on an existing streaming recogniser, which is what is available for a language the size of Estonian and what can be executed inside a browser.

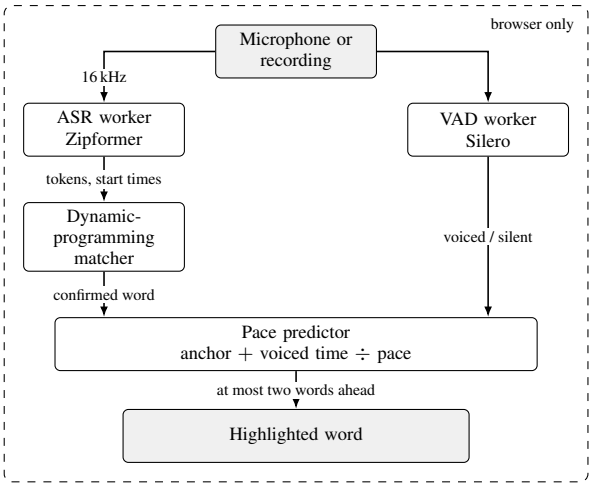


Figure 1: Architecture. Recognition and voice activity detection run as WebAssembly workers inside the browser. The matcher yields the confirmed position; the pace predictor carries it forward by the voiced time that has passed since the confirmed word began.

The task is also related to score following in music, where a system tracks a performer's position in a known score in real time (Orio et al., 2003). A common assumption in score following is that the performance follows a tempo that can be estimated from recent observations and extrapolated forward. Our approach applies a similar idea to oral reading: we estimate the reader's current pace and use it to predict their position between recognition updates. To our knowledge, this principle has not previously been applied to real-time tracking of oral reading.

3 System

The application is a single page in a Next.js web application. The user loads a reading text, presses record, and reads aloud; the word the system believes is being read is highlighted. Everything described below runs in the browser: ASR and voice activity detection are WebAssembly modules in dedicated workers, and no audio or transcript is sent anywhere. Figure 1 shows how the parts fit together. Source code is available at `https://github.com/Danbog32/estonian-reading-tracker` and a demo video at `https://github.com/Danbog32/estonian-reading-tracker#demo`.

The reading tracking system will be integrated into a reading tutor project funded by the Estonian Ministry of Education and Research (EKTR29).

3.1 Speech recognition

For ASR, we use a general-purpose Estonian streaming model[1] based on the Zipformer neural transducer architecture (Yao et al., 2024), trained using the Icefall toolkit[2]. The model has about 67 million parameters and was trained on roughly 1334 hours of manually transcribed Estonian speech from the TalTech Speech Dataset 1.0[3], around 4000 hours of automatically transcribed Estonian broadcast data, and a 500-hour subset of the Gigaspeech dataset (Chen et al., 2021), which was included so that the model can transcribe English terms often embedded in Estonian sentences. The training data contain adult speech only; the model has not been adapted to children's speech.

Audio is captured through the Web Audio API, resampled to 16 kHz and passed to the recogniser in 100 ms chunks. Recognition is executed with sherpa-onnx[4] compiled to WebAssembly, with greedy decoding. The recogniser emits partial hypotheses as audio arrives and a final hypothesis when an utterance ends.

For each hypothesis the worker exposes the underlying token sequence together with the start time of every token and its log probability. The token start times are the anchor the prediction layer needs: they say when a sound was produced, not when the recogniser reported it.

3.2 Matching the transcript to the text

The reading text is known in advance, so matching is a monotonic alignment problem rather than open recognition. On every recogniser event the matcher aligns the last ten transcript words against a window of the reading text that extends three words behind the current position and up to the end of the sentence ahead of it, using dynamic programming.

Word similarity combines ordinary edit distance with a phonetic key that collapses the contrasts Estonian beginning readers most often blur: voiced and voiceless stop pairs, doubled letters, and vowel quality. A child who produces *leib* for *loeb* scores 0.5 on letters but 1.0 on the phonetic key. Each candidate alignment receives a confidence score combining the quality of the matched pairs, how much of the transcript window found a match, the total alignment cost and a small term (weight 0.06) that favours alignments advancing the position further, with penalties for skipped and unmatched words. The position advances only when that score clears a threshold, and never moves backwards. This matcher is the system's confirmed position and is used unchanged as the baseline in Section 4.

3.3 Pace prediction

The matcher can confirm a word only after the recogniser has produced evidence for it. The prediction layer operates on top of the matcher and does not modify its output.

Anchor. The prediction is anchored at the most recently confirmed word and at the estimated time when the child began pronouncing it. This time is obtained from the ASR token start times after correcting for the systematic latency of the recogniser's timestamps (Section 4.4).

Pace. For each confirmed word, the layer estimates the reading pace as the voiced duration of the word divided by its number of letters. Estonian spelling is close to phonemic, so the number of letters approximates the number of speech sounds. Within a child, word duration in our data is approximately proportional to word length in letters. The current pace is the median of the eight most recent samples. A median is used because a single word on which the child hesitates would substantially increase a mean and delay all subsequent predictions. The history length of eight words was fixed before the evaluation as a compromise between adapting to changes in pace and stability of the estimate. Words on which the child lingers, for example by sounding them out or prolonging them, are referred to as held words. They are detected as samples exceeding 2.5 times the current pace and are excluded from the history.

Projection. Starting from the anchor, the layer uses the voiced time elapsed since the anchor word began to estimate how far the child has progressed. The expected duration of each following word is computed as the current pace multiplied by its length in letters. The elapsed voiced time is then allocated to successive words, and the word in which this time budget is exhausted is selected as the predicted position. The projection is recomputed from the anchor after every recogniser event

[1] https://huggingface.co/TalTechNLP/streaming-zipformer.et-en

[2] https://github.com/k2-fsa/icefall

[3] https://cs.taltech.ee/staff/tanel.alumae/data/est-pub-asr-data/

[4] https://github.com/k2-fsa/sherpa-onnx

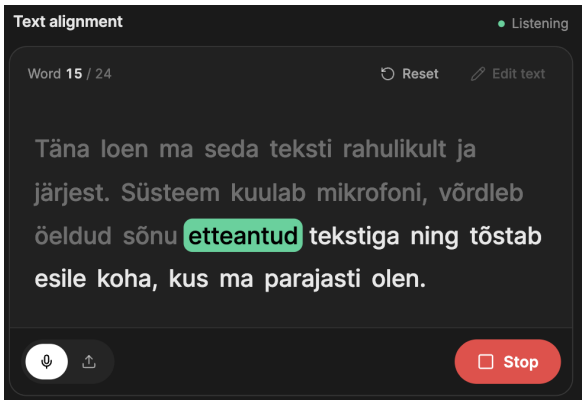


Figure 2: The reading view during live microphone input. Words already read are dimmed, the current word is highlighted, and the source switch and the start/stop control sit below the text.

and every 100 ms rather than updated incrementally. As a result, if a word is confirmed later than expected, the next prediction already takes into account all voiced time since the onset of the anchor word. The prediction is limited to at most two words beyond the anchor, that is, at most one word ahead of the matcher's current position, and is never placed behind the position confirmed by the matcher.

Pauses. Only voiced time contributes to the projection. A separate Silero voice activity detector (Silero Team, 2021) runs alongside the recogniser using a 0.15 s minimum silence duration. This is short enough to capture between-word pauses, whose median duration in our data is 259 ms. During detected silence, the elapsed-time budget remains unchanged and the highlight therefore does not advance.

Repetitions and self-corrections. The matcher aligns the transcript against a window that begins three words before the current position, so repeated words are absorbed as insertions and do not move the confirmed position backwards. If the child repeats or prolongs the anchor word, the recogniser produces further evidence for it; the layer then treats the word as held, extends its expected duration, and the prediction returns to the matcher's position. This is the main source of rollbacks. After a self-correction, the matcher may confirm several words in a single update, in which case the highlight advances by more than one word at once.

3.4 Interface

Figure 2 shows the interface. The user pastes the reading text, chooses between the microphone and an uploaded recording, and presses start. The current word is highlighted. A word to which the prediction layer has moved ahead of the matcher is shown in a lighter shade, so that a prediction that later proves wrong is withdrawn less conspicuously. The panel also reports the measured reading speed in words per minute. If voice activity detection fails to load, the panel reports that the speed is unavailable.

4 Evaluation

4.1 Data

Estonian read-speech data were collected from children aged 6 to 11, irrespective of their native language. The participants' native languages include Estonian, Russian, English, Nepali, and Yoruba. To date, 101 children (54 girls and 47 boys) have been recorded, yielding 2,551 recordings and approximately 14 hours of speech. The recordings were conducted in a small, quiet room at the children's schools using a web-based application and a close-talking headset microphone. Audio was stored in uncompressed mono WAV format at a sampling rate of 44.1 kHz and 16-bit resolution. The ongoing annotation process covers both phoneme- and word-level annotation, as well as the classification of speech errors, including substitutions, omissions, insertions, and repetitions. Written informed consent was obtained from a parent prior to recording. Speech data collection was conducted as part of a reading tutor project. The data collection protocol was approved by the ethics committee of Tallinn University of Technology.

We evaluate on all recordings for which manual word-level annotation was available at the time of the experiments: 134 recordings from nine children, five girls and four boys, reading Estonian texts aloud, 43.8 minutes in total. Word boundaries were annotated manually in Praat TextGrids, yielding 2,072 labelled word intervals. Since manual annotation at the word level is time-consuming, the remaining recordings have not yet been annotated at this level. The recordings cannot be released, as the consent covers research use only; the tracking algorithms are available with the source code.

	Matcher	+ pace
Exact accuracy	49.7%	**57.7%**
Within one word	79.0%	**83.5%**
Median delay	264 ms	**87 ms**
False jumps / min	1.33	**0.64**
Rollbacks / min	0.00	0.36
Ahead of the child	6.0%	13.5%
Behind the child	33.4%	17.9%

Table 1: Tracking quality over 134 recordings from nine children.

Variant	Exact	Delay
Full system	57.5%	87 ms
One pace for every child	48.6%	80 ms
No voice activity detection	53.5%	90 ms
No latency correction	53.1%	198 ms
Lead capped at one word	49.2%	257 ms
Lead capped at three words	57.1%	79 ms
Also held behind the matcher	51.9%	161 ms

Table 2: Ablations, one component removed or modified at a time. These runs predate a later refinement of the matcher, which changes two recordings; the full system therefore reads 57.5% here and the matcher alone 49.2%.

To characterise recognition quality on this subset, we computed the word error rate (WER) and character error rate (CER) of the recogniser's final hypotheses. Measured against the words the children actually produced, as transcribed in the annotation and excluding false starts, the WER is 22.0% and the CER 11.2%; the per-child WER ranges from 8.5% to 37.0%. Measured against the prompt text, the WER is 25.5% and the CER 13.3%. The difference mainly reflects reading errors, as the children's productions differ from the prompt text in 7.0% of words.

The recordings are replayed through the same recogniser worker the application uses, and all systems consume the same captured recogniser events, so differences between them come only from the tracking logic and not from variation in recognition. Positions are sampled every 100 ms while a labelled word is being spoken.

4.2 Metrics

Following Li and Mostow (2012) we report *exact accuracy*, the proportion of sampled timesteps at which the highlighted word is the word being read, and *within one word* (near accuracy in Li and Mostow, 2012), the share at which it is off by at most one word. We add three measures that matter for a live interface. *Median delay* is the time from the onset of a word until the highlight is at or past it. *False jumps* per minute counts movements to a word the child is not reading. *Rollbacks* per minute counts movements of the highlight to an earlier word; the matcher alone never rolls back, while a predictor can.

4.3 Results

Table 1 shows the effect of the prediction layer. Exact accuracy rises by 8.0 points and median delay falls from 264 ms to 87 ms. False jumps fall by half: predicting the next word from pace is more reliable than accepting a weak acoustic match for it. The cost is 0.36 rollbacks per minute, roughly one every three minutes.

Exact accuracy, within-one-word accuracy, and median delay improve for all nine children. At the level of individual recordings, performance improves for 102 recordings, degrades for 21, and remains unchanged for 11. The largest gains are observed for the children for whom the baseline matcher performs worst: among the three children with baseline exact accuracy below 30%, the prediction layer improves accuracy by 14–22 pp.

With pace prediction, the highlight is ahead of the child's true position more often (13.5% *vs* 6.0%), but behind substantially less often (17.9% *vs* 33.4%). Since the delay metric considers the highlight to have arrived once it reaches or passes the current word, part of the reduction in delay reflects this increased forward bias. However, exact accuracy penalizes predictions that are ahead and behind equally, and its improvement therefore provides separate evidence that the prediction layer tracks the reading position more accurately overall.

4.4 Ablation study

Table 2 reports the results when individual components are removed or modified. The largest decrease occurs when a single constant pace is used for all children: in this condition, exact accuracy falls below that of the matcher alone. This is consistent with the variation in reading pace in our data, where the time per letter ranges from 88 to 257 ms across recordings. Removing voice activity detection reduces exact accuracy by 4.0 percentage points, and removing the latency correction reduces it by 4.4 points and increases the median delay by 111 ms.

The latency correction is based on a compari-

son of token start times with the manual annotations, which showed that the onsets reported by the recogniser follow the annotated onsets by approximately half a second. For each child, the layer subtracts an offset estimated from the other eight children only, so that no correction is derived from the child's own data. Without this correction, the onset of the anchor word is estimated approximately half a second too late, and most of the improvement is lost.

When the lead is limited to one word, the results are identical to those of the matcher (49.2% and 257 ms in the same run), which indicates that the layer does not alter the output when it cannot advance beyond the matcher. Allowing a lead of three words does not improve accuracy compared with two words but increases the rate of rollbacks from 0.36 to 2.97 per minute; we therefore use a limit of two words.

An earlier version of the layer could also keep the highlight on a word behind the matcher when the estimated pace indicated that the child had not yet finished it, with the aim of handling long words that are read slowly. This variant reduced exact accuracy by 5.6 percentage points. A likely explanation is that onsets estimated from token times are accurate only to within a few hundred milliseconds, which is sufficient to retain the wrong word. The observed improvement therefore appears to result from advancing the position rather than from holding it.

5 Limitations

The evaluation covers one language, one recogniser and one group of readers. The approach itself assumes only a streaming recogniser that reports token times, but the values used here, the length of the pace history, a lead of two words and the latency offset, were established for this configuration and would have to be re-established for another. The configuration of the layer was developed on the same recordings that were used for evaluation, and its effect should therefore be confirmed on new data. The recogniser was trained on adult speech only, which is a likely contributor to the variation in word error rate across children.

Pace prediction relies on the assumption that reading speed can be extrapolated from recent observations. When a reader pauses to decode a difficult word, the predicted position continues to advance according to the estimated pace and is corrected only after the recogniser produces sufficient evidence. Limiting the prediction lead to two words and using a lighter highlight for predicted words reduce the visibility of such corrections but do not eliminate them. A related limitation occurs at the beginning of a text, when insufficient observations are available to establish a reliable pace estimate. An unusually slow first word may therefore bias the estimate until additional observations become available.

The present study does not examine how real-time highlighting of the reading position affects learners, particularly when the highlighted position is occasionally inaccurate. Such feedback may introduce distractions, encourage over-reliance on visual cues, or lead to unnatural reading habits. Future work will investigate whether real-time highlighting improves reading performance and learning outcomes, and how these benefits are affected by highlighting errors.

6 Conclusion

Most of the delay in a reading tracker originates from ASR rather than from the matching process itself. A streaming ASR system typically emits a word several hundred milliseconds after the child begins or completes saying it, so any matcher that waits for this evidence will necessarily lag behind, regardless of how accurate the matching is. Much of this delay can be recovered by estimating the child's reading pace and using the elapsed voiced time to predict the current position. The same mechanism also improves tracking accuracy, since advancing based on the estimated reading pace can be more reliable than accepting a weak acoustic match. Three components are particularly important: a voice activity detector with a short silence threshold, which prevents pauses from advancing the predicted position; a measured correction for the recogniser's systematic timestamp latency; and a median over recent words, which makes the pace estimate robust to individual words that are read unusually slowly.

Acknowledgments

This work was supported by the Estonian Ministry of Education and Research under project EKTR29, and by the Estonian Centre of Excellence in Artificial Intelligence (EXAI), funded by the Estonian Ministry of Education and Research grant TK213.

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: AI assistants (Claude, Anthropic) were used to discuss the reviewers' comments, to check the description of the system against its source code, and to improve grammar, style and clarity of the revised text. The method, experiments and results are the authors' own, and all text was reviewed and edited by the authors.

References

Stuart K. Card, Thomas P. Moran, and Allen Newell. 1983. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*. ISCA.

Yuanpeng Li and Jack Mostow. 2012. Evaluating and improving real-time tracking of children's oral reading. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*.

Robert B. Miller. 1968. Response time in man-computer conversational transactions. In *Proceedings of the AFIPS Fall Joint Computer Conference*, pages 267–277.

Jack Mostow. 2006. Is ASR accurate enough for automated reading tutors, and how can we tell? In *Proceedings of Interspeech*, pages 837–840.

Jack Mostow and Gregory S. Aist. 1999. Giving help and praise in a reading tutor with imperfect listening because automated speech recognition means never being able to say you're certain. *CALICO Journal*, 16(3):407–424.

Nicola Orio, Serge Lemouton, and Diemo Schwarz. 2003. Score following: state of the art and new developments. In *Proceedings of the Conference on New Interfaces for Musical Expression (NIME)*, pages 36–41.

Morten Hjelm Rasmussen, Jack Mostow, Zheng-Hua Tan, Børge Lindberg, and Yuanpeng Li. 2011. Evaluating tracking accuracy of an automatic reading tutor. In *Proceedings of the ISCA SIG Workshop on Speech and Language Technology in Education (SLaTE)*.

Silero Team. 2021. Silero VAD: pre-trained enterprise-grade voice activity detector. https://github.com/snakers4/silero-vad.

Vishal Sunder, Beulah Karrolla, and Eric Fosler-Lussier. 2023. End-to-end real time tracking of children's reading with pointer network. *arXiv preprint arXiv:2310.11486*.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *International Conference on Learning Representations (ICLR)*.

ABC Diagnostics: A Web-Based Tool for Explainable CEFR Classification

Ricardo Muñoz Sánchez[1] **Kostiantyn Kucher**[2] **Filip Kayar**[2] **Elena Volodina**[1]

[1] Gothenburg University
{ricardo.munoz.sanchez,elena.volodina}@gu.se

[2] Linköping University
{kostiantyn.kucher,filip.kayar}@liu.se

Abstract

A core concept in the field of second language assessment is that of construct validity. That is, ensuring we are testing for what we claim to be testing for. The introduction of neural networks for automated assessment clashes with this concept due to their lack of transparency. Moreover, explainability methods tend to require technical expertise, increasing the risk of automated assessment systems being used to replace teachers rather than to support them. In this paper we introduce the ABC Diagnostics Tool, a web-based system that provides an interactive user interface to display feature importance for essay classification systems. We showcase it by exploring the role of given names and names of places when doing CEFR level classification of Swedish learner essays. We notice that altering this information has no noticeable impact in terms of feature importance or classification results.

1 Introduction

Proof of language proficiency is often requested in contexts such as when studying or working abroad. However, determining a person's proficiency in a given language is not a trivial task as the concept of 'language proficiency' is not well defined (Spolsky, 2007). This has led to a focus on construct validity in second language assessment, that is, verifying that language tests actually evaluate what they claim to be evaluating (Messick, 1989; Im et al., 2019).

As with many other endeavours, AI has been proposed as a way to either automate or semi-automate language assessment. This is not hypothetical: several large corporations dedicated to standardized testing already do so (e.g. the Duolingo English Test, Cardwell et al., 2023).

When using NLP-based systems, this task is called automated essay scoring (AES). In the European context, this is commonly done using the CEFR[1] scale as a reference, which consists of six levels ranging from level A1 (elementary communication skills) to level C2 (advanced communication skills). While older systems used handcrafted features (e.g. Pilán et al., 2016b), current approaches tend to use encoder-only models (e.g. Wilkens et al., 2023). Of note is that generative models do not tend to perform this task well (Benedetto et al., 2025), despite them establishing the state of the art in many other NLP tasks nowadays.

This raises a key concern: while construct validity is a major focus in the field of second language assessment, neural networks are infamous for their lack of transparency (hence the term 'black box models'). Moreover, most current approaches assume (either implicitly or explicitly) an end-to-end approach. This opens up the question of whether these systems are meant to support or to substitute highly qualified professionals.

In this paper we introduce the *ABC Diagnostics Tool*. This tool uses explainability features to determine which words are important for the underlying model when assigning CEFR levels. It allows for versions of the same text with minimal changes to be evaluated next to each other (see Figure 1). As such, it can be used as a tool to explore the dynamics between learner language and things unrelated to language assessment such as names or other pieces of personally identifiable information (PII). This can be done through the use of artificially biased corpora[2] to help identify cases where the underlying model is making spurious correlations by focusing on features that have nothing to do with the construct validity, which can

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

[1] Common European Framework of Reference, originally introduced by the Council of Europe (2001) as a means to standardise teaching and evaluation of second language learning.

[2] Hence the *ABC* in its title.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 295–305, November 19–20, 2026, University of Tartu Library

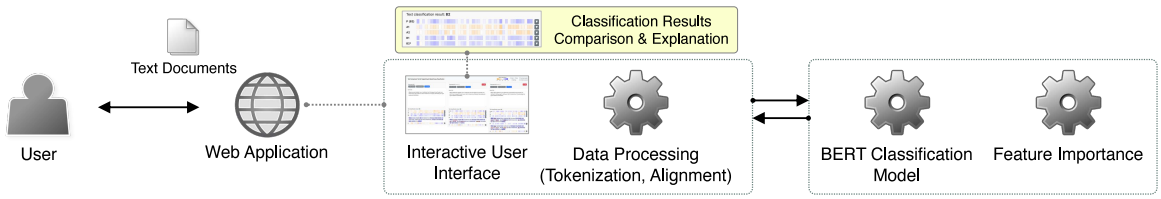


Figure 1: High-level overview of the design of our interactive visualization tool, implemented as a web-based application that processes input text documents (learner essay texts) in two versions: original and modified by users.

lead to unfair assessment. Furthermore, the ABC Diagnostics Tool has an easy-to-interpret visual interface to allow for qualitative analysis, making it accessible for non-technical people (such as teachers or students). We make the code necessary to run the ABC Diagnostics Tool locally available on GitHub,[3] as well as a video demo on how to use it.[4]

In order to showcase the tool, we explore how a Swedish BERT model fine-tuned for CEFR classification (Muñoz Sánchez et al., 2024a) reacts to PII within the text, more specifically, names of people and places. This aims to provide useful insights into the viability of using methods such as pseudonymization to protect the privacy of the people whose writing is being assessed.

