Investigating the Generalizability of Portuguese Readability Assessment Models Trained Using Linguistic Complexity Features

Soroosh Akef[1,2] **Detmar Meurers**[3,4,2] **Amália Mendes**[1] **Patrick Rebuschat**[5,2]
[1]Center of Linguistics of the University of Lisbon
[2]LEAD Graduate School and Research Network
[3]Knowledge Media Research Center
[4]University of Tübingen
[5]Lancaster University
sorooshakef@edu.ulisboa.pt dm@sfs.uni-tuebingen.de
amaliamendes@letras.ulisboa.pt p.rebuschat@lancaster.ac.uk

Abstract

This paper discusses our effort to build an automatic Portuguese readability assessment system aimed at Portuguese language learners. We demonstrate that using linguistic complexity features combined with traditional machine learning techniques allows for more control over selecting features that are more informative, resulting in models that generalize better to texts in authentic language learning environments. Using 489 linguistic complexity measures automatically extracted from a corpus of 500 texts annotated according to their level, we train random forest and LogitBoost classifiers to predict the CEFR level of a given text. Subsequently, we investigate the models' generalizability by testing them on an independently collected and annotated corpus. We conclude that through using more informative features, the models' capacity to generalize to novel data can increase even if performance on the test set extracted from the training data decreases.

1 Introduction

The crucial role of comprehensible input, as hypothesized by Krashen (1985), in the process of second language acquisition is widely agreed upon. In order for the input to be comprehensible to the learner, the complexity of the input must be in accordance with the proficiency level of the learner. Such an endeavor would require an accurate estimate of the complexity of the text and the proficiency level for which it is suitable, as well as an accurate estimate of the proficiency level of the learner.

In the context of a language class, a teacher would require significant experience to develop the intuition required to determine whether a text which is to be presented as reading material to the students is at the appropriate level, and even for an experienced teacher, it would be a burden to search for and find such a text.

Consequently, a system capable of automatically analyzing the complexity of texts and selecting a text at an appropriate level of difficulty for the learner, a task known as automatic readability assessment, can not only facilitate the teacher's job, but can also enhance the language learning experience of the learner. The need for such a system has already inspired the development of systems such as FLAIR (Chinkina and Meurers, 2016) (Chinkina et al., 2016) and Syb (Chen and Meurers, 2017) for English, KANSAS (Weiss et al., 2018) for German, and LX-Proficiency (Santos et al., 2021) for Portuguese.

However, in order for an intelligent system powered by a machine learning model to be effective in real-world settings, the ecological validity of the model must be established through tests of generalizability, such as cross-corpus validation, which attempts to test the performance of a model trained on a specific corpus on a different, independently collected corpus. For supervised tasks requiring expert-annotated data, such as automatic readability assessment, such an experiment is especially challenging, as the amount of available data may be barely sufficient for the training algorithm. Despite this challenge, previous attempts have been made to perform cross-corpus validation as a test of generalizability in particular for the task of automatic readability assessment, for instance by Vajjala and Meurers (2016) and Chatzipanagiotidis et al. (2021).

This paper discusses our attempt at the task of automatic Portuguese readability assessment using an array of linguistic complexity features and the subsequent cross-corpus validation performed in order to investigate whether more informative features result in improved generalizability. Linguistic complexity features are considered to be predictive for text readability considering that the conceptualization of linguistic complexity includes constructs such as ease or difficulty of processing, which in the context of reading passages, closely correlates with readability. The predictiveness of linguistic complexity features of text readability has also been demonstrated by previous attempts of this task for different languages (Weiss et al., 2021a) (Chatzipanagiotidis et al., 2021).

In the subsequent sections, some background on the task of automatic readability assessment, with particular focus on Portuguese, is presented; the linguistic complexity features used in the current study are discussed; the corpora used in the experiments are described; and the training, testing, and cross-corpus validation experiments are outlined. Finally, a discussion of the implications of the results and the future avenues to be explored are presented.

