Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches

Eugénio Ribeiro[1] and **Nuno Mamede**[1,2] and **Jorge Baptista**[1,3]
[1] INESC-ID Lisboa, Portugal
[2] Instituto Superior Técnico, Universidade de Lisboa, Portugal
[3] Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Portugal
{eugenio.ribeiro,nuno.mamede,jorge.baptista}@inesc-id.pt

Abstract

The automatic assessment of text readability and the classification of texts by levels is essential for language education and language-related industries that rely on effective communication. In European Portuguese, most of the studies on this subject focus on identifying the level of texts used for proficiency evaluation purposes according to the Common European Framework of Reference for Languages (CEFR). However, the ordinal nature of the levels is not considered by the classification models used in those studies. In this paper, we address the problem as a regression task in an attempt to leverage that information. Our experiments using fine-tuned versions of a state-of-the-art foundation model for Portuguese show that addressing the problem as a regression task leads to improved performance in terms of adjacent accuracy and improved generalization ability to different kinds of textual data.

1 Introduction

Identifying the readability or complexity level of a text is relevant across diverse domains, encompassing not only language education but also various language-related industries and many other human activities, in order to adjust it according to the target audience. However, automatically determining the readability level of texts presents its own set of challenges, particularly when working with languages that have limited annotated resources, as is the case of the European variety of Portuguese.

Most of the research on this subject in the context of European Portuguese has focused on the automatic assessment of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) level of texts used for proficiency evaluation purposes by Camões, I.P. [1], the official Portuguese language institute. However, even though the levels have an ordinal nature, recent studies have approached the problem as a classification task in which the ordinal relations between the levels are not considered by the models (Curto et al., 2015; Santos et al., 2021; Ribeiro et al., 2024).

In this study, we explore the use of regression approaches that consider the ordinal nature of CEFR levels and assess how they perform in comparison to a classification approach based on the same foundation model (Bommasani et al., 2021). More specifically, we compare the performance of fine-tuned versions of the Albertina PT-PT model (Rodrigues et al., 2023) that address the problem as either a classification or regression task. Furthermore, we also explore the adaptation of the classification model to the regression task, by leveraging the predicted class probability distributions.

We start by providing an overview on related work on automatic text readability level assessment in Section 2. Then, in Section 3, we describe our experimental setup, including the dataset, the foundation model, and the methodologies employed for fine-tuning and evaluation. Next, in Section 4, we present and discuss the results of our experiments. Finally, in Section 5, we summarize the contributions of this study and provide pointers for future research in the area.

2 Related Work

Automatic readability assessment is a problem that has been widely explored over the years. Traditionally, it was addressed by creating readability formulas or indexes based on statistical information and/or domain knowledge (Kincaid et al., 1975; DuBay, 2004; Crossley et al., 2017). However, considering the developments in Natural Language Processing (NLP), research shifted towards following the trends in that area (McNamara et al., 2014), from the pairing of handcrafted features with traditional machine learning algorithms (e.g. Aluisio et al., 2010; François and Fairon, 2012; Karpov

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

et al., 2014; Curto et al., 2015; Pilán and Volodina, 2018; Forti et al., 2020; Leal et al., 2023) to the fine-tuning of large transformer-based foundation models (e.g. Santos et al., 2021; Yancey et al., 2021; Martinc et al., 2021; Mohtaj et al., 2022).

Although several studies have addressed text readability or complexity assessment as a regression task (e.g. Marujo et al., 2009; Cha et al., 2017; Nadeem and Ostendorf, 2018; Martinc et al., 2021; Wilkens et al., 2022; Mohtaj et al., 2022), only a few explored the differences between regression and classification approaches to the task.

Heilman et al. (2008) compared linear regression with the Proportional Odds Model (McCullagh, 1980) and multiclass logistic regression. The second achieved the best performance in terms of correlation, Root Mean Squared Error (RMSE), and adjacent accuracy in a cross-validation scenario. However, the simpler linear regression model generalized better to a left-out test set.

Aluisio et al. (2010) compared the performance of Support Vector Machines (SVMs) trained for classification, regression, and ordinal classification with the Proportional Odds Model. The models performed similarly, but each had a slight advantage in terms of one of the evaluation metrics, with classification achieving the highest F_1 score, regression the highest correlation, and ordinal classification the lowest error.