The novelty of this work is three-fold:

1. We apply token-level feature importance methods to determine the effects of personally-identifiable information (PII) appearing within the texts. This allows us to conduct bias testing following a counterfactual reasoning workflow beyond the usual minimal pairs approaches with potentially multiple comparisons across several text versions (Kusner et al., 2017; Hofmann et al., 2024; Sorin et al., 2025).

2. We provide a visual possibility to explore cumulative effects of text perturbations in language learner essays in a step-by-step manner, gradually adding the intensity of changes applied to the text under controlled conditions — the task that has not been previously explored for language learner essays to the best of our knowledge.

3. We conduct qualitative analysis of a Swedish essay grading model with respect to the effects of pseudonymising learner essays, which contrasts the risks of privacy breaches versus the risks of unfair grading due to biases regarding personally-identifiable information (PII) appearing within the essays.

Proof-of-concept study. There are indications that names appearing within student essays might affect how humans grade and provide feedback (Aldrin, 2017). Muñoz Sánchez et al. (2024b) looked at whether similar patterns surfaced when using automated grading tools. That is, given the original essay containing one given name (e.g. "My name is Anna"), they replaced that name with one from another cultural context (e.g. 'Fatima'). The authors asked the question: would the model grade the essay differently depending on the pseudonyms used instead of the original name? Their results showed that their system was stable in its classification outputs despite such changes.

We take this scenario further and explore whether replacing several mentions of personal information may have cumulative effect on the system output. We are primarily interested to inspect model behaviour in more detail, zooming into the change of weights assigned by the model to individual tokens in the essay in the course of performed substitutions.

In this respect, pseudonymisation of learner essays is a double-edged sword: On the one hand, it is ethically and legally required to protect privacy of a leaner/writer (Official Journal of the European Union, 2016; European Data Protection Supervisor, 2025; Volodina et al., 2025a). On the other hand, replacements may harm a learner through unfair grading, as shown in multiple studies with human graders (e.g. Tyrefors Hinnerich et al., 2015; Aldrin, 2017; Menegatti et al., 2017).

The ABC Diagnostics Tool is aimed at methodological exploration of a system behaviour in response to changes introduced to the original version, so that potential unfair assessment can be dis-

[3] https://github.com/spraakbanken/ABC-Diagnostics-Tool.

[4] Video demo link is provided from the repository.



Figure 2: Interface of the web application for two versions of the same text (see Appendix A for a transcription). The tool shows the CEFR levels assigned both versions of the text, marked with an asterisk (original on the left, modified on the right). It also includes the level assigned to the original version of the essay along with the modified version (denoted by **P(A2)**), as well as color-coded feature importance for each level given the current essay. Shades of blue mean that specific token makes it more likely for the model to select that level, while shades of orange make it less likely to do so. Darker shades mean that the effect is more marked. Note that, even though this example only contains two versions of the same text text, the tool allows for an arbitrary number of versions to be compared in a tree-like manner.

covered before a system deployed in a real-world context. In our particular proof-of-concept study, we explore whether we can protect learners' identity through pseudonymisation without influencing assessment results in a negative way. The ABC tool, thus, allows the users to explore the effects of intensity of perturbations in the input text, examining both the model output in the form of the CEFR grades, as well as through visualisation and focused inspection of changing feature weights in response to such perturbations.

2 The ABC Diagnostics Tool

2.1 Testing Environment

Essay analysis tools have previously been used to visualise the complexity of a text being evaluated (e.g. Pilán et al., 2016a, see Appendix C). However, such tools tend to do CEFR level classification and word-level complexity separately. This means that there is no actual correlation between the word-level complexity labels and the overall level of the input text. This is true even if both tasks were trained concurrently, as complexity labels would not serve as explainability features even in this case.

The ABC Diagnostics Tool aims to fill this gap by providing insights into which *features*[5] are the most important for the resulting level. As such, we focused on the following requirements for the tool:

R1 allow users to provide and modify their own texts,

R2 support comparison between the original text and subsequent modified versions of it,

R3 use explainability tools to determine token-based feature importance across CEFR levels,

R4 represent feature alignment across the original and modified versions of the text, and

[5]As determined by the embedding layer of the model used. In the context of text data and BERT models, features correspond to word/subword-level tokens.

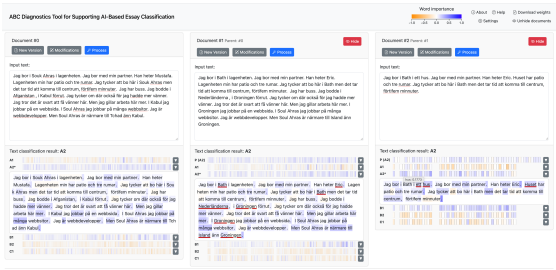


Figure 3: The complete user interface of the tool, including the results for three documents.

R5 allow users to download weights for further inspection.

To address these requirements, the ABC Diagnostics Tool consists of a computational backend and a web-based interactive tool, implemented in Python and JavaScript. Figure 1 provides an overview of the tool and its components. Even though the visualizations are the central component of the tool, users can also download the numerical values for further inspection.

The computational backend exposes a RESTful API for a BERT-based classification model while enriching the classification results with additional metadata such as feature importance weights. The feature importance functionality relies on the `transformers-interpret` library,[6] which itself uses the layer integrated gradients technique (Sundararajan et al., 2017) from Captum (Kokhlikyan et al., 2020).[7]

In order to explore the essay classification results, the user accesses the web-based frontend served from either their local machine or a server. The backend of this interactive tool performs text preprocessing, dispatching text classification requests to the computational backend and the analyses related to alignment when provided with several versions of the same text. For the latter task, we make use of the `pytokenizations` and `sequence_align` Python packages. These backend analyses ensure that the interactive interface represents the alignment of the latest results in line with the respective parent document.

The web-based user interface is organized to represent the computational results for a series of documents. These documents consist of the original input document, as well as any modified child versions. This forms a tree structure where each modified document can be compared to its parent version. After the user provides the initial input text, they can request classification and/or create a new version of the document. The platform allows for child versions to be modified manually or by applying a modification strategy through a modal dialog.

The classification and feature importance results are represented in a user interface for each document (see Figures 2 and 3). Details about the top predicted class are displayed for the document by default, while those for other classes are folded to avoid visual clutter.

A sequence-aligned heatmap allows for feature importance comparison across multiple CEFR levels for a given version of the text. Meanwhile, the alignment of features and respective feature importance values between a document and both its parent and children versions allows for determining whether specific changes affect the levels the model assigns to the text. This in turn opens the gate to do counterfactual analysis on how certain characteristics of the text affect classification (for more details see Barocas et al., 2023, Chapter 5). This heatmap representation was inspired by prior work on text alignment visualization (Yousef and Jänicke, 2021), as well as prior work on feature importance representation for text classification (e.g. Štrumbelj and Kononenko, 2014; Ribeiro et al., 2016). Shades of blue are used to denote that a specific token increases the likelihood of the model selecting that level, while shades of orange mean the opposite. The intensity of a shade represents how marked this effect is.

2.2 The ABC Diagnostics Tool in Practice

When developing a tool, it is important to keep in mind concrete real world scenarios in which it could be used. Throughout this section we present some potential practical uses of the ABC Diagnostics Tool, both by researchers and by non-technical users (such as teachers and students).

Students can use the tool in order to self-asses their CEFR level in a preliminary manner. Furthermore, they can test whether specific vocabulary choices or grammatical errors would have lead to a different level.

Teachers can use the system in order to identify areas of opportunity for further development of the students by looking at which features of the texts led to the given level. They can also use these

[6] https://github.com/cdpierse/transformers-interpret (Accessed: 2026-09-25)

[7] https://captum.ai/ (Accessed: 2026-09-25)

systems to determine whether specific aspects of the texts could have led to another result, be them linguistic or otherwise.

Researchers can use the tool to explore whether the underlying models focus on the same aspects of the texts as human assessors would. It can also be used to explore whether the systems present biases or whether techniques used for privacy preservations (such as pseudonymisation) affect the levels assigned by the models. The proof-of-concept study included in Section 3 shows how the latter can be done.

2.3 Language Resources

The ABC Diagnostics Tool itself is in principle model- and language-agnostic (within the scope of transformer-based approaches). However, it requires a model to be loaded when deploying it. We focus on Swedish as there are several non-overlapping language-learner datasets available, one of which is annotated with PII information.

SweLL-gold is a dataset of pseudonymized Swedish learner essays (Volodina et al., 2019, 2025b). It contains annotations with approximate proficiency levels according to course advancement (beginner, intermediate, or advanced) and with PII labels. This allows us to explore the effects of altering pieces of PII in Swedish learner texts.

We use the model by Muñoz Sánchez et al. (2024a)[8] in the backend. This model is a version of Swedish BERT (Malmsten et al., 2020) fine-tuned for CEFR classification on Swedish learner texts.[9] This model was originally trained on SweLL-pilot (Volodina, 2024), a dataset of Swedish learner essays annotated with CEFR labels. Of note is that this model has not seen any of the essays from SweLL-gold during training, thus avoiding potential contamination from having previously seen the texts.

3 Proof-of-Concept Study

3.1 Methodology

In order to assert the role that pseudonymization and PII play in CEFR classification, we take two essays for each of the three proficiency levels present in the SweLL-gold dataset. These essays are chosen such that each of them have at least three (in the case of beginner and intermediate essays) or two (in the case of advanced-level essays) distinct pieces of PII mentioned in the text. The IDs of the chosen essays can be found in Appendix B.

For each of these essays we create a parallel version where only PII is changed, leaving all other linguistic content intact. These intermediate versions between these texts are generated by partially substituting individual pieces of PII in a lattice manner. That is, at the ends of the lattice we have both the original pseudonymized text and the version with all pieces of PII changed. In between we have versions where only one or two kinds of PII have been changed. We would expect that the level predicted by a fair model should not change between versions of the same text.

Of note is that SweLL-gold was pseudonymized using names of fictionalized places, with the original texts not being available for the general public. In order to substitute the names, we use a combination of the name lists used by Muñoz Sánchez et al. (2024b) and rare names (e.g. less than 10 people with that given name in Sweden). We generate substitutions of place names by using real-world places fitting the background of the given names chosen (e.g. matching typical Finnish names with typical Finnish cities within the same text).

Given the challenges involved in identifying suitable substitutions and the large number of replacements required, we limit the present analysis to a qualitative study of a small number of essays.

To illustrate the scale of the task, the sample essay on Figure 4 contains 20 instances of PII, comprising six first names, one country, one city, and three other pieces of PII. Analysing the effect of perturbing PII in a cumulative manner in this essay would require generating 2^{20} variants of this essay in a lattice manner. Besides, for each cultural context considered, we would need to identify appropriate and reliable substitutions for culturally dependent entities such as first names and course names. Given these methodological and practical challenges, we restrict the present work to a small-scale qualitative analysis and leave a systematic quantitative investigation for future work.

3.2 Results and Discussion

A pattern shared by all the chosen essays is that the underlying model shows a clear preference for

[8] `sprakbanken-se/KB-bert-swedish_CEFR-classification` (Accessed: 2026-09-25)

[9] Note that we do not use generative models as they underperform for CEFR classification (Benedetto et al., 2025). Even though BERT-like models are considered to be outdated in various NLP tasks, they still represent the current state-of-the-art across multiple languages (Imperial et al., 2025).

Hi Anna!

Welcome to Torrelaguna.

My name is Aylin! I am from Turkey. I speak Kurdish and a little Spanish. I am married to Alex. We have three children. Their names are Aysu, Gizem, and Elif. Aysu is 11 years old, Gizem is 9 years old, and Elif is 6 years old.

I like Torrelaguna because my family is in Torrelaguna and my friends are also in Torrelaguna. I am writing now in this place. I want to learn Spanish. I study Spanish at CEPA.

Best wishes,
Aylin

Figure 4: Sample essay from SweLL-gold, translated into English and pseudonymised.

a given level. This is visualised in the tool as a clear cloud of blue appearing at that level but not at the others (see Figure 2 for an example). This is a good sign as we would expect CEFR classification systems to be confident about their predictions. However, it may also mean that they are confident even when misclassifying, which is less than desirable.

Of note is that none of the edits led to flips in terms of classification regardless of whether common or uncommon names were used. This points to PII not being a deciding factor for CEFR classification. Having said that, it would be necessary to do causal counterfactual analysis in a systematic manner across the whole dataset in order to be able to assure that this is the case.

Both of these results corroborate the findings of Muñoz Sánchez et al. (2024b), where they found that given names within learner essays did not have a major impact on CEFR classification systems. They further note that these kind of systems seem to draw upon the more basic linguistic knowledge encoded in the models (Muñoz Sánchez et al., 2024a).

An important future step would be to ensure that these systems are not ignoring content words by solely focusing on function words. Naismith et al. (2023) suggest that BERT-like models are not good at evaluating the coherence of a text, especially when compared to generative models. However, Benedetto et al. (2025) note that the current generation of generative models does not work well for CEFR classification. There are other potential sources of bias that can arise even in this case. For instance, Uluslu and Schneider (2025) point out that, while content words are important, function words alone can help us determine a language learner's first language.

Yet another future direction is to extend the tool to allow users to easily plug in other models for inspection on a range of different text-based tasks. As mentioned in Section 2.3, the design is in principle agnostic to a specific language or model: the current implementation relies on explainability techniques available for transformers, and thus providing access to further models (BERT or potentially other approaches) would be of interest. Additional explainability techniques might also complement the current approach. For instance, NeuroX (Dalvi et al., 2019) allows analysing individual neurons rather than feature importance for the whole model. Such functionality would also open up opportunities for interactive comparison across the results predicted by several models.

4 Conclusions

In this paper we have presented the ABC Diagnostics Tool. This is a web-based tool meant to assist non-technical professional graders in exploring which pieces of a text are important for automated CEFR level classification systems; as well as a tool for methodological inspection of effects of gradual text perturbation on model outputs.

In order to showcase its usage, we have explored the impact of changing PII in Swedish learner essays by gradually altering this information within the texts. We have noticed that these changes have little effect on the overall feature importance map displayed by the tool. Moreover, these changes did not influence the final label predicted by the model. This is further supported by the fact that no essay got assigned a different CEFR level because of these changes.

Future experiments using the ABC Diagnostics Tool should extend the replacements from PII-only to other linguistic characteristics. This would help us determine which linguistic features are important for the different CEFR levels according to the model. Not only that, but it would allow us to compare whether this matches the aspects that human assessors focus on.

A particularly interesting case would be that of gender markers in languages with male/female

grammatical gender, such as German or French. This would allow us to study the importance that gender agreement has for CEFR level classification on those languages. Not only that, but we could also explore biases related to this. For instance, the role that gender-based stereotypes play in how essays are classified or whether these models work with gender-neutral language (e.g. singular *they* in English or *hen* in Swedish).

We consider that this is an important step towards the development of tools aimed to help teachers with language assessment, rather than those meant to replace them. Furthermore, this tool can be potentially used to explore the presence of biases in automated essay assessment systems. In the future, we intend to continue developing this kind of tools under the umbrella name of the ABC Diagnostics Educational suite.

Limitations

In this paper we present a pilot version of the ABC Diagnostics tool. It aims to provide a basis for model behaviour explanations using *feature* importance methodologies. This means that it is not evaluating for more complex linguistic structures. For instance, if a word with a conjugation error leads to a given CEFR level, it should not be taken to mean that conjugation is crucial for that specific level. While it can be indicative that such is the case, further analysis would be required in order to assert so. As such, we encourage not to take the results as being authoritative over those of a professional grader.

Moreover, the tool works on a small scale basis. That is, it uses variations of a single text. Even though this can be indicative of the presence of biases or other undesired correlations, a more systematic approach across the whole dataset would be necessary in order to properly ascertain so.

Ethical Considerations

An important thing to note is that the ABC Diagnostics tool should not be used in lieu of professional human language assessment. Instead, it is meant to be used in conjunction in order to help teachers and graders understand how automated scoring works, study which characteristics of the texts CEFR classification models are focusing on, or to guide students and/or teachers on which areas to focus on in order to improve in the target language.

Acknowledgements

The authors acknowledge support from *InfraVis* for providing application expertise for visualization through the Swedish Research Council grant 2021-00181.

This work has been supported by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161).

It has also been supported by Grandma Karl is 27 years old: Automatic pseudonymization of research data through the Swedish Research Council grant 2022-02311 for the years 2023-2029.

AI Disclosure

☒ Not applicable (no AI tools used)
☐ Assistive use
☐ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: No AI tools were used in preparing this work.

References

Emilia Aldrin. 2017. Assessing names? Effects of name-based stereotypes on teachers' evaluations of pupils' texts. *Names*, 65(1):3–14.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Ramsey Cardwell, Ben Naismith, Geoffrey LaFlair, and Steven Nydick. 2023. Duolingo English Test: Technical Manual. Technical Manual, Duolingo.

Council of Europe. 2001. *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Fahim Dalvi, Avery Nortonsmith, D Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019. NeuroX: A toolkit for analyzing individual neurons in neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

European Data Protection Supervisor. 2025. AI act regulation (EU) 2024/1689 – regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) no 300/2008, (EU) no 167/2013, (EU) no 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (text with EEA relevance). *Publications Office of the European Union*.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Gwan-Hyeok Im, Dongil Shin, and Liying Cheng. 2019. Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia*, 9(1):14.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for PyTorch. *Preprint*, arXiv:2009.07896.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of Sweden – making a Swedish BERT. *Preprint*, arXiv:2007.01658.

Michela Menegatti, Elisabetta Crocetti, and Monica Rubini. 2017. Do gender and ethnicity make the difference? Linguistic evaluation bias in primary school. *Journal of Language and Social Psychology*, 36(4):415–437.

Samuel Messick. 1989. Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2):5–11.

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Irena Szawerna, and Elena Volodina. 2024a. Jingle BERT, frozen all the way: Freezing layers to identify CEFR levels of second language learners using BERT. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 137–152, Rennes, France. LiU Electronic Press.

Ricardo Muñoz Sánchez, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann, and Elena Volodina. 2024b. Did the names I used within my essay affect my score? diagnosing name biases in automated essay scoring. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 81–91, St. Julian's, Malta. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Official Journal of the European Union. 2016. Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance). *Official Journal*, (Document 02016R0679-20160504).

Ildikó Pilán, David Alfter, and Elena Volodina. 2016a. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 120–126, Osaka, Japan. The COLING 2016 Organizing Committee.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016b. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Vera Sorin, Panagiotis Korfiatis, Jeremy D Collins, Donald Apakama, Mahmud Omar, Benjamin S Glicksberg, Mei-Ean Yeow, Megan Brandeland, Girish N Nadkarni, and Eyal Klang. 2025. Socio-demographic modifiers shape large language models' ethical decisions. *Journal of Healthcare Informatics Research*, 9(4):567–586.

Bernard Spolsky. 2007. On second thoughts. In *Language Testing Reconsidered*, pages 9–18. University of Ottawa Press.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Bjorn Tyrefors Hinnerich, Erik Höglin, and Magnus Johannesson. 2015. Discrimination against students with foreign backgrounds: Evidence from grading in Swedish public high schools. *Education Economics*, 23(6):660–676.

Ahmet Yavuz Uluslu and Gerold Schneider. 2025. Investigating linguistic abilities of LLMs for native language identification. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 81–88, Tallinn, Estonia. University of Tartu Library.

Elena Volodina. 2024. On two SweLL learner corpora—SweLL-pilot and SweLL-gold. In *Proceedings of the Huminfra Conference*, HiC 2024, pages 83–94. Linköping University Electronic Press.

Elena Volodina, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Maria Irena Szawerna, Lisa Södergård, and Xuan-Son Vu. 2025a. Towards shared standards for pseudonymization of research data. In *Proceedings of the 2nd Huminfra Conference*, pages 101–114. NEALT Proceedings Series, University of Tartu.

Elena Volodina, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice, Dan Rosén, Lisa Rudebeck, Carl-Johan Schenström, Gunlög Sundberg, and Mats Wirén. 2019. The SweLL language learner corpus: From design to annotation. *Northern European Journal of Language Technology*, 6:67–104.

Elena Volodina, Arianna Masciolini, Beáta Megyesi, Julia Prentice, Lisa Rudebeck, Gunlög Sundberg, and Mats Wirén. 2025b. SweLL with pride: How to put a learner corpus to good use. *Huminfra handbook: Empowering digital and experimental humanities*.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.

Tariq Yousef and Stefan Jänicke. 2021. A survey of text alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1149–1159.

Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665.

A Example Essay

Below we include the example text used in Figure 2 and a rough translation into English. Note that the pieces of PII have been replaced with tags.

> Jag bor i <city_1> i lagenheten. Jag bor med min partner. Han heter <given_name>. Lagenheten min har patio och tre rumar. Jag tycker att bo här i <city_1> men det tar tid att komma till centrum, förtifem minnuter. Jag har buss. Jag bodde i <country>, i <city_2> förrut. Jag tycker om där också för jag hadde mer vänner. Jag tror det är svart att få vänner här. Men jag gillar arbeta här mer. I <city_2> jag jobbar på en webbsida. I <city_1> jag jobbar på många webbsitor. Jag är webbdevelopper. Men <city_1> är närmare till <country> änn <city_2>.

An English translation of the essay is as follows. Note that we have attempted to reproduce similar errors as in the original Swedish version.

> I live in <city_1> in the apartment. I live with my partner. His name is <given_name>. The apartment mine has patio and three roomes. I like to live here <city_1> but it takes time to go to the city centre, fortyfive minnutes. I have bus. I lived in <country>, in <city_2> beffore. I like it there too because I hadd more friends. I think it is difficalt to get friends here. But I like to work here more. In <city_2> i work on a website. In <city_1> I work on many webbsites. I am a web-developper. But <city_1> is closer to <country> thann <city_2>.

B Essay IDs

This appendix presents the IDs of the essays chosen as part of the experiments listed in this paper.

- Beginner-level essays
  - G24GT3
  - G36GT2
- Intermediate-level essays
  - K5KT1
  - K48KT1
- Advanced-level essays
  - F25FT3
  - M42MT1

C The Lärka Tool

As mentioned in Section 2.1, Lärka (Pilán et al., 2016a) is a previously existing tool that visualises CEFR levels at both the document and word levels. However, this is done by two separate systems. As such, the word-level labels cannot be used for explainability purposes for the CEFR level classification system. Figure 5 provides an image of the tool for illustrative purposes, using the same example from Figure 2 (full text in Appendix A).