2 Related Work

Text readability, broadly defined in terms of the comprehensibility of a text for target readers (Klare, 1974), is an interdisciplinary line of research going back to late 19th century (DuBay, 2004). Its interdisciplinary nature is due to the various factors contributing to how "readable" a text is, ranging from features intrinsic to the text to the individual differences of the readers and the objective for which the text is read (Vajjala, 2022) (Valencia et al., 2014).

The approach taken toward readability assessment in L1, however, is different from that of L2 or heritage language. While the primary concern in the former is to maximize readability for objectives revolving around the maximum uptake of information (for instance in Aluisio et al. (2010)), for the purposes of reading for language acquisition, the readability of a text must be tuned according to the proficiency level of the learner (Xia et al., 2019). While individual differences are a factor, most computational methods of measuring readability automatically focus on intrinsic features of the text.

Earliest methods to automatically assess the readability of a text were based on readability formulae (Vajjala, 2022). However, as the fields of computational linguistics and machine learning evolved and developed more sophisticated techniques, these techniques began to yield more accurate results. An overview of the utilization of such techniques in particular for automatic Portuguese readability assessment follows.

Often framed as a supervised machine learning task, automatic readability assessment can be treated as a classification, regression, or ranking task (Xia et al., 2019). Often more important than how the task is framed, however, are the features used for this task and the size and quality of the available corpora, with most resources being available for the English language. However, noteworthy efforts have also been made in other languages, including German (Weiss and Meurers, 2022), Swedish (Pilán et al., 2016), French (Wilkens et al., 2022), Italian (Dell'Orletta et al., 2011), Arabic (Nassiri et al., 2018), Greek (Chatzipanagiotidis et al., 2021) etc.

To the best knowledge of the authors, the first work attempting automatic readability assessment for Portuguese was conducted by utilizing lexical features to train an SVM classifier over a corpus of 47 textbooks, exercise books, and national exams, designed for students of grades five to twelve, divided into eight classes according to the grade and containing a total of 6,862,024 tokens. This approach resulted in an adjacent accuracy of 0.8760 (Marujo et al., 2009).

The first attempt at the task of automatic readability assessment specifically targeting Portuguese L2 learners is by Branco et al. (2014), who used the Flesch reading ease index, along with other so-called surface features, with 125 excerpts annotated according to their CEFR level, ranging from A1 to C1, which resulted in an accuracy of 0.2182 obtained by the Flesch index, highlighting the difficulty of this task and the need for more informative features.

Another attempt at this task used a larger corpus of 237 texts categorized into five classes according to their CEFR level, and by taking advantage of a set of 52 linguistic complexity features extracted from the text using the hybrid statistical and rule-based NLP chain STRING (Mamede et al., 2012), attained an accuracy of 0.7511 using the LogitBoost machine learning algorithm (Curto et al., 2015).

Exploring deep learning approaches for this task, Correia and Mendes (2021) and Santos et al. (2021) fine-tuned neural networks to classify texts according to their CEFR label in a five-class classification task, with Correia and Mendes (2021) attaining an accuracy of 0.73 and Santos et al. (2021) attaining an accuracy of 0.7562, demonstrating the range of tasks the transformer architecture can be applied to. However, despite favorable results, lack of interpretability remains an important downfall of these models, potentially resulting in models that fail to generalize to authentic settings.

3 Linguistic Complexity Features

Often defined as the variety and sophistication of structures and words in a text (Wolfe-Quintero et al., 1998) or simply, use of more challenging and difficult language (Ellis and Barkhuizen, 2005), linguistic complexity is a construct which has been quite prevalent in various disciplines of linguistics, ranging from phonology, to psycholinguistics, and computational linguistics.

This prevalence, however, has also contributed to disagreements over how this construct should be conceptualized (Pallotti, 2015), with syntactic and lexical complexity features dominating the features used to operationalize complexity. Nonetheless, features informed by research in the fields of discourse analysis and psycholinguistics have also shown to be informative predictors for the task of automatic readability assessment (Weiss et al., 2021b) (Weiss and Meurers, 2018).

