LLM-Powered but Rule-Grounded: Pedagogically Relevant Grammatical Error Characterization for Learner Model Construction

Soroosh Akef[1,2,3] **Amália Mendes**[2] **Patrick Rebuschat**[3,4] **Detmar Meurers**[1,3]
[1]Leibniz Institute für Wissensmedien (IWM), Germany
[2]Center of Linguistics, Faculty of Arts and Humanities, University of Lisbon, Portugal
[3]LEAD Graduate School and Research Network, University of Tübingen, Germany
[4]Lancaster University, United Kingdom
s.akef@iwm-tuebingen.de d.meurers@iwm-tuebingen.de
mendes@edu.ulisboa.pt p.rebuschat@lancaster.ac.uk

Abstract

Grammatical error correction approaches rarely characterize the pedagogical value of corrected errors. We propose a framework that combines LLM-based second-language writing correction with a rule-based characterization module to identify pedagogically relevant, fine-grained grammatical properties in learner texts. The characterization module targets 252 European Portuguese properties which are categorized by the CEFR level at which they are taught according to an authoritative curriculum, and property accuracy is inferred from contrasts between the learner and corrected texts. We evaluate the framework extrinsically by training interpretable automatic proficiency assessment models on accuracy features extracted from characterized errors in a Portuguese learner corpus. Across different prompting strategies, we show that models trained on features derived from LLM-corrected texts perform similarly to those trained on features derived from annotator-corrected texts and comparably to models trained on linguistic complexity features. Feature importance overlap is likewise high, and similar predictive patterns are observed in both annotator-based and LLM-based models, further supporting the validity of the proposed framework.

1 Introduction

Grammatical error correction (GEC) has received significant attention from the ACL community in recent years, as evidenced by the organization of multiple shared tasks (Ng et al., 2013; Bryant et al., 2019; Masciolini et al., 2025). The attention this task has received is well justified, as a system capable of proofreading free prose is not only of interest to second language (L2) learners but also to native (L1) speakers.

Identifying errors in learners' production is particularly relevant in the context of L2 learning, as it can be used not only to give feedback to learners, but also to construct learner models that can be used in intelligent tutoring systems (ITSs) and intelligent computer-assisted language learning (ICALL) systems to deliver personalized language education. However, while GEC has moved toward the direction of evaluation methods with greater emphasis on characterizing errors at a finer level of granularity, the current granularity level of error annotation is insufficient for a learner model built based on the identification of such errors to be actionably informative. For instance, while the ERRANT scorer (Bryant et al., 2017) used to evaluate models submitted to the BEA-2019 shared task (Bryant et al., 2019) distinguishes between 25 main error categories, the information that an error represents the VERB:TENSE category is not fine-grained enough to build and update a learner model that can personalize L2 learning.

Evaluation methods used in GEC, and how validly they assess whether a corrected text contains corrections that are informative about an L2 learner's language development, also need to be reconsidered in the context of learner modeling. While reference-based evaluation methods compare the corrected text with a reference corrected by an expert, they do not account for the reality that there could be multiple valid approaches to correct a text (Bryant et al., 2023). To mitigate the limited nature of reference-based evaluation methods, various reference-free evaluation methods have been proposed (Asano et al., 2017; Choshen and Abend, 2018; Yoshimura et al., 2020; Islam and Magnani, 2021; Maeda et al., 2022; Östling et al., 2025). However, the existing reference-free evaluation methods also do not approximate the pedagogical relevance of the corrections in L2 learning contexts. Relatedly, Masciolini et al. (2025) distinguish between minimal edits and fluency edits in the shared task MultiGEC-2025, the latter of which emphasizes rephrasing the original text in favor of idiomaticity. The preference of one GEC approach

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 851–862

over the other in an L2 context likely hinges on the aspects of the L2 which are to be monitored and acted upon in a learner model.

The present work thus seeks to address the twofold challenge of identifying pedagogically relevant grammatical errors from texts produced by L2 European Portuguese learners and establishing the extent to which the identified errors are informative about L2 language development and can therefore be validly used to construct and update a learner model representing the learner's knowledge of various grammatical properties.

2 Related Work

A learner model represents a learner's current knowledge and mastery of a particular domain, inferred based on the data collected from the learner's interaction with a learning system (Xu and Bull, 2010). Learner models are thus vital to the construction of any adaptive educational system, as only through a detailed and accurate understanding of learners' competencies can they be matched to developmentally proximal learning material.

ICALL systems capable of identifying learners' errors and constructing or updating learner models accordingly have traditionally depended on mal-rules, symbolic representations of various ill-formed productions by learners according to an error taxonomy (Schneider and McCoy, 1998).

To build the BELLOC system, intended for English-speaking learners of French, Chanier et al. (1992) took advantage of what they referred to as "applicable rules," which explored rules underlying a learner's misconceptions if a correct parse for the sentence produced by the learner was not found. A limitation of the system was that when multiple potential misconceptions were matched, the system would prompt the learner to assess the grammaticality of test cases intended to isolate the learner's misconception, consequently exposing learners to further ill-formed constructions, which could reinforce the misconceptions.