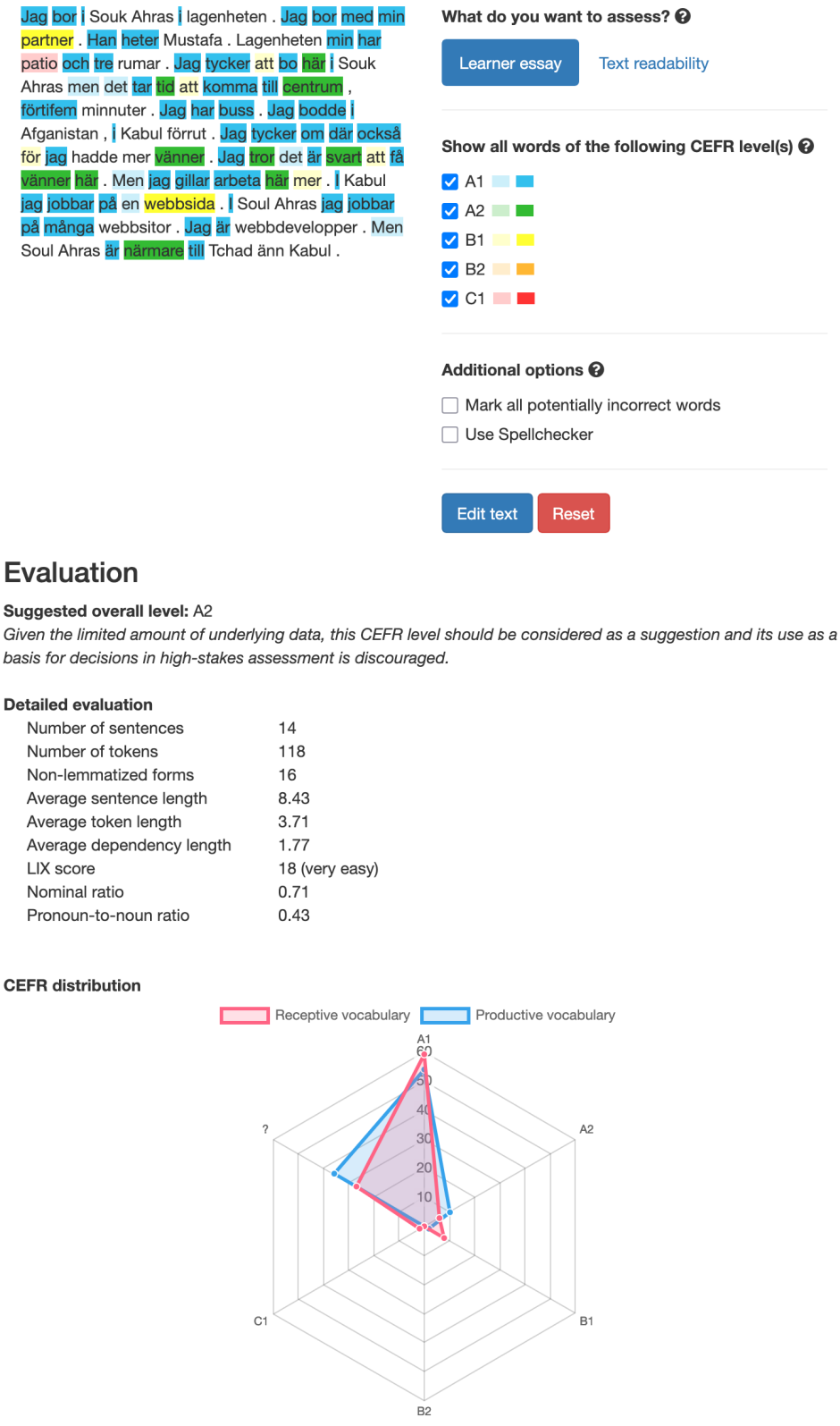


Figure 5: The Lärka tool visualises distribution of vocabulary across CEFR levels. The overall CEFR level is also calculated. However, there is no direct connection between the vocabulary-level labels and the document-level one.

Acoustic and Lexical Signals for Native Language Identification of Language Learners

Ahmet Yavuz Uluslu[1,2] **Mengjie Qian**[1] **Kate Knill**[1] **Mark Gales**[1]

[1]ALTA Institute, Department of Engineering, University of Cambridge
[2]Department of Computational Linguistics, University of Zurich

{ayu20, mq227, mjfg, kmk1001}@cam.ac.uk

Abstract

Spoken native language (L1) identification for second language learners has traditionally relied on acoustic attributes of speech. Comparatively little attention has been paid to analysing speech transcripts for L1 traces. In this paper, we present a study of L1 identification using both speech and corresponding transcripts from a spoken learner English corpus. We compare a WavLM-based audio L1 classifier with a zero-shot GPT-4o classifier operating on transcripts and analyse their predictions across proficiency levels using a CEFR-balanced evaluation set. Our experiments show that the audio classifier outperforms the transcript-based approach overall when full transcripts are available, although their relative strengths vary with learner proficiency. Audio performs better at A1, whereas text performs better from A2 to C1 with full transcripts. Transcript-based accuracy generally increases with proficiency, while acoustic performance remains comparatively stable. Restricting transcript length substantially reduces text-based accuracy, whereas audio is less sensitive to input length and performs relatively well with limited input. These findings provide insights into L1 identification from acoustic and textual representations of learner speech and how performance varies with proficiency and the amount of available language material.

1 Introduction

Native language identification (NLI) aims to infer a language user's first language (L1) from their production in a second language (L2). Much of the existing work has focused on learner writing, using lexical choices, grammatical patterns, and spelling errors as evidence of L1 background (Malmasi et al., 2017). In spoken language, cross-linguistic influence may also appear in pronunciation, rhythm, and intonation (Piske et al., 2001). Learner speech therefore offers potentially complementary sources of evidence for NLI: audio retains phonetic and prosodic information, while transcripts provide more direct access to lexical and grammatical patterns.

The usefulness of these sources of evidence may vary with learner proficiency. Some grammatical errors become less frequent at higher proficiency levels, although developmental patterns differ across error types (Thewissen, 2013). L1-related pronunciation patterns may also change with L2 learning and experience (Derwing and Munro, 2013). More proficient learners may additionally produce longer and more varied responses (Alzahrani, 2024), providing different opportunities for identifying their L1. These changes make it unclear whether L1 identification from audio and transcripts follows similar patterns across proficiency levels.

To investigate this question, we compare an audio L1 classifier based on WavLM (Chen et al., 2022) with zero-shot large language models (LLMs) prompted for textual NLI (Uluslu et al., 2026), operating on transcripts of the same learner English responses. We create BULATS-L1, a new dataset for spoken NLI with an evaluation set balanced jointly across six L1 backgrounds and five manually assigned CEFR levels, from A1 to C1. The transcripts were generated by an automatic speech recognition (ASR) model adapted to learner speech (Qian et al., 2025). This dataset allows us to examine overall identification performance and differences across L1 groups, and to compare proficiency-related patterns while holding the distribution of L1 backgrounds constant. Our results show that transcript-based NLI outperforms the audio classifier overall when full transcripts are available. However, this advantage depends on proficiency and transcript length. Audio

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 306–313, November 19–20, 2026, University of Tartu Library

performs better at A1, while text performs better from A2 to C1 with full transcripts. Restricting transcript length reduces overall text-based accuracy whereas audio is less sensitive to input length and performs relatively well with limited input.

2 Related Work

Shared tasks have established benchmarks for NLI from both written and spoken learner language. The 2017 NLI Shared Task evaluated identification from essays, spoken responses, and their combination (Malmasi et al., 2017). Strong essay-based systems used support vector machines with character and word n-grams and syntactic features, often combined through ensembles or classifier stacking. In the speech track, the top-ranked systems combined transcript-based features with acoustic i-vectors. Transcript-only systems performed considerably worse, with the relatively short spoken responses identified as a possible contributing factor. The report consequently highlighted the need to investigate NLI using longer transcripts.

More recent research has investigated pretrained language models and prompted LLMs as alternatives to feature-based NLI systems (Ng and Markov, 2025; Goswami et al., 2025). Exposure to multilingual text (Kew et al., 2024) and examples of translation (Zhu et al., 2024) and grammatical error correction (Masciolini et al., 2025) during pretraining may equip LLMs with knowledge of cross-linguistic correspondences and non-standard language patterns that can subsequently support NLI. Zero-shot approaches then infer L1 through task instructions without training a dedicated classifier on labelled NLI examples (Uluslu and Schneider, 2025).

On the audio side, self-supervised learning has enabled transferable speech representations learned from unlabelled recordings (Baevski et al., 2020). Their applicability to speaker identification and verification has been examined through benchmarks such as SUPERB (Yang et al., 2021), while WavLM demonstrates strong performance across speech and speaker-related tasks (Chen et al., 2022). Building on these developments, we compare WavLM-based audio classification with zero-shot transcript-based NLI on the same learner responses, examining performance across proficiency levels and the effect of transcript length.

3 Datasets

We use two learner English speech corpora with complementary roles. BULATS-L1 is the primary corpus for our L1 identification experiments, providing self-reported L1 labels and a CEFR-balanced evaluation set, but no manual transcriptions. The Speak & Improve Corpus 2025 (Knill et al., 2024, 2025) provides manual transcriptions for learner speech, allowing us to assess the sensitivity of transcript-based L1 predictions to automatic transcription.

CEFR level	n	Avg. Length
A1	54	190 ± 88
A2	54	286 ± 88
B1	54	390 ± 94
B2	54	464 ± 80
C1	54	555 ± 101
Overall	270	377 ± 157

Table 1: BULATS-L1 dataset composition by CEFR level. Each level includes nine submissions from each of six L1 backgrounds. Transcript lengths are measured in words using the ASR generated transcripts. The values are reported as mean $\pm$ standard deviation.

3.1 BULATS-L1

We construct the BULATS-L1 dataset from BULATS (Business Language Testing Service)[1], a discontinued online English language assessment service developed by Cambridge English. The dataset includes learners with six self-reported L1 backgrounds: Arabic, Chinese, French, Italian, Japanese, and Spanish. These L1 groups were selected based on the availability of sufficient test submissions across the included CEFR levels to construct a balanced evaluation set.

Audio classifier training and development data. The BULATS-L1 training set comprises 16,149 utterances from 769 speakers across the six target L1 backgrounds, totalling approximately 86.4 hours of speech. The training and evaluation sets are speaker-disjoint. Unlike the balanced evaluation set, the training data could not be perfectly balanced across L1 backgrounds because of differences in data availability. We used as much available data as possible for the lower-resource L1s, particularly Arabic and Mandarin,

[1] https://en.wikipedia.org/wiki/Business_Language_Testing_Service

while downsampling higher-resource L1s such as Spanish to reduce the disparity. As a result, the amount of training data ranges from 32 Arabic-speaking learners (3.6 hours) to 213 Spanish-speaking learners (24.0 hours). The BULATS-L1 development set has a similar distribution as the training set, with a total of 15.05 hours of speech across six target L1 backgrounds. The development set is used for hyperparameter turning and model selection only.

Balanced evaluation set. Each BULATS test submission contains responses to five distinct speaking tasks, each targeting specific aspects of speaking proficiency. We exclude Part 2, a read-aloud task with prescribed linguistic content, and concatenate the remaining four spoken responses within each test submission for audio-based analysis and their corresponding transcripts for transcript-based analysis. The evaluation set is balanced jointly across L1 backgrounds and CEFR levels A1–C1, with nine test submissions per L1–CEFR combination, yielding 54 submissions per proficiency level and 270 submissions overall. C2 was excluded because insufficient submissions were available to meet this balancing criterion. CEFR proficiency levels are manually assigned by human experts.

Manual transcriptions are not available for this corpus. We generate transcripts using the Whisper-based ASR model adapted to learner speech by Qian et al. (2025), and use these transcripts as input to the transcript-based L1 classifiers. Table 1 summarizes the evaluation set by CEFR level. Mean transcript length per submission increases from 190 words at A1 to 555 words at C1. Although submission counts are balanced, the amount of available text therefore varies across proficiency levels.

3.2 Speak & Improve 2025

The Speak & Improve Corpus 2025 is a public dataset of second-language (L2) learner speech (Knill et al., 2024, 2025). We use the fluent manual transcriptions from the evaluation subset of the Speak & Improve Corpus 2025, comprising 300 test submissions. These transcripts retain learner language errors but exclude disfluencies. Since reference L1 labels are unavailable in the data used here, we assess L1 prediction consistency between manual and automatic transcriptions rather than L1 identification accuracy.

4 Experimental Setup

4.1 Speech-based NLI

For speech-based NLI, we train a classifier on representations from a frozen pretrained WavLM-Large [2] encoder (Chen et al., 2022). Audio is sampled at 16 kHz. We learn a weighted combination of the encoder's hidden representations rather than using only the final transformer layer:

$$\mathbf{H} = \sum_{l=0}^{L} \alpha_l \mathbf{H}^{(l)}, \tag{1}$$

where $\mathbf{H}^{(l)}$ denotes the representation at layer l, L is the number of transformer layers, and α_l are softmax-normalised trainable layer weights.

We apply attention pooling over valid encoder frames to obtain a clip-level embedding. The classification head consists of a fully connected layer with ReLU activation and dropout, followed by a linear output layer. A softmax converts the output logits into probabilities over the six L1 classes. Only the layer weights and classification head are trained, using AdamW with cross-entropy loss and label smoothing. Training hyperparameters are provided in Appendix B.

At inference time, we divide each concatenated test submission into consecutive, non-overlapping 30-second clips, retaining any shorter final clip. We process all clips and average their predicted class probabilities to obtain a submission-level distribution:

$$p(y \mid x) = \frac{1}{K} \sum_{k=1}^{K} p(y \mid x_k), \tag{2}$$

where K is the number of clips in submission x. The class with the highest average probability is selected as the predicted L1.

4.2 Transcript-based NLI

For transcript-based NLI, we follow the zero-shot classification procedure of Uluslu et al. (2026), using OpenAI's `gpt-4o`. We formulate NLI as a multiple-choice task, mapping the six candidate L1 labels to letter identifiers. The prompt templates are provided in Appendix A. Constrained decoding restricts the response to valid identifiers, and their token-level log probabilities are normalised to obtain a distribution over the candidate classes.

[2] https://huggingface.co/microsoft/wavlm-large

	Accuracy		Macro-F1	
CEFR	Audio	Text	Audio	Text
A1	81.48	48.15	81.03	45.31
A2	83.33	87.04	81.31	87.06
B1	88.89	90.74	88.45	90.58
B2	92.59	94.44	92.46	94.30
C1	81.48	92.59	81.25	92.67
Overall	85.56	82.59	85.37	82.60

Table 2: Submission-level NLI accuracy and macro-F1 (%) on BULATS-L1 by CEFR level. Text results are unrestricted transcripts.

To reduce sensitivity to label position, we evaluate each transcript using three randomised candidate orderings (Liusie et al., 2024). We map the resulting probabilities back to their corresponding L1 labels and average the distributions. The highest-probability class is selected as the prediction. Before classification, we mask named entities, including locations, nationalities, and personal names, as well as non-English words. This approach aims to reduce reliance on explicit identity cues and topic-related shortcuts (Uluslu et al., 2025).

4.3 Evaluation

Sensitivity to automatic transcription. On Speak & Improve, we compare fluent manual transcripts and ASR-generated transcripts. We calculate WER separately for each test submission and report the mean across submissions. To assess the effect on transcript-based NLI, we report predicted-label agreement, defined as the proportion of paired transcripts receiving the same predicted L1, and mean Jensen–Shannon divergence (JSD) (Lin, 1991) between their predicted L1 distributions. Agreement captures changes in the selected label, while JSD captures differences across the full distribution. Since reference L1 labels are unavailable, these metrics assess prediction consistency rather than L1 identification accuracy.

L1 identification. On BULATS-L1, we evaluate both pipelines against self-reported L1 labels at the test-submission level. We report accuracy and macro-averaged F1. Macro-averaged metrics assign equal weight to each of the six L1 classes. Macro-F1 is computed by averaging the class-specific F1-scores.

5 Results

5.1 NLI across Proficiency Levels

Table 2 compares the speech-based and transcript-based systems on BULATS-L1. Using unrestricted transcripts, GPT-4o achieves an overall accuracy of 82.59%, compared with 85.56% for the WavLM-based classifier, a difference of 2.97 percentage points. The corresponding macro-F1 scores are 82.60% and 85.37%, respectively, showing a similar overall pattern.

The relative performance of the two systems varies across proficiency levels. At A1, the audio classifier clearly outperforms the transcript-based system, achieving 81.48% accuracy compared with 48.15%, a different of 33.33 percentage points. From A2 onwards, however, the transcript-based system achieves higher accuracy. Its accuracy rises sharply from 48.15% at A1 to 94.44% at B2, before decreasing slightly to 92.59% at C1. Audio accuracy varies over a narrower range, increasing from 81.48% at A1 to 92.59% at B2 before dropping to 81.48% at C1. Macro-F1 closely follows the same trend as accuracy across proficiency levels, suggesting that the observed differences are not driven by a small subset of L1 classes.

These results show different proficiency-related patterns for the two pipelines. Transcript-based identification improves substantially between A1 and B2 by 46.29% accuracy, whereas audio performance is comparatively stable. Both systems decline at C1, although the decrease is considerably larger for the audio classifier. This pattern alone does not establish that L1-related acoustic information diminishes at higher proficiency, since other factors may also contribute. In particular, transcript length increases with proficiency,

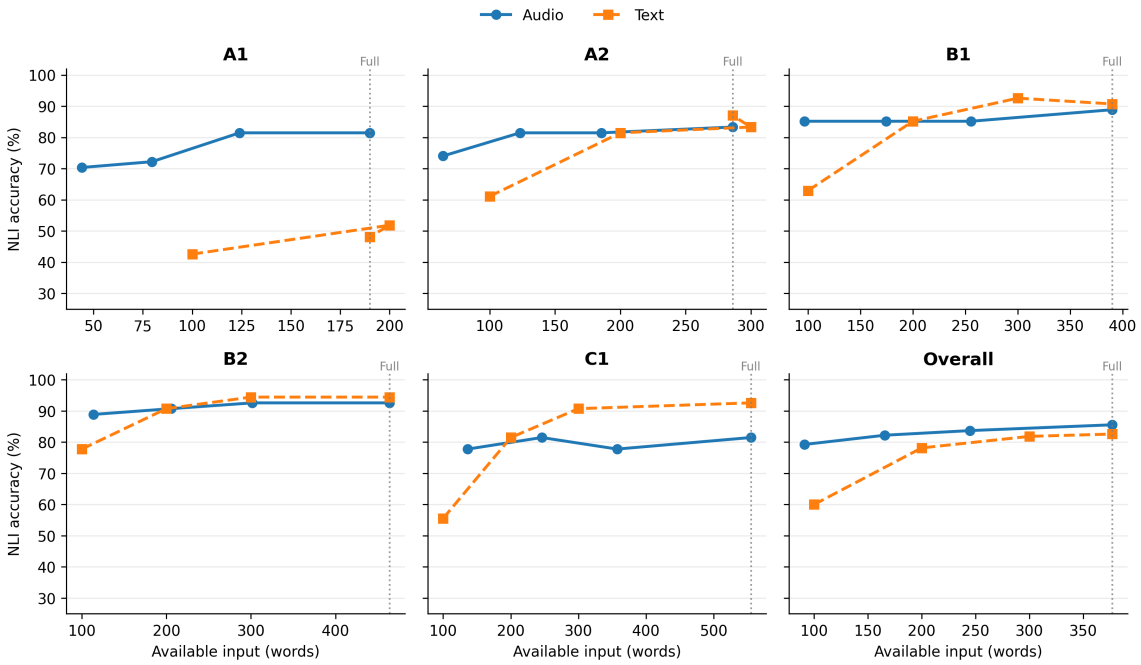


Figure 1: Effect of available input on audio- and text-based NLI

motivating a closer examination of the effect of available text on identification performance.

5.2 Effect of Transcript Length

To examine the contribution of available text to transcript-based NLI, we evaluate transcripts capped at 100, 200, and 300 words alongside the unrestricted condition. For each capped condition, we retain the first N words of each concatenated transcript, leaving shorter transcripts unchanged. Table 3 reports accuracy across these conditions.

Len.	A1	A2	B1	B2	C1	All
100	42.59	61.11	62.96	77.78	55.56	60.00
200	51.85	81.48	85.19	90.74	81.48	78.15
300	48.15	83.33	92.59	94.44	90.74	81.85
Full	48.15	87.04	90.74	94.44	92.59	82.59

Table 3: Transcript-based NLI accuracy (%) by CEFR level and maximum transcript length in words.

Overall accuracy increases with the word cap, from 60.00% at 100 words to 78.15% at 200 words and 81.85% at 300 words, with a further small increase to 82.59% for the full transcript. The largest gain occurs between 100- and 200-word conditions, while performance starts to saturate beyond 200 words.

Proficiency-related differences remain under a common word cap. With a 200-word cap, accuracy is 51.85% at A1 and 81.48% at A2, compared with 85.19% at B1, 90.74% at B2, and 81.48% at C1. Thus, the substantially lower performance at A1 cannot be explained solely by the shorter responses typically produced by lower-proficiency learners. From A2 onwards, the classifier achieves relatively strong performance once sufficient text is available, suggesting that the availability or distinctiveness of lexical and grammatical L1 cues may also vary with proficiency.

The strong dependence of transcript-based NLI on input length motivates a corresponding analysis of the audio system. Although the full audio system achieves higher overall accuracy, the two modalities may differ in how performance depends on input length.

5.3 Effect of Audio Length

We examine the effect of available audio by progressively adding the retrained task responses: P1, P1 + P3, P1 + P3 + P4, and P1 + P3 + P4 + P5. Table 4 reports accuracy across these conditions.

Parts	Len	A1	A2	B1	B2	C1	All
P1	91	70.37	74.07	85.19	88.89	77.78	79.26
P1,3	166	72.22	81.48	85.19	90.74	81.48	82.22
P1,3,4	245	81.48	81.48	85.19	92.59	77.78	83.70
P1,3,4,5	377	81.48	83.33	88.89	92.59	81.48	85.56

Table 4: Audio NLI accuracy (%) on restricted audio parts by CEFR level and average length in words.

Figure 1 plots audio- and text-based model accuracy overall and at each CEFR as the amount of available input words increases. Per-part performance is in Appendix C.