To extract the linguistic complexity measures from texts, we utilized CTAP [1], a freely available linguistic complexity analyzer initially developed by Chen and Meurers (2016) for English and later expanded to include other languages, including Portuguese (Ribeiro-Flucht, 2023). However, as of this writing, the version of the tool supporting Portuguese is not yet online, and the authors were granted local access for the current work.

A total of 489 complexity features for Portuguese are currently extractable, with the majority of the features being lexical features, as demonstrated in Table 1.

Count-based features, referring to features indicating the raw count of constituents, are sometimes considered as syntactic complexity features owing to the fact that longer linguistic units are often more syntactically complex. However, for the purposes of the current task of automatic readability assessment, they are categorized in a class of their own. Examples of this class of features include number of agent modifiers, number of complex noun phrases, among others.

Lexical features, the most populous class of features in the current study, capture the sophistication and richness of the vocabulary used in a given text. The most typical examples of this class of features are variations of type-token ratio (root, logarithmic, corrected, standard) and word frequency per million.

The other class commonly used in studies involving linguistic complexity analysis is syntactic features, which are indicators of the sophistication of the structures used in the text, including the rate of subordination or embedding. Examples of syntactic features used in the current work include prepositional phrase types per token and mean length of clause.

Another class of features contributing to the complexity of a text is morphological features, which capture the inflections and derivations of lexical items, such as first person per word token or indicatives per word token.

A relatively under-utilized class of features used in this study is discourse features, which can be regarded as a metric of the coherence and cohesion of the text. Examples of this class of features used in this study include temporal connectives per token and single-word connectives per token.

Finally, psycholinguistic features draw on the research in this field to extract measures such as age of acquisition and imageability, which could be considered as a subset of lexical features.

4 Data

Two independently collected corpora were used in the current study. The first corpus is identical to the corpus dubbed c500 in Santos et al. (2021), which contains 500 excerpts of books, newspaper articles, etc., annotated by teachers of the Camões Institute[2] according to their CEFR level, ranging from A1 (elementary) to C1 (advanced), excluding C2 (proficient). These texts have been used as part of exams administered to heritage language learners of Portuguese aged six to eighteen in the countries of Switzerland, Spain, Germany, and Andorra.

Smaller subsets of this corpus, dubbed c237, c225bal, and c114, have also been used in previous studies (Santos et al., 2021) (Branco et al., 2014) (Curto et al., 2015). c114, is a subset of the later expanded c237, which is in turn a subset of c500. c225bal is a balanced version of c500 in which the number of texts in each proficiency class has been truncated to match that of the smallest class, i.e. B2 with 45 texts.

Importantly, the corpus c114 was later deemed poorly annotated (Santos et al., 2021), prompting the introduction of a new subset of c500, dubbed

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

[2] https://www.instituto-camoes.pt/

Class	Count-Based	Lexical	Syntactic	Discourse	Morphological	Psycholinguistic
Count	98	226	74	42	32	17

Table 1: Count of features by class.

c386, which excludes the 114 poorly annotated texts in c500. The use of a smaller but higher quality corpus and how the performance of models on it compare to the original c500 corpus would also be investigated in this study.

The distribution of texts among the classes of c500 and its subsets is outlined in Table 2, where class imbalance in all corpora, save c225bal, is visible, with the vast majority of texts belonging to level B1 in the corpora c237 and c114, and the distribution of the texts being skewed toward A2 and B1 in c500 and c386.

Corpus	A1	A2	B1	B2	C1
c500	80	135	184	45	56
c386	69	124	112	37	44
c237	29	39	136	14	19
c225bal	45	45	45	45	45
c114	11	11	72	8	12

Table 2: Corpora distribution.

The second corpus, which was not used in previous studies, contains 157 texts distributed among six CEFR classes (A1-C2), which were extracted from reading activities in Portuguese L2 textbooks using optical character recognition (OCR) technology and shared with the authors. In order to fix the mistakes resulting from OCR, the authors utilized GPT-4 (OpenAI, 2023), which proved quite capable of this task, owing to its understanding of the context.

