# Explainable AI in Language Learning: Linking Empirical Evidence and Theoretical Concepts in Proficiency and Readability Modeling of Portuguese

**Luisa Ribeiro-Flucht, Xiaobin Chen, Detmar Meurers**
LEAD Graduate School and Research Network
University of Tübingen / Germany
luisa.ribeiro-flucht@uni-tuebingen.de
xiaobin.chen@uni-tuebingen.de
dm@sfs.uni-tuebingen.de

## Abstract

While machine learning methods have supported significantly improved results in education research, a common deficiency lies in the explainability of the result. Explainable AI (XAI) aims to fill that gap by providing transparent, conceptually understandable explanations for the classification decisions, enhancing human comprehension and trust in the outcomes. This paper explores an XAI approach to proficiency and readability assessment employing a comprehensive set of 465 linguistic complexity measures. We identify theoretical descriptions associating such measures with varying levels of proficiency and readability and validate them using cross-corpus experiments employing supervised machine learning and Shapley Additive Explanations. The results not only highlight the utility of a diverse set of complexity measures in effectively modeling proficiency and readability in Portuguese, achieving a state-of-the-art accuracy of 0.70 in the proficiency classification task and of 0.84 in the readability classification task, but they largely corroborate the theoretical research assumptions, especially in the lexical domain.

## 1 Introduction

As technology evolves at a rapid pace, the field of education undergoes continuous adaptation. Particularly in language learning, numerous tools are being developed with the goal of facilitating the practice of a second language and providing tailored materials. In order to effectively model natural language, it's crucial to identify and empirically validate the relevant linguistic properties to use. Linguistic modelling with complexity measures has been proven to be highly effective in providing evidence-based insight into the assessment of both proficiency and readability (Benjamin, 2012; Crossley et al., 2017).

Second language proficiency and text readability are often associated concepts in language learning. Proficiency is usually equated to the notions of mastery and ability of understanding and producing another language (Hulstijn, 2015). Readability, in turn, encompasses the degree of reading difficulty which a text may exert on a reader (Dale and Chall, 1949). While widely acknowledged as multidimensional and dynamic constructs, proficiency and readability are commonly assessed using standardized scales. The Common European Framework of Reference for Languages (CEFR, Council of Europe) stands out as one of the most prominent scales for measuring proficiency, while readability is usually estimated according to different education, proficiency, and literacy levels.

In this context, it is essential to note the limited empirical evidence supporting the categorization of the mentioned constructs into levels and the precise definition of each level (Hulstijn, 2015). The English Profile Programme (EGP, Hawkins and Buttery, 2008) is a notable effort to clarify proficiency levels by identifying linguistic features whose presence or absence corresponds to specific English CEFR levels. However, the success of such an initiative heavily relies on the availability of abundant data and specialized manpower for data annotation and analysis, which may not be readily accessible for languages other than English.

In this paper, our objective is to propose an automatic method that comprehensively captures the nuanced characteristics defining language proficiency and text readability in Portuguese, as well as to provide a robust multilingual text analysis platform which will be made freely available online. By performing linguistic modeling and applying explanatory methods, we seek to validate theoretical postulations and enhance our understanding of the linguistic properties which are crucial for language learning, by answering the following research questions:

1. How well does a broad set of linguistic com-



*Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209

plexity measures model proficiency and readability levels in Portuguese?

2. What are the most discriminative measures for each proficiency and readability level? Do they coincide with theoretical suggestions?

3. Can Explainable AI be used with the purpose of describing proficiency and readability levels?

To address these questions, we use two distinct datasets: One consisting of European Portuguese learner productions and another consisting of Brazilian school materials. The constructs under investigation will be modeled as classification tasks. By leveraging such measures, language learning tools and generative models may better capture the nuances of written language, leading to a deeper understanding of the intricacies of readability and proficiency assessment (Housen et al., 2012).

## 2 Related Work

The extraction and analysis of linguistic complexity measures have been extensively explored in research. While many studies on automatic proficiency and readability assessment have primarily investigated the English language (e.g. Ortega, 2003; Lu, 2010; Bulté and Roothooft, 2020), advancements in Natural Language Processing (NLP) have facilitated the extension of research to other languages.

In European Portuguese, del Río (2019b) employed a supervised-learning approach using a learner corpus, achieving a 0.72 accuracy and 0.71 F-score by combining 39 linguistic complexity measures with other types of features, such as n-grams and readability formulas. Similarly, in Brazilian Portuguese, Evers (2013) used a corpus from a Brazilian Portuguese proficiency exam, extracting 48 linguistic measures for a binary classification task distinguishing between beginner and advanced learners, achieving an accuracy of 0.70 with a J48 classifier.

Automatic readability assessment has also been explored in the context of the Portuguese language. For instance, Curto et al. (2014) analyzed a corpus of L2 Portuguese texts, extracting 52 linguistic complexity measures. Their experiments achieved accuracy scores of 0.86 and 0.79 for three-level and five-level classification tasks, respectively. Additionally, Akef et al., 2024 extracted 489 linguistic complexity measures using the platform herein presented with machine learning algorithms for readability classification. This study demonstrated that models which incorporated informative features exhibited the highest generalization rate across various samples.