Audio-based NLI is less sensitive to input length than transcript-based NLI in overall performance. Audio accuracy increases from 79.26% using only P1 to 82.22% after adding P3, and reaches 85.56% with all four retained parts. In comparison, text accuracy increases from 60.00% at 100 words to 78.15% at 200 words and 82.59% with the full transcript. The audio system therefore achieves relatively strong performance with limited input, while the text system benefits more substantially from additional text material.

The same general pattern is observed across proficiency levels, although the effect varies by CEFR level. At A1, audio consistently outperforms text across the available input range. At A2, the initial advantage of audio narrows as more text becomes available, and the text system slightly exceeds audio with the full input. A similar crossover occurs for B1 and B2, where the text model outperforms the audio model once sufficient transcript material is available. At C1, text accuracy increases substantially with additional input and substantially outperforms the audio model, whereas the audio model performance varies within a comparatively narrow range.

These results suggest different sensitivities to input length across the two modalities. Acoustic L1 cues appear to be informative from relatively short speech samples, with additional audio providing modest gains. In contrast, transcript-based NLI benefits more from having longer transcripts, especially at shorter input lengths. The remaining differences across CEFR levels indicate that input length alone does not explain the proficiency-related differences; the availability and strength of acoustic and lexical L1 cues may also vary with proficiency.

5.4 Fusion of Audio and Text Predictions

We further examine whether the audio and text modalities provide complementary information by linearly interpolating the predicted L1 probability distributions at the submission-level,

$$p_{\text{fuse}} = \lambda p_{\text{audio}} + (1 - \lambda) p_{\text{text}}, \qquad (3)$$

where λ controls the weight of the audio model. Figure 2 shows accuracy as λ varies from 0 to 1.

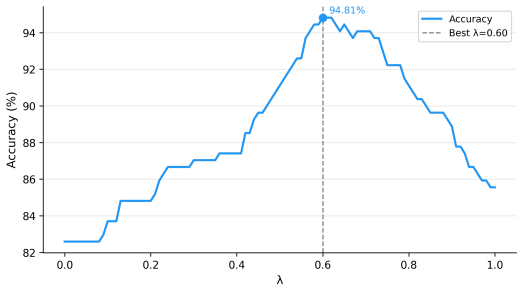


Figure 2: NLI accuracy on BULATS-L1 for linear fusion of audio and text predictive distributions

Fusion consistently improves over either individual system for a broad range of weights. The best performance is obtained at $\lambda = 0.60$, reaching 94.81% accuracy, compared with 85.56% for audio alone and 82.59% for text alone. Performance peaks when slightly more weight is assigned to the audio model, but remains high around the optimum. These results indicate that the audio- and transcript-based systems capture complementary L1 information, and that combining their predictions substantially improves identification performance.

5.5 Prediction Consistency on Speak & Improve

We use Speak & Improve to examine prediction consistency in the absence of reference L1 labels. We first assess the sensitivity of transcript-based NLI to automatic transcription. The ASR transcripts have a mean submission-level WER of 9.2% against fluent manual references. The transcript-based NLI classifier produces the same predicted L1 under both the manual and automatic transcription conditions for 89.7% of submissions, with a mean JSD of 0.075 between the predictive distributions. Thus, the predicted label remains unchanged for most submissions, although approximately one in ten predictions differs. The high agreement suggests that automatic transcription retains information used by the classifier for L1 prediction in most cases, although it does not establish the accuracy of those predictions.

We then compare the audio- and transcript-based classifiers. The two models produced the same top-ranked L1 for 77.7% of the 300 submissions, with a mean Jensen–Shannon divergence (JSD) of 0.221 bits between their six-class predictive distributions. This indicates substantial label agreement alongside moderate differences in prediction probabilities.

6 Conclusion

We compared a WavLM-based audio classifier with a zero-shot GPT-4o transcript-based classifier for NLI using BULATS-L1, whose evaluation set is balanced across six L1 backgrounds and five manually assigned CEFR levels. The audio system achieves higher overall accuracy than the transcript-based system, while their relative performance varies with proficiency. Audio performs substantially better at A1, whereas text is stronger at the higher proficiency levels once sufficient transcript material is available. Both modalities are affected by input length, although the audio system is less sensitive to shorter input than the transcript-based system.

The results also show that the two modalities provide complementary information. Linear fusion of their predictive distributions improves performance substantially, reaching 94.81% accuracy, compared with 85.56% for audio alone and 82.59% for text alone. Overall, the findings indicate that learner proficiency, input length, and modality all influence NLI performance, and that combining acoustic and lexical evidence provides the strongest identification results.

Acknowledgments

This work reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge. It was also supported by Innosuisse innovation project 103.188 IP-ICT (AUCH).

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: GPT-6 Luna was used for editorial paraphrasing, grammatical error correction, and generating LaTeX table structures.

References

Alaa Alzahrani. 2024. Utility of Kolmogorov complexity measures: Analysis of L2 groups and L1 backgrounds. *PLoS One*, 19(4):e0301806.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Tracey M Derwing and Murray J Munro. 2013. The development of L2 oral language skills in two L1 groups: A 7-year study. *Language learning*, 63(2):163–185.

Dhiman Goswami, Marcos Zampieri, Kai North, Shervin Malmasi, and Antonios Anastasopoulos. 2025. Multilingual native language identification with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 193–199.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2024. Turning English-centric LLMs into polyglots: How much multilinguality is needed? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13097–13124.

Kate Knill, Diane Nicholls, Mark JF Gales, Mengjie Qian, and Pawel Stroinski. 2024. Speak & Improve Corpus 2025: an L2 English speech corpus for language assessment and feedback.

Kate M. Knill, Diane Nicholls, Mark J.F. Gales, Mengjie Qian, and Pawel Stroinski. 2025. Introducing the Speak & Improve Corpus 2025: an L2 English Speech Corpus for Language Assessment and Feedback. In *10th Workshop on Speech and Language Technology in Education (SLaTE)*, pages 167–171.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In *Proceedings of the 14th workshop on natural language processing for computer assisted language learning*, pages 1–33.

Yee Man Ng and Ilia Markov. 2025. Leveraging open-source large language models for native language identification. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–28.

Thorsten Piske, Ian RA MacKay, and James E Flege. 2001. Factors affecting degree of foreign accent in an L2: A review. *Journal of phonetics*, 29(2):191–215.

Mengjie Qian, Rao Ma, Stefano Bannò, Mark JF Gales, and Kate M Knill. 2025. End-to-End Spoken Grammatical Error Correction. *arXiv preprint arXiv:2506.18532*.

Jennifer Thewissen. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1):77–101.

Ahmet Yavuz Uluslu, Mark Gales, Kate Knill, and Gerold Schneider. 2026. The impact of editorial intervention on detecting native language traces. *arXiv preprint arXiv:2605.10216*.

Ahmet Yavuz Uluslu, Tannon Kew, Tilia Ellendorff, Gerold Schneider, and Rico Sennrich. 2025. Robust native language identification through agentic decomposition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8409–8425.

Ahmet Yavuz Uluslu and Gerold Schneider. 2025. Investigating linguistic abilities of LLMs for native language identification. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 81–88.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.

A L1 Classification Prompt

For the L1 attribution task, the model was cast in an expert persona and provided with a multiple-choice selection of the six candidate languages (represented by $\{options\}$). The exact prompt template was:

> "You are a forensic linguistics expert that reads essays from non-native speakers of English in order to identify their native language. Use clues such as word choice, syntactic patterns, and grammatical errors to decide. Analyze the text and choose the most likely native language from the following options:
>
> {options}
>
> (X) Other
>
> Reply with only the single letter (A, B, C, etc.) representing your choice."

B Audio Model Parameters

The audio classifier was trained for 10 epochs with the WavLM-Large encoder frozen throughout training; only the layer-combination weights and classification head were updated. We used AdamW with a learning rate of 1×10^{-4} and weight decay of 10^{-2}. The learning rate followed a cosine schedule with 500 warm-up steps. Cross-entropy loss with label smoothing of 0.1 was used as the training objective. The batch size was 2, with gradients accumulated over four steps, resulting in an effective batch size of 8. Training used mixed-precision computation, and gradients were clipped to a maximum norm of 1.0. The random seed was set to 42. Model performance was evaluated on the development set after each epoch, and the checkpoint achieving the highest development-set accuracy was retained for evaluation.

C Audio Model Per-part Performance

Part	A1	A2	B1	B2	C1	All
P1	70.37	74.07	85.19	88.89	77.78	79.26
P3	75.93	79.63	87.04	92.59	87.04	84.44
P4	79.63	85.19	92.59	87.04	79.63	84.81
P5	74.07	79.63	92.59	90.74	79.63	83.33

Table 5: Audio NLI accuracy on per-part audios

Poor Generalisation of Proficiency Classifiers Trained on Linguistic Complexity Features to New Writing Tasks

Emre Güçer[1] **Soroosh Akef**[2,3,4] **Detmar Meurers**[1,2,3]
[1]University of Tübingen, Germany
[2]Leibniz-Institut für Wissensmedien (IWM), Germany
[3]LEAD Graduate School and Research Network, University of Tübingen, Germany
[4]Center of Linguistics, Faculty of Arts and Humanities, University of Lisbon, Portugal
emre.guecer@student.uni-tuebingen.de s.akef@iwm-tuebingen.de
d.meurers@iwm-tuebingen.de

Abstract

CALL systems and level-adaptive language models need proficiency estimates whose linguistic basis can be inspected and that hold for new writing tasks. Starting from 1,784 features extracted with six analysis tools, we train Explainable Boosting Machines to predict the course level (1–15) and CEFR band (A1–C1) of 406,062 EFCAMDAT texts. For new learners on familiar tasks, the all-feature classifier reaches 97.7% CEFR accuracy (macro F_1 0.956) and 94.4% exact-level accuracy, and ten-fold cross-validation confirms these results. Excluding designated length measures (word and other raw counts, and measures with an absolute mean within-level Spearman correlation of at least 0.3 with text length) leaves performance virtually unchanged (97.6% CEFR accuracy), although predictions still follow length within levels. In this main model, lexical sophistication contributes 53% of the grouped importance, followed by syntactic complexity (22%) and grammatical constructions (18%); grammatical constructions have the largest share for the lowest-level output and lexical sophistication for all others, and the leading individual features are stable across folds. Every EFCAMDAT writing task belongs to a single level, however, the models generalise poorly to new tasks: for texts on tasks withheld from training, exact-level macro F_1 falls to 0.19 and CEFR accuracy to 71% (78% for a regression model). Since new learners still write for the tasks seen in training, high accuracy on new learners is not the most accurate measurement.

1 Introduction

Automatically assessing a learner's language level has long been an important challenge for computer-assisted language learning (CALL) systems, which need such estimates to place learners at the right level and to assign them appropriate materials.

CALL covers the use of computers in language teaching and learning, from vocabulary and grammar exercises to complete online courses such as the one from which our data come. Intelligent CALL (ICALL) adds automatic language analysis and learner modelling: it analyses what learners produce, keeps a model of their current abilities and uses this model to give individual feedback and to select or adapt material (Chen and Meurers, 2017; Chen et al., 2022; Akef et al., 2026). Because learners progress, such a learner model has to be updated from the texts they write.

Linguistic complexity can help select input that challenges individual learners at an appropriate level (Chen and Meurers, 2019; Chen et al., 2022). Such evidence should also be interpretable: a predicted proficiency label is more useful when a teacher can inspect the linguistic patterns behind it.

The same need arises when large language models (LLMs) generate learning materials or conversational responses. Knowledge of CEFR descriptors does not guarantee their reliable application in educational tasks (Benedetto et al., 2025). Grammar-controlled generation offers one route towards pedagogically appropriate output (Glandorf and Meurers, 2024; Glandorf et al., 2025; Ribeiro-Flucht et al., 2025). Another possibility is to use lexical, syntactic and grammatical feature profiles to guide generation. Before using such profiles, however, we need to establish what they actually describe and how important each part is for classification.

This question connects to the identification of *criterial features*: properties of learner language that help distinguish proficiency levels (Hawkins and Buttery, 2010). A feature can predict a level without being a reliable indicator of proficiency across contexts. Word counts may distinguish learners partly because higher-level assignments elicit longer answers, so we analyse models with and without designated length measures (word and other raw counts, and measures with an absolute

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 314–333, November 19–20, 2026, University of Tartu Library

mean within-level Spearman correlation of at least 0.3 with length; Section 3.2.2), and test how far their predictions still follow length. Vocabulary and grammatical constructions may identify the writing task rather than an ability that transfers to other tasks. Predictive accuracy and linguistic interpretability therefore need to be evaluated alongside generalisation to new tasks.

We investigate these issues using the cleaned version (Shatz, 2020) of the EF-Cambridge Open Language Database (EFCAMDAT), which contains learner writing produced in an online English course where each writing task belongs to one level (Geertzen et al., 2014; Huang et al., 2017). We use Explainable Boosting Machines (EBMs), whose additive structure exposes the contribution of every feature (Nori et al., 2019). Our feature inventory spans five dimensions: lexical sophistication, lexical diversity and density, syntactic complexity, grammatical constructions, and accuracy. We address three research questions:

1. How accurately do complexity and accuracy features predict CEFR levels and the more detailed course levels?

2. Which linguistic dimensions contribute to the models' predictions, and how do these contributions vary across level outputs?

3. How well do the models generalise to writing tasks excluded from training?

Our contributions are as follows:

- **Identifying a transfer gap.** On writing tasks left out of training, exact-level macro F_1 drops from 0.90 to 0.19 and CEFR accuracy to 71%. Tests with new learners but familiar tasks therefore overstate how well such models work on new tasks.

- **Accurate, interpretable models.** EBMs are interpretable by design, since each prediction is the sum of one contribution per feature. For new learners on familiar tasks, they predict the CEFR band with 97.7% accuracy and the course level with a macro F_1 of 0.90. Removing the designated length measures hardly changes accuracy but changes which features the models use.

- **Level profiles.** We show how much five linguistic dimensions contribute at each of the 15 levels: lexical sophistication contributes most overall, and grammatical constructions lead only at level 1.

- **Implications.** We discuss what these profiles and models can offer as references for level-aware LLMs, the next step of our project.

Our code and aggregated results are available at `https://github.com/gucere/Poor-Generalisation-of-Proficiency-Classifiers-to-New-Writing-Tasks`.

2 Related Work

2.1 Proficiency prediction and linguistic profiles

Automated scoring and proficiency classification of learner writing combine lexical, syntactic, discourse and error features with supervised models (Yannakoudakis et al., 2011; Crossley and McNamara, 2012), and models trained on complexity features predict CEFR levels in several languages (Pilán et al., 2016; Allkivi, 2026). For English, Yannakoudakis et al. (2018) predict CEFR levels across the full scale for the *Write & Improve* placement system, and Vajjala (2018) finds that the most useful linguistic features differ between two essay datasets. Hawkins and Buttery (2010) connect proficiency levels to characteristic linguistic properties, including patterns of grammatical use and error, and the English Grammar Profile (EGP) provides an inventory of level-associated grammatical abilities (O'Keeffe and Mark, 2017). On EFCAMDAT, Arnold et al. (2018) predict proficiency from linguistic metrics and full texts and find word-token and word-type counts among the most influential features, and Kerz et al. (2021) model sequences of complexity measurements with recurrent neural networks. Ballier et al. (2020) report a shared task on predicting the CEFR levels of EFCAMDAT texts by French learners of English and describe a lexical bias: because each task prompt belongs to one level, prompt words such as *robot* and *Eiffel* became strong predictors of C2. They therefore recommend testing on tasks not seen in training, as our task holdout does. Gaillat et al. (2022) reach a balanced accuracy of 82% on their own data, but their best features reach only 59.2% on external corpora, a transfer problem related to the one we study. The corpus is also widely used in second language acquisition research (Murakami, 2016; Alexopoulou et al., 2017).

2.2 Interpretable models

Generalised additive models learn one shape function per feature and are interpretable by construction. EBMs fit these functions with gradient boosting and approach the accuracy of black-box models on tabular data (Lou et al., 2013; Nori et al., 2019). This makes it possible to examine the linguistic evidence a model uses without approximating its predictions with a separate explanation model. In language learning research, Akef et al. (2025) use an EBM on linguistic complexity measures to identify patterns that distinguish English- and German-influenced texts of Portuguese heritage language learners, controlling for topic and proficiency. Akef et al. (2026) train EBMs for Portuguese proficiency assessment on accuracy features, which perform comparably to models trained on linguistic complexity features, and Loefflad et al. (2026) use EBMs to relate German Grammar Profile features to the CEFR levels of graded texts. Our study extends this line of work to a broad English feature inventory and examines how performance changes when evaluation tasks are excluded from training.

2.3 Grammatical use and accuracy

Our accuracy features come from ERRANT (Bryant et al., 2017), which aligns a learner text with its correction (Felice et al., 2016) and labels each edit with an error type. ERRANT is a standard tool for annotating and evaluating grammatical error correction (GEC), for example in the BEA-2019 shared task (Bryant et al., 2019), and Bryant et al. (2023) survey the field. Manually annotated learner errors alone have also been used to predict CEFR levels (Ballier et al., 2019). Based on automatically identified grammatical constructions, Verratti-Souto et al. (2025) compare corpus-derived CEFR assignments for EGP structures in EFCAMDAT with the EGP levels and find the closest agreement for beginner levels. Akef et al. (2026) combine LLM-based correction of learner writing with a rule-based module that characterises 252 CEFR-linked grammatical properties of European Portuguese, so that construction-specific accuracy can be measured from the contrast between a learner text and its correction. Their representation distinguishes the accuracy of grammatical properties from the frequency of error categories used here, and provides a methodological comparison for our correction-derived feature group.

2.4 Complexity measures and text length

Complexity is multidimensional, spanning lexical, syntactic and grammatical aspects (Housen and Kuiken, 2009; Norris and Ortega, 2009; Bulté and Housen, 2012). Automatic tools operationalise many of these aspects: syntactic complexity with L2SCA (Lu, 2010) and TAASSC (Kyle, 2016; Kyle and Crossley, 2018), lexical richness with LCA (Lu, 2012), and lexical sophistication with TAALES (Kyle and Crossley, 2015; Kyle et al., 2018). Many measures depend on text length by construction. Type-token ratios fall as texts get longer, which motivated length-corrected and moving-window diversity measures (McCarthy and Jarvis, 2010; Covington and McFall, 2010), although a fixed window also requires sufficiently long texts. Raw counts of words, clauses or structures grow with the amount written, and in essay scoring, length alone predicts a large part of human scores (Perelman, 2014).

2.5 Task effects and shortcuts

Writing tasks shape which linguistic features learners use: Alexopoulou et al. (2017) show that EFCAMDAT tasks influence linguistic complexity and accuracy, and Arnold et al. (2018) split EFCAMDAT by topic so that their models could not predict levels from topics rather than proficiency. In machine learning, models are known to exploit such incidental correlations between the data collection and the label, known as shortcuts (Gururangan et al., 2018; Geirhos et al., 2020). Generalisation beyond the training data is also a concern for complexity-based models: Akef et al. (2024) show that Portuguese readability models built on more informative complexity features generalise better to an independently collected corpus, even when their performance on the original test set decreases. We quantify the performance gap associated with withholding writing tasks for a broad-coverage proficiency model.

2.6 Adaptive CALL and level-controlled generation

Complexity measures can drive adaptive CALL. Chen and Meurers (2017) propose syntactic-development benchmarks for selecting input just beyond a learner's current production, Chen and Meurers (2019) relate text readability and learner proficiency through distances between complexity feature vectors, and in an ICALL experiment Chen et al. (2022) find that learners' writing aligns

more readily with low and medium complexity challenges than with high ones. With LLMs, the question becomes how to make generated language fit a level. LLMs encode some knowledge of the CEFR but do not apply it reliably (Benedetto et al., 2025). Work on grammar control grounds generation in the EGP: Glandorf and Meurers (2024) control grammatical complexity in educational texts, Glandorf et al. (2025) extend this to dialogue responses, and Ribeiro-Flucht et al. (2025) combine task-anchored, grammar-targeted prompting with post-generation validation for proficiency-aligned dialogue practice. Our study covers a broader set of linguistic dimensions that might later complement such control, but we do not evaluate a generation system here.

3 Methodology

3.1 Corpus and splits

We use the cleaned EFCAMDAT subcorpus developed by Shatz (2020), specifically its main-prompt sample. EFCAMDAT contains texts written by learners of English in an online course, each labelled with the course level at which it was written (Geertzen et al., 2014). We use 406,062 texts by 89,589 learners from levels 1–15. Following the course design, three consecutive levels form one CEFR band (Council of Europe, 2001): levels 1–3 correspond to A1, 4–6 to A2, 7–9 to B1, 10–12 to B2 and 13–15 to C1 (Geertzen et al., 2014). These labels indicate placement and progression in the course; they are not independent proficiency ratings of each text. The cleaned release restricts the corpus to eleven nationality groups and removes texts shorter than 20 words, texts with extensive non-English or duplicated material, and task-specific word-count outliers.

Every level has eight writing tasks (120 in total), and every task belongs to exactly one level. Each task is a course unit with its own prompt, and prompts change in topic and genre across levels, for example from *Introducing yourself by email* (level 1) and *Writing a weather guide for your city* (level 2) to *Claiming back your security deposit* (level 8) (Shatz, 2020). Texts are short: the median text has 57 words, and 41% have fewer than 50 words (Table 1). The distribution is strongly imbalanced: C1 accounts for about 1.3% of the texts. Learners were assigned to training (70%), validation (15%) and test (15%) sets by a SHA-256 hash of their identifier (seed 42), so all texts of a learner fall into one set. Feature filtering and model development used the training and validation data; the test set was evaluated once, after all models were frozen.

3.2 Linguistic analysis tools

We use six tools that capture complementary properties of learner writing. TAALES describes the sophistication of vocabulary using lexical reference resources such as frequency lists and psycholinguistic norms, while LCA measures vocabulary diversity, density and sophistication. L2SCA measures the size and organisation of syntactic units, and TAASSC adds finer-grained information about phrases, clauses and syntactic sophistication. POLKE (Sagirov and Chen, 2026) identifies pedagogically defined grammatical constructions from the EGP; in a validation on native-speaker sentences, it extracts structures with an average precision of 0.91 and recall of 0.84 (Verratti-Souto et al., 2025). ERRANT identifies the edits between an original and a corrected text and assigns error categories, providing the basis for our accuracy features.