A similar learner modeling strategy was adopted for the ICALL system Mr. Collins (Bull, 1994), which utilized pronoun placement tasks targeting European Portuguese clitic usage that required students to position pronouns and apply necessary phonetic contractions. The system addressed the issue of isolating a learner's misconceptions by taking into account the acquisition order of the structures. When a grammatical parse of a sentence failed, the system first attempted to identify target-language misconceptions, such as overgeneralization. If that failed, the system investigated negative transfer, prioritizing background languages based on the student's proficiency and linguistic distance to Portuguese. The limited scope of the grammatical structures targeted, as well as the controlled nature of the tasks, is among the limitations of this system.

Another relevant system was German Tutor (Heift and McFetridge, 1999), later renamed to E-Tutor (Heift, 2005, 2008), which incorporated a learner model that tracked a student's performance across various grammatical properties. Each property was assigned a score from 0 to 30, where 0 indicated an expert and 30 a novice. Unlike previous systems that relied on mal-rules after a parse failure, this system utilized a grammar that freely generated parses for both grammatical and ungrammatical input by relaxing specific constraints. With every successful production, the score representing the property in the learner model was decremented, and with every failed production, it was incremented. Finally, an error priority queue ensured that the learner received feedback on only the most significant error depending on their performance history or the instructor's decision to avoid pedagogical overload.

The ICICLE system (Michaud et al., 2005) further extended these approaches by modeling the evolution of interlanguage in deaf learners of English who use American Sign Language. ICICLE utilized a grammar that included mal-rules to allow the parser to generate multiple interpretations for ill-formed input. To select the most plausible parse, the system consulted the existing learner model to determine which parse was more likely.

To mitigate the computational bottlenecks of parsing ill-formed learner productions at runtime, Rudzewitz et al. (2018) proposed a novel architecture that utilized an offline hypothesis generation mechanism. Starting from the well-formed target responses expected for specific exercises, the system pre-computed a constrained search space of both well-formed and ill-formed response variations, storing them alongside their corresponding error diagnoses. At runtime, an online component matched the actual student input to these pre-computed hypotheses.

Fine-grained, pedagogically relevant grammatical error identification, therefore, has often been constrained by either the exercise type or by the

range of grammatical properties covered, as writing mal-rules that can reliably detect the full spectrum of potential learner errors for all grammatical properties in free writing is an unfeasible task.

3 Methodology

The proposed framework for identifying pedagogically relevant grammatical properties consists of a GEC module and a subsequent rule-based error characterization module. In this sense, the framework resembles that of ERRANT, with the key distinction that the grammatical properties characterized are at a finer level of granularity and are designed to align with L2 learning curricula, making them appropriate for learner modeling. To evaluate whether the identified errors are meaningful for the process of language acquisition, we adopt an extrinsic evaluation approach via a downstream task (Resnik and Lin, 2010), whereby the predictiveness of accuracy features extracted based on the identification of errors for an automatic proficiency assessment model is used as a proxy for whether the errors can be used to track language development. By utilizing interpretable machine learning models, the predictive behavior of each feature can be visualized, and a determination about the sensibility of this behavior can be made by human experts.

In the remainder of this section, we first present the rule-based tool developed to identify European Portuguese grammatical properties, dubbed SABER[1]. We then describe the learner corpus used in this study and the LLM-powered GEC approach. Next, we describe the adopted text alignment strategy and the accuracy inference algorithm devised for this study. Finally, we describe the experiment used for extrinsic evaluation.

3.1 Grammatical Property Identification

SABER is a rule-based NLP tool designed to identify formal grammatical properties in European Portuguese, which are aligned with the *Referencial Camões*[2], an authoritative reference mapping grammatical properties to the CEFR levels in which they are taught (A1–C2). SABER facilitates the extraction of 252 grammatical properties, comprising 147 word-level and 105 sentence-level properties.

The tool integrates both Stanza (Qi et al., 2020) and spaCy's medium Portuguese NLP pipelines to leverage their respective strengths at annotating the texts with linguistic information. Stanza serves as the primary pipeline due to its superior robustness in Portuguese lemmatization, while spaCy is retained for its specialized capabilities in named-entity recognition and multi-word token processing. This dual pipeline integration allows each pattern matching function, written using spaCy's token, phrase, and dependency matchers, to utilize the annotations of the pipeline most relevant to the targeted grammatical property. To systematically mitigate upstream NLP annotation errors, SABER employs alternative patterns for each grammatical property designed to handle such preprocessing errors. At the time of development, these identification rules were iteratively tested and fine-tuned using synthetic, property-dense texts generated via GPT-4, optimizing the precision and recall of each rule. To evaluate the tool, five positive and negative test cases were generated for each property using Gemini 3.1, which were subsequently manually validated, resulting in an F_1 score of 85.70%, a recall of 80.48%, and a precision of 96.29%. SABER is available online as a Streamlit application[3], and features extracted from an earlier version of it were used in Akef et al. (2025).