Regarding the use of explainable AI in Portuguese studies, Oliveira et al. (2023) explores the estimation of textual cohesion across essays in both Portuguese and English. The study found that although a deep learning-based model demonstrated superior performance, conventional machine learning models showed stronger potential in explainability.

The mentioned studies represent a crucial advancement in the automatic classification of proficiency and readability in Portuguese; however, they have limitations. Except for a few, most of the studies in Portuguese readability and proficiency assessment suffer from either a lack of a comprehensive set of measures, which might not fully capture the complexity and nuances of proficiency and readability, or from the absence of interpretability and detailed insight into feature importance. As a result, the depth of understanding regarding the constructs themselves and their categorization into separate levels may be limited.

## 3 Data

Two corpora were selected for our experiments and analyses: NLI-PT (Gayo et al., 2018) and Corpus de Complexidade Textual para Estágios Escolares do Sistema Educacional Brasileiro (Gazzola et al., 2019). The former comprises 3069 L2 Portuguese learner texts, categorized into three general levels: A (consisting of the CEFR levels A1 and A2), B (B1 and B2), and C (C1). The distributions of the texts in this corpus can be found in Table 1.

| Proficiency Level | Number of Texts |
|---|---|
| A - Beginner | 1,388 |
| B - Intermediate | 1,215 |
| C - Advanced | 466 |
| **Total** | **3,069** |

Table 1: Distribution of texts across proficiency levels in NLI-PT.

Regarding the latter corpus, it is a collection of 2076 Portuguese texts taken from Brazilian public school materials, and are separated into four school levels (elementary school, middle school,

high school and university education). Their distribution is displayed in Table 2.

| Education Level | Number of Texts |
|---|---|
| Elementary School | 297 |
| Middle School | 325 |
| High School | 628 |
| University Education | 826 |
| **Total** | **2,076** |

Table 2: Distribution of texts across school levels in the Córpus de Complexidade Textual para Estágios Escolares do Sistema Educacional Brasileiro corpus.

It is important to note that the distribution of texts into the separate categories in both corpora is imbalanced. That means that some levels are better represented than others, possibly influencing the classification results.

## 4 Methods

Our experiments consist of three main steps: first, extracting linguistic complexity measures from the chosen corpora; then conducting two types of classification experiments, one for proficiency and one for readability; and finally, analyzing the results using an explainable AI method to understand feature importance.

### 4.1 Automatic Complexity Measure Extraction

The Common Text Analysis Platform (CTAP, Chen and Meurers, 2016),[1] which already supports other languages, was extended to accommodate the extraction of 465 Portuguese complexity measures covering superficial counts and the linguistic domains of lexicon, syntax, morphology and discourse.[2] Table 3 displays the current distribution of measures across these domains.

The extension to Portuguese analysis was made possible via the integration of the Stanza pipeline (Qi et al., 2020), which provides a pipeline for tokenization, lemmatization, sentence segmentation, part-of-speech tagging, morphological annotation, dependency and constituency parsing, followed by specific methods based on extraction rules and word frequencies. While the analysis tool is available online and is free to use, the Portuguese analysis feature is not yet online as of this publication.

[1] https://sifnos.sfs.uni-tuebingen.de/ctap/

[2] The complete list of measures can be found as a supplementary material.

The selection of which measures should be added to the Portuguese complexity measure set in this work was based on previous related works (Weiss and Meurers, 2019). Additionally, in order to understand which of these measures are associated with the different proficiency and readability levels, a detailed study was performed of the Camões Institute's Reference Level Descriptions (RLD, Referencial Camões, 2017; Vaz et al., 2019), which outlines the discursive notions, grammatical structures, and lexical items expected of learners based on their placement in the CEFR proficiency scale. The Manual for Syntactic Simplification for Portuguese (Specia et al., 2008) and the SIMPLEX-PB 3.0 database (Hartmann et al., 2018) were also consulted. These are resources which categorize vocabulary and linguistic structures as easy or difficult based on their occurrence in different readability levels.

| Domain | Number of Measures |
|---|---|
| Superficial | 26 |
| Lexical | 235 |
| Syntactic | 108 |
| Morphological | 52 |
| Discourse-based | 44 |
| **Total** | **465** |

Table 3: Distribution of linguistic complexity measures across the five domains.

#### 4.1.1 Superficial Measures

Superficial aspects of the text are some of the most traditionally analyzed ones. Although these measures require minimum computational power, they have consistently shown good discriminative capabilities (Bulté and Roothooft, 2020; Housen et al., 2012; Norris and Ortega, 2009). These consist of linguistic element counts, lengths, normalizations and ratios.