3.2.1 Features

The initial inventory contains 1,784 values extracted by the six tools, including auxiliary counts and coverage values:

- TAALES contributes 485 lexical-sophistication indices and 352 coverage values (Kyle and Crossley, 2015; Kyle et al., 2018). The indices include frequency, range, psycholinguistic norms and n-gram measures. Coverage records how much of a text is represented in an index's underlying resource.
- LCA contributes 33 outputs covering lexical density, diversity and sophistication, including counts (Lu, 2012).
- L2SCA contributes 23 outputs: syntactic-unit counts and complexity indices (Lu, 2010).
- TAASSC contributes 149 clausal, phrasal and syntactic-sophistication measures (Kyle, 2016).
- POLKE contributes rates for 667 EGP constructions, expressed per 100 words, and three text-size measures (Sagirov and Chen, 2026).

CEFR	Levels	Texts per level	Total	Train	Valid.	Test	Learners	Median words
A1	1–3	96,762 / 56,839 / 38,062	191,663	133,980	28,825	28,858	48,819	36
A2	4–6	70,470 / 37,277 / 21,844	129,591	91,122	18,953	19,516	28,235	68
B1	7–9	37,460 / 14,157 / 9,889	61,506	43,052	9,143	9,311	17,908	91
B2	10–12	10,458 / 5,010 / 2,719	18,187	12,823	2,546	2,818	5,702	127
C1	13–15	3,264 / 1,194 / 657	5,115	3,775	763	577	1,278	173
All	1–15		406,062	284,752	60,230	61,080	89,589	57

Table 1: Texts per course level and CEFR band, learner-disjoint training, validation and test splits, number of learners (who may appear in several bands) and median number of words per text.

- ERRANT contributes 72 correction-derived outputs, primarily error rates per 100 words by error type, obtained by aligning learner texts with the corrected versions supplied with EFCAMDAT (Bryant et al., 2017).

We group TAALES and its coverage values as lexical sophistication, LCA as lexical diversity and density, L2SCA and TAASSC as syntactic complexity, POLKE as grammatical constructions, and ERRANT as accuracy. This grouping follows the tools for practical reasons, and some tools overlap in what they measure; LCA, for example, also includes sophistication measures.

Implementation and preprocessing. We use TAALES 2.2, L2SCA (release 2023-08-15), TAASSC 2.0.0.58, ERRANT 3.0.2, Lu's LCA script and the published POLKE implementation.[1] LCA, TAASSC and ERRANT use spaCy 3.8.14 with `en_core_web_sm` 3.8.0. Before Stanford parsing, L2SCA input is sentence-segmented and remaining long segments are split into chunks of at most 45 words; POLKE falls back to chunking when a sentence reaches 120 words. ERRANT rates divide edit counts by the number of words in the original text and multiply by 100. All 406,062 text pairs were processed. Missing feature values are passed to the EBM without imputation; InterpretML assigns them to a separate bin of each shape function, so no texts were removed.

3.2.2 Length-based and duplicate features

We define 45 *designated length measures*: nine direct length measures, comprising the word-token and word-type counts reported by the tools, and 36 additional ones. The latter include 22 raw counts of other units and 14 measures selected by their association with text length: we compute Spearman correlations with word count within each course level on the training texts and average them across levels, weighting each level by its number of training texts, and measures whose mean correlation has an absolute value of at least 0.3 are designated length-dependent. The selected measures include type-token ratios and diversity measures computed over 50-word windows or samples.

Features with fewer than 20 non-missing training observations or no variation are removed; the all-feature classifier uses 1,719 features after filtering. The *main* feature set excludes the designated length measures and retains 1,675 features (44 of the 45 pass the filter; ERRANT's count of other edits is constant). The *strictest* set also excludes near-duplicate features, using a training-set absolute Spearman correlation of at least 0.95 with a retained feature, and contains 1,158 features. The exclusions remove only these listed sources of length dependence and redundancy; the remaining features can still be related to length or task, individually below the threshold or in combination (Section 4.2).

3.3 Models

An EBM is a generalised additive model that learns a separate, potentially nonlinear function for each feature through gradient boosting (Lou et al., 2013; Nori et al., 2019). The model builds each feature's shape function through many small updates, each a shallow decision tree on a single feature fitted to the current errors, so that the learned functions are built up gradually. A lexical-frequency measure, for example, can have different effects in different parts of its range without requiring a single linear coefficient. For a classifier output k, the additive score is

$$s_k(\mathbf{x}) = \beta_k + \sum_{j=1}^{p} f_{jk}(x_j), \qquad (1)$$

[1] https://sites.psu.edu/xxl13/lca/; https://github.com/chxiaobin/polke; https://github.com/chrisjbryant/errant.

where β_k is the intercept of output k and f_{jk} the learned shape function of feature j. The scores are transformed into class probabilities, so additivity applies to the scores rather than directly to the probabilities; the regressor has the corresponding single-output form. Because every prediction is the sum of the feature contributions $f_{jk}(x_j)$, the contribution of each feature and of each group of features can be read off exactly. They explain how the fitted model computes its predictions and say nothing about causal effects of linguistic features on proficiency.

We use the InterpretML implementation of EBMs (version 0.7.8) without pairwise interactions: 200 boosting rounds, learning rate 0.04, 128 bins per feature, 50 smoothing rounds, one outer bag, no early stopping and random seed 42. Classifiers use balanced sample weights, a cost-sensitive treatment of the class imbalance. The main analysis uses two models on the main feature set: a level classifier over the 15 course levels and a level regressor that predicts the level as a number. CEFR predictions are derived from the classifier by summing the predicted probabilities of the three levels of each band, and from the regressor by clipping its prediction to 1–15 and rounding it to the nearest level. For comparison, we train the same models on further feature sets: all features; all features without the direct length measures only; the strictest set; POLKE constructions alone; ERRANT error features alone; the 200 features ranked highest by mutual information on the training texts; the nine direct length measures alone; and majority-class and median-level baselines. Appendix A reports all 16 configurations.

3.4 Evaluation

Our primary metric is macro F_1 over the 15 course levels, which gives each class equal weight, whereas accuracy reflects the frequency distribution of the texts. We also report exact-level accuracy, accuracy within one level, quadratic weighted kappa (QWK), macro precision, balanced accuracy (macro recall), CEFR macro F_1 and accuracy, and the regressor's mean absolute error (MAE) in levels. The MAE uses the regressor's clipped continuous predictions; its other metrics use the rounded levels. For test results, we estimate uncertainty with a learner-clustered bootstrap: test learners are resampled with replacement 2,000 times, each bringing all of their texts, and we report 95% percentile intervals.

To estimate how much the results depend on the training data, we run ten-fold cross-validation (CV) of the main and all-feature models over the combined training and validation data (344,982 texts), grouping folds by learner and stratifying them by course level (about 310,000 training and 34,500 evaluation texts per fold). The minimum-observation and constant-feature filters are fitted on each training fold, whereas the designated length measures and the near-duplicate list are fixed from the original training split. Differences between the two feature sets are tested with the corrected resampled t-test of Nadeau and Bengio (2003), which accounts for the overlap between the folds' training sets. The other configurations are compared on the validation and test splits.

3.4.1 Feature group importance

For a group G and classifier output k, we sum the group's feature contributions for each evaluation text and average their absolute values:

$$I_{G,k} = \frac{1}{n} \sum_{i=1}^{n} \left| \sum_{j \in G} f_{jk}(x_{ij}) \right| . \qquad (2)$$

Summing before taking absolute values allows opposing contributions within a group to cancel. A group's share for output k is its importance divided by the sum over all groups for that output. For the overall classifier shares, we sum each group's importance over the 15 outputs and divide by the total over all groups, so outputs with larger contributions weigh more. We report the mean and standard deviation over the ten CV folds. The shares describe how the model uses the feature groups and should not be read as percentages of explained variance.

4 Results

All results in this section use the learner-disjoint evaluation, in which every writing task also occurs in the training data; Section 5 examines what this implies.

4.1 Overall performance

Table 2 presents the principal results. On the held-out test set, the all-feature classifier reaches 97.66% accuracy and 0.956 macro F_1 for CEFR bands, and 94.39% accuracy and 0.897 macro F_1 for the 15 course levels. The main classifier, which excludes the designated length measures, retains 97.59%

CEFR accuracy and 0.956 macro F_1 (95% CI 0.951–0.960), with 94.31% level accuracy, 0.896 level macro F_1 (0.881–0.905) and 97.01% within-one-level accuracy. The small fold-to-fold variation and the close agreement between CV and test results support the consistency of prediction for new learners completing familiar tasks.

Table 3 shows the results for every CEFR band. The main model's test F_1 ranges from 0.989 at A1 to 0.917 at C1 (95% CI 0.896–0.935). For the 15 course levels (Appendix B), errors mostly stay close to the true level: of the main model's test errors, 47% fall on a neighbouring level, 73% lie within two levels, and for 57% the predicted level still belongs to the correct CEFR band. The pooled CV confusion matrices in Figure 1 show the same pattern, and they also show that texts of the rarest levels are often assigned to more frequent lower levels: 10.4% of level-15 texts are predicted as level 13 and 9.1% as level 11, and 6.8% of C1 texts are assigned to B2. The rarest levels are the hardest and the least precisely estimated. Of the 657 level-15 texts, only 54 (8% rather than the nominal 15%, because the split assigns whole learners) fall into the test set, written by 20 learners; there, level 15 reaches an F_1 of 0.679 (95% CI 0.52–0.77). The other 603 are in the training and validation data, and ten-fold CV over these gives 0.743 ± 0.048, inside the test interval. Level 1, by contrast, has 14,613 of its 96,762 texts (15%, by 6,421 learners) in the test set and reaches 0.971 (0.969–0.973). These finer distinctions explain why the course-level macro F_1 is lower than the CEFR macro F_1. The main regressor obtains an MAE of 0.613 levels on validation, 0.619 in CV and 0.616 on test, but its rounded predictions reach a test CEFR macro F_1 of only 0.718, so the classifier is the better choice for assigning bands.

Published results on EFCAMDAT are not directly comparable to ours. Kerz et al. (2021) report 75.4% accuracy on six CEFR classes, using 152,314 texts of at least 100 words and random ten-fold cross-validation; their filtering removes most short A1 texts, and a random split, like our learner-disjoint one, does not hold writing tasks out of training. Arnold et al. (2018) instead separate topics between training and test data and report test AUCs of 0.895, 0.897, 0.778, 0.821 and 0.587 for binary classifiers between successive CEFR bands. Their setting is closer to our task-transfer analysis, but binary AUCs cannot be ranked against multi-class accuracy or macro F_1. Verratti-Souto et al. (2025) also predict CEFR levels from EFCAMDAT, but they derive level mappings for EGP structures from corpus statistics and use these mappings to classify texts.

4.2 Effects of feature exclusion

Table 4 compares the feature sets. Excluding the direct length measures leaves test level macro F_1 unchanged at 0.897, and excluding all designated length measures gives 0.896. In CV, the all-feature and main classifiers reach 0.902 and 0.900; the difference of 0.0014 is not statistically significant ($p = 0.102$), although this is not a formal equivalence result. The CEFR macro F_1 decreases by 0.0018 (0.955 to 0.953, $p < 0.001$), and the regressor is somewhat more affected: its CV MAE increases from 0.604 to 0.619 levels ($p < 0.001$). Removing the near-duplicates as well (the strictest set) costs little more. Single feature families are much weaker than their combination: on the test set, POLKE constructions alone reach a level macro F_1 of 0.72, the 200 individually most informative features 0.74, and ERRANT error features alone only 0.24. Conversely, removing all ERRANT features from the full set leaves the test level macro F_1 at 0.897 (Appendix A). Deriving CEFR bands from the level classifier is also better than training a classifier directly on the five bands: summed level probabilities give a test CEFR macro F_1 of 0.956 for both the all-feature and the main set, compared with 0.936 and 0.933 for direct CEFR classifiers.

Text length on its own is a strong but limited predictor: the nine direct length measures reach a CEFR macro F_1 of 0.653 on test, far above the majority-class baseline (0.128) but far below the main classifier (0.956). Figure 2 shows similar performance for the all-feature, main and strictest models across text-length bins, whereas the length-only classifier is useful mainly for short texts and becomes particularly weak for longer ones. Because text length, level and class frequency vary together, these bins are descriptive rather than controlled comparisons. Excluding the designated length measures does not, however, make the model independent of length. Among test texts of the same course level, the main classifier's expected level (the probability-weighted mean of the 15 levels) still rises with word count (Spearman $\rho = 0.33$, averaged over levels and weighted by their number of texts; all features 0.35, length-only classifier 0.94):

Target	Features	Evaluation	Accuracy (%)	Balanced acc. (%)	Macro P	Macro F_1
CEFR	All	Test	97.66	95.50	0.958	0.956
CEFR	Main	Test	97.59	95.40	0.957	0.956
CEFR	All	CV	97.54 ± 0.12	95.24	0.957	0.955 ± 0.003
CEFR	Main	CV	97.46 ± 0.11	95.07	0.955	0.953 ± 0.003
Course level	All	Test	94.39	89.83	0.896	0.897
Course level	Main	Test	94.31	89.65	0.895	0.896
Course level	All	CV	94.43 ± 0.16	89.92	0.906	0.902 ± 0.006
Course level	Main	CV	94.37 ± 0.16	89.83	0.904	0.900 ± 0.006

Table 2: Classifier results for all features and the main feature set on the held-out test set (61,080 texts) and in ten learner-grouped CV folds ($\pm$ fold standard deviation; CV precision and balanced accuracy are fold means). Balanced accuracy is macro recall; P is precision. CEFR predictions sum course-level probabilities within bands.

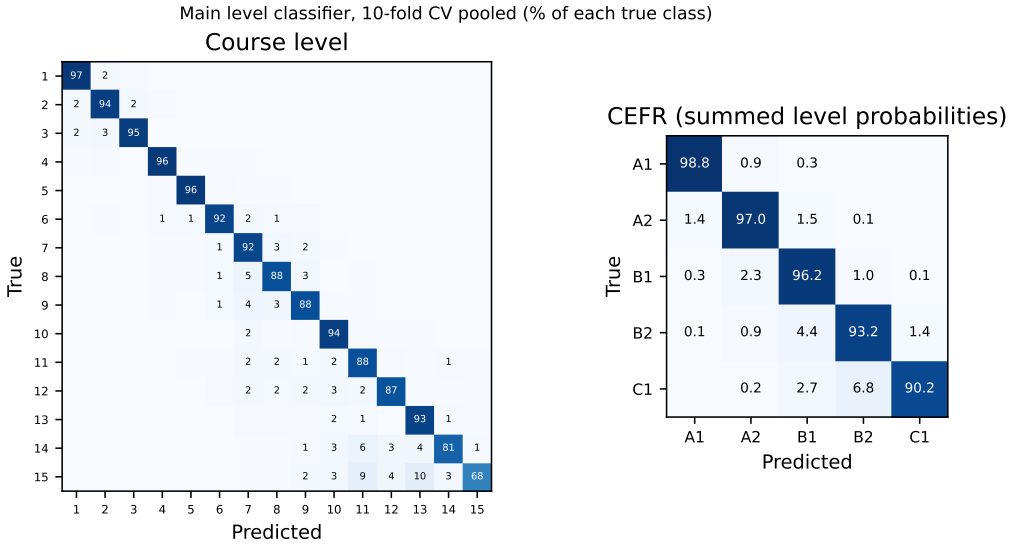


Figure 1: Confusion matrices of the main level classifier, pooled over the ten CV folds: course levels (left) and CEFR bands from summed level probabilities (right). Each cell gives the percentage of the true class, so the diagonal is the recall.

		All features			Main		
CEFR	Support	P	R	F_1	P	R	F_1
A1	28,858	.990	.990	.990	.989	.989	.989
A2	19,516	.975	.972	.973	.974	.970	.972
B1	9,311	.951	.962	.957	.949	.963	.955
B2	2,818	.948	.938	.943	.951	.938	.944
C1	577	.925	.913	.919	.924	.910	.917
Macro	–	.958	.955	.956	.957	.954	.956

Table 3: Per-band results on the held-out test set. P: precision; R: recall, equal to the accuracy within the true band. The macro recall is the balanced accuracy.

its mean error is -0.11 levels for the shortest fifth of each level's texts and $+0.18$ for the longest fifth. This may reflect length information spread over many weakly length-related features, or genuine proficiency differences within a level that go with longer texts.

4.3 Contributions of linguistic dimensions

Lexical sophistication has the largest share of the main classifier's grouped importance (53.4%), followed by syntactic complexity (22.4%) and grammatical constructions (18.1%); lexical diversity and density account for 4.5% and accuracy for 1.6% (Table 5). The regressor distributes its importance more evenly: syntactic complexity accounts for 34.7%, lexical sophistication for 33.8% and grammatical constructions for 25.0%. The shares are very stable, with standard deviations of at most 0.52 percentage points across the ten folds. The contributions are largest for the outputs of the rarest levels, and levels 13–15 account for 56% of the summed importance; weighting all 15 outputs equally gives the same ranking, but lexical sophistication falls to 45.2% and grammatical constructions rise to 23.2% (Table 5), so the size of the lexical lead depends on how the outputs are ag-

Feature set	Features	Level macro F_1 ↑			CEFR macro F_1 ↑			Regressor MAE ↓		
		Valid.	CV	Test	Valid.	CV	Test	Valid.	CV	Test
All features	1,719	.904	.902	.897	.956	.955	.956	.597	.604	.602
All − direct length	1,710	.902	–	.897	.955	–	.957	.601	–	.605
Main: all − length-related	1,675	.900	.900	.896	.953	.953	.956	.613	.619	.616
Strictest: main − near-duplicates	1,158	.900	–	.894	.953	–	.954	.622	–	.626
POLKE constructions only	618	.722	–	.716	.869	–	.867	.726	–	.735
ERRANT error features only	69	.242	–	.240	.608	–	.590	1.289	–	1.284
Top 200 by mutual information	200	.746	–	.736	.880	–	.869	.726	–	.735
Direct length measures only	9	.255	–	.250	.673	–	.653	1.083	–	1.096
Majority class / median level	0	.026	.026	.026	.129	.128	.128	2.229	2.243	2.221

Table 4: Level classification (15 levels), CEFR classification through summed level probabilities (5 bands) and level regression on the validation set, in ten-fold CV (run for the all-feature and main models and the baselines) and on the held-out test set. Features: number of features used by the level classifier.

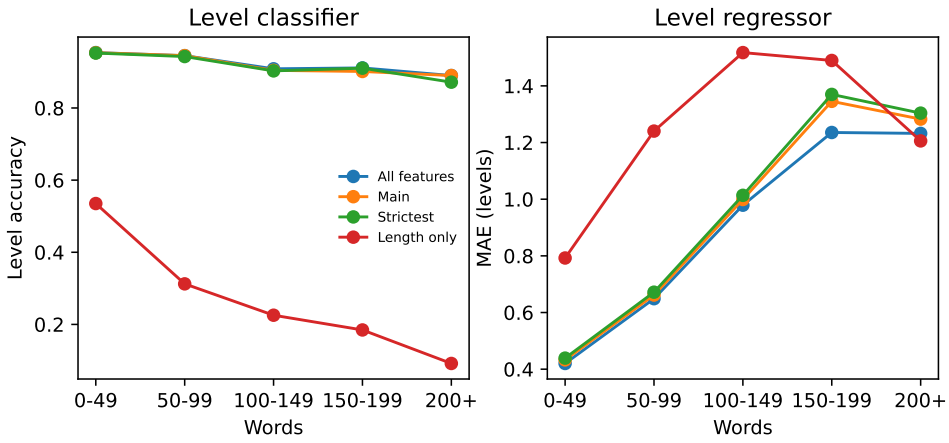


Figure 2: Validation performance by text length. Left: exact-level accuracy of the level classifiers. Right: MAE of the level regressors. The share of texts per bin is 42%, 46%, 9%, 2% and 0.2%. Longer texts tend to occur at higher, less frequent levels, so the bins also differ in level composition.

gregated. At the level of tools (Table 6), TAALES indices have the largest share in the main classifier (42.6%), followed by TAASSC (19.8%), POLKE (17.1%) and TAALES coverage values (11.9%); in the regressor, TAASSC leads (32.0%), followed by TAALES indices (27.7%) and POLKE (23.6%). Appendix C shows the tool shares for each course-level output. A large share means that the fitted model relies on a tool, which does not necessarily mean that the tool would perform best alone or cause the largest loss if removed.

The classifier's profile also varies across outputs (Figure 3). For the level-1 output, grammatical constructions have the largest share (36%), ahead of lexical sophistication (30%). Towards the highest levels, lexical sophistication gains weight, reaching about 66% for the level-15 output, while grammatical constructions fall to 11%; syntactic complexity remains between about 19% and 29%. Level 1 is the only output for which grammatical constructions contribute most; from level 2 upwards, lexical sophistication has the largest share, and its lead grows towards the highest levels. This parallels the finding of Verratti-Souto et al. (2025) that level assignments of EGP structures derived from EFCAMDAT agree with the EGP best at A1 and

Dimension	Classifier		
	Summed	Equal	Regr.
Lexical sophistication	53.4	45.2	33.8
Syntactic complexity	22.4	23.5	34.7
Grammatical constructions	18.1	23.2	25.0
Lexical diversity/density	4.5	6.0	5.5
Accuracy	1.6	2.1	0.9

Table 5: Share (%) of the main models' contributions per linguistic dimension, mean over the ten CV folds (standard deviations at most 0.52 points). Summed: share of the classifier's importance summed over its 15 outputs, as used in the text; Equal: mean of the 15 per-output shares. The regressor has a single output.

Tool output group	Classifier	Regressor
TAALES indices	42.63 ± 0.60	27.66 ± 0.24
TAASSC	19.79 ± 0.25	32.02 ± 0.23
POLKE	17.11 ± 0.23	23.63 ± 0.21
TAALES coverage	11.94 ± 0.62	8.95 ± 0.12
LCA	4.25 ± 0.14	5.20 ± 0.09
L2SCA	2.81 ± 0.17	1.69 ± 0.05
ERRANT	1.48 ± 0.04	0.85 ± 0.04

Table 6: Share (%) of the main models' contributions per tool output group, mean ± SD over the ten CV folds. Because contributions are summed within a group before taking absolute values, tool shares do not add up exactly to the dimension shares in Table 5.

Feature	Tool	Mean rank	Top 20
Determiners per nominal	TAASSC	1.4	10/10
Nonfinite clause proportion	TAASSC	1.6	10/10
Content-word age of acquisition	TAALES	3.1	10/10
First-person pronoun rate	TAASSC	4.1	10/10
All-word meaningfulness	TAALES	5.1	10/10
Possessives per nominal	TAASSC	6.3	10/10

Table 7: The main classifier's six highest-ranked features by mean absolute contribution. One model is trained in each of the ten CV folds. Mean rank: the feature's rank averaged over these ten models (e.g. 1.4 = first in six models and second in four); Top 20: in how many of the ten models the feature is among the 20 most important, so 10/10 means it is in the top 20 of every model.

