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.

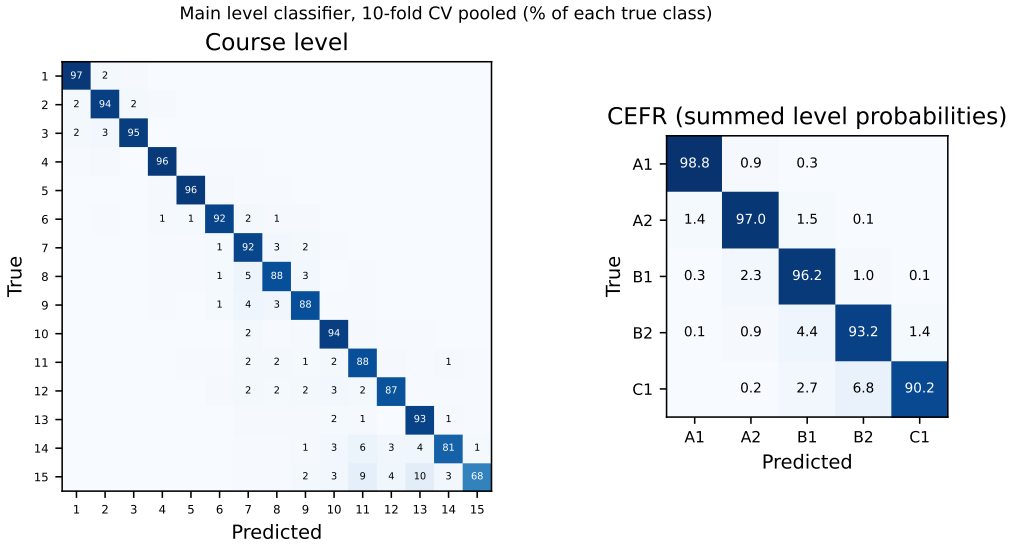


Figure 1: Confusion matrices of the main level classifier, pooled over the ten CV folds: course levels (left) and CEFR bands from summed level probabilities (right). Each cell gives the percentage of the true class, so the diagonal is the recall.

		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.

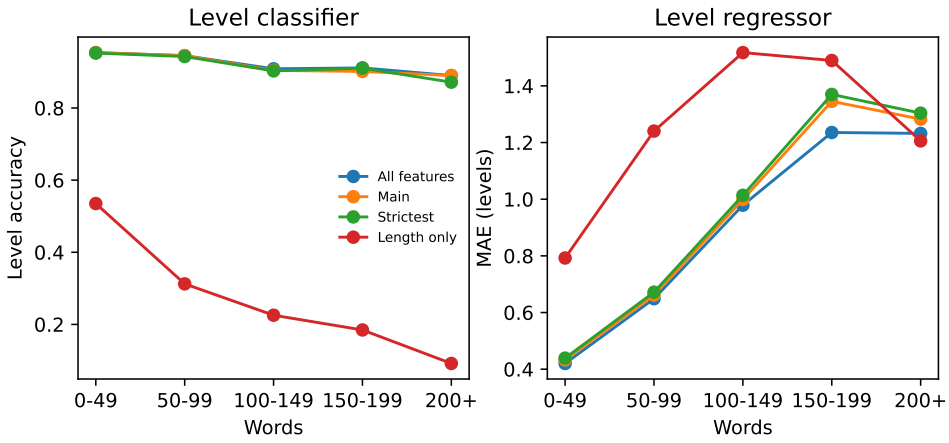


Figure 2: Validation performance by text length. Left: exact-level accuracy of the level classifiers. Right: MAE of the level regressors. The share of texts per bin is 42%, 46%, 9%, 2% and 0.2%. Longer texts tend to occur at higher, less frequent levels, so the bins also differ in level composition.

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

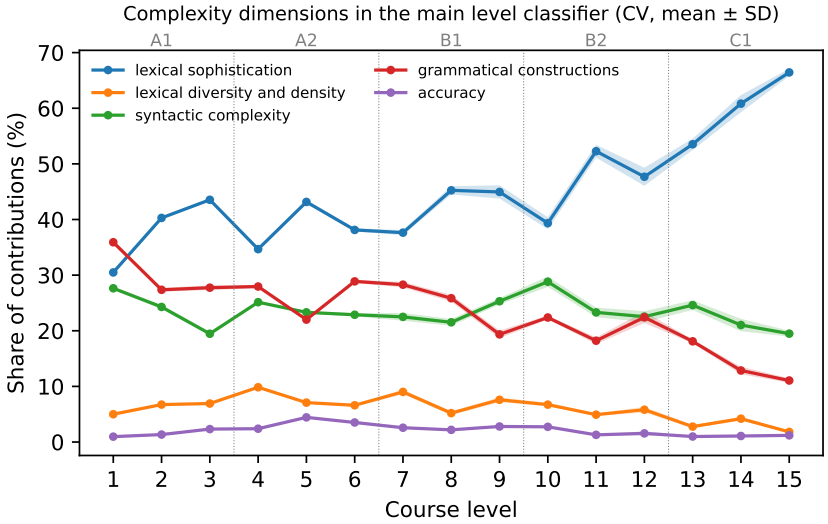


Figure 3: Share of the main level classifier's contributions per linguistic dimension for each course-level output (mean ± SD over the ten CV folds).

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.

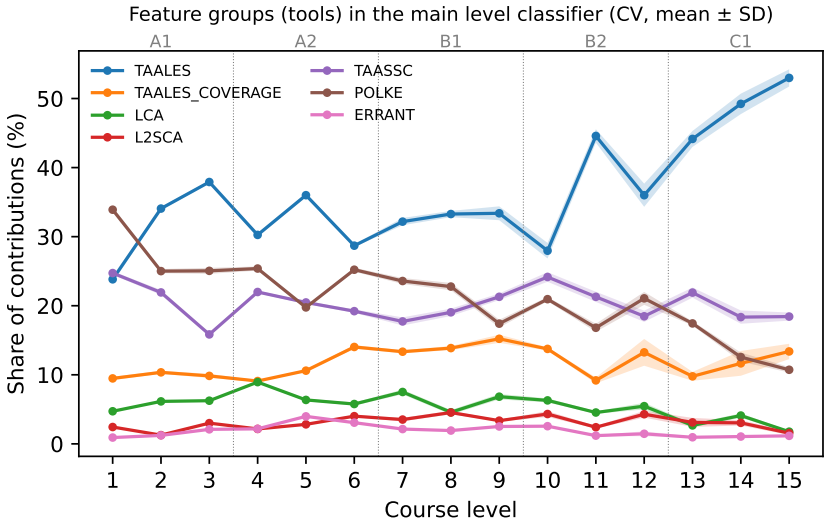


Figure 4: Share of the main level classifier's contributions per tool output group for each course-level output (mean ± SD over the ten CV folds).