Xia et al. (2016) compared SVM classification with a pairwise ranking approach and achieved a better correlation with the former.

Focusing on Portuguese, there are a few studies covering the Brazilian variety of the language (e.g. Scarton and Aluísio, 2010; Aluisio et al., 2010; Leal et al., 2023). However, in this study, we will focus on the European variety.

The Portuguese version of the REAP tutoring system (Marujo et al., 2009) included a readability level classifier trained on school textbooks. The model was based on SVMs applied to lexical features and used the Proportional Odds Model to capture the ordinal nature of the levels.

The remaining studies mainly focused on the automatic assessment of the CEFR-level of texts used for proficiency evaluation purposes. Branco et al. (2014a,b) explored the use of four independent features: Flesch Reading Ease index, lexical category density, average word length, and average sentence length. Curto et al. (2015) explored the use of several traditional Machine Learning (ML) algorithms for the task. The algorithms were applied to 52 features split into 5 different groups: Part-of-Speech (POS), chunks, sentences and words, verbs, averages and frequencies, and extras. The highest performance was achieved using LogitBoost (Friedman et al., 2000). Santos et al. (2021) explored the fine-tuning of Portuguese versions of the GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) foundation models. The highest performance was achieved by the former. We have performed a more thorough study (Ribeiro et al., 2024) covering several additional foundation models. The highest performance in a cross-validation scenario was achieved using a fine-tuned version of the Albertina PT-PT model (Rodrigues et al., 2023). However, considering the reduced amount of training data, using a smaller model as a foundation leads to better generalization ability.

3 Experimental Setup

In this section, we describe our experimental setup. We start by describing the dataset used in our experiments in Section 3.1. Then, in Section 3.2, we shortly describe the foundation model used in our study. In Section 3.3, we describe the methodology used for fine-tuning that model and evaluate its performance on the task. Finally, in Section 3.4, we provide implementation details that enable the future reproduction of our experiments.

3.1 Dataset

Similarly to most of the previous studies on automatic text readability assessment in European Portuguese, our dataset is comprised of texts extracted from the Portuguese exams performed by Camões, I.P.. The texts cover the CEFR levels A1 to C1, as defined in the Portuguese version of the framework (Grosso et al., 2011; Direção de Serviços de Língua e Cultura, Camões, I.P., 2017). We use the same version of the dataset used in our previous study (Ribeiro et al., 2024), consisting of a training set of 598 texts extracted from exams that are not publicly available and a test set of 32 texts extracted from the publicly available model exams. Table 1 shows the distribution of the texts across levels.

3.2 Models

As a foundation model, we use the base version of the Albertina PT-PT model (Rodrigues et al., 2023), as it led to the best results in our previous study on text readability assessment (Ribeiro et al., 2024). We fine-tune this model for both classification and regression tasks. For the former, each CEFR level

	A1	A2	B1	B2	C1	Total
Train	92	157	240	49	60	598
Test	8	12	5	3	4	32

Table 1: Distribution of the texts in the dataset of Camões, I.P. exams across CEFR levels.

Approach	RMSE	Acc	Adj	F_1
Classification	0.5491	80.02	96.96	73.76
Regression	0.5236	79.10	**97.68**	72.93
Softmax Reg.	**0.5190**	**80.07**	97.27	**73.86**

Table 2: Results in the cross-validation scenario.

is considered an independent class, while for the latter the levels are converted to numerical values. Additionally, we explore the adaptation of the classification model to the regression task, by computing the weighted average of the class probability distribution obtained using the softmax function. We refer to this approach as softmax regression.

3.3 Evaluation Methodology

Starting with the evaluation metrics, considering that we are addressing the problem as a regression task, we report the RMSE. Additionally, we adopt accuracy (Acc), adjacent accuracy (Adj), and the macro F_1 score, which are some of the most common across previous studies. To compute these metrics for the regression approaches, we convert the numerical prediction to the closest level.

We rely on two evaluation scenarios. First, 10-fold cross-validation is used to perform hyperparameter tuning and assess the highest performance that can be achieved in a scenario similar to those of previous studies. In each fold, the model is fine-tuned for 20 epochs. The best epoch is selected according to the accuracy of the model. Second, we apply the models to the test set to assess their generalization ability. Considering that the cross-validation process generates one model per fold, we use them as an ensemble to generate the predictions for the test set by averaging their predictions.