The simplification manual suggests a uniform increase in the length of texts one can read as they advance in literacy. Regarding Portuguese as a second language, while the RLD does not explicitly mention an increase in superficial aspects of texts, it suggests that different noun and verb inflections as well as syntactic constituents are acquired as the learner moves to more advanced levels, which may consequently influence the length of their words, phrases and clauses.

#### 4.1.2 Lexical Measures

Lexical complexity has been shown to be very relevant in linguistic complexity studies. For instance, Crossley et al. (2011) reported that older writers seem to produce more infrequent, less diverse, and more abstract words. It has been also demonstrated that the use of infrequent words may exert a negative impact on reading comprehension (Nation and Coady, 1988). In addition, McCarthy and Jarvis (2010) suggest that low values of lexical density may be indicative of a smaller propositional complexity. McNamara et al. (2010) also report on the positive correlation between lexical diversity and linguistic competence.

In this study, we extracted measures related to lexical density, variation, and sophistication. Lexical density was computed by scaling lexical and function words by the total number of tokens. Variation was assessed by dividing the number of lexical types by the number of lexical tokens, and through edit distance calculations for lemmas, parts of speech, and tokens. To measure lexical sophistication, we considered aspects such as age of acquisition (Cameirao and Vicente, 2010), concreteness, imageability, and familiarity (Soares et al., 2017). Contextual diversity and word frequencies were derived from the SUBTLEX-PT lexical database (Soares et al., 2015). Additionally, we included frequency-based measures from the Portuguese Vocabulary Profile project (Torigoe, 2017) and a list of complex words (Hartmann et al., 2018).

#### 4.1.3 Morphological Measures

The morphological measures implemented in this work prioritize the inflectional system of Portuguese, as they can be easily generalized to other languages. Measures of derivational and compositional morphology will be eventually added to the system.

The RLD suggests that verb forms are learned incrementally starting from the simple present. The past participle verb form, for instance, is described as being incrementally learned, starting at A2 level, with its regular and irregular forms, and is consolidated at level B1 with the double participle with gender and number inflection. Given that this verb form is prevalent in constructions deemed as advanced, like passive sentences and the present perfect tense, it is expected to be more common in texts for advanced learners rather than beginners. Both the RLD and the simplification manual affirm that inflections like the present perfect tense, simple future tense, present subjunctive, conditional mood, and passive voice are typically found in advanced texts.

#### 4.1.4 Syntactic Measures

The syntactic measures are based on both syntactic element counts and ratios. Clauses, phrases, complements, T-Units, modifiers, subjects and clefts were taken into consideration. We have measured clausal elaborateness, by taking clausal subordination and coordination into account. More specifically, regarding coordination, we calculate copulative, disjunctive and asyndetic coordinate clauses. In addition, we measure phrasal elaborateness by accounting for noun and verb phrases, as well as different types of subject, such as null and clausal subjects. Lastly, measures based on the Dependency Locality Theory (DLT, Gibson et al., 2000) were also included.

Most studies done in English have shown that measures like sentence length, clausal elaborateness, number of clauses and dependent clauses per T-unit increase throughout proficiency levels ((Norris and Ortega, 2009; Ortega, 2003; Lu, 2010). Moreover, the simplification manual suggests that sentences with a high rate of embeddedness are more challenging to read, as well as the inverse order of verb-subject, instead of subject-verb. The latter, being learned only at the B1 level, according to the RLD.

#### 4.1.5 Discourse-based Measures

The measures implemented concerning discourse are based on the list of connectives developed by Mendes and Río (2018). Additionally, we measured the use of single and multi-word connectives, as well as easy and difficult connectives. The latter are based on two lists created by Leal et al. (2021). In terms of referential cohesion, measures regarding argument overlap, lemma overlap and lexical word overlap were calculated. For all these features, their mean and standard deviation values were also calculated and included as separate features.

In the simplification manual, it is suggested that discourse connectives improve comprehension, meaning they should occur often in earlier levels and be replaced incrementally by more advanced linguistic devices. The RLD also suggests that constructs like anaphora are acquired by intermediate learners.

### 4.2 Classification Experiments

Based on prior research findings (del Río, 2019a), we conducted classification tasks implementing three distinct supervised learning classifiers: Support Vector Machine, Linear Regression and Random Forest. In addition, a Multi-Layer Perceptron classifier was used in order to verify whether a simple neural network architecture performs significantly better or worse.[3]

For the sake of consistency and in order to warrant fair comparisons, all of the experiments herein performed were implemented in the Python programming language, using algorithms provided by the Scikit-learn library (Pedregosa et al., 2011).

The values for each measure were scaled using the library's method StandardScaler, in order to avoid the effect of high cardinality, due to differing range sizes among measures. Additionally, Scikit-learn's method GridSearch was applied alongside 10-fold cross-validation in order to optimize the models' performance and avoid overfitting. Lastly, to evaluate model performance, separate testing sets were created using an 80/20 split for training and testing purposes. Results from both the 10-fold cross-validation and held-out test sets are presented below.