Among the noteworthy differences between this corpus, henceforth referred to as the validation corpus, and previously described corpora is the different distributions of texts across classes.

Table 3 demonstrates the imbalance of the validation corpus in favor of the B2 level. This is in direct contrast to c500 and its subsets (with the exception of c225bal), in which B2 was the minority class. This drastic difference in distribution poses a significant challenge to models trained on one corpus and tested on the other. Furthermore, as c500 and its subsets did not include texts from level C2, texts at this level were also excluded from the validation corpus at the time of testing.

Additionally, the fact that c500 and the validation corpus come from different sources poses a challenge with regard to the annotation scheme. While the texts in the validation corpus were from textbooks and were therefore intended to aid the language acquisition of adolescent or adult L2 learners of Portuguese, the texts in c500 were intended for examination of heritage language learners of Portuguese.

Level	A1	A2	B1	B2	C1	C2
Count	12	23	38	67	8	9

Table 3: Text distribution in the validation corpus.

The complications arising from these factors make it justifiable to also use a laxer metric of performance, namely adjacent accuracy, which considers a prediction correct as long as it falls in or within one class of the true class.

5 Experiments and Discussion

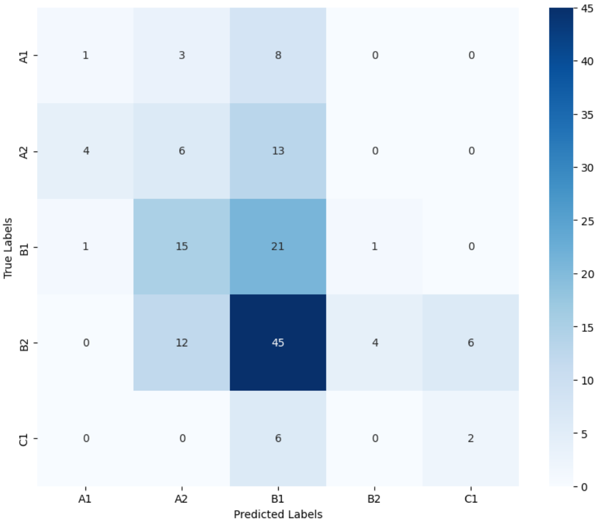
Two classification algorithms of random forest and LogitBoost were used to train models using the 489 linguistic complexity measures extracted from the texts in the c500 corpus and its subsets. As the primary interest of this investigation was to study how the utilization of subsets of a broad range of linguistic features impact generalizability as opposed to necessarily optimize the performance of the trained model, we opted to use the full feature set without feature selection despite the high dimension of the features compared to the data size. The random forest algorithm was selected primarily because of the insight it would be possible to gain by extracting the importance of the features according to the reduction in Gini impurity, but also because as an ensemble model, it is less prone to noise in the data, which considering the inherent complexity of the readability assessment task, is an important quality for the model to have. The LogitBoost algorithm was primarily selected to allow for comparability with the previous attempts of this task, in particular Curto et al. (2015).

5.1 Using all features

In order to train the models, c500 and its subsets were each divided into five folds for hyperparame-

	c500		c386		c237		c225bal		c114	
	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy
Random Forest (our model)	0.6117	0.7200	0.6284	0.6969	0.5040	0.7299	0.5834	0.5867	0.6557	0.8241
LogitBoost (our model)	0.5866	0.6800	**0.6394**	**0.7020**	**0.5641**	0.7427	0.5556	0.5556	0.5774	0.8071
LogitBoost (Curto et al., 2015)	0.643	0.6860	-	-	0.553	0.7412	0.595	0.5970	**0.737**	**0.8684**
GPT-2 (Santos et al., 2021)	**0.689**	**0.7562**	-	-	0.556	**0.7623**	**0.649**	**0.6548**	0.675	0.8421
RoBERTa (Santos et al., 2021)	0.589	0.725	-	-	0.510	0.7545	0.562	0.6319	0.615	0.8532