Importantly, SABER is designed to identify grammatical properties in well-formed texts, and may not be as robust on ill-formed texts. While this limitation may be considered as a drawback of the tool, it is in fact exploited to characterize learners' errors in the accuracy inference algorithm described in Section 3.5.

3.2 Dataset

The dataset used to evaluate our framework is the COPLE2 learner corpus (Mendes et al., 2016), which contains 1,634 texts produced by L2 Portuguese learners as part of class examinations at the Instituto de Cultura e Língua Portuguesa (ICLP). The corpus is annotated with error tags and metadata, such as the learner's L1, task prompt, and fine-grained CEFR levels (A1–C1). These metadata are particularly relevant for the current study, as we aim to test whether LLM prompts providing additional information about the task and the learner's linguistic background result in corrections that are more informative about the learner's proficiency.

For this purpose, three additional subsets of the

[1] Sistema de Análise e Busca de Estruturas Relevantes

[2] https://www.instituto-camoes.pt/activity/centro-virtual/referencial-camoes-ple

[3] https://saber-online.pt

corpus were extracted for L1-specific proficiency assessment experiments. Based on the availability of texts for each L1, these subsets include L1 Chinese (both Mandarin and Cantonese), L1 Romance (Spanish, Italian, French, and Romanian), and L1 Germanic (English, German, and Dutch). The distribution of the texts in the entire corpus and each subcorpus across the five proficiency levels is presented in Table 1.

Subcorpus	A1	A2	B1	B2	C1
COPLE2	327	419	377	315	196
L1 Chinese	45	104	101	88	100
L1 Romance	171	132	151	117	62
L1 Germanic	71	95	67	60	11

Table 1: Distribution of texts across CEFR levels in COPLE2

3.3 Grammatical Error Correction

Leveraging LLMs' ability to perform GEC in a zero-shot fashion (Kobayashi et al., 2024; Davis et al., 2024), we performed preliminary experiments by providing a detailed prompt to a range of open-source and proprietary models (including different versions and sizes of Gemma, Llama, Mistral, and Gemini accessed through an inference hosting provider and Google AI Studio) and by examining each model's corrections of one text from COPLE2. The temperature, a proxy for the degree of creativity in LLMs, was set to the relatively low value of 0.2 (in a range of 0 to 2) for all models, as a measure of preventing over-correction by the model. The most cost-efficient model that satisfied the prompt instructions in the preliminary experiments was Gemini 2.0 Flash, and it is thus the main LLM used in this study, as the objective of this study is not to establish the state of the art in GEC, but rather to demonstrate the feasibility of leveraging corrections produced by even a relatively modest LLM for pedagogically informed learner modeling. For this purpose, we used three prompt variants to generate three corrected versions of each text in COPLE2: a minimalistic prompt (see Appendix A), a detailed prompt including the task instructions, and a detailed prompt including both the task instructions and the learner's L1 (see Appendix B), to test whether the LLM could leverage this additional information to produce more informative corrections. Additionally, we used the annotator-provided corrections in COPLE2 as a reference for evaluating the LLM-generated corrections through an extrinsic automatic proficiency assessment task. Moreover, to ensure that this framework is not specific to any one LLM, we utilized corrections made using Gervásio 70B PTPT (Santos et al., 2024), an open-source LLM for European Portuguese, with the detailed prompt including task instructions, as an alternative LLM to compare Gemini's performance to.

3.4 Text Alignment

The alignment algorithm used is the linguistically enhanced Damerau-Levenshtein algorithm introduced by Felice et al. (2016) and used in ERRANT (Bryant et al., 2017). The key advantage of this alignment algorithm is its consideration for words' lemmas and parts-of-speech, facilitating alignment in cases of word order errors. As two distinct NLP annotation pipelines are used in SABER, each corrected text also needs to be aligned twice due to the two pipelines using different multi-word tokenization strategies. At the time of error characterization based on the accuracy inference algorithm, each grammatical property utilizes the alignment corresponding to the NLP annotation pipeline used in SABER for that property.

3.5 Accuracy Inference Algorithm

For every property identified in the corrected version, the following scenarios are explored: If the same property is identified in the same aligned span of the learner-produced text, this is considered as a signal that the learner produced the property correctly. There is, however, a caveat: There may be cases where the same property is identified despite the form produced by the learner being different from the corrected form. For instance, the property *Subordinada adverbial (temporal/causal/final)* is matched for both the learner-produced form "*para banharmo*" and the corrected from "*para banhar*." To account for these inaccuracies and at the same time allow for minor orthographical errors, the two aligned spans must have a normalized Levenshtein edit distance of θ or smaller to be considered as accurate, which is a threshold we determined empirically by comparing the performance of automatic proficiency assessment models trained on features extracted using the annotator-corrected version of the texts and based on different values of θ (see Section 4.1). On the other hand, if the property identified in the corrected version in a particular span is not equivalent to the property identified in the

corresponding span of the learner-produced text (or if no property is identified in the learner-produced text in that span), this is inferred as an inaccuracy, unless the forms of the two spans are identical. This could indeed happen in cases where other tokens that are not within the span of the identified property are different and thus affect the dependency parser's annotations. In such cases, mismatch between the properties identified is attributed to the NLP pipeline's error on the learner-produced text and the property identified in the corrected version is marked as accurate.