To enhance robustness, we performed 10 independent experimental runs, each with a different random seed for the cross-validation splitting process. The evaluation metrics are reported as the average across these runs. All non-error metrics are reported in percentage form.

3.4 Implementation Details

To train our models, we relied on the functionality offered by the HuggingFace's Transformers library (Wolf et al., 2020). We used the default values for most of the hyperparameters. However, we performed a grid search to identify appropriate values for the batch size and learning rate. In our experiments, the best results were achieved using a batch size of 32 and a learning rate of 5×10^{-5}.

4 Results

In Section 4.1, we start by presenting and discussing the results achieved in the cross-validation scenario. Then, in Section 4.2, we assess the generalization ability of the multiple approaches by analyzing their performance on the test set.

4.1 Cross-Validation

Table 2 shows the cross-validation results achieved using the different approaches to the task. First of all, similarly to what was observed by Aluisio et al. (2010), we can see that the performance differences between approaches are small. More specifically, the differences between the highest and lowest average performance are 0.03 in terms of RMSE and around 1 percentage point in terms of the remaining metrics. This suggests that the foundation model and the data used for fine-tuning are more relevant than capturing the ordinal nature of the readability levels. Still, the differences may become more evident if more diverse data is considered.

Comparing the results in terms of specific metrics, as expected, the regression approaches have a lower RMSE than the classification approach. However, softmax regression achieved a lower RMSE than pure regression, in spite of the model not being specifically trained to minimize that loss. This suggests that the higher number of neurons in the output layer improves the ability of the model to capture the specific characteristics of each level, which can then be used to obtain a closer approximation of the actual level of a text. Additionally, although softmax regression only slightly outperforms the classification approach in terms of accuracy and macro F_1, it more significantly outperforms it in terms of adjacent accuracy. This suggests that weighting the probability attributed to each level instead of simply selecting the one with the highest probability is an appropriate approach to capture some information regarding the ordinal nature of the levels. Still regarding adjacent

Class.		Predicted				
		A1	**A2**	**B1**	**B2**	**C1**
Actual	**A1**	75	14	3	0	0
	A2	25	128	4	0	0
	B1	1	7	218	13	2
	B2	0	0	7	34	8
	C1	0	0	11	9	40

Reg.		Predicted				
		A1	**A2**	**B1**	**B2**	**C1**
Actual	**A1**	71	18	3	0	0
	A2	34	122	1	0	0
	B1	0	7	221	10	2
	B2	0	0	17	31	1
	C1	0	1	11	14	34

Smax		Predicted				
Reg.		**A1**	**A2**	**B1**	**B2**	**C1**
Actual	**A1**	74	16	2	0	0
	A2	24	130	3	0	0
	B1	0	7	218	14	1
	B2	1	0	9	37	2
	C1	0	0	11	10	39

Table 3: Confusion matrices of the best runs of the different approaches in the cross-validation scenario.

Approach	RMSE	Acc	Adj	F_1
Classification	1.1067	43.13	78.13	51.27
Regression	**0.8022**	**49.06**	**92.81**	**53.50**
Softmax Reg.	1.0129	43.75	80.00	51.05

Table 4: Results achieved on the test set.

accuracy, pure regression leads to the best results. However, it comes at the cost of a significant drop in performance in terms of accuracy and macro F_1 in comparison to softmax regression.

To obtain additional insight regarding the performance of the approaches, Table 3 shows the confusion matrices of the best run of each of them. We can see that all approaches have their highest recall for level B1, which is both the one in the middle and the most prominent level in the dataset. This might suggest some bias towards the prediction of that level. However, that is also one of the levels with higher precision, only surpassed by level C1 for the regression approaches. On the other hand, the models seem to have some difficulties in distinguishing between the A levels, especially the one obtained using the pure regression approach. There are also issues at the other end of the spectrum. First, there is a set of C1 texts that are classified as B1 by every model. The recognition of level B2 is that which varies the most among approaches. When using the classification approach, misclassifications fall on both of its neighbors. On the other hand, regression approaches seem to be more biased towards the B1 class. Still, softmax regression is significantly more accurate than pure regression.

