Cross-Language Text Readability Assessment: Leveraging Multilingual Models for Improved Performance in CEFR-Level Classification

Eugénio Ribeiro
INESC-ID Lisboa, Portugal
Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal

Jorge Baptista
INESC-ID Lisboa, Portugal
Faculdade de Ciências Humanas e Sociais, Universidade do Algarve, Faro, Portugal

Abstract

The automatic assessment of text readability and the classification of texts by levels is essential for language education and language industries that rely on effective communication. This study explores cross-language automatic readability level classification using the levels defined by the Common European Framework of Reference for Languages (CEFR). We investigate the potential of using data in one language to improve classification performance in different languages and, thus, optimize the utilization of the limited labeled resources available for each language. We rely on a pre-trained multilingual Transformer-based language model, by fine-tuning it on annotated data in one language or in a combination of languages, and then assessing its ability to generalize even to unseen languages. In an additional scenario, we further fine-tune the models on data in the target language, to assess whether the models trained on data in different languages can capture generic information regarding text readability and then be further specialized to capture the specific characteristics of the target language. Our experiments covering the English, Dutch, and German languages revealed that direct generalization to unseen languages is challenging. However, when paired with data in the target language, multilingual data can be leveraged to capture cross-language aspects of text readability, leading to more robust and better-performing models.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Applied computing → Education

Keywords and phrases Readability, Text Complexity, CEFR, Multilinguality

Digital Object Identifier 10.4230/OASIcs.SLATE.2026.3

Funding This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: `https://doi.org/10.54499/UID/50021/2025`), UID/PRR/50021/2025 (DOI: `https://doi.org/10.54499/UID/PRR/50021/2025`), and UID/04466/2025 (DOI: `https://doi.org/10.54499/UID/04466/2025`); and by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: `https://doi.org/10.3030/101094837`).

1 Introduction

The Common European Framework of Reference for Languages (CEFR) [9] provides a widely recognized framework for classifying language proficiency levels, ranging from A1 (beginner) to C2 (proficient). Additionally, the CEFR has been used to assess the proficiency of individuals as learners of different languages. However, it can also be used from a readability perspective, shifting the focus to understanding the proficiency required to read specific pieces of text. Identifying the readability 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 education, assessing the readability level allows educators

and curriculum designers to match texts to the learners' abilities, fostering effective language development and personalized learning experiences. Moreover, outside the education domain, readability level classification finds applications in different sectors. For instance, in the banking industry, presenting financial information and policies at an appropriate readability level ensures that clients can understand terms and conditions, enabling well-informed decision-making. Similarly, in healthcare, accessible and understandable medical instructions, consent forms, and patient information materials are crucial for individuals with varying levels of language proficiency. Furthermore, legal information, government communications, user manuals, and many others, benefit from accurately assessing the readability level of written materials, facilitating effective communication, content transparency, and general comprehension. Therefore, by exploring the readability perspective of the CEFR, we can make a significant contribution to enhancing the understanding of text comprehension factors and their far-reaching implications for both education and language-related industries, seeking to convey information to learners or clients in a manner that is clear, concise, and easily understood.

Determining the readability level of texts presents its own set of challenges, particularly when working with languages that have limited annotated resources. Annotating large amounts of text data with CEFR levels is a labor-intensive and time-consuming task, often requiring expert domain knowledge. Consequently, the scarcity of labeled data in multiple languages hinders the development of robust and accurate models for automatic readability level classification. To address this challenge, we look into the problem from a cross-language perspective, which offers a promising avenue for leveraging the limited annotated resources available for each language.

Starting with a pre-trained multilingual Transformer-based language model [48], we explore its fine-tuning for CEFR-level classification on data in three different languages – English, Dutch, and German – and assess the ability of the obtained fine-tuned models to generalize across languages. The aim of this study is not to achieve the highest possible performance on text readability level classification, but rather to assess whether data in one language can help improve the classification performance in different languages. This point is highly relevant. If data from various languages can be utilized to enhance performance in a specific target language, it suggests the existence of shared linguistic patterns and structures that exhibit a similar perceived difficulty level across languages. Moreover, this indicates that the limited availability of labeled resources can be effectively combined and leveraged in future model development for automatic readability level classification.

In the following sections, we will proceed as follows: Section 2 will provide an overview of related work on automatic text readability level classification. Then, in Section 3, we will describe our experimental setup, including the datasets used and the methodology employed for training and evaluating various models. Next, Section 4 will present and discuss the findings from our experiments. Finally, in Section 5, we will summarize the contributions of this study, discuss its limitations, and suggest directions for future research in this area.

2 Related Work

Readability assessment is a problem that has been widely explored over the years. Traditionally, the problem is addressed by creating readability formulas or indexes based on statistical information and/or domain knowledge [13, 10]. Among these, the most widely used are the Flesch Reading Ease index and the Flesch-Kincaid Grade Level [23].