It is important to note that the final property stored (whether marked as accurate or inaccurate) is the one identified in the corrected version, even when a different property is identified in the same span of the learner-produced version. The rationale for this decision is that the learner's failure to produce the target property correctly is more informative about the developmental stage of that target property than the inaccurately produced property itself. For instance, if an L2 learner of Portuguese uses the indicative mood instead of the subjunctive mood with *talvez* (maybe), this is a stronger signal that they have not yet (fully) acquired the subjunctive mood than that they have not acquired the indicative mood. In fact, the learner may be fully aware that the indicative mood is inappropriate in that utterance and choose it only to achieve their communicative objective. The full flowchart of the proposed accuracy inference algorithm is presented in Figure 1.

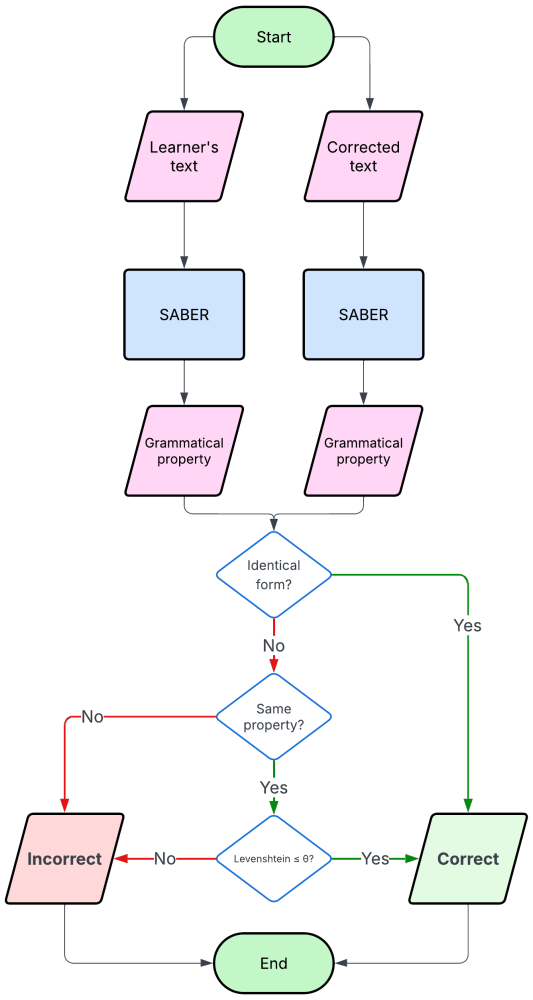

Figure 1: Flowchart representing the property accuracy inference algorithm

3.6 Evaluation

To perform extrinsic evaluation of whether accuracy features extracted using the accuracy inference algorithm can be used to track L2 development, 28 automatic proficiency assessment models were trained via five-fold cross-validation on various feature sets resulting from different text correction strategies. Additionally, models trained on broad linguistic complexity features, which have been shown to be predictive of learner proficiency (Ribeiro-Flucht et al., 2024), were included to test the contribution of accuracy features extracted in this study against.

3.6.1 Features

Following the characterization of grammatical errors in learners' texts, accuracy features are calculated by dividing the number of accurate occurrences of a property by all attempts of that property, represented as the sum of all accurate and inaccurate occurrences, including occurrences of inaccuracies resulting from the omission of the property in a text. Crucially, in cases where a property is appropriately absent from the text (meaning the counts of both accurate and inaccurate occurrences are zero), the value of the accuracy feature for that property is left as a missing value. The rationale for this decision is that not needing a property or successfully avoiding it is a qualitatively different signal from any accuracy value in the range of 0 to 1. Therefore, while the value of the property accuracy feature would be a missing value in such cases, it would be an informative missing value. This is analogous to not having a particular medical test's results when trying to predict the condition of a patient: The fact that the test result is missing indicates that it was not prescribed by a medical professional, which itself is an informative signal.

This would, however, mean that the machine learning algorithm that is selected must indeed be able to treat missing values as informative. Moreover, as many properties in a text are sparse, the overall accuracy of properties at the same CEFR level were also included as additional features.

The system used to extract broad linguistic complexity features is CTAP (Chen and Meurers, 2016; Weiss and Meurers, 2019) which can extract a total of 489 complexity features for Portuguese and has previously been used for automatic proficiency assessment of Portuguese (Ribeiro-Flucht et al., 2024). The features can be classified into the five classes of count-based, lexical, syntactic, morphological, and discourse features.

3.6.2 Training Algorithm

Aiming for not only robust performance but also innate interpretability, we opted for explainable boosting machines (EBMs; Nori et al., 2019) as the training algorithm used for automatic proficiency assessment. EBMs combine generalized additive models (GAMs) with modern machine learning techniques such as bagging and boosting to achieve performance on par with the highest performing feature-based machine learning models while affording inherent interpretability.

