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

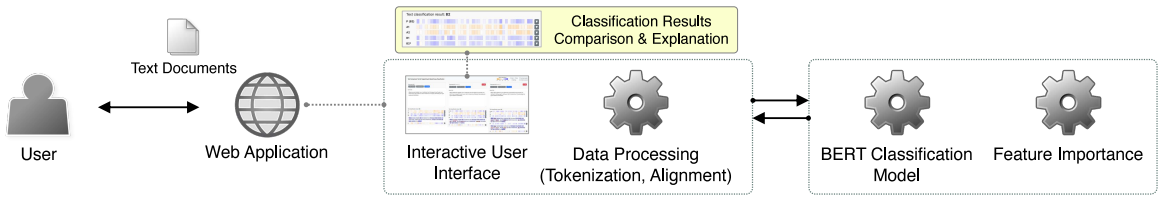


Figure 1: High-level overview of the design of our interactive visualization tool, implemented as a web-based application that processes input text documents (learner essay texts) in two versions: original and modified by users.

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.

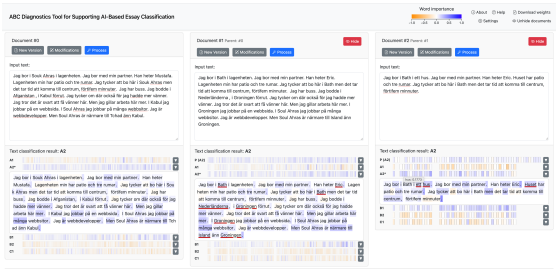


Figure 3: The complete user interface of the tool, including the results for three documents.

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).

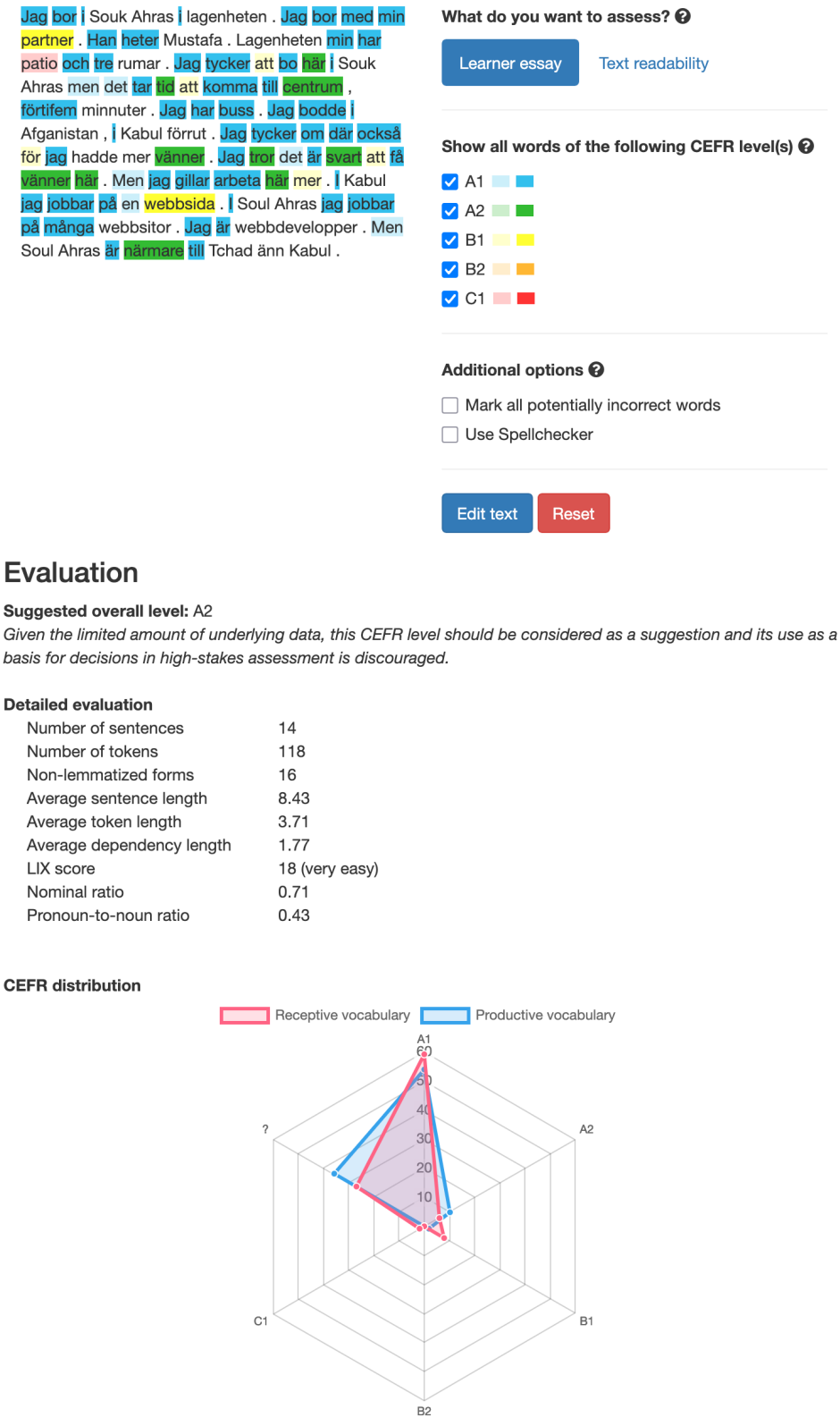


Figure 5: The Lärka tool visualises distribution of vocabulary across CEFR levels. The overall CEFR level is also calculated. However, there is no direct connection between the vocabulary-level labels and the document-level one.