### 4.3 Explainable Artificial Intelligence with SHAP

Explainable Artificial Intelligence has seen significant development, offering various approaches for understanding model outputs (Došilović et al., 2018). To gain insight into feature contributions, we adopted the Shapley Additive Explanations (SHAP, Lundberg and Lee, 2017) framework, which has been recently applied in proficiency and readability studies (e.g. Korniichuk and Boryczka, 2021; Nguyen and Wintner, 2022).

SHAP was selected over alternative interpretation methods such as LIME (Ribeiro et al., 2016) due to its ability to offer insights into feature importance both locally and globally, irrespective of the underlying model's complexity. This flexibility was crucial for our study, given the diverse linguistic measures and the use of non-linear SVMs with RBF kernels, where interpretation can be challenging (Sanz et al., 2018). In contrast, SHAP allows us to delve into each prediction, offering a deeper understanding of how specific features influence model outcomes.

Specifically, we employed the KernelExplainer method from the SHAP package. This method estimates the importance of each feature in making a particular prediction. It calculates the SHAP values, which represent the marginal contribution of each feature to the prediction across all possible combinations of features. Positive SHAP values indicate a feature's contribution to increasing the model's prediction, whereas negative values signify a decrease in the prediction. These values are then combined using a weighted sum to determine the overall importance of each feature.

## 5 Proficiency Classification Results

While all classifiers showed similar performance, the SVM classifier exhibited slightly better results compared to the others, as shown in Table 4. Conversely, the sole neural network architecture included in the analysis performed the worst. With 10-fold cross-validation, the best-performing classifier achieved a mean accuracy score of 0.70 and a weighted F-score of 0.68. Furthermore, on evaluation with the held-out test set, it achieved an accuracy of 0.73 and a weighted F-score of 0.72.

| | 10-Fold CV | | Test Set | |
|---|---|---|---|---|
| | F1 | Acc | F1 | Acc |
| Logistic Regression | 0.68 | 0.68 | 0.70 | 0.69 |
| Multi Layer Perceptron | 0.64 | 0.66 | 0.66 | 0.67 |
| Random Forest | 0.68 | 0.68 | 0.67 | 0.63 |
| Support Vector Machine | 0.68 | 0.70 | 0.73 | 0.72 |

Table 4: 10-fold cross-validation and test set accuracy and F1-scores achieved in proficiency classification experiments with all features.

In the confusion matrix (Table 5), proficiency level A had the highest accuracy, with 219 true positives, but 41 were misclassified as B. Notably, 51 texts from level B were misclassified as level A, and 19 as level C. Level C had the fewest true positives, possibly due to class imbalance, as discussed by (del Río, 2019b) or due to factor which have not been currently accounted for.

Figure 1 shows the mean SHAP values for the top 20 features with the most impact on the model's output for each proficiency level, listed in descending order. Among these features, 10 are related to the lexical domain. The most impactful feature is complex word frequency, particularly influential

[3] All experiment resources can be accessed through the following link: https://osf.io/ehdc9/?view$_only$ = 2e7ee278d187417c82219dc6eab6e29e

| | A | B | C |
|---|---|---|---|
| A | **219** | 41 | 3 |
| B | 51 | **178** | 19 |
| C | 19 | **32** | 29 |

Table 5: Confusion matrix of the test set, obtained from the classification performed using the SVM classifier on all features.

for levels A and C. Additionally, two Portuguese Vocabulary Profile features, the A1 and B1 word lists, strongly influenced the prediction of level A.

Phrasal and clausal elaboration significantly influenced the model's output. The measures of relative clauses per clause and per T-unit were influential for distinguishing levels A and C, while the mean length of noun phrases is most impactful for predicting level B. The nominative case inflection emerges as the sole highly discriminative morphological measure. Additionally, word frequency-based features, clausal elaborateness, and lexical sophistication measures contribute to the list.

## 6 Readability Classification Results

Consistently with the proficiency classification experiments, Logistic Regression, Random Forest, Support Vector Machine and Multy-Layer Perceptron classifiers were implemented. The results achieved with 10-fold cross-validation and held-out test sets for each classifier are displayed in Table 6. Similarly to the proficiency experiments, the SVM classifier showed the best results, achieving an accuracy of 0.84 from 10-fold cross-validation, and an accuracy 0.85 with the held-out test set.

| | 10-Fold CV | | Test Set | |
|---|---|---|---|---|
| | F1 | Acc | F1 | Acc |
| Logistic Regression | 0.81 | 0.83 | 0.81 | 0.81 |
| Multi Layer Perceptron | 0.83 | 0.83 | 0.82 | 0.82 |
| Random Forest | 0.74 | 0.79 | 0.76 | 0.76 |
| Support Vector Machine | 0.85 | 0.86 | 0.87 | 0.87 |

Table 6: 10-fold cross-validation and test set accuracy and F1-scores achieved in the readability classification experiments with all features.