An important advantage of EBMs is that due to the round-robin nature of training, the effect of collinearity is mitigated (Nori et al., 2019). This is especially important in this project, where a wide array of features is utilized, many of which are expected to correlate. Another advantage is EBMs' handling of missing feature values by treating missing values as a separate value. For the purposes of visualization, EBM's library recommends that missing features be set to an extreme value (InterpretML Team, 2021). The value of missing accuracy features was thus set to -1. This assignment would not interfere with the model's prediction for the other values in the range of 0 to 1 due to the EBM's binning strategy, which places values in bins, with -1 being in its own separate bin. Finally, due to their handling of collinearity, EBMs do not require external feature selection. Caruana (2020) posits that feature selection could result in better performance in EBMs but that the best feature selection strategy is based on the features contributing the most to the prediction of an already trained EBM model using the full feature set.

Threshold	Macro F_1
0.0	52.55% ± 2.56%
0.05	52.72% ± 1.63%
0.1	51.77% ± 2.57%
0.15	52.86 ± 2.51
0.2	52.62% ± 1.79%
0.25	52.91% ± 1.76%
0.5	53.54% ± 2.41%
1	53.47% ± 3.19%

Table 2: Performance of EBM models trained on accuracy features extracted using different threshold values based on the annotator-corrected version of COPLE2 texts

4 Results and Discussion

4.1 Orthographical Error Threshold

As introduced in Section 3.5, various values for θ were used to infer grammatical accuracy in learners' texts. Specifically, the relevant spans share an identical grammatical property with their counterpart in the annotator-corrected version, despite differing in surface form. These values were 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, and 1, where a θ value of 0 represented extreme strictness and a value of 1 represented extreme leniency. The results of automatic proficiency assessment using property-based features extracted using each threshold value are presented in Table 2.

The results indicate that the relatively high threshold of 0.5 results in the features most capable of modeling L2 proficiency. While a normalized Levenshtein distance of 0.5 is unlikely to represent mere orthographical errors in longer spans, the better results achieved from the thresholds of 0.5 and 1 compared to the lower thresholds suggest that as long as the learner has produced the target grammatical property correctly, other lexical (or distinct grammatical) errors made within that production are likely not to undermine the learner's knowledge of the target property.

4.2 Framework Evaluation

The results of the various models trained via five-fold cross-validation reported according to macro F_1 are presented in Table 3.

It can be observed that accuracy features extracted using the framework described in this paper demonstrate predictiveness of language proficiency with a macro F_1 in the range of 52% to 55%. While

Feature set	General	L1 Chinese	L1 Romance	L1 Germanic
CTAP	58.67 ± 3.45	63.74 ± 4.32	50.35 ± 2.63	50.05 ± 1.68
Annotator-based	53.47 ± 2.46	60.55 ± 3.66	47.39 ± 4.40	50.71 ± 6.15
Gemini-based – minimalistic	52.00 ± 2.26	55.42 ± 4.68	47.36 ± 4.60	47.77 ± 5.39
Gemini-based – detailed + task	53.44 ± 1.49	57.02 ± 3.65	45.95 ± 6.08	48.86 ± 6.65
Gemini-based – detailed + task + L1	54.92 ± 1.40	56.11 ± 2.52	47.49 ± 6.24	50.02 ± 5.49
Gervásio-based – detailed + task	53.03 ± 1.14	57.11 ± 5.03	44.07 ± 4.82	47.69 ± 3.75
CTAP and Gemini-based – detailed + task + L1	59.71 ± 1.73	63.09 ± 3.93	50.38 ± 4.92	51.34 ± 2.11

Table 3: Performance of EBM models according to macro F_1 trained on different feature sets and corrected subcorpora

CTAP complexity features demonstrate superior predictiveness with a macro F_1 score of 58.67%, we observe an improved performance up to 59.71% with the inclusion of accuracy features.

Moreover, we can observe that using a more detailed prompt, along with providing information about the task, consistently resulted in better performance than the minimalistic prompt, often approaching (and in some cases surpassing) the results based on the annotator's corrections. Moreover, the inclusion of information about the learner's L1 resulted in gains compared to cases when this information was excluded, but these cases were not consistent across all L1s, with a slight drop observed from L1 Chinese learners. Nevertheless, these results may be a promising indication of LLMs' ability to capture traces of cross-linguistic influence in a zero-shot fashion.

Relatedly, while L1-specific Chinese models resulted in consistent improvement over the general model, L1-specific Romance and Germanic models consistently resulted in worse performance. This result may reflect class distribution effects from severe C1 imbalance in the L1 Romance and L1 Germanic datasets. (While the L1 Chinese dataset also has fewer samples at the A1 level, the distinction between learners at lower levels is often easier to model than at more advanced levels, where there is more variability among learners.)

The results obtained from the Gervásio-based features also support the applicability of this framework to other LLMs besides Gemini and supports the choice of Gemini 2.0 Flash as a low-cost but well-performing LLM for this particular task.