4.2 Generalization to the Test Set

Table 4 shows the results achieved when the models trained for the cross-validation scenario are applied to the test set. Similarly to what was observed in our previous study (Ribeiro et al., 2024), the performance of the models is significantly impaired when they are applied to this test set. In terms of accuracy, it decreases to nearly half for the classification and softmax regression approaches. The performance of the pure regression approach also decreases significantly, but not as much as that of the others, making it the top performer in this scenario, similarly to what was observed by Heilman et al. (2008). Overall, the regression approaches seem to generalize better than the classification approach, as they are less impacted by the higher uncertainty of the predicted class distributions.

Table 5 shows the confusion matrices of the best run by each approach on the test set. We can see that there are two main reasons for the higher performance achieved by the pure regression approach. On the one hand, it achieves a higher recall for level A2, leading to improved accuracy. On the other hand, the larger difference in terms of adjacent accuracy is mainly justified by the several examples of level A1 that it classifies as A2, while the other approaches classify them as B1.

The examples of the A levels that are misclassified as B1 correspond to short texts that are exclusive to a type of exercise that only appears in the model exams of the A levels. The classification approach classifies all of those examples as B1 because, even though they are significantly longer, the shortest texts on the training data are of that level. On the other hand, the regression approaches are able to accurately classify a reduced set of those examples, with an average accuracy of 2.31% by the softmax regression approach and 13.08% by the pure regression approach.

If the problematic short texts are not considered, the average accuracy of the three approaches is much closer: 72.63%, 73.68%, and 72.11% for classification, regression, and softmax regression, respectively. In this case, pure regression still significantly outperforms the others in terms of adjacent accuracy, achieving a perfect score, while classification achieves 89.47% and softmax regression 91.05%. However, it is important to remember that the classification of texts by readability level is a task that is subjective and difficult even for

Class.		Predicted				
		A1	A2	B1	B2	C1
Actual	A1	3	0	5	0	0
	A2	2	2	8	0	0
	B1	1	0	4	0	0
	B2	0	0	0	3	0
	C1	0	0	1	0	3

Reg.		Predicted				
		A1	A2	B1	B2	C1
Actual	A1	2	6	0	0	0
	A2	1	6	5	0	0
	B1	0	1	4	0	0
	B2	0	0	0	3	0
	C1	0	0	0	2	2

Smax		Predicted				
Reg.		A1	A2	B1	B2	C1
Actual	A1	3	1	4	0	0
	A2	2	3	7	0	0
	B1	1	0	4	0	0
	B2	0	0	0	3	0
	C1	0	0	1	1	2

Table 5: Confusion matrices of the best runs of the different approaches on the test set.

humans (Branco et al., 2014a; Curto, 2014).

5 Conclusion

In this paper, we have addressed the automatic assessment of text readability level in European Portuguese as a regression task in an attempt to leverage the ordinal nature of CEFR levels. Our experiments in a cross-validation scenario revealed that by computing the weighted average of the class probability distributions predicted by a fine-tuned version of the Albertina PT-PT model instead of simply selecting the level with the highest probability, we can obtain more robust predictions that lead to improved adjacent accuracy while maintaining similar accuracy and macro F_1 score. Furthermore, the regression approaches, and especially a model fine-tuned specifically for the regression task, generalize better to unseen kinds of textual data.

Considering the difficulty in obtaining additional annotated data for training more robust models, as future work, it is important to assess how large language models like ChatGPT (OpenAI, 2023) and LLaMa (Touvron et al., 2023) perform on this task in zero or few-shot scenario. Furthermore, considering the subjectivity of readability level assessment and its potential applications, it is important to make an effort towards the development of interpretable models that provide insight regarding the proposed classifications.

Acknowledgments

This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) (Reference: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020) and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837).

We would like to thank Camões, I.P. - Language Services Directorate for granting us access to the texts used in their exams and allowing us to use them to train our models.