Table 4: Comparison of the models with previous works on macro F1 and exact accuracy



	A1	A2	B1	B2	C1
A1	1	3	8	0	0
A2	4	6	13	0	0
B1	1	15	21	1	0
B2	0	12	45	4	6
C1	0	0	6	0	2

Figure 1: Confusion matrix heatmap for cross-corpus validation of random forest trained on c500 using all features.

ter fine-tuning through grid search and were subsequently trained and tested on each corpus through 5-fold cross-validation using all 489 features.

Using all the features, the random forest model attained an accuracy of 0.72 and macro F1 score of 0.6117 on c500, demonstrating a comparable, albeit slightly poorer performance, to that of the transformer-based models in Santos et al. (2021). The LogitBoost model also performed similarly to the LogitBoost model trained using 52 features consisting of mostly length-based features in Curto et al. (2015) and re-implemented by Santos et al. (2021) by attaining an accuracy of 0.68 and a macro F1 of 0.5866 despite the much larger number of features. A comparison of the performance of the models on the different subsets is presented in Table 4.

Subsequently, the models were trained on all the samples from c500 and its subsets to perform cross-corpus validation on the validation corpus. Despite the accuracy of 0.2297 and macro F1 of 0.1992 not showing promising results, inspecting the confusion matrix heatmap of the cross-corpus validation indicated a systematic underestimation of the elementary and lower-intermediate levels of A1, A2, and B1 (Figure 1). Consequently, the adjacent accuracy score for cross-corpus validation of the same model stood at 0.8176, which is a considerable improvement over the random guess baseline of 0.52 for adjacent accuracy among five classes.

Upon closer inspection of the most important features to the random forest model, it was observed that 14 out of the top 20 most important features to the model are raw count features, which were either identical or closely resembled the 52 features used in Curto et al. (2015) (Table 5), leading the model to draw the conclusion that the length of the text has a correlation with its difficulty, an assumption that could result in poor generalizability of the model. Subsequently, the hypothesis that more informative and theoretically-supported features would lead to better generalizability was tested by removing all length-based features, including raw counts and type-token ratio, which is heavily correlated with the length of the text, and training the models again.

5.2 Excluding length-based features

By training the models again using the 332 remaining length-independent features, it was observed that even though the performance of the models

Features	**Category**
Number of Word Types (excluding Punctuation and numbers)	Count-based
Number of syllables	Count-based
Lexical Richness: Type Token Ratio (Corrected TTR)	Lexical
Number of POS Feature: Noun Lemma Types	Count-based
Number of POS Feature: Lexical word Tokens	Count-based
Number of POS Feature: Noun Tokens	Count-based
Number of POS Feature: Lexical word Lemma Types	Count-based
Number of Word Types (including Punctuation and Numbers)	Count-based
Number of Word Tokens (including Punctuation and Numbers)	Count-based
Number of POS Feature: Noun Types	Count-based
Lexical Richness: Type Token Ratio (STTR Lexical Words)	Lexical
Lexical Richness: Type Token Ratio (Corrected TTR Lexical Words)	Lexical
Number of Word Tokens (excluding punctuation and numbers)	Count-based
Number of Tokens with More Than 2 Syllables	Count-based
Number of Word Types with More Than 2 Syllables	Count-based
Lexical Sophistication Feature: SUBTLEX Word Frequency per Million (AW Type)	Lexical
Number of Syntactic Constituents: Prepositional Phrase	Count-based
Number of Tokens	Count-based
Lexical Richness: Type Token Ratio (Root TTR)	Lexical
Lexical Richness: Type Token Ratio (STTR Nouns)	Lexical

Table 5: Top 20 most important features for the random forest model when trained and tested on c500.