Upon reviewing the confusion matrix presented in Table 7, it becomes evident that the classifier effectively distinguished the elementary school school level from the others, with only 6 misclassifications as middle school texts. For the last three levels, there were minimal misclassifications into adjacent levels.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | **49** | 6 | 0 | 0 |
| 2 | 4 | **58** | 12 | 1 |
| 3 | 0 | 1 | **78** | 12 |
| 4 | 0 | 1 | 14 | **142** |

Table 7: Confusion matrix of the test set, obtained from the classification performed using the SVM classifier on all features.

Figure 2 displays the mean SHAP values for the top twenty influential features. Thirteen of these features pertain to the lexical domain, four to morphology, two to surface features, and one to syntax.

The imageability of lexical word types had the strongest impact on the model's output. Familiarity, age of acquisition, the lexical density of articles and determiners and frequency-based measures were also highly discriminative. Additionally, superficial measures like the standard deviation of token length in syllables and letters were predictive. Morphological measures were also influential, with inflections in case, mood, person, and number showing strong impacts, particularly in differentiating the first and last levels. Notably, phrasal and clausal elaborateness seemed less significant in predicting school levels compared to proficiency classification.

## 7 Feature Selection

During the analysis of the measures, we found that some linguistic features were highly correlated with each other, aligning closely with expectations, for example, the correlation between the number of letters and the number of syllables, or the number of determiners and the number of articles. However, other correlations were less anticipated, such as those between subordinate clauses and corrected Type-Token Ratio (TTR) of verbs. Although removing correlated features is important for enhancing a model's performance, appreciating their interactions remains crucial for interpretation. Thus, a trade-off between interpretability, model complexity and performance emerges as a central consideration.

To maintain model interpretability, we refrained from employing feature engineering or dimensionality reduction techniques, opting instead for the

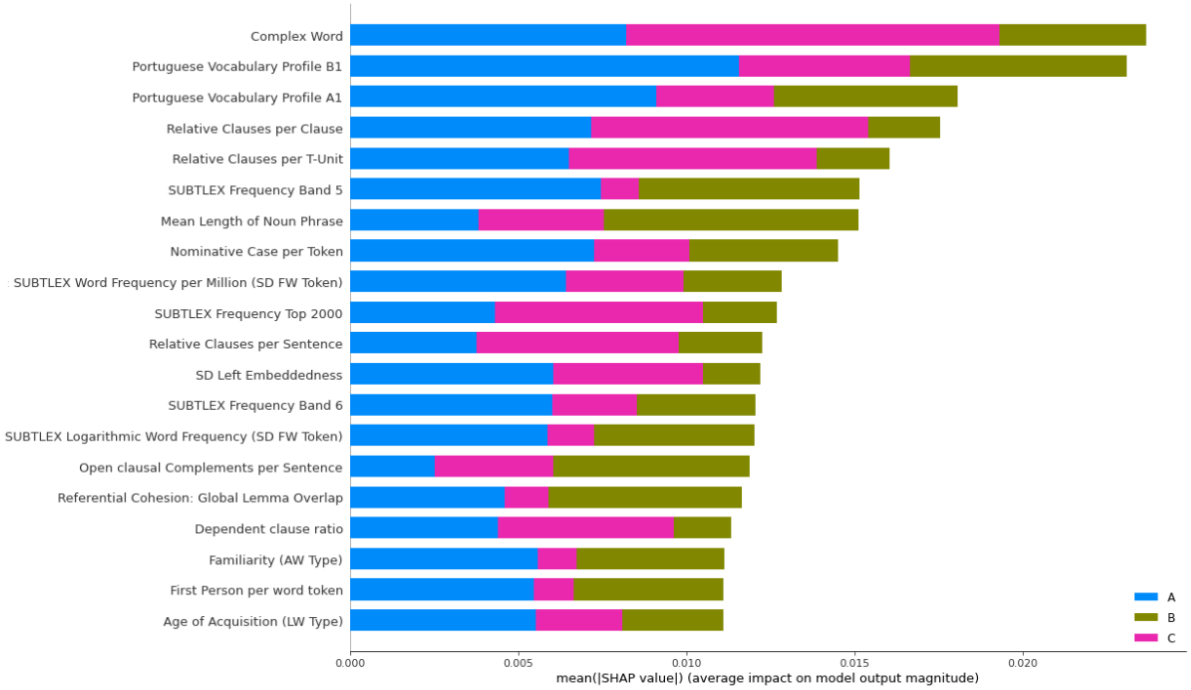


Figure 1: Distribution of the mean SHAP values of the 20 most discriminative features for proficiency level prediction (in descending order vertically). This figure also shows to what extent each measure, be its presence or absence, impacted the prediction of each level (on the horizontal axis).