To further test the LLM-based accuracy feature extraction approach, we compared the top 30 features of the best Gemini-based model with those of the annotator-based model. The overlap was high (29/30), indicating consistent feature prioritization. Because overlap does not guarantee similar informativeness, we compared contribution plots for three shared features: Accuracy of *Pretérito imperfeito do indicativo – verbos regulares*, *Presente do conjuntivo – verbos irregulares*, and *Subordinada adverbial (condicional/concessiva)* (Figures 2, 4, and 6 for Gemini; Figures 3, 5, and 7 for the annotator model).

While differences can be observed between the corresponding graphs, general patterns regarding how the use and accuracy of each feature develops from one proficiency level to the next are similar between the two models. Investigating the cases where noticeable differences exist between the two models (such as the prediction that inaccurate productions of *Presente do conjuntivo – verbos irregulares* are likely to occur at level C1 according to the annotator-based model in Figure 5, but not so according to the Gemini-based model in Figure 4), we observed differences in how the learner's production was rendered grammatical, an example of which is as follows:

Learner: *Que seja com o telemóvel ou com o computado, podemos ficar em contacto permanente com os amigos ou com a família,* ***assim que temos*** *pouco para dizer uns aos outros.*

Annotator: *Quer seja com o telemóvel ou com o computador, podemos estar em contacto permanente com os amigos ou com a família,* ***pelo que tenhamos*** *pouco para dizer uns aos outros.*

Gemini: *Quer seja com o telemóvel ou com o computador, podemos ficar em contacto permanente com os amigos ou com a família,* ***de tal forma***

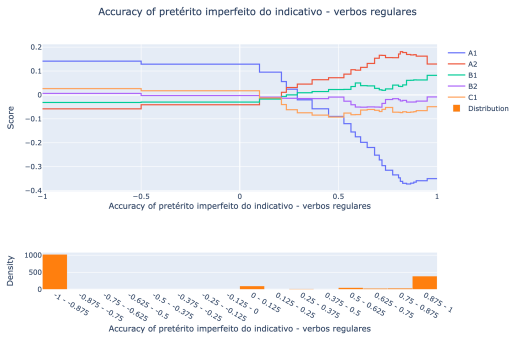

Figure 2: Contribution of Accuracy of *Pretérito imperfeito do indicativo – verbos regulares* to the Gemini-based model

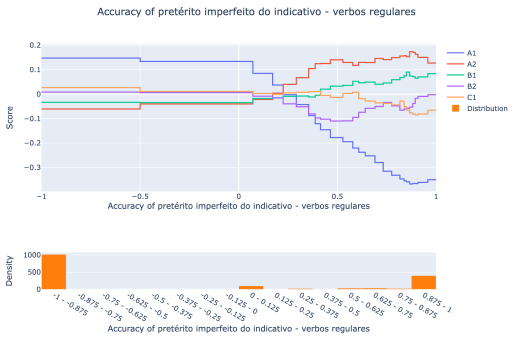

Figure 3: Contribution of Accuracy of *Pretérito imperfeito do indicativo – verbos regulares* to the annotator-based model

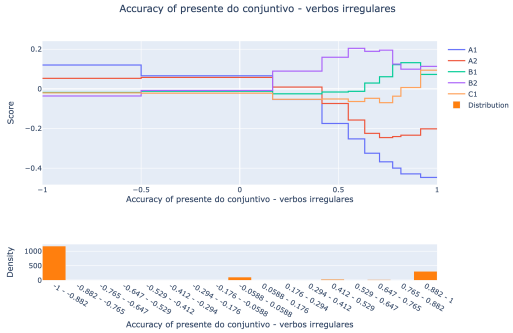

Figure 4: Contribution of Accuracy of *Presente do conjuntivo – verbos irregulares* to the Gemini-based model

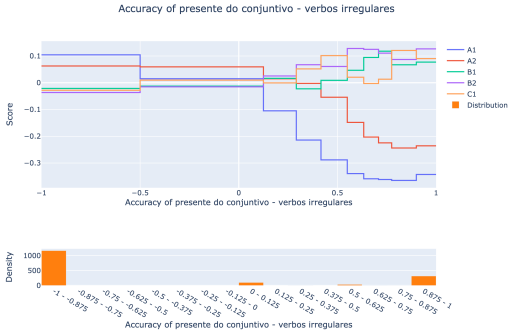

Figure 5: Contribution of Accuracy of *Presente do conjuntivo – verbos irregulares* to the annotator-based model

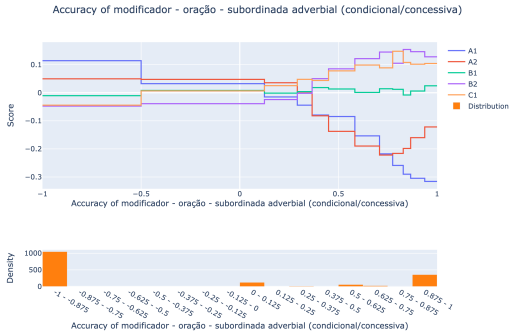

Figure 6: Contribution of Accuracy of *Modificador – oração – subordinada adverbial (condicional/concessiva)* to the Gemini-based model