	c500		**c386**		**c237**		**c225bal**		**c114**	
	F1	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**
Random Forest (all features)	0.61	0.72	0.63	0.70	0.50	0.73	0.59	0.59	0.66	0.82
Cross-corpus validation	**0.20**	**0.23**	**0.28**	0.30	0.18	**0.21**	**0.23**	**0.22**	0.10	**0.14**
Random Forest (length-based removed)	0.52	0.62	0.59	0.65	0.36	0.67	0.51	0.51	0.48	0.73
Cross-corpus validation	**0.30**	**0.24**	**0.30**	0.24	0.17	**0.26**	**0.27**	**0.28**	0.08	**0.26**
LogitBoost (all features)	0.59	0.68	0.64	0.70	0.56	0.74	0.56	0.56	0.58	0.81
Cross-corpus validation	0.22	0.26	0.34	0.33	0.16	0.25	**0.23**	**0.23**	0.17	0.27
LogitBoost (length-based removed)	0.54	0.62	0.53	0.59	0.43	0.69	0.53	0.53	0.52	0.78
Cross-corpus validation	0.21	0.26	0.25	0.26	0.16	0.25	**0.27**	**0.27**	0.14	0.24

Table 6: Comparison of the performance of the models in cross-corpus validation on macro F1 and exact accuracy with the instances of better performance on cross-corpus validation when excluding shallow features highlighted in boldface.

Experiment	**c500**	**c386**	**c237**	**c225bal**	**c114**
Random Forest (all features)	0.94	0.94	0.92	0.95	0.94
Cross-corpus validation	**0.82**	**0.84**	**0.79**	**0.82**	**0.75**
Random Forest (length-based removed)	0.90	0.91	0.84	0.89	0.85
Cross-corpus validation	**0.89**	**0.95**	**0.89**	**0.85**	**0.86**
LogitBoost (all features)	0.93	0.92	0.90	0.93	0.97
Cross-corpus validation	0.86	0.85	**0.80**	**0.83**	0.84
LogitBoost (length-based removed)	0.89	0.91	0.88	0.87	0.92
Cross-corpus validation	0.80	0.81	**0.85**	**0.86**	0.80

Table 7: Adjacent accuracy metrics for each model across different corpora with cross-corpus validation with the instances of better performance on cross-corpus validation when excluding shallow features highlighted in boldface.

when trained and tested on c500 and its subsets decreased, the generalizability of the models in many instances improved. Table 6 includes an overview of the accuracy and macro F1 scores calculated for the two models on different corpora and their cross-corpus validation results.

Table 7 displays the results for the same models and corpora according to adjacent accuracy.

The better generalizability of the random forest model trained on more informative linguistic complexity features is particularly visible in Table 7, in which adjacent accuracy of cross-corpus validation for the higher quality c386 corpus has increased from 0.84 to 0.95. This is also true for the poorly annotated c114 corpus, which despite the below random generalization results when using accuracy and macro F1, managed to attain an improved adjacent accuracy of 0.86 when using more informative features compared to the 0.75 when using more shallow features. This is plausible, as even if a corpus is poorly annotated, the human annotator is unlikely to stray farther than one class away from the true label.

The same pattern, however, is not consistently observed with LogitBoost, with c225bal and c237 resulting in a better performance in cross-corpus validation and the other corpora resulting in a worse generalization for this model. This may be attributed to the different training mechanism of this model, which warrants further investigation with other classification algorithms to identify the underlying cause of this difference in behavior between the two algorithms.

5.3 Fine-tuning GPT-3.5 Turbo

In an attempt to investigate how state-of-the-art large language models compare with regard to generalizability to the feature-based models used in this work, GPT-3.5 Turbo was fine-tuned using 320 of the texts in c500 as the training set, 80 texts as the validation set, and 100 texts as the test set by respecting the distribution of the texts among levels in the entire corpus for each set. OpenAI's API was used to fine-tune the base model gpt-3.5-turbo-1106 in three epochs while maintaining the recommended values for the hyperparameters.