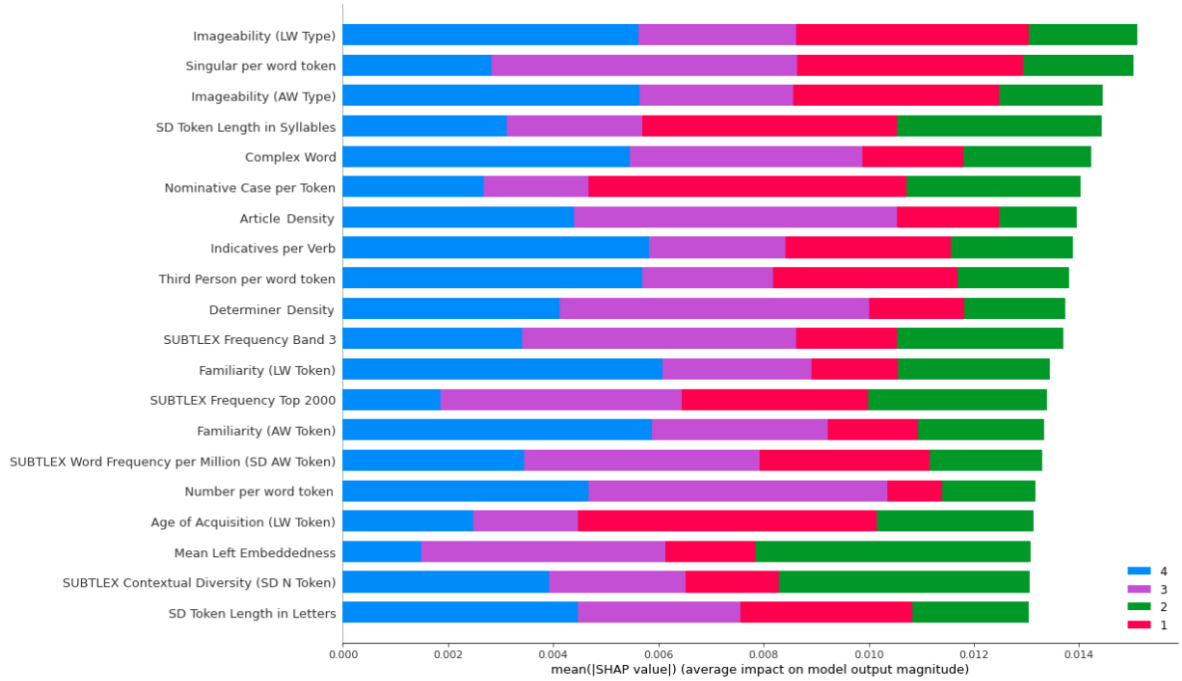


Figure 2: Distribution of the mean SHAP values of the 20 most discriminative features for readability level prediction.

CfsSubsetEval and InfoGain methods implemented by the Waikato Environment for Knowledge Analysis (WEKA, Hall et al., 2009). CfsSubsetEval identifies an informative yet uncorrelated subset of features. Similarly, InfoGain evaluates each feature's contribution to reducing entropy, aiding in the selection of the most informative features.

Reducing the number of features resulted in only a minor decline in performance, indicating that fewer features are adequate to achieve satisfactory classification results. Tables 8 and 9 provide a comprehensive overview of the SVM classifier's performance with both selected and full feature sets.

| | 10-Fold CV | | Test Set | |
|---|---|---|---|---|
| | F1 | Acc | F1 | Acc |
| All features | 0.68 | 0.70 | 0.73 | 0.72 |
| CfsSubsetEval | 0.65 | 0.67 | 0.68 | 0.68 |
| InfoGain | 0.66 | 0.65 | 0.68 | 0.67 |

Table 8: 10-fold cross-validation and test set accuracy and F1-scores achieved in the proficiency classification experiments with all the features and the selected feature sets.

| | 10-Fold CV | | Test Set | |
|---|---|---|---|---|
| | F1 | Acc | F1 | Acc |
| All features | 0.85 | 0.86 | 0.87 | 0.87 |
| CfsSubsetEval | 0.83 | 0.84 | 0.83 | 0.83 |
| InfoGain | 0.85 | 0.86 | 0.86 | 0.86 |

Table 9: 10-fold cross-validation and test set accuracy and F1-scores achieved in the readability classification experiments with all the features and the selected feature sets.

## 8 Discussion

In addition to obtaining the mean SHAP values of the most discriminative features, the SHAP values associated with each level were also inspected,[4] the measures referring to the superficial and lexical aspects exhibited the strongest discriminative power. Advanced learners produced more and longer words and sentences than beginners. A uniform increase was present concerning most of these features. The same is true for the school materials corpus herein utilized. Texts from the highest education levels demonstrated a higher incidence of words considered complex, abstract, infrequent and generally unfamiliar when compared to the lower ones. The same pattern was identified in terms of lexical variation. This is in line with the postulations in the simplification manual.

Regarding the syntactic domain, our data also corroborates most of the remarks. In the productions of L2 Portuguese learners, it was verified that clausal subjects, passive sentences, subordinate and relative clauses, as well as asyndetic coordinated clauses are indicative of more advanced levels. These grammatical constructions only arise after the general level B in the analyzed data. Comparatively, texts from the different educational levels demonstrated more homogeneity regarding syntactic measures. Although sentences and clauses are shorter in the early levels, constructions like subordinate and relative clauses as well as clausal subjects remained relatively constant across the levels. More pronounced contrasts regarding this domain were only found in terms of passive sentences and left embeddedness, which is in line with both the Portuguese RLD and the simplification manual.