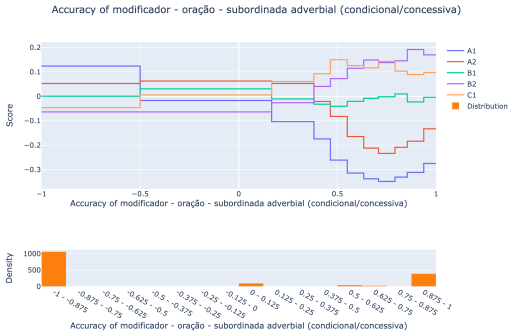

Figure 7: Contribution of Accuracy of *Modificador – oração – subordinada adverbial (condicional/concessiva)* to the annotator-based model

que temos *pouco para dizer uns aos outros.*

We can observe that while Gemini's correction diverges from that of the annotator in this instance, it is in fact the preferable correction, which could explain why the Gemini-based model with the most detailed prompt including both the task and the learner's L1 outperformed the annotator-based model.

5 Conclusion

To be used as a basis for making determinations about a learner's competencies, GEC must be performed by hypothesizing what the learner intended to produce and what possible factors led them to produce the ill-formed output. While LLMs may be suboptimal for pedagogically-informed grammatical error correction in every context, their capacity to render texts grammatical may still be leveraged for L2 learning applications.

For this purpose, we proposed a framework for characterizing pedagogically relevant grammatical errors corrected by an LLM using a rule-based tool capable of identifying European Portuguese grammatical properties in a fine-grained manner. With the development of similar tools for other languages, such as English (Sagirov and Chen, 2025) and German (Löfflad et al., 2025), the same framework can be applied to learner-produced texts of those languages to construct and update learner models.

Additionally, to evaluate our framework, we opted for an extrinsic evaluation approach by training automatic proficiency assessment models on accuracy features extracted based on either annotator-corrected texts or LLM-corrected texts and demonstrated not only largely similar performance between the two models, but also performance comparable to models trained on linguistic complexity features, which have previously been shown to be highly predictive of proficiency.

This framework enables the construction of highly detailed learner models by allowing learners' free-writing productions to be analyzed and monitored over time, thereby supporting adaptive and personalized language education through ITSs and ICALL systems. Consequently, while more powerful LLMs offer considerable potential for adaptive language education, realizing this potential requires sustained effort to develop NLP tools that pedagogically ground LLM outputs.

Limitations

Owing to the rule-based nature of SABER, it primarily covers formal grammatical properties and only a few use-based features. While a form may be taught at a lower CEFR level, certain uses (or functions) of that same form may correspond to higher CEFR levels, which SABER cannot currently distinguish.

Moreover, because the so-called accuracy features also encode information about the presence or absence of various grammatical properties, much of their predictiveness in the automatic proficiency assessment model is likely attributable to this information, making "accuracy" an operational label rather than a simple binary correctness score.

Furthermore, because the LLMs selected in this study were intended to demonstrate the feasibility of our proposed framework, more powerful reasoning-enabled LLMs could improve both overall and L1-specific model performance; however, this possibility was not explored in the present study.

Acknowledgments

This work was developed within the scope of the project *Promoção da Aquisição e ensino do Português como Língua de Herança através de Ferramentas Digitais Inteligentes*, financed by the Foundation for Science and Technology – FCT of the Republic of Portugal and the Camões Institute. We would like to thank anonymous reviewers for their insightful comments on a previous version of this paper.

References

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.

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based Metrics can be Replaced with Reference-less Metrics in Evaluating Grammatical Error Correction Systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Susan Bull. 1994. Student modelling for second language acquisition. *Computers & Education*, 23(1):13–20.

Rich Caruana. 2020. Comment on issue #179: Importance of feature selection for EBM. GitHub issue comment. Accessed: 2026-02-23.

Thierry Chanier, Michael Pengelly, Michael Twidale, and John Self. 1992. Conceptual Modelling in Error Analysis in Computer-Assisted Language Learning Systems. In Merryanna L. Swartz and Masoud Yazdani, editors, *Intelligent Tutoring Systems for Foreign Language Learning*, pages 125–150. Springer Berlin Heidelberg, Berlin, Heidelberg.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 113–119, Osaka, Japan. The COLING 2016 Organizing Committee.

Leshem Choshen and Omri Abend. 2018. Reference-less Measure of Faithfulness for Grammatical Error Correction. 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 124–129, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Davis, Andrew Caines, Øistein E. Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of English learner text. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11952–11967, Bangkok, Thailand. Association for Computational Linguistics.

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.

Trude Heift. 2005. Inspectable learner reports for web-based language learning. *ReCALL*, 17(1):32–46.

Trude Heift. 2008. Modeling learner variability in CALL. *Computer Assisted Language Learning*, 21(4):305–321.