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability Assessment for Text Simplification. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the Opportunities and Risks of Foundation Models. *Computing Research Repository*, arXiv:2108.07258.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014a. Assessing Automatic Text Classification for Interactive Language Learning. In *Proceedings of the International Conference on Information Society (i-Society)*, pages 70–78.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014b. Rolling out Text Categorization for Language Learning Assessment Supported by Language Technology. In *Proceedings of the International Conference on the Computational Processing of the Portuguese Language (PROPOR)*, pages 256–261.

Miriam Cha, Youngjune Gwon, and H.T. Kung. 2017. Language Modeling by Clustering with Word Embeddings for Text Readability Assessment. In *Proceedings of the Conference on Information and Knowledge Management (CIKM)*, pages 2003–2006.

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

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. 2014. Classificador de Textos para o Ensino de Português como Segunda Língua. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa.

Pedro Curto, Nuno Mamede, and Jorge Baptista. 2015. Automatic Text Difficulty Classifier. In *Proceedings of the International Conference on Computer Supported Education (CSEDU)*, volume 1, pages 36–44.

Direção de Serviços de Língua e Cultura, Camões, I.P. 2017. *Referencial Camões Português Língua Estrangeira*. Camões, Instituto da Cooperação e da Língua I.P., Lisboa.

William H. DuBay. 2004. *The Principles of Readability*. Impact Information.

Luciana Forti, Giuliana Grego Bolli, Filippo Santarelli, Valentino Santucci, and Stefania Spina. 2020. MALT-IT2: A New Resource to Measure Text Difficulty in Light of CEFR Levels for Italian L2 Learning. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 7204–7211.

Thomas François and Cédrick Fairon. 2012. An "AI Readability" Formula for French as a Foreign Language. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 466–477.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics*, 28(2):337–407.

Maria José Grosso, António Soares, Fernanda de Sousa, and José Pascoal. 2011. QuaREPE: Quadro de Referência para o Ensino Português no Estrangeiro – Documento Orientador. Technical report, Direção-Geral da Educação (DGE).

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the Workshop on Innovative use of NLP for Building Educational Applications (BEA)*, pages 71–79.

Nikolay Karpov, Julia Baranova, and Fedor Vitugin. 2014. Single-sentence Readability Prediction in Russian. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST)*, pages 91–100.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Institute for Simulation and Training, University of Central Florida.

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 and Evaluation*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository*, arXiv:1907.11692.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

Luís Marujo, José Lopes, Nuno Mamede, Isabel Trancoso, Juan Pino, Maxine Eskenazi, Jorge Baptista, and Céu Viana. 2009. Porting REAP to European Portuguese. In *Proceedings of the International Workshop on Speech and Language Technology in Education (SLaTE)*, pages 69–72.

Peter McCullagh. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. Overview of the GermEval 2022 Shared Task on Text Complexity Assessment of German Text. In *Proceedings of the GermEval Workshop on Text Complexity Assessment of German Text*, pages 1–9.

Farah Nadeem and Mari Ostendorf. 2018. Estimating Linguistic Complexity for Science Texts. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55.

OpenAI. 2023. ChatGPT. https://chat.openai.com/.

Ildikó Pilán and Elena Volodina. 2018. Investigating the Importance of Linguistic Complexity Features Across Different Datasets Related to Language Learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. OpenAI Blog.

Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. 2024. Automatic Text Readability Assessment in European Portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. *Computing Research Repository*, arXiv:2305.06721.

Rodrigo Santos, João Rodrigues, António Branco, and Rui Vaz. 2021. Neural Text Categorization with Transformers for Learning Portuguese as a Second Language. In *Proceedings of the Portuguese Conference on Artificial Intelligence (EPIA)*, pages 715–726.

Carolina Evaristo Scarton and Sandra Maria Aluísio. 2010. Análise da Inteligibilidade de Textos via Ferramentas de Processamento de Língua Natural: Adaptando as Métricas do Coh-Metrix para o Português. *Linguamática*, 2(1):45–61.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Computing Research Repository*, arXiv:2302.13971.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin Yancey, and Thomas François. 2022. FABRA: French Aggregator-Based Readability Assessment Toolkit. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1217–1233.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 38–45.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Kevin Yancey, Alice Pintard, and Thomas Francois. 2021. Investigating Readability of French as a Foreign Language with Deep Learning and Cognitive and Pedagogical Features. *Lingue e Linguaggio*, 20(2):229–258.