Morphological measures also demonstrated contributions in differentiating the levels. For instance, it was observed that L2 learners placed at proficiency level A produced a distinctively higher amount of nominative case inflections, and, on the other hand, they exhibited low amounts of the accusative case inflection. In terms of verbal mood, it was observed that beginners also produce high amounts of indicative mood. This corroborates suggestions from the Portuguese RLD which suggests most verb tenses in the subjunctive mood are learned at levels B1 and B2. The same trend regarding the high use of the nominative case was observed in the Brazilian corpus. The elementary school texts contained a distinctively higher amount of this inflection when compared to high school or university ones; however the incidence of accusative case inflection was not as pronounced.

Concerning discursive measures, it has been suggested that as L2 learners progress, they tend to use fewer explicit cohesive devices (Crossley and McNamara, 2012). This trend was observed specifically for causal connectives: Their absence indi-

[4]The generated plots for each separate level can be found in the following link: https://osf.io/ehdc9/?view$_o nly$ = 2e7ee278d187417c82219dc6eab6e29e

cated advanced L2 proficiency. On the other hand, regarding the school texts, this was the case for temporal connectives, which may very well point to a diminished presence of narrative discourse and higher amounts of expository discourse. Finally, in terms of referential cohesion, its low values were decisive for the prediction of the beginning proficiency level, but no impact was identified for the prediction of other levels or for the school texts.

## 9 Conclusion

In this paper, we explored Portuguese broad linguistic modeling in relation to L2 proficiency and text readability. Employing an elaborate NLP pipeline, we extracted 465 measures of linguistic complexity from two corpora. Our ultimate objective was to understand which measures exerted the most impact in each level's prediction and assess the extent to which these measures support the concept of holistic, static, ascending categories of proficiency and readability by implementing classification experiments and applying explainable AI methods.

Our results show that the consistent performance across different evaluation metrics suggests that the SVM classifier, trained on a broad set of linguistic complexity measures, provides a robust framework for modeling proficiency and readability levels in Portuguese texts. In particular, lexical features were found to have strong discriminative capabilities between different proficiency and readability levels. These findings provide evidence as to validate these measures and confirm the feasibility of modeling natural language using a diverse range of linguistic features. It also shows that XAI methods can be applied to linguistic complexity analysis.

In line with the Portuguese RLD and the simplification manual, the texts herein analyzed exhibited a uniform increase in the use of longer, more abstract, less familiar and less frequent words across both proficiency and readability levels. Moreover, an increase in sentence embeddedness and coordination, as well as tense and voice inflection was also positively confirmed in our findings. Additionally, trends in discursive measures suggest shifts in cohesive device usage as proficiency progresses, with possible implications for different discourse types.

These findings offer valuable insights for the refinement of language learning tools and assessment techniques. Specifically, they emphasize the significance of certain linguistic characteristics, such as vocabulary type, morphological and syntactic complexity, in modeling learner language and assessing proficiency and readability. Additionally, our intention to make CTAP's Portuguese analysis feature openly accessible online aims to support the development of more linguistically informed analyses through an accessible platform. This initiative is expected to facilitate the integration of linguistic insights into educational technologies.

## Limitations

Although SHAP offers valuable insights, multicollinearity among highly correlated features may inflate or diminish feature importance, affecting SHAP interpretation. Despite potential changes in absolute SHAP magnitudes, relative importance rankings remain informative. SHAP values evaluate each feature's marginal contribution, taking into account feature interactions. Additionally, linguistic analyses lend credibility to SHAP interpretation.

The imbalance in both corpora underscores the necessity of balanced datasets to ensure reliable results in proficiency and readability assessment. An imbalanced corpus may lead to an overemphasis on dominant class characteristics, neglecting those of minority classes and affecting model performance. Another important observation is the influence that distinct topic and tasks may inflict in the emergence of specific grammar structures and lexical elements. These aspects have not been accounted for in these corpora's metadata, suggesting a need for future corpus creation that considers these aspects.

## References

Soroosh Akef, Amália Mendes, Detmar Meurers, 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*, pages 332–341.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Bram Bulté and Hanne Roothooft. 2020. Investigating the interrelationship between rated l2 proficiency and linguistic complexity in l2 speech. *System*, 91:102246.

Manuela L Cameirao and Selene G Vicente. 2010. Age-of-acquisition norms for a set of 1,749 portuguese words. *Behavior research methods*, 42(2):474–480.

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*, pages 113–119, Osaka, Japan. COLING.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

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.

Scott A Crossley, Tom Salsbury, Danielle S McNamara, and Scott Jarvis. 2011. Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28(4):561–580.

Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Pedro Curto, Nuno Mamede, and Jorge Baptista. 2014. Automatic readability classifier for european portuguese. *system*, 5:6.

Edgar Dale and Jeanne Chall. 1949. The concept of readability. *Elementary English*, 26(3).

Iria del Río. 2019a. Automatic proficiency classification in l2 portuguese. *Procesamiento del Lenguaje Natural*, 63:67–74.

Iria Iria del Río. 2019b. Linguistic features and proficiency classification in l2 spanish and l2 portuguese. In *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), September 30, Turku Finland*, 164, pages 31–40. Linköping University Electronic Press.

Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE.

Aline Evers. 2013. Processamento de língua natural e níveis de proficiência do português: um estudo de produções textuais do exame celpe-bras.

Iria Del Río Gayo, Marcos Zampieri, and Shervin Malmasi. 2018. A portuguese native language identification dataset. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 291–296.

Murilo Gazzola, Sidney Evaldo Leal, and Sandra Maria Aluisio. 2019. Prediç ao da complexidade textual de recursos educacionais abertos em português.

Edward Gibson et al. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Nathan S Hartmann, Gustavo H Paetzold, and Sandra M Aluísio. 2018. Simplex-pb: A lexical simplification database and benchmark for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer.

J Hawkins and Paula Buttery. 2008. Using learner language from corpora to profile levels of proficiency: Insights from the english profile programme. In *Language testing matters: Investigating the wider social and educational impact of assessment–proceedings of the ALTE Cambridge conference*, pages 158–175.

Alex Housen, Folkert Kuiken, and Ineke Vedder. 2012. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, volume 32. John Benjamins Publishing.

Jan H Hulstijn. 2015. Language proficiency in native and non-native speakers. *Language Proficiency in Native and Non-native Speakers*, pages 1–206.

Ruslan Korniichuk and Mariusz Boryczka. 2021. Averaging and boosting methods in ensemble-based classifiers for text readability. *Procedia Computer Science*, 192:3677–3685.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2021. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *arXiv preprint arXiv:2201.03445*.

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

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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.

Danielle S McNamara, Scott A Crossley, and Philip M McCarthy. 2010. Linguistic features of writing quality. *Written communication*, 27(1):57–86.

Amália Mendes and Iria del Río. 2018. Using a discourse bank and a lexicon for the automatic identification of discourse connectives. In *International Conference on Computational Processing of the Portuguese Language*, pages 211–221. Springer.

Paul Nation and James Coady. 1988. Vocabulary and reading. *Vocabulary and language teaching*, 97:110.

Isabelle Nguyen and Shuly Wintner. 2022. Predicting the proficiency level of nonnative hebrew authors. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5356–5365.

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.

Hilário Oliveira, Rafael Ferreira Mello, Bruno Alexandre Barreiros Rosa, Mladen Rakovic, Pericles Miranda, Thiago Cordeiro, Seiji Isotani, Ig Bittencourt, and Dragan Gasevic. 2023. Towards explainable prediction of essay cohesion in portuguese and english. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 509–519.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied linguistics*, 24(4):492–518.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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*.

PLE Referencial Camões. 2017. Direção de serviços de língua e cultura.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Hector Sanz, Clarissa Valim, Esteban Vegas, Josep M Oller, and Ferran Reverter. 2018. Svm-rfe: selection and visualization of the most relevant features through non-linear kernels. *BMC bioinformatics*, 19:1–18.

Ana Paula Soares, Ana Santos Costa, João Machado, Montserrat Comesaña, and Helena Mendes Oliveira. 2017. The minho word pool: Norms for imageability, concreteness, and subjective frequency for 3,800 portuguese words. *Behavior Research Methods*, 49(3):1065–1081.

Ana Paula Soares, João Machado, Ana Costa, Álvaro Iriarte, Alberto Simões, José João de Almeida, Montserrat Comesaña, and Manuel Perea. 2015. On the advantages of word frequency and contextual diversity measures extracted from subtitles: The case of portuguese. *Quarterly Journal of Experimental Psychology*, 68(4):680–696.

Lucia Specia, Sandra Maria Aluísio, and Thiago A Salgueiro Pardo. 2008. Manual de simplificação sintática para o português. *Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional (NILC-TR-08-06)*.

Shintaro Torigoe. 2017. Portuguese vocabulary profile: uma lista de vocabulário a aprendentes do pl2/ple, baseada nos corpora de aprendentes e de livros de ensino. *Revista da Associação Portuguesa de Linguística*, (3):387–400.

Rui Vaz, Fátima Mendes, and Joana Batalha. 2019. Referencial camões ple. *VI Jornadas de Português Língua Estrangeira-Aquisição e Didática*.

Zarah Weiss and Detmar Meurers. 2019. Broad linguistic modeling is beneficial for german l2 proficiency assessment. In *Widening the Scope of Learner Corpus Research, Selected Papers from the Fourth Learner Corpus Research Conference, Louvain-la-Neuve: Presses Universitaires de Louvain*, pages 419–435.