	All	Half split		Task
	tasks	seen	unseen	holdout
Level classifier				
Level macro F_1	.900	.943	.176	.187
Level accuracy	.944	.970	.315	.274
Within one level	.970	.984	.651	.658
Level QWK	.977	.987	.787	.806
CEFR macro F_1	.953	.975	.529	.566
CEFR accuracy	.975	.987	.662	.710
Level regressor				
MAE (levels)	.613	.556	1.037	1.008
Level QWK	.942	.951	.862	.874
CEFR macro F_1	.716	.738	.597	.623
CEFR accuracy	.860	.871	.750	.782

Table 8: Main models on validation texts: trained on all tasks, on half of each level's tasks (texts of seen and unseen tasks, mean of both halves), and one-task-per-level holdout (unseen tasks, pooled over eight folds). The majority-class CEFR accuracy is .479.

A2. Each output is scored on all evaluation texts, so these profiles describe how the model scores a level rather than how learners at that level develop.

The most important individual features are stable across folds (Table 7): the six highest-ranked features of the main classifier are among the 20 most important features of all ten models trained in the CV folds. Determiner dependents per nominal and the proportion of nonfinite clauses lead the ranking, followed by the age of acquisition of content words (Kuperman et al., 2012), the rate of first-person pronouns, meaningfulness and possessive dependents per nominal. Although lexical sophistication has the largest total share, the two strongest individual features are syntactic: a dimension's share combines many features. In the all-feature classifier, the lexical-token count and the TAALES word count rank first and second; their replacement by syntactic features in the main model, with little loss of accuracy, shows how exclusion changes the explanation while preserving performance. The regressor places the mean length of T-units first in all ten fold models. Importance values show which measures matter, but whether higher values raise or lower a predicted level can only be read from the learned shape functions.

5 Generalisation to New Writing Tasks

Because every task in our data belongs to one level, recognising task-related patterns can help predict the label, even when training and evaluation learners are disjoint. We evaluate this with two task-withholding procedures on the validation texts. In the *half-task split*, we train on the texts of four of the eight tasks at each level (140,000–145,000 training texts), evaluate separately on texts of seen and unseen tasks, repeat the procedure with the complementary four tasks and average the two evaluations; we run this for the all-feature, main and strictest models. In the *one-task-per-level holdout*, we train eight versions of the main models, each withholding one task at every level (15 tasks per fold, 213,000–264,000 training texts), so that every validation text receives a prediction from the model that did not encounter its task, and we pool these predictions.

The half-task split produces a large difference between familiar and unfamiliar tasks (Table 8). The main classifier reaches a level macro F_1 of 0.943 on seen tasks but 0.176 on unseen tasks, and its CEFR accuracy falls from 0.987 to 0.662. On the same seen-task texts, the model trained on all tasks reaches only 0.898, so a model that has to distinguish fewer tasks per level does better on exactly those tasks. The all-feature and strictest classifiers

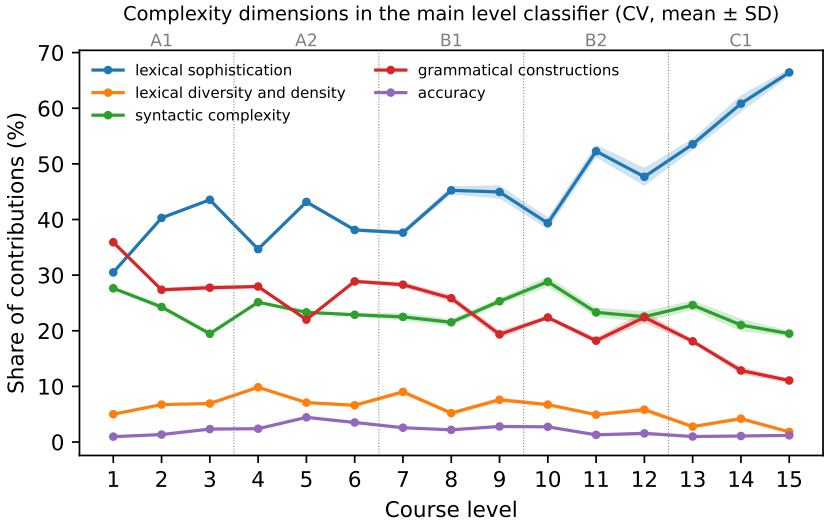


Figure 3: Share of the main level classifier's contributions per linguistic dimension for each course-level output (mean ± SD over the ten CV folds).

show the same pattern (unseen-task level macro F_1 of 0.181 and 0.179), so removing the designated length measures does not remove the transfer problem. Training on seven tasks per level improves the unseen-task level macro F_1 only slightly, from 0.176 to 0.187, far below the 0.900 of the model trained on all tasks; CEFR macro F_1 and accuracy rise from 0.529 and 0.662 to 0.566 and 0.710, compared with 0.953 and 0.975 on familiar tasks. The CEFR accuracy still exceeds the majority-band accuracy of 0.479, so predictive information does transfer across tasks. The regressor transfers better on the CEFR metrics (macro F_1 0.623, accuracy 0.782), mainly at A2 and B1, although its MAE increases from 0.613 to 1.008 levels and its C1 accuracy is lower than the classifier's (0.22 compared with 0.43).

Errors on unseen tasks remain largely ordinal: pooled over the eight folds, 86% of the classifier's predictions (94% of the regressor's rounded predictions) fall within two levels of the true level. The classifier's predictions are, however, pulled towards the middle of the scale: its mean signed error (predicted minus true level) is $+0.5$ levels for texts from A1 tasks and -0.9 and -1.9 levels for texts from B2 and C1 tasks, so that its CEFR accuracy falls from 0.88 at A1 to 0.40 at B2 and 0.43 at C1. Transfer also varies strongly between tasks: the classifier's median per-task CEFR accuracy is 0.63 (interquartile range 0.38–0.80); 14 of the 120 tasks reach at least 0.90, whereas 40 remain below 0.50. The least transferable are advanced tasks such as *Delivering a punch line* (level 14, 16 texts) or *Turning down an invitation* (level 12, 90 texts), whose level the classifier underestimates by 3.9 and 3.8 levels on average.

Removing designated length measures and withholding tasks therefore address different questions: the first retains strong familiar-task performance, whereas the second exposes a substantial transfer gap. Lexical and grammatical features may reflect both proficiency and the course's elicitation conditions; in EFCAMDAT, grammatical structures often occur in identical sentences across texts written for the same unit, probably copied from the task description (Verratti-Souto et al., 2025). Because each task fixes a topic, a genre and the structures its unit practises, and belongs to a single level, the holdout measures transfer to new prompts as a whole: it cannot isolate topic effects, and the topic cues of a level's tasks add up to cues for the level itself. The models therefore partly learn the feature patterns of the eight tasks at each level, and a withheld task brings patterns they have not seen. The same half-split models perform much worse on unseen than on seen tasks, so this contrast cannot be explained by different training-set sizes. A substantial transfer gap also remains when training includes seven tasks per level.

6 Discussion

Our main finding is that the models generalise poorly to new writing tasks: they assess new learners on familiar tasks accurately, but on withheld

tasks they only order texts roughly by level. Unlike the prompt words that drove the bias described by Ballier et al. (2020), our features are derived linguistic measures rather than words, so task cues seem to reach the models through the complexity measures themselves (Alexopoulou et al., 2017). A learner-disjoint evaluation cannot reveal this gap, because every task it tests also occurs in training; proficiency models should therefore also be tested on withheld tasks.

Beyond accuracy, the configurations show what the EBM models rely on. Grammatical constructions contribute most to the level-1 output (36%) but only 11% at level 15, while lexical sophistication rises from 30% to about 66% (Figure 3): the model relies most on constructions to score the lowest level and on vocabulary to score the highest. The two strongest single features are nevertheless syntactic, and once the length measures are excluded, syntactic features replace word counts at the top of the ranking at almost no cost in accuracy, so equally accurate models can rest on different evidence.

For adaptive CALL and level-aware LLMs, the profiles suggest candidate assessment features, text-selection criteria and targets for level-conditioned generation. These candidates still have to be validated across tasks: a feature that carries task cues would push a generation model towards the typical topics of a level instead of its language. Learner writing is also not the same as good input for learners. A generated response has to be grammatical, comprehensible and pedagogically suitable, and matching the feature distribution of learner texts is not enough. Finally, the regressor transfers better to unseen tasks on the CEFR metrics, which suggests that treating proficiency as an ordered scale is worth pursuing.

7 Conclusion and Future Work

In this study, we used a broad set of linguistic complexity and accuracy features to predict CEFR proficiency levels in EFCAMDAT, leveraging EBMs due to their inherent interpretability, and tested how well the models generalise to new writing tasks. For new learners on familiar tasks, the all-feature classifier achieves 97.66% CEFR test accuracy and 0.956 macro F_1, with similar ten-fold results, and the main model retains this performance after excluding the designated length measures, although its predictions still follow text length. The main model's contribution analysis identifies lexical sophistication as the largest dimension, grammatical constructions as the largest contributor to the lowest-level output, syntactic features as the two strongest individual predictors, and stable rankings across folds. The task holdout shows the limits of this result: this accuracy does not carry over to new tasks. We therefore see the profiles as a starting point whose transfer to new tasks still has to be improved and tested.

The next stage of our project builds on these results: we will post-train LLMs to first estimate a user's language level from their language production and then generate a response appropriate to that level. The feature profiles identified here can inform assessment criteria and response analysis, subject to validation across tasks. Evaluation will distinguish the two stages: the accuracy of the inferred level, and the linguistic and pedagogical suitability of the generated response. Generation from a supplied reference level will be compared with generation from the predicted level, which separates assessment errors from generation errors. Our models already estimate the level of a new learner from a single text, and without the ERRANT features they need no corrected version; a related cold-start question is how short that text can be before the estimate becomes unreliable. This can be tested by truncating EFCAMDAT texts or by using the texts below the 20-word minimum from the unfiltered EFCAMDAT, once non-text content such as strings of symbols is removed. Very short texts may contain too few of the structures our features measure for an accurate estimate; if they behave as edge cases, for example because most of their measures are zero, they can be routed to a separate model or flagged as too short for a reliable estimate. Comparing our EBM results with other models, especially models with different architectures, could show how they handle this task differently and which of our findings hold beyond EBMs. An even more strictly filtered feature set, which also removes features that are largely predictable from several others, could test how much broader multicollinearity affects the shares. Independent task and corpus tests, with feature selection repeated within each training partition, will strengthen the evidence before this approach is evaluated with learners.

Limitations

Three limitations matter most: the high accuracy holds only for familiar tasks (Section 5), the main model still follows text length (Section 4.2), and the highest levels have few learners.

Data. We use one English corpus from one online course, and course levels stand in for proficiency. All six tools are built for English, and profiles for other languages rely on different tools (Pilán et al., 2016; Allkivi, 2026; Akef et al., 2026; Loefflad et al., 2026), so dimension shares can be compared across languages only with comparable inventories; languages with richer morphology and writing from exams or classrooms with free topics may give different profiles. The cleaned release removes texts under 20 words, the minimum learners were asked to write, and texts that are unusually short or long for their task (Shatz, 2020). We kept this minimum because very short texts leave little room for the structures our features count. The filter removed 68,976 texts (median 13 words), three quarters of them from the first three tasks of level 1. A1 is still by far the largest band: 191,663 texts against 5,115 for C1, and 96,762 at level 1 against 657 at level 15. The filter nevertheless disproportionately excludes very short beginner responses and may therefore bias the evaluated distribution of beginner writing. Our results do not cover the very short answers a CALL system often receives. The cleaned release also sorts texts into prompt sets automatically, and its release note reports one level-7 task whose texts seem to answer different prompts, so a task label does not guarantee a single prompt.

Tasks. Each task belongs to one level, so we can test transfer to withheld tasks but cannot fully separate a learner's proficiency from task difficulty and course design. The task experiments are also evaluated on the validation set and train on less data than the main models.

Length. Removing the designated length measures excludes raw counts and the measures most associated with length, but the model's predictions still follow text length (Section 4.2). Stronger controls remain to be tested, such as features adjusted for length within each level, comparisons of adjacent levels at matched lengths, or word count as a fixed covariate. Because the cleaned release narrows the range of lengths within each task, length effects may be larger in unfiltered learner writing.

Uncertainty. Balanced sample weights give rare levels equal weight in training but cannot add data, so results for the highest levels are much less precise: the 95% interval of the test F_1 is 0.25 wide at level 15 but 0.004 at level 1. These intervals resample learners, not tasks, so they apply to new learners on the same tasks; Section 5 shows how results vary across tasks. The designated length measures and the near-duplicate list were fixed on the original training split, which covers all tasks, and were not chosen again in the CV folds or task experiments; the all-feature model, which uses neither list, shows the same transfer gap.

Features and interpretation. We use only EBMs because they are interpretable and efficient: a single EBM both classifies texts and gives the exact contribution of each feature, which makes its predictions easy to understand. These contributions show how this model weights the features; other models may weight them differently, so the profiles describe this model rather than proficiency in general. Comparing EBMs with other models is part of our future work (Section 7). The analysis tools can make errors on short or non-standard learner texts, and splitting long segments for L2SCA and POLKE may change sentence- and clause-based measures; we did not measure these effects. Correlated features can spread credit across groups of different size, so the reported shares depend on the feature inventory and on how importance is defined. We removed only near-duplicates, but did not test broader multicollinearity. Removing the near-duplicates cost little accuracy (Section 4.2). Further reducing redundant features is expected to have little effect on predictive accuracy while allowing faster training. Its effect on grouped contribution shares remains to be evaluated. About 90% of the 1,105,310 ERRANT edits concern spelling, orthography or spacing, so our accuracy features mainly capture surface form; construction-specific accuracy measures (Akef et al., 2026) may say more about proficiency.

AI Disclosure

☐ Not applicable (no AI tools used)
☒ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation

☒ Compression/expansion
☐ Generation from prompt

Details: AI assistants (Claude by Anthropic and ChatGPT by OpenAI) were used during the revision of this paper to suggest improvements in wording, phrasing, conciseness, and tone, and to identify repetitive language. They were also used to draft revisions addressing reviewers' comments, check the consistency of numerical values and citations, and generate minor portions of the code. All AI-generated code was reviewed and verified by the authors. The authors also reviewed and revised all AI-suggested text and take full responsibility for the final content of the paper.

References

Soroosh Akef, Amália Mendes, Patrick Rebuschat, and Detmar Meurers. 2026. LLM-powered but rule-grounded: Pedagogically relevant grammatical error characterization for learner model construction. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 851–862, San Diego, California, USA. Association for Computational Linguistics.

Soroosh Akef, Detmar Meurers, Amália Mendes, and Patrick Rebuschat. 2024. Investigating the generalizability of Portuguese readability assessment models trained using linguistic complexity features. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 332–341, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Soroosh Akef, Detmar Meurers, Amália Mendes, and Patrick Rebuschat. 2025. Interpretable machine learning for societal language identification: Modeling English and German influences on Portuguese heritage language. In *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 50–62, Tallinn, Estonia. University of Tartu Library.

Theodora Alexopoulou, Marije Michel, Akira Murakami, and Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1):180–208.

Kais Allkivi. 2026. Towards interpretable models for language proficiency assessment: Predicting the CEFR level of Estonian learner texts. *arXiv preprint arXiv:2602.13102*.

Taylor Arnold, Nicolas Ballier, Thomas Gaillat, and Paula Lissón. 2018. Predicting CEFRL levels in learner English on the basis of metrics and full texts. *arXiv preprint arXiv:1806.11099*.

Nicolas Ballier, Stéphane Canu, Caroline Petitjean, Gilles Gasso, Carlos Balhana, Theodora Alexopoulou, and Thomas Gaillat. 2020. Machine learning for learner English: A plea for creating learner data challenges. *International Journal of Learner Corpus Research*, 6(1):72–103.

Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, and Manel Zarrouk. 2019. A supervised learning model for the automatic assessment of language levels based on learner errors. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning (EC-TEL 2019)*, Lecture Notes in Computer Science, pages 308–320, Cham. Springer.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the Common European Framework of Reference for Languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, 49(3):643–701.

Bram Bulté and Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, pages 21–46. John Benjamins, Amsterdam.

Xiaobin Chen and Detmar Meurers. 2017. Challenging learners in their individual zone of proximal development using pedagogic developmental benchmarks of syntactic complexity. In *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 8–17, Gothenburg, Sweden. LiU Electronic Press.

Xiaobin Chen and Detmar Meurers. 2019. Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, 32(4):418–447.

Xiaobin Chen, Detmar Meurers, and Patrick Rebuschat. 2022. ICALL offering individually adaptive input: Effects of complex input on L2 development. *Language Learning & Technology*, 26(1):1–21.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

Scott A. Crossley and Danielle S. McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2014. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCamDat). In *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines*, pages 240–254, Somerville, MA. Cascadilla Proceedings Project.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Dominik Glandorf, Peng Cui, Detmar Meurers, and Mrinmaya Sachan. 2025. Grammar control in dialogue response generation for language learning chatbots. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9820–9839, Albuquerque, New Mexico. Association for Computational Linguistics.

Dominik Glandorf and Detmar Meurers. 2024. Towards fine-grained pedagogical control over English grammar complexity in educational text generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 299–308, Mexico City, Mexico. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

John A. Hawkins and Paula Buttery. 2010. Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1(1):e5.

Alex Housen and Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.

Yan Huang, Jeroen Geertzen, Rachel Baker, Anna Korhonen, and Theodora Alexopoulou. 2017. The EF Cambridge open language database (EFCAMDAT): Information for users. EF Research Lab, University of Cambridge.

Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. 2021. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, Online. Association for Computational Linguistics.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4):978–990.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.D. thesis, Georgia State University.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3):1030–1046.

Kristopher Kyle and Scott A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4):757–786.

Kristopher Kyle and Scott A. Crossley. 2018. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2):333–349.

Denise Loefflad, Sofia Kathmann, Heiko Holz, and Detmar Meurers. 2026. Criterial features in German: Towards interpretable NLP in readability assessment. In *Proceedings of the 21st Workshop on Innovative*

Use of NLP for Building Educational Applications (BEA 2026), pages 63–77, San Diego, California, USA. Association for Computational Linguistics.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 623–631. Association for Computing Machinery.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Akira Murakami. 2016. Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, 66(4):834–871.

Claude Nadeau and Yoshua Bengio. 2003. Inference for the generalization error. *Machine Learning*, 52(3):239–281.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

John M. Norris and Lourdes Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4):555–578.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Les Perelman. 2014. When "the state of the art" is counting words. *Assessing Writing*, 21:104–111.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2025. A framework for proficiency-aligned grammar practice in LLM-based dialogue systems. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 978–987, Vienna, Austria. Association for Computational Linguistics.

Nelly Sagirov and Xiaobin Chen. 2026. POLKE: A system for comprehensively annotating pedagogically-oriented grammatical structure use in language production. *Research Methods in Applied Linguistics*, 5(3):100336.

Itamar Shatz. 2020. Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, 6(2):220–236.

Sowmya Vajjala. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

Daniela Verratti-Souto, Nelly Sagirov, and Xiaobin Chen. 2025. NLP-powered quantitative verification of the English Grammar Profile's structure-level assignment. *Annual Review of Applied Linguistics*, 45:109–130.

Helen Yannakoudakis, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

A Results for All Feature Configurations

Table 9 reports all 16 feature configurations on the validation and test sets. The error-composition configuration adds four proportions: missing, unnecessary, replacement and other edits divided by all edits, with zero assigned when no edits occur. ERRANT recorded no edits of the other type, so that proportion is constant and removed by the feature filter. The mutual-information and information-gain rankings were computed on the training texts, and each selects the top 200 eligible features.

	Validation					Test				
Feature set	Lvl F_1	QWK	CEFR	$CEFR_d$	MAE	Lvl F_1	QWK	CEFR	$CEFR_d$	MAE
All features	.904	.979	.956	.937	.597	.897	.979	.956	.936	.602
All − direct length	.902	.978	.955	.937	.601	.897	.979	.957	.938	.605
All − length-related (main)	.900	.977	.953	.935	.613	.896	.978	.956	.933	.616
All − near-duplicates	.900	.978	.955	.934	.607	.896	.979	.956	.934	.611
All − direct length − near-duplicates	.900	.978	.955	.935	.609	.895	.978	.955	.932	.613
All − length-related − near-duplicates	.900	.976	.953	.932	.622	.894	.977	.954	.929	.626
All + error composition	.902	.979	.956	.937	.596	.897	.979	.957	.937	.601
Complexity (without ERRANT)	.902	.979	.956	.937	.597	.897	.979	.956	.936	.603
Complexity − direct length	.901	.978	.955	.936	.601	.896	.979	.955	.936	.605
Complexity − length-related	.899	.977	.953	.935	.613	.895	.978	.956	.933	.617
Top 200 by mutual information	.746	.941	.880	.855	.726	.736	.940	.869	.841	.735
Top 200 by information gain	.729	.939	.874	.847	.722	.716	.936	.864	.836	.730
POLKE constructions only	.722	.941	.869	.847	.726	.716	.939	.867	.839	.735
ERRANT error features only	.242	.759	.608	.609	1.289	.240	.756	.590	.592	1.284
Direct length measures only	.255	.825	.673	.674	1.083	.250	.818	.653	.654	1.096
Majority class / median level	.026	.000	.129	.129	2.229	.026	.000	.128	.128	2.221

Table 9: All 16 feature sets on the validation and test sets. Lvl F_1 and QWK: level classifier; CEFR: macro F_1 of CEFR bands from summed level probabilities; $CEFR_d$: macro F_1 of a classifier trained directly on the five CEFR bands; MAE: level regressor.

B Per-Class Results

Table 10 reports all 15 course levels on the test set, and Tables 11 and 12 the corresponding cross-validation results. P: precision; R: recall, equal to the accuracy within the true class; the macro recall is the balanced accuracy.