However, considering the developments in Machine Learning (ML), and especially in Natural Language Processing (NLP), part of the research on automatic readability assessment shifted towards following the trends in the NLP area [33]. Early approaches (and many recent ones for low-resource languages) relied on handcrafted features, such as word frequency, sentence length, and syntactic complexity, combined with traditional machine learning algorithms, such as decision trees and support vector machines (e.g. [4, 16, 22, 11, 34, 15, 3]). Then, Deep Learning (DL) approaches relying on pre-trained word embeddings, such as those generated by Word2Vec [30], emerged (e.g. [8, 32, 14]). More recently, research in the area shifted towards the fine-tuning of pre-trained Transformer-based language models, such as BERT [12], GPT [35], and RoBERTa [27] (e.g. [44, 53, 28, 31, 40]). Still, some studies revealed that model fine-tuning and handcrafted features are actually complementary for the task and that the best performance can be achieved by combining both in hybrid models (e.g. [25, 26, 50, 1]). Prompt-based approaches relying on the intrinsic knowledge of Large Language Models (LLMs), such as those of the GPT [7], LLaMA [47], Gemma [18], and EuroLLM [29] families, have also been explored in both zero-shot and few-shot settings. However, they tend to underperform in comparison to fine-tuned approaches (e.g. [39, 20, 2]).

The previous studies on automatic text readability level assessment cover several languages, such as English (e.g. [52, 8, 32, 14, 28]), French (e.g. [16, 17, 53, 49, 19, 2]), Chinese (e.g. [46]), German (e.g. [31]), Brazilian (e.g. [4, 24]) and European Portuguese (e.g. [5, 11, 44, 24, 40, 42, 3]), Italian (e.g. [15, 45]), Russian (e.g. [22, 38]), Swedish (e.g. [21, 34]), Spanish (e.g. [1, 36]), Galician (e.g. [37]), and Slovenian (e.g. [28]). In this context, it is important to mention the UniversalCEFR initiative [20], which aggregated CEFR-annotated data in 13 languages and explored several automatic assessment approaches. However, most of the languages are still under-resourced and, although several languages are covered in the studies, to the best of our knowledge, cross-language models have not been explored for the task. Thus, our study fills a gap that may define a path for future research in this area.

3 Experimental Setup

In this section, we describe our experimental setup. As we intend to explore cross-language text readability assessment, we need datasets in multiple languages. These are described in Section 3.1. Then, in Sections 3.2 and 3.3, we describe our training and evaluation procedures. Finally, in Section 3.4, we provide implementation details that enable the future reproduction of our experiments.

3.1 Datasets

In our experiments, we use the datasets collected in the context of the CEFR Labelling and Assessment Services (CEFRSERV) project [6]. These cover three languages – English (EN), Dutch (NL), and German (DE). We opted for using these datasets as they were collected in the context of the same project and cover languages of the same family – west-germanic –, which provides an interesting starting point for this cross-lingual analysis. Still, as the datasets were collected by different teams, some differences can be found in the annotation and the format in which the annotated texts are provided. In order to improve reproducibility, we describe these differences below and discuss how we address them in our experiments.

Firstly, the German dataset is annotated with the six CEFR levels, between A1 and C2, while the English and Dutch datasets also include intermediate levels, such as A1+ or B1–B2. For consistency and simplicity, we do not consider intermediate levels in this study and the texts annotated with those levels are considered as belonging to the lower level. The only exception is the A1- level, which only occurs in the English dataset and is considered as A1.

Table 1 Level distribution of texts across datasets.

Language	A1	A2	B1	B2	C1	C2	Total
English	40	154	216	174	97	69	750
Dutch	31	155	520	410	73	7	1,196
German	31	90	90	115	88	95	509
Total	102	399	826	699	258	171	2,455

Also, the annotation of each text in the English and German datasets refers to an aggregate of multiple annotators, while the Dutch dataset includes all the annotations for each text. For consistency, we aggregate the latter as an average.

Finally, while the German and Dutch datasets are provided in a single file featuring all the text-annotation pairs, the English dataset is provided as multiple files. Besides, the dataset description refers to approximately 1,200 texts, but there are only 750 files. This means that some files contain more than one text with the same CEFR level. However, there is no straightforward way to automatically identify these files nor to separate the texts. Thus, we considered each file to be a single text.

Table 1 shows the distribution of the texts across languages and CEFR levels. We can see that there is a bias towards the B levels. This was also observed on many of the datasets used in previous research on automatic text readability level classification (e.g. [11, 38, 34, 15]). On the German dataset, this bias is not as pronounced as in the other two languages. However, this is also the smallest dataset, with only 509 annotated texts. Overall, there are 2,455 annotated texts, most of which (48.72%) are in Dutch.

3.2 Training Methodology

The aim of this study is to assess whether data in one language can help improve automatic text readability classification performance in different languages and, thus, whether scarce labeled resources can be used more efficiently.

To do so, we use as a base a pre-trained multilingual Transformer-based language model and then train multiple models by fine-tuning it for CEFR-level classification on data in each of the three languages and on each combination of those languages. The performance of these models can then be assessed across languages.