The fine-tuned model attained an accuracy of 0.79 and macro F1 score of 0.7011 on the test set, expectedly outperforming the model based on GPT-2 used in Santos et al. (2021). In cross-corpus validation, the fine-tuned GPT-3.5 model appeared to perform notably better than all the other feature-based models according to the accuracy and macro F1 metrics by attaining an accuracy of 0.3581 and a macro F1 score of 0.3463. The fine-tuned model's adjacent accuracy score of 0.9391 was also better than all but one of the feature-based models.

Despite this apparently better performance, the problem at the heart of all models based on artificial neural networks, i.e. their lack of interpretability, casts doubt on whether the fine-tuned GPT-3.5 model indeed bases its assessment of the readability of texts on theoretically sound characteristics of the text contributing to readability. For instance, it is plausible that texts written for beginners share themes concerning daily routine and general topics while texts written for more advanced learners are more specialized and cover complex topics such as politics or philosophy. Indeed, such a correlation between the topics and the level of proficiency exists in learner corpora containing texts produced by learners (for instance in Mendes et al. (2016)), which makes it all the more likely that reading passages selected for learners follow a similar pattern. Therefore, a model highly capable of encoding the semantic information of a text can exploit a correlation between this information and the levels of readability, similar to how feature-based models exploited length-dependent features. When such a correlation is prevalent enough in diverse datasets, cross-corpus validation can also fail to demonstrate the shortcomings of such a model.

It could of course be argued that the topic of a text is an ecologically valid indicator of its readability, as a text containing complex concepts is inherently more difficult to read regardless of its linguistic complexity. However, the possible exploitation of text topics was only one imaginable scenario in which the fine-tuned large language model takes advantage of an unanticipated quality of the texts. It remains within the realm of possibility that the model may have found a correlation between the letter P and the readability of a text, which also happens to perform relatively well in cross-corpus validation. As ludicrous as such a claim may be, there is no way to falsify all such possibilities, which renders the utilization of large language models in scientific settings generally undesirable.

6 Conclusion

This study attempted to showcase the importance of cross-corpus validation to test the ecological validity of machine learning models before they are deployed.

It was demonstrated that by using linguistic complexity features combined with traditional machine learning algorithms, one could be more confident about the model using more informative features, which result in models that generalize better to samples different from the training set.

We also demonstrated why despite the model trained on c114 having attained the highest accuracy and macro F1 score of 0.82 and 0.66 respectively, it should not be considered as the best model for deployment, as it has the worst generalization capability among subsets of c500.

Moreover, in an attempt to compare the generalizability of feature-based models to that of large language models, GPT-3.5 Turbo was fine-tuned for the task of automatic readability assessment and demonstrated a superior performance to most other models in all the metrics. However, a case for why such a superior performance must not be taken at face value was presented.

The future directions of this work include comparing how features extracted using other linguistic complexity feature extractors available for Portuguese, such as Leal et al. (2023) perform on this task.

Furthermore, considering the demonstrated impact of using more informative features, the development of features based on criterial features, which are grammatical constructs whose appearance in a text could be indicative of the level of that text could open the path for the development of more generalizable automatic readability assessment systems.

Acknowledgements

We would like to thank the Camões Institute and the Centre for Evaluation of Portuguese as a Foreign Language (CAPLE) for graciously sharing with us the data used in this work. 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 (UIDP/00214/2020) and the Camões Institute.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014. Assessing automatic text classification for interactive language learning. In *International Conference on Information Society (i-Society 2014)*, pages 70–78. IEEE.

Savvas Chatzipanagiotidis, Maria Giagkou, and Detmar Meurers. 2021. Broad linguistic complexity analysis for Greek readability classification. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–58, Online. Association for Computational Linguistics.

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.

Xiaobin Chen and Detmar Meurers. 2017. Challenging learners in their individual zone of proximal development using pedagogic developmental benchmarks of syntactic complexity. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 8–17, Gothenburg, Sweden. LiU Electronic Press.