		All features			Main			
Course level	Support	P	R	F_1	P	R	F_1	95% CI
1	14,613	0.975	0.968	0.971	0.975	0.968	0.971	0.969–0.973
2	8,492	0.934	0.941	0.937	0.934	0.942	0.938	0.934–0.942
3	5,753	0.940	0.943	0.942	0.941	0.942	0.942	0.936–0.947
4	10,679	0.969	0.962	0.966	0.969	0.961	0.965	0.962–0.968
5	5,586	0.959	0.961	0.960	0.957	0.959	0.958	0.954–0.962
6	3,251	0.907	0.910	0.909	0.904	0.911	0.907	0.897–0.916
7	5,602	0.932	0.920	0.926	0.928	0.916	0.922	0.916–0.928
8	2,148	0.846	0.873	0.859	0.840	0.872	0.856	0.843–0.867
9	1,561	0.827	0.876	0.851	0.827	0.876	0.851	0.835–0.865
10	1,680	0.946	0.938	0.942	0.946	0.941	0.943	0.934–0.952
11	800	0.899	0.881	0.890	0.900	0.879	0.889	0.871–0.906
12	338	0.864	0.899	0.881	0.873	0.893	0.883	0.855–0.907
13	416	0.918	0.921	0.920	0.918	0.916	0.917	0.893–0.937
14	107	0.833	0.794	0.813	0.840	0.785	0.812	0.725–0.875
15	54	0.685	0.685	0.685	0.673	0.685	0.679	0.521–0.774
Macro	–	0.896	0.898	0.897	0.895	0.896	0.896	0.881–0.905

Table 10: Per-level results on the held-out test set. Support is the number of texts with the true label. 95% CI: learner-clustered bootstrap percentile interval of the main model's F_1 (2,000 resamples; Section 3.4).

		All features			Main		
CEFR band	Support	P	R	F_1	P	R	F_1
A1	16280.5	0.990	0.989	0.989 ± 0.001	0.990	0.988	0.989 ± 0.001
A2	11007.5	0.975	0.970	0.973 ± 0.001	0.974	0.970	0.972 ± 0.001
B1	5219.5	0.947	0.964	0.955 ± 0.003	0.946	0.962	0.954 ± 0.002
B2	1536.9	0.938	0.934	0.936 ± 0.006	0.934	0.932	0.933 ± 0.006
C1	453.8	0.937	0.906	0.921 ± 0.011	0.934	0.902	0.917 ± 0.012
Macro	–	0.957	0.952	0.955	0.955	0.951	0.953

Table 11: Per-band results in ten-fold cross-validation. Entries are fold means; F_1 also shows its fold standard deviation. Support is the mean number of texts per held-out fold.

		All features			Main		
Course level	Support	P	R	F_1	P	R	F_1
1	8214.9	0.975	0.967	0.971 ± 0.001	0.975	0.966	0.970 ± 0.001
2	4834.7	0.936	0.942	0.939 ± 0.002	0.936	0.941	0.939 ± 0.002
3	3230.9	0.945	0.947	0.946 ± 0.003	0.945	0.947	0.946 ± 0.004
4	5979.1	0.971	0.960	0.966 ± 0.002	0.971	0.960	0.965 ± 0.002
5	3169.1	0.963	0.963	0.963 ± 0.002	0.962	0.963	0.962 ± 0.003
6	1859.3	0.908	0.922	0.915 ± 0.004	0.904	0.921	0.912 ± 0.004
7	3185.8	0.927	0.920	0.923 ± 0.005	0.927	0.916	0.922 ± 0.005
8	1200.9	0.841	0.882	0.861 ± 0.007	0.839	0.882	0.860 ± 0.007
9	832.8	0.835	0.882	0.857 ± 0.005	0.827	0.882	0.854 ± 0.007
10	877.8	0.943	0.939	0.941 ± 0.007	0.941	0.938	0.940 ± 0.007
11	421.0	0.871	0.884	0.877 ± 0.015	0.863	0.880	0.871 ± 0.016
12	238.1	0.878	0.866	0.871 ± 0.019	0.874	0.866	0.870 ± 0.018
13	284.8	0.923	0.927	0.925 ± 0.007	0.918	0.925	0.921 ± 0.009
14	108.7	0.850	0.804	0.825 ± 0.026	0.855	0.807	0.829 ± 0.021
15	60.3	0.829	0.683	0.746 ± 0.051	0.826	0.680	0.743 ± 0.048
Macro	–	0.906	0.899	0.902	0.904	0.898	0.900

Table 12: Per-level results in ten-fold cross-validation. Entries are fold means; F_1 also shows its fold standard deviation. Support is the mean number of texts per held-out fold.

C Tool Shares by Course Level

Figure 4 breaks the dimension profile of Figure 3 down by tool output group. TAALES indices gain weight towards the highest levels, while the share of POLKE constructions falls from about a third at level 1 to about a tenth at level 15.

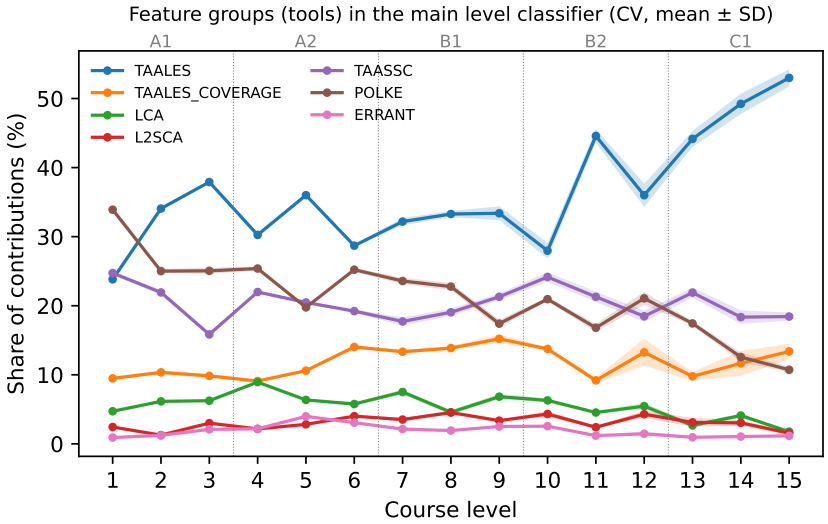


Figure 4: Share of the main level classifier's contributions per tool output group for each course-level output (mean ± SD over the ten CV folds).

Automated L2 Speech Assessment with Audio-Based Self-Supervised Neural Networks

Eli Stafford[1,2*], **Sylvain Coulange**[2,3,4]

[1]ALMAnaCH, INRIA Paris, Paris, France

[2]Service des Langues, Université Grenoble Alpes, Grenoble, France

[3] Laboratory of Linguistics and Didactics of Foreign and Mother Tongues (LIDILEM), Université Grenoble Alpes, France

[4] CNRS, Institute of Engineering, Grenoble Computer Science Laboratory (LIG), Université Grenoble Alpes, France

eli.stafford@inria.fr, sylvain.coulange@univ-grenoble-alpes.fr

Abstract

Automated L2 speech assessments traditionally use vectors of acoustic, lexical, and syntactic metrics as representations of speech for statistical learning. This work explores the use of audio-based self-supervised neural networks to supplement existing scoring strategies with robust computational representations. We explore architectural considerations of neural L2 scoring, probe their effects in the context of a reliable and equitable examination, and measure their linguistic coverage compared to heuristic-based and neural text-based systems. We find that their use in automated speech scoring corresponds to a measurable improvement in performance, while revealing evidence of major drawbacks in terms of explainability and negligible increases in robustness compared to traditional systems, especially in low-resource contexts.

1 Introduction

L2 automatic speaking proficiency assessment is an important advancement in second language education that enhances the practicality and deliverability of speech assessments facing ever-growing demand (Van Moere and Downey, 2016). Private testing bodies that employ automated speech assessments, such as Pearson (with the Test of English Academic), Cambridge University Press & Assessment (with the Cambridge English Testing System), or Duolingo, are now widely used. While publicly available descriptions of their assessment methods cite the usage of statistical learning with vectors of a variety of computed heuristics on large, professionally graded student corpora (Saito et al., 2023; Xu et al., 2021; Park et al., 2025), they do not mention the usage of neural methods in text or speech.

In the academic literature, machine learning models show a strong correlation with human judgments on this task (Qian et al., 2025a; Zechner and Evanini, 2019). The majority of competitive systems utilize cascaded architectures, with transcriptions of student speech (produced by an automatic speech recognition model like Whisper (Radford et al., 2023)) fed into text-based classifiers built on pre-trained text encoders or decoder-only Large Language Models (Qian et al., 2025a; Bannò et al., 2025; Wang et al., 2021). Although there have been explorations of multimodal models (Ma et al., 2025) and correlations of transcription model outputs with pronunciation accuracy (Ballier and Méli, 2024; Liu et al., 2025), the use of large pre-trained audio encoders that consider only speech without a transcription, such as Wav2Vec2.0 (Baevski et al., 2020) and WavLM (Chen et al., 2022), remains understudied, with few experiments investigating their application (Bannò and Matassoni, 2023).

These pre-trained audio encoders are encoder-only deep neural networks built on the transformer architecture (Vaswani et al., 2017), trained on unlabeled audio inputs to produce high-dimensional representations of speech. These self-supervised learning models (SSLs), also called foundation models, can be applied to a variety of tasks by adding specialized classification heads to their base architecture. They achieve high performance across a wide variety of spoken language processing benchmarks (Yang et al., 2024), including speaker identification, emotion recognition, speaker diarization, and low-resource automatic speech recognition (Yi et al., 2021). L2 speaker recognition, as an utterance-level classification task, is a prime use case for audio SSLs and has the advantage of targeting the speech quality of

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

*This work was carried out during an internship at Service des Langues.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 334–345, November 19–20, 2026, University of Tartu Library

an L2 language learner, which is unavailable in the text-only setting. Compared to state-of-the-art speech-based heuristics (Coulange, 2025), which are calculated directly from audio to augment text-only systems and are based on linguistic intuitions of speech quality, audio SSLs can learn the most pertinent features of speech automatically.

However, the usage of any neural model comes with a host of ethical concerns regarding validity, practicality, and transparency (De Jong et al., 2025; Vajjala et al., 2026). Audio SSLs suffer both from classic questions of validity and dataset overfitting due to their black-box nature (Tsai et al., 2025), as well as questions of speaker group bias (Herron et al., 2026, 2025). Utilizing these systems for reliable and equitable examinations requires close examination of these factors.

Further, while other tasks have well-defined architectures to leverage pertinent encoded features, such as graph-based mechanisms for synthetic voice detection (Jung et al., 2022) and temporal average pooling for speaker identification (Wang et al., 2024), there is no agreed-upon architecture for L2 spoken language assessment when using audio encoder models.

In this work, we explore a wide range of possible architectures for audio SSLs based on their application to other tasks, with the goal of optimizing the aggregation of information over time (phoneme level to speaker level) and the abstractness of representation at each model layer (low-level acoustic information to abstract speech representation). We then compare the performance of these models to heuristic-based and text-only neural baselines, and combine their usage to see if audio SSLs are a promising supplement to existing pipelines by targeting synergistic aspects of speech. Additionally, we explore the information encoded in our best-trained model via layerwise probing to determine possible avenues of exploitation of non-linguistic information, as well as the model's encoding of linguistic features like pragmatics, pronunciation, and grammar.

2 Architectures

While audio SSLs can sometimes be used end-to-end by feeding the final layer as input into a classifier head, many tasks require preliminary aggregation of model representations. As SSLs are deep neural networks given audio as input (a sequential medium), each input produces an embedding vector for each time slice (usually ∼20 ms of audio) at each hidden layer. A model like Wav2Vec2.0, with 13 layers, given an input of 4,000 time slices (∼80 seconds of audio), produces 52,000 available embeddings, each representing a specific slice of audio at a specific hidden layer (level of abstraction (Rogers et al., 2020)). We separate possible aggregation approaches into *Layerwise Aggregation* approaches, which aggregate each model layer into a single dimension, and time-wise *Pooling* approaches, which aggregate a sequential audio input into a single representation. Together, these approaches produce a single embedding vector, which is fed to a feed-forward regression head to predict student level.

Layerwise Aggregation strategies seek to combine disparate representational information across layers into a single output. Embeddings in lower layers of an SSL generally contain raw acoustic information, while higher layers contain higher-level learned abstractions like word classes or intonation. We experiment with three major aggregation strategies, which can be applied independently of pooling strategies along the time dimension:

- Layerwise: The simplest of our approaches trains a classifier head on each layer of our frozen SSL to output a global score for that layer, and then trains a linear regression classifier on these scores.

- Weighted Mean: This approach, standard for frozen SSLs in utterance classification (Yang et al., 2024), trains a vector with length equal to the number of layers in our SSL during fine-tuning, and performs weighted mean averaging based on these weights. The weights for each layer are learned at training time.

- FT/PEFT: In this approach, we feed only the final layer of our SSL into a classification head and leave all previous layers unfrozen for training; in a generic sense, this can be called *full fine-tuning* (FT). Conventional full fine-tuning of audio SSLs is prone to catastrophic forgetting, especially in low-resource settings, so we opt for a Parameter-Efficient Fine-Tuning (PEFT) method, namely Low-Rank Adaptation (LoRA) (Hu et al., 2022) on the query and key matrices of each transformer layer.

Pooling approaches aggregate sequences of vectors (along the time dimension) into a single embedding representing the entire utterance. We experiment with three pooling approaches, applied after layerwise aggregation (or directly on layers in the case of the multilayer strategy):

- Mean Pooling: The simple average of each embedding across the time or token dimension. This erodes information specific to a time slice (such as particular mispronunciations or accent features) in favor of a generalized representation.
- Attention Pooling: Each vector is assigned an attention value by a linear feed-forward layer, and we compute a weighted average based on the softmax of these values. This allows the model to select pertinent time slices or tokens based on their learned relevance.
- Four-Head Pooling: The third approach is a specific variant of attention pooling utilizing a more complex attention mechanism with four heads instead of one, based on the baseline approach of the Speak and Improve Challenge (Qian et al., 2025b). This approach considers the entire context of the phrase when determining the weight of a time slice, doing so with four distinct heads. The resulting weighted means are concatenated, and a non-linear regression head utilizes this vector to predict student level. This allows us to have multiple specialized heads that consider different non-linear relationships across parts of the utterance.

Probing models (probes) are secondary, auxiliary models trained on top of a pre-trained primary model to test whether specific information is encoded within the primary model's internal representations. Probes can identify the presence of specific information in model outputs by training on secondary classification objectives, where high performance suggests the presence of the information necessary to complete that task. Previous explorations of audio encoder layers have found that these models generally do not encode non-acoustic linguistic information like word meaning (Pasad et al., 2021), therefore any spoken language assessment system solely comprised of these models will inherit this acoustic bias. To determine which linguistic features of speech our best system encodes, we probe model layers for phonetic, grammatical, and pragmatic fluency according to holistic scoring data provided in our dataset. Furthermore, to quantify speaker group bias in our best-performing model, we train probes for each model layer on a speaker identification task, where each input is classified with the student's speaker ID. Finally, we probe for task identification, which identifies the specific task being performed by the student since our dataset includes recordings of different tasks. From a pedagogical standpoint, student speaker features and the task being performed have a non-causal correlation with the proficiency level of an utterance and should not be prominently present in our model's outputs.

3 Experiments

3.1 Dataset

Models were evaluated on a novel dataset collected for the SELF Production Orale project[*]. It consists of 873 recordings of spontaneous speech from 339 L2 learners of English responding to 10 diverse tasks designed for this project. Each recording is rated on a continuous scale between 1.0 and 7.0, with integer values representing discrete CEFR levels[†]. Fourteen raters, comprising CLES[‡] certified raters and teachers of English as a second language, were recruited to create gold ratings for these recordings. At 150 ratings per rater, this resulted in 2,100 ratings (an average of 2.4 ratings per recording), which were then normalized and averaged to produce continuous, reliable gold ratings.

It is important to note that this data suffers from class imbalance at extreme ranges, with $\sim 10\%$ of all data at the A1 and C1+ levels (Appendix A). Determining if audio SSLs can deal with this imbalance is a goal of this work, as statistical learning of heuristics do not handle this imbalance sufficiently.

In a *holistic scoring* paradigm, students were also rated on three *sub-scores* in addition to their overall score. These sub-scores, on a A1–C2 categorical scale, measure more specific axes of linguistic competency, including Pragmatics, Vocabulary, and Pronunciation.

[*] https://self.univ-grenoble-alpes.fr/self-production-orale-1664940.kjsp

[†] $1.0 \leq x < 2.0$ for A1, $2.0 \leq x < 3.0$ for A2, up to $5.0 \leq x$ for C1+

[‡] https://www.certification-cles.fr/

3.2 Architecture and Hyperparameters

We choose model and architecture combinations for evaluation on this dataset based on best-performing previous explorations of their application to automated L2 assessment and, where literature is absent, their performance on related tasks.

We test the proposed architectures on Wav2Vec2.0, the only previously used SSL for this task, and WavLM-Large, a much larger and more performant pre-trained audio encoder. We explore all possible combinations of the aggregation strategies previously mentioned on Wav2Vec2.0. WavLM-Large is prohibitively expensive for full fine-tuning, even with PEFT methods, so we experiment only with the multilayer and weighted-mean layer aggregation methods, while testing all pooling methods.

After aggregation and pooling, audio models were passed to a two-layer regression head with a 256-dimensional intermediate layer and a dropout rate of 0.1. Dropout, learning rate, optimizer choice, and intermediate layer size were determined via layerwise prediction results on the target task using mean pooling, comparing results for the best-performing layer under each configuration.

3.3 Baseline System

Our heuristic baseline system represents student speech with a vector of 51 linguistic metrics, input into a Linear Mixed Model predicting student level. These 51 metrics come from a wide range of literature on automatic L2 spoken language assessment and include surface-level heuristics such as pause frequency and position, word count and word per second, measurements of vowel stress contrast, intelligibility scores derived from ASR system confidences, syntactic complexity, and the proportion of vocabulary present in the A1 Oxford Learner's Dictionary. This wide range of scores, covering pronunciation, intelligibility, grammar, and lexical context, serves as a strong baseline against which to compare our audio-based neural system, specifically its acoustic bias.

Additionally, we provide a BERT-based baseline trained on our data using the architecture from Qian et al. (2025a). This system is text-only, using BERT-base (Devlin et al., 2019) as a backbone model and the same four-head classifier head as in our audio model experiment.

3.4 Evaluation Methodology

Models are evaluated on continuous outputs using the standard spoken language assessment metrics of RMSE (Root Mean Squared Error) and PCC (Pearson Correlation Coefficient) against continuous gold predictions. We additionally evaluate accuracy and macro F1 after binning continuous outputs into discrete CEFR classes. While continuous metrics are standard and better represent overall model performance, our binning metrics highlight specific weaknesses across CEFR levels, specifically performance on outlier classes with macro F1 score. We consider the linguistic heuristic system to be our primary baseline with respect to these metrics.

Due to our small dataset, we evaluate each of our models using a stratified 10-fold cross-validation strategy, where ten 10% data samples are withheld during training for evaluation. Each of these 10 samples is balanced in terms of CEFR level (binned rating). All metrics proposed are the average of performance on each fold. As student level and recording level are highly correlated, we ensure that recordings featuring the same student are never split across training and test sets.

3.5 Probing

All probing tasks were carried out with a two-layer linear probe featuring an intermediate layer size of 256. Each task was formulated as a classification task, with target classes being sub-score CEFR levels, task IDs, or speaker IDs. Sub-score (Pragmatics, Pronunciation, and Grammar) and task ID probes were trained on a single train/test split (fold) taken from other experiments, whereas speaker ID probes were trained on a modified split ensuring each student present in the training set had at least one recording in the test set.

3.6 Combination

To determine whether pre-trained SSLs predict based on the same information as the baselines—and are therefore alternative rather than synergistic approaches—we experiment with combinations of scores from the heuristic-based baseline model, the neural text-based model, and the best audio SSL model. The combinations of neural text-based baseline + audio SSL, our heuristic baseline + neural text-based baseline, and all 3 scores combined, we train a linear regression model with the combined scores as vector

Model	Aggregation	Pooling	RMSE	PCC	Acc	Macro F1
wav2vec2.0	lora	meanpool	.626	.760	51.3	.425
		attnpool	.632	.757	51.5	.409
		fourhead	.651	.739	51.6	.413
	wmean	meanpool	.695	.691	50.3	.394
		attnpool	.704	.683	49.0	.371
		fourhead	.640	.748	53.6	.466
	multilayer	meanpool	.606	.778	55.4	.471
		attnpool	.586	.794	54.2	.482
		fourhead	.611	.774	55.0	.497
wavlm	**wmean**	meanpool	**.564**	**.814**	**57.2**	**.499**
		attnpool	.633	.756	50.1	.460
	multilayer	meanpool	.643	.745	51.3	.424
		attnpool	.693	.694	47.8	.357

Table 1: Audio SSL Architecture Comparison

input (either 2 or 3 scores) and the gold continuous scores as targets. Each regression model is fit on the same training data as our main models and evaluated on the unseen test set.

4 Results

4.1 Audio SSL Architecture

Here we present results for Wav2Vec2.0 and WavLM on our primary task across architecture combinations.

We find that while there are clearly optimal strategies for each model, there is no single best layerwise aggregation strategy across models, nor a best pooling strategy across layerwise aggregation strategies (Table 1). WavLM showed a 0.11 decrease in RMSE (and a corresponding improvement across other metrics) when using the weighted-mean strategy compared to the multilayer strategy, whereas Wav2Vec2.0 achieved a 0.1 decrease in RMSE when using the multilayer strategy over the weighted-mean strategy. These architectures are conceptually distinct: weighted mean targets utterance-level features, while multilayer selects specific time slices and layers for aggregation. This points to a fundamental difference in necessary architectural considerations across different model sizes.

WavLM-Large with the weighted-mean and mean-pooling architecture achieved the highest performance, reaching 0.499 macro F1, 0.564 RMSE, and 0.814 PCC, representing a 0.038 decrease in RMSE compared to Wav2Vec2.0's best combination. While the RMSE values for the mean pooling and attention pooling architectures were comparable (a 0.016 difference in RMSE), gains in macro F1 were more significant (0.05).

4.2 Inter-Approach Comparisons

Included Scores	RMSE	PCC	Acc %	Macro F1
Heuristic LMM	0.654	0.758	50.5	0.524
WavLM-Large	**0.564**	**0.814**	**56.6**	0.499
Neural text Baseline	0.694	0.704	51.6	0.400
WavLM-Large + Neural Text	0.617	0.784	50.5	0.514
LMM + Neural Text	0.637	0.768	51.8	0.535
All	0.590	0.808	56.0	**0.567**

Table 2: Score Comparison/Combination

Here we compare the results of our heuristic-based baseline model, the neural text-based model, and the best audio SSL model, as well as all possible combinations. All models were evaluated on a single kfold iteration. The analysis of consistency over multiple iterations, with different splits, can be found in Appendix B.