Trude Heift and Paul McFetridge. 1999. Exploiting the Student Model to Emphasize Language Pedagogy in Natural Language Processing. In *Computer Mediated Language Assessment and Evaluation in Natural Language Processing*.

InterpretML Team. 2021. Comment on issue #18: Doing some treatment for missing values. GitHub issue comment. Accessed: 2026-02-23.

Md Asadul Islam and Enrico Magnani. 2021. Is this the end of the gold standard? A straightforward reference-less grammatical error correction metric. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large Language Models Are State-of-the-Art Evaluator for Grammatical Error Correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Denise Löfflad, Benedikt Beuttler, and Detmar Meurers. 2025. German Grammar Profile for Learners: Pedagogical Feature Definition and Automated Extraction. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 212–223, Hannover, Germany. HsH Applied Academics.

Koki Maeda, Masahiro Kaneko, and Naoaki Okazaki. 2022. IMPARA: Impact-Based Metric for GEC Using Parallel Data. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3578–3588, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Arianna Masciolini, Andrew Caines, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 Shared Task on Multilingual Grammatical Error Correction at NLP4CALL.

Amália Mendes, Sandra Antunes, Maarten Janssen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3207–3214, Portorož, Slovenia. European Language Resources Association (ELRA).

Lisa N. Michaud, Kathleen F. McCoy, and Rashida Z. Davis. 2005. A Model to Disambiguate Natural Language Parses on the Basis of User Language Proficiency: Design and Evaluation. *User Modeling and User-Adapted Interaction*, 15(1):55–84.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 Shared Task on Grammatical Error Correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint*. ArXiv:1909.09223 [cs].

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv preprint*. ArXiv:2003.07082 [cs].

Philip Resnik and Jimmy Lin. 2010. Evaluation of NLP Systems. In *The Handbook of Computational Linguistics and Natural Language Processing*, pages 271–295. John Wiley & Sons, Ltd.

Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2024. Explainable AI in language learning: Linking empirical evidence and theoretical concepts in proficiency and readability modeling of Portuguese. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 199–209, Mexico City, Mexico. Association for Computational Linguistics.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. Generating Feedback for English Foreign Language Exercises. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–136, New Orleans, Louisiana. Association for Computational Linguistics.

Nelly Sagirov and Xiaobin Chen. 2025. POLKE: A system for comprehensively annotating pedagogicallyoriented grammatical structure use in language production. *Manuscript submitted for publication to Behavior Research Methods*.

Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, and António Branco. 2024. Advancing generative AI for Portuguese with Open Decoder Gervásio PT-*. *Preprint*, arXiv:2402.18766.

David Schneider and Kathleen F. McCoy. 1998. Recognizing Syntactic Errors in the Writing of Second Language Learners. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 1198–1204, Montreal, Quebec, Canada. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2019. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 380–393, Florence, Italy. Association for Computational Linguistics.

Jing Xu and Susan Bull. 2010. Encouraging advanced second language speakers to recognise their language difficulties: a personalised computer-based approach. *Computer Assisted Language Learning*, 23(2):111–127.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less Sub-Metrics Optimized for Manual Evaluations of Grammatical Error Correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robert Östling, Murathan Kurfali, and Andrew Caines. 2025. LLM-based post-editing as reference-free GEC evaluation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 213–224, Vienna, Austria. Association for Computational Linguistics.

A Minimalistic Prompt

Correct the following text according to standard European Portuguese grammar. Output only the corrected text.

B Detailed Prompt

You are an expert proofreader and teacher of European Portuguese, specializing in correcting texts written by language learners. Your primary goal is to correct the "Learner's Text" you will be given.

Adhere strictly to these instructions:

1. **Correction Scope:** Focus on fixing clear errors in European Portuguese grammar, spelling, punctuation, and word choice.

2. **Minimal Edits:** Implement the fewest changes necessary. Preserve the learner's original meaning, voice, and sentence structure as much as possible. Only alter what is incorrect or hinders comprehension. Do not rephrase for stylistic preference if the original is grammatically acceptable. That being said, cases that would still be acceptable in Brazilian Portuguese but not so in European Portuguese must also be corrected. Also, you must take the whole sentence into account for grammaticality. If you need to change the word form the learner produced to make the corrected sentence grammatical, you are allowed to do it.

3. **Intent and Context:** The learner wrote their text in response to a specific "Task Prompt". You will receive this "Task Prompt" along with the "Learner's Text". Interpret and correct errors based on what the learner likely intended to communicate to fulfill that "Task Prompt".

4. **L1 Interference (Negative Transfer):** You will be provided with the learner's native language (L1). Actively utilize this information to identify instances of negative transfer. When determining the correct form, base your decision on what the learner likely intended to produce given typical L1 interference patterns.

5. **Language Variant:** All corrections must use standard European Portuguese, not Brazilian Portuguese.

6. **Output Format:** Your response MUST BE ONLY the fully corrected European Portuguese

text. Do NOT include any explanations, apologies, greetings, or any other text before or after the corrected version.

You will receive the "Learner's L1", the "Task Prompt", and the "Learner's Text" clearly labeled in the user message.