Maria Chinkina, Madeeswaran Kannan, and Detmar Meurers. 2016. Online information retrieval for language learning. In *Proceedings of ACL-2016 System Demonstrations*, pages 7–12, Berlin, Germany. Association for Computational Linguistics.

Maria Chinkina and Detmar Meurers. 2016. Linguistically aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–198, San Diego, CA. Association for Computational Linguistics.

João Correia and Rui Mendes. 2021. Neural complexity assessment: A deep learning approach to readability classification for european portuguese corpora. In *Intelligent Data Engineering and Automated Learning – IDEAL 2021*, pages 300–311, Cham. Springer International Publishing.

Pedro Curto, Nuno Mamede, and Jorge Baptista. 2015. Automatic text difficulty classifier. In *Proceedings of the 7th International Conference on Computer Supported Education*, volume 1, pages 36–44.

Felice Dell’Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of Italian texts with a view to text simplification. In

Proceedings of the second workshop on speech and language processing for assistive technologies, pages 73–83.

William H DuBay. 2004. The principles of readability. *Online Submission*.

Rod Ellis and Gary Patrick Barkhuizen. 2005. *Analysing learner language*. Oxford applied linguistics. Oxford University Press.

George R. Klare. 1974. Assessing readability. *Reading Research Quarterly*, 10(1):62–102.

Stephen D. Krashen. 1985. *The input hypothesis : issues and implications*. Longman.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources Evaluation*.

Nuno Mamede, Jorge Baptista, Cláudio Diniz, and Vera Cabarrão. 2012. String: An hybrid statistical and rule-based natural language processing chain for Portuguese. In *Computational Processing of the Portuguese Language, Proceedings of the 10th International Conference, PROPOR*, pages 17–20.

Luis Marujo, José Lopes, Nuno Mamede, Isabel Trancoso, Juan Pino, Maxine Eskenazi, Jorge Baptista, and Céu Viana. 2009. Porting REAP to European Portuguese. In *International Workshop on Speech and Language Technology in Education*.

Amália Mendes, Sandra Antunes, Maarten Jansseen, and Anabela Gonçalves. 2016. The COPLE2 corpus: a learner corpus for Portuguese. In *Proceedings of the Tenth Language Resources and Evaluation Conference–LREC-16*, pages 3207–3214. European Language Resources Association.

Naoual Nassiri, Abdelhak Lakhouaja, and Violetta Cavalli-Sforza. 2018. Arabic readability assessment for foreign language learners. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 480–488. Springer.

OpenAI. 2023. GPT-4 technical report.

Gabriele Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A readable read: Automatic assessment of language learning materials based on linguistic complexity. *arXiv preprint arXiv:1603.08868*.

Luisa Ribeiro-Flucht. 2023. Assessment of text readability and learner proficiency with linguistic complexity. Master's thesis, University of Tübingen.

Rodrigo Santos, João Rodrigues, António Branco, and Rui Vaz. 2021. Neural text categorization with transformers for learning portuguese as a second language. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, pages 715–726. Springer.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Walt Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *ArXiv*, abs/1603.06009.

Sheila W Valencia, Karen K Wixson, and P David Pearson. 2014. Putting text complexity in context: Refocusing on comprehension of complex text. *The Elementary School Journal*, 115(2):270–289.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021a. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54, Online. LiU Electronic Press.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021b. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54, Online. LiU Electronic Press.

Zarah Weiss, Sabrina Dittrich, and Detmar Meurers. 2018. A linguistically-informed search engine to identifiy reading material for functional illiteracy classes. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 79–90, Stockholm, Sweden. LiU Electronic Press.

Zarah Weiss and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 303–317, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153, Seattle, Washington. Association for Computational Linguistics.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey, and Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. In *Proceedings of* *the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. European Language Resources Association.

Kathryn Elizabeth Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second language development in writing : measures of fluency, accuracy, complexity*. Technical report. Second Language Teaching Curriculum Center, University of Hawai'i at Mānoa ; Distributed by University of Hawai'i Press.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2019. Text readability assessment for second language learners. *arXiv preprint arXiv:1906.07580*.