The performance of our audio SSL strategy, WavLM-Large with weighted mean layerwise aggregation and mean pooling, outperforms the baseline system (Table 2) by 0.9 RMSE and 0.046 PCC, a significant margin. This is expected when using such a large model, but the gains are less dramatic compared to our binning statistics— with an under 3-percentage-point increase in accuracy, and a 0.025 decrease in macro F1 score. While the decrease in RMSE is promising, the lack of significant improvement in F1 score (which indicates fairness across CEFR levels) suggests that this performance increase may stem from better modeling of the unbalanced dataset distribution, or other statistical artifacts, rather than improved representa-

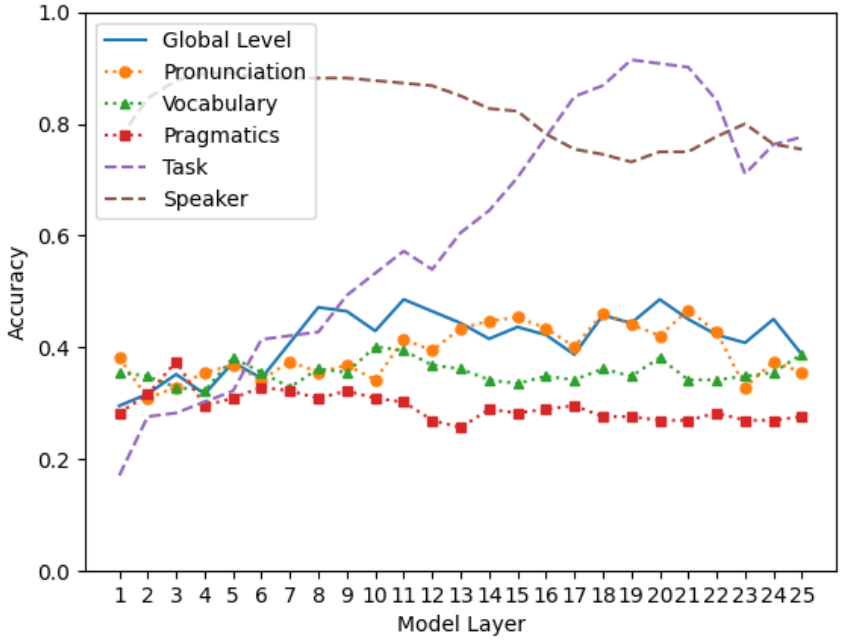


Figure 1: Layerwise holistic scoring, speaker, and task probing compared to base task: WavLM-Large

tional generalizability.

All approaches, except for the neural text-based baseline, fall in the range of 0.56–0.65 RMSE, indicating an average error of just over half a CEFR level. Similarly, accuracies range between 50% and 56%, and macro F1 scores between 0.50 and 0.57. This narrow performance range indicates either that baseline and neural models are functionally equivalent (leaving little room for improvement) or that dataset size imposes a performance ceiling.

When combining the heuristic and neural text-based baselines, we observe that despite its weak standalone performance, the BERT model adds appreciable value (a 0.017 decrease in RMSE).

WavLM-Large alone achieves the lowest RMSE at 0.564, whereas combining all scores leads to the highest absolute accuracy and macro F1 score, alongside a 0.026 increase in RMSE compared to WavLM-Large alone. We believe this demonstrates that none of the three evaluated approaches are redundant, offering measurable performance gains; however, tradeoffs between CEFR coverage and RMSE remain necessary due to dataset imbalance. Analysis of predictor estimates (relative importance of model scores/heuristics in the all scores model) can be found in Appendix C.

5 Speech Encoder Probing Results

Figure 1 displays layerwise probing accuracy results on WavLM-Large for our target CEFR predictions, holistic score classification, task identification, and speaker identification.

Holistic score probing yields results aligned with SSL intuitions. WavLM-Large probes predict Pronunciation accuracy with performance similar to overall level prediction (both Pronunciation accuracy and Global Level accuracy hover around 50% at layer 18), while poorly predicting semantic or discourse-level information (Vocabulary and Pragmatics remain below 40% accuracy). This suggests that the model does not account for a broad range of linguistic competencies when predicting recording levels, focusing instead primarily on acoustic properties of speech.

Our speaker identification probe aligns with previous findings. On the VoxCeleb speaker identification benchmark (Nagrani et al., 2017), WavLM-Large achieves accuracies exceeding 90%. This indicates that if any correlation exists between speaker identity and CEFR level, classifiers built on large models like WavLM-Large can exploit non-linguistic speaker characteristics to predict CEFR scores. This risks poor generalization to unseen speakers and is pedagogically undesirable.

Interestingly, our task probes (question ID) were able to accurately predict the task a stu-

dent was performing in the final layers of the model, achieving up to 90% accuracy. This is a notable result: although WavLM-Large excels at keyword identification (a key signal for task type) under other classifier configurations, the time-wise mean-pooling approach evaluated here would theoretically remove identifying task information. This points to a potential prediction strategy where the model identifies student level (approximated via speaker characteristics) alongside the task being performed—features that correlate highly with recording level but non-causally with the underlying linguistic proficiency.

6 Discussion

By continuous metrics, WavLM-Large is our best performing model by a significant margin with 0.9 RMSE below baseline, and 0.4 RMSE below the linear combination of all scores. However, its usage comes with a host of issues that make it difficult to justify in this context, including lack of improvements at extreme class ranges (macro F1 score), speaker bias, predictor coverage, and computational efficiency.

As far as our data suggests, there is no definitive optimal architecture for SSLs. Our results, only examining two models, give contradictory evidence for each, where Wav2Vec2.0's optimal architecture targets extremely specific production information (specific layers at specific timesteps), whereas WavLM-Large aggregates all utterance information into a specific vector. Without experiments on more models, we can only tenuously say that smaller models (Wav2Vec2.0) work better with attention-based architectures, and larger models (WavLM-Large), with hypothetically better or more consistent representations, are better with more severe aggregation.

Our probing results follow from previous explorations in audio SSL tendencies, and suggest that the information they consider at prediction time may not be entirely linguistic. These results, in addition to the lack of large macro F1 improvement indicate that our best audio SSL may be capitalizing on statistical artifacts related to task type and speaker distribution to find their improvements, rather than discovering more pertinent linguistic data that is not covered by our heuristic or text-based models.

We hoped that the introduction of deep neural networks would easily contend with extreme CEFR and levels and lead to a large increase in macro F1 score. It is clear now that SSLs suffer from the same dataset dependence as heuristic based models. Simply collecting more A1/C1+ levels, or perhaps native-level responses to the same tasks, would probably improve performance for all approaches at these levels, and at this time it is not clear if neural approaches would perform significantly better than traditional approaches if given more data.

7 Conclusion

In this work we have explored the possible uses of SSLs in the context of an automated assessment of L2 learners of English. We determined the best architectures for leveraging pre-trained audio encoder models for scoring spontaneous speech utterances, the fact that this decision changes between base models, and found slight improvements over traditional heuristic baseline and neural text-based approaches. Additionally, we find evidence that learned representations of text and speech can be used to improve heuristic-based systems, suggesting that their coverage of linguistic competencies differs in some way from traditional systems, that this coverage would be limited to acoustic phenomena, but that these models have the capacity for utilizing non-causal correlations between both the task being performed and speaker bias to determine recording level. Overall, audio SSLs represent a promising avenue of work in the creation of automated speech assessments, but their merely minor performance improvement still leave much room for debate about their nature as black-box systems, and their adherence to principles of language education.

8 Further Work

The use of zero-shot systems (text-based/multimodal LLMs) are a possible avenue of improvement with respect to extreme class levels, which could identify these outliers without more annotated A1/C1+ data, which is scarce in publicly available corpora.

While a major goal of this work was to see if audio SSLs could improve performance on an unbalanced dataset, our results regarding architecture choices must be verified on balanced data. For our conclusions to generalize across datasets, it will be important to attempt the same experiments in a more balanced setting, as the one utilized in the

Speak and Improve Challenge (Knill et al., 2025).

On the black-box nature of these systems, there is of course a large body of literature in interpretability which could apply in this setting, especially input-output oriented approaches with existing disfluency annotated corpora. Certain architectures covered (namely weighted mean layerwise pooling and timewise attention pooling) are ripe with data for close analysis on which layers and specific utterance contribute most to a recordings rating. This type of interpretability is important for justifying their use to educators, and can give insights to students without more complex error annotating models.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☐ Generation of minor elements
☐ Translation
☐ Compression/expansion
☐ Generation from prompt

Details: Generative AI was utilized for grammar correction and spellchecking for the production of this document.

Acknowledgments

This work was carried out in the context of the SELF Production Orale project, headed by Monica Masperi, hosted between Service des Langues and Laboratoire d'Informatique de Grenoble of Université Grenoble Alpes. This project was funded by a FITInnovE grant provided by the Pôle Universitaire d'Innovation of Université Grenoble Alpes, with additional substantial funding from the Service des Langues of Université Grenoble Alpes.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Nicolas Ballier and Adrien Méli. 2024. Investigating acoustic correlates of whisper scoring for L2 speech using forced alignment with the Italian component of the ISLE corpus. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 20–32, Rennes, France. LiU Electronic Press.

Stefano Bannò, Rao Ma, Mengjie Qian, Siyuan Tang, Kate Knill, and Mark Gales. 2025. Natural language-based assessment of L2 oral proficiency using LLMs. Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE).

Stefano Bannò and Marco Matassoni. 2023. Proficiency assessment of L2 spoken English using wav2vec 2.0. In *2022 IEEE Spoken Language Technology Workshop (SLT)*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sylvain Coulange. 2025. *Évaluation automatique de la parole spontanée en anglais langue étrangère : le rôle des pauses et de l'accent lexical dans la compréhensibilité du locuteur*. Theses, Université Grenoble Alpes.

Nivja De Jong, Stephan Raaijmakers, and Dineke Tigelaar. 2025. Developing high-quality, practical, and ethical automated L2 speaking assessments. *System*, 134:103796.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Felix Herron, Solange Rossato, Alexandre Allauzen, Benoit Favre, and François Portet. 2025. *Speaker Group Encoding in Self-supervised Speech Recognition Models*, pages 121–132. Springer Nature Switzerland.

Felix Herron, Solange Rossato, Alexandre Allauzen, and François Portet. 2026. Identifying and typifying demographic unfairness in phoneme-level embeddings of self-supervised speech recognition models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33603–33620, San Diego, California, United States. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6367–6371.

Kate Knill, Diane Nicholls, Mark J. F. Gales, Mengjie Qian, and Pawel Stroinski. 2025. Speak and improve corpus 2025: an L2 English speech corpus for language assessment and feedback. Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE).

Yao Liu, Faizahani binti Ab Rahman, and Farah binti Mohamad Zain. 2025. A systematic literature review of research on automatic speech recognition in EFL pronunciation. *Cogent Education*, 12(1):2466288.

Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Assessment of L2 oral proficiency using speech large language models. In *Proc. Interspeech 2025*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In *Proc. Interspeech 2017*, pages 2616–2620. ISCA.

Yena Park, Ramsey Cardwell, Sarah Goodwin, Ben Naismith, Geoffrey T. LaFlair, Kevin P. Yancey, and Xiaowan Zhang. 2025. Assessing speaking on the Duolingo English Test. Whitepaper, Duolingo.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Mengjie Qian, Kate Knill, Stefano Bannò, Siyuan Tang, Penny Karanasou, Mark J. F. Gales, and Diane Nicholls. 2025a. Speak and improve challenge 2025. In *Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE 2025)*. ISCA.

Mengjie Qian, Kate Knill, Stefano Bannò, Siyuan Tang, Penny Karanasou, Mark J. F. Gales, and Diane Nicholls. 2025b. Speak and improve challenge 2025: Tasks and baseline systems. Proc. 10th Workshop on Speech and Language Technology in Education (SLaTE).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Kazuya Saito, Konstantinos Macmillan, Magdalena Kachlicka, Takuya Kunihara, and Nobuaki Minematsu. 2023. Automated assessment of second language comprehensibility: Review, training, validation, and generalization studies. *Studies in Second Language Acquisition*, 45(1):234–263.

Fong-Chun Tsai, Kuan-Tang Huang, Bi-Cheng Yan, Tien-Hong Lo, and Berlin Chen. 2025. Mitigating data imbalance in automated speaking assessment. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 688–693. IEEE.

Sowmya Vajjala, Bashar Alhafni, Stefano Bannò, Kaushal Maurya, and Ekaterina Kochmar. 2026. Opportunities and challenges of LLMs in education: An NLP perspective. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 368–404. Association for Computational Linguistics.

Alistair Van Moere and Ryan Downey. 2016. Technology and artificial intelligence in language assessment. In Dina Tsagari and Jayanti Banerjee, editors, *Handbook of Second Language Assessment*, pages 341–358. De Gruyter Mouton.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Shuai Wang, Zhengyang Chen, Kong Aik Lee, Yanmin Qian, and Haizhou Li. 2024. Overview of speaker modeling and its applications: From the lens of deep speaker representation learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4971–4998.

Xinhao Wang, Keelan Evanini, Yao Qian, and Matthew Mulholland. 2021. Automated scoring of spontaneous speech from young learners of english using transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712.

Jing Xu, Edmund Jones, Victoria Laxton, and Evelina Galaczi. 2021. Assessing L2 English speaking using automated scoring technology: Examining automarker reliability. *Assessment in Education: Principles, Policy & Practice*, 28:411–436.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu-hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Shang-Wen Li, Abdelrahman Mohamed, Shinji Watanabe, and Hung-yi Lee. 2024.

A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2884–2899.

Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2021. Applying wav2vec2.0 to speech recognition in various low-resource languages. arXiv preprint arXiv:2012.12121.

Klaus Zechner and Keelan Evanini. 2019. Overview of automated speech scoring. In *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*, pages 3–20. Routledge.

Appendix A Dataset Imbalance

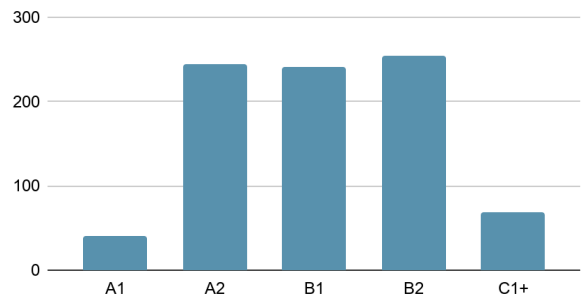


Figure 2: Distribution of CEFR levels in the final dataset by recording

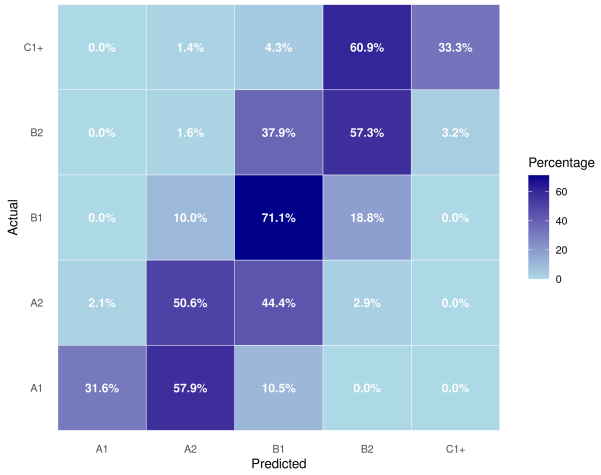


Figure 3: Confusion matrix of All Scores Model

Appendix B Effect of Train/Test Splits

Included Scores	RMSE	PCC	Acc %	Macro F1
Heuristic LMM	±0.003	±0.003	±0.3	±0.004
WavLM-Large	±0.020	±0.023	±1.9	±0.032
All	±0.034	±0.12	±3.6	±0.014

Table 3: Score Comparison/Combination

3 presents the variance of each approach by our four metrics. These metrics were collected from 5 different 10-fold iterations, where we combine all 10 test folds into a single set of predictions for comparison with our gold labels. Each value represents a 95% confidence interval. We find that while the the overall variance is not enough to affect our conclusions, there is a stark difference between the dataset dependence of our Heuristic LMM and neural speech approaches. The 10 times increase in variance indicates that neural speech models are much more susceptible to train/test splits than the simple regression model used by our heuristic LMM, at least in our low-resource setting.

Appendix C Predictor Estimates

Figures 4 and 5 show the measurable effect of each heuristic on prediction outcomes, firstly on the original heuristic baseline, and then when combining the heuristic baseline, the neural text-based model (`bert_score`) and the audio SSL model (`wavlm_score`). The scores produced by our deep neural models consistently receive the highest prediction estimates, often attenuating the weights of metrics previously estimated as highly significant. This suggests that while other scores remain important, our neural models implicitly capture their information during prediction. This does not make non-neural scores redundant, and performance could potentially improve further with a non-linear combination model.

While deep neural models exhibit very strong predictor estimations compared to most traditional scores, we believe it is not likely that their scores are strictly adding to coverage. When introducing SSL based scoring to the baseline model, we find that some of the scores which have the most *reduction* in predictor estimates are surface level metrics like number of words and articulation rate. These two metrics, the 2nd and 10th most significant scores in Figure 4, are reduced to the 11th and 24th most significant in Figure 5. While replacing these high estimators with more pedagogically pertinent metrics is always preferable, it seems that the scores produced by deep neural models rely on the same surface level information during prediction. It is true however, that some metrics retain high predictor estimates, like pause frequency.

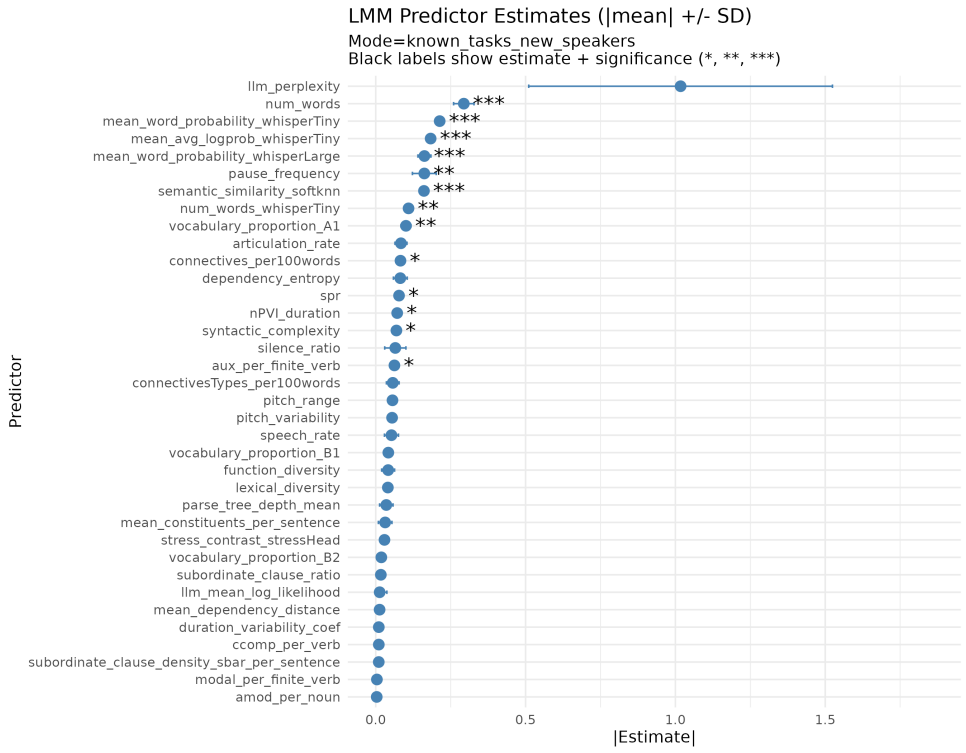


Figure 4: Predictor estimates of the heuristic based baseline

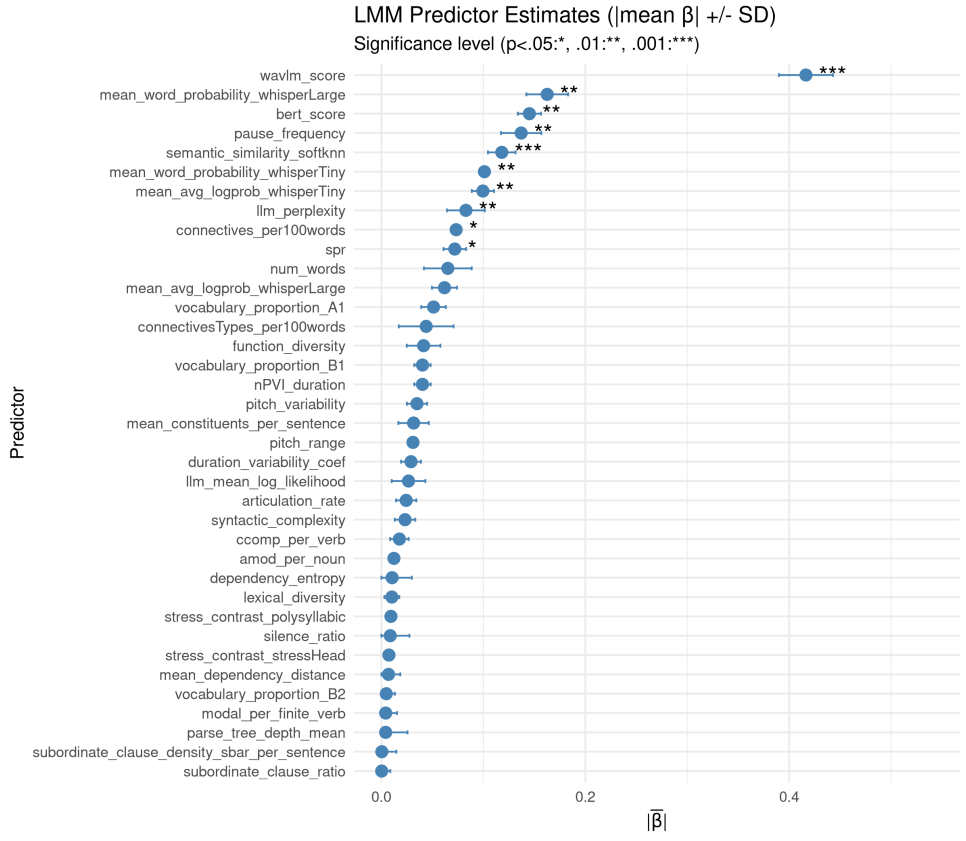


Figure 5: Predictor estimates of a model utilizing all baseline heuristics, WavLM predictions, and neural text-based predictions

University of Tartu Library

NEALT Proceedings Series Number: 61
ISBN: 978-9985-4-1522-1
ISSN: 1736-8197
eISSN: 1736-6305

2026