In our experiments, we explored two training approaches: single-step and two-step. In single-step training, each model is obtained by fine-tuning the base model on the training data of the corresponding language or combination of languages. This allows the assessment of the direct generalization ability of the models across languages. In two-step training, each of the models generated using single-step training is further fine-tuned on the training data for each of the target languages. In this scenario, the first training step can be seen as an adaptation of the base pre-trained model to the text readability classification task in general, while the second step is used to further refine the models so as to capture the particular characteristics of the target language in the context of the task.

Furthermore, in line with other studies on automatic CEFR-level classification (e.g., [11]), we train models for both six levels (A1–C2) and three levels (A–C). This approach enables us to assess whether a classification with lower granularity can leverage the information available from data in other languages more effectively.

Table 2 Accuracy (Acc.) and adjacent accuracy (Adj. Acc.) results (%) using single-step training for six CEFR levels.

	English		Dutch		German	
	Acc.	**Adj. Acc.**	**Acc.**	**Adj. Acc.**	**Acc.**	**Adj. Acc.**
EN	**62.89±3.01**	97.11±1.39	43.06±0.64	94.31±0.24	34.64±4.84	80.06±0.57
NL	23.78±2.70	74.22±4.53	64.58±1.10	99.17±0.42	25.16±5.40	59.80±7.40
DE	26.22±5.43	65.11±10.21	18.19±4.45	54.17±5.12	**67.32±1.50**	87.91±4.42
EN+NL	62.67±3.06	96.67±1.34	64.86±0.86	**99.86±0.24**	27.78±4.53	68.30±1.50
EN+DE	**62.89±2.14**	**97.33±0.00**	26.53±1.58	79.44±0.24	63.72±1.70	89.22±2.60
NL+DE	24.00±0.67	70.22±1.02	63.75±2.21	99.44±0.24	65.69±3.93	86.93±1.14
All	59.33±3.06	97.11±0.77	**65.00±3.25**	99.44±0.24	65.36±2.47	**89.55±1.13**

3.3 Evaluation Methodology

Regarding evaluation metrics, we adopt *accuracy* and *adjacent accuracy*, which are consistently used in previous studies on automatic CEFR readability level classification (e.g., [16, 11, 53, 41, 1]). Accuracy evaluates the precise identification of a text's readability level, while adjacent accuracy also considers neighboring levels, offering further insight into the identification of texts slightly easier or harder than the assigned level. As the coarse-grained setting considers only three levels, we rely solely on *accuracy* for assessing the performance in that context. Both metrics are reported as percentages.

Due to the absence of a standard partition in the datasets, we conduct a random 80/20 train/test split. To enhance robustness and mitigate the impact of randomness, we perform three independent experimental runs, each with different random seeds for the splitting process. The evaluation metrics are reported as both the average and standard deviation across these runs.

3.4 Implementation Details

As our objective is a comparative analysis rather than achieving the utmost performance on the task, we use the multilingual cased version of DistilBERT [43] as a base model. This choice enables faster experiments and reduces resource consumption. Each model undergoes 50 epochs of training and the weights from the best epoch are loaded in the end. For model training and evaluation, we rely on the functionality offered by HuggingFace's Transformers library [51].

4 Results

In this investigation, we adopt a two-step training approach that builds on the models obtained through single-step training. Thus, to begin, in Section 4.1, we present and discuss the results obtained using the single-step training approach. Next, in Section 4.2, we focus on the results achieved with the two-step approach. Finally, in Section 4.3, we produce an overall discussion, draw further conclusions, and address certain limitations encountered during the study.

Table 3 Accuracy (%) results using single-step training for three CEFR levels.

	English	Dutch	German
EN	79.11±1.92	72.50±3.63	52.94±1.96
NL	45.11±13.04	84.72±1.58	42.16±8.03
DE	48.44±1.39	49.03±12.31	**81.04±5.74**
EN+NL	**81.11±2.04**	**85.28±0.48**	46.08±1.96
EN+DE	79.56±0.77	61.53±4.09	79.08±2.04
NL+DE	55.78±7.73	84.58±1.91	76.47±5.88
All	78.22±2.04	85.00±2.21	78.43±3.53

4.1 Single-Step Training

The accuracy and adjacent accuracy results obtained through single-step training, considering the six CEFR levels, are presented in Table 2. Notably, the models demonstrate limited generalization capabilities on unseen languages. The highest achieved accuracy is 43.06% when applying the model trained on English texts to Dutch. This performance stands 20.69 percentage points below the lowest-performing model trained on Dutch data and 12.27 percentage points lower than the least performing model trained on the target language in general. These discrepancies may arise due to various factors, including distinct annotation guidelines for different languages, variations in nature of the text, or potential misalignment of cross-language representations in the base model. Nonetheless, these findings suggest a significant language dependence concerning the perceived difficulty level of texts.

Upon examining the model performance for those trained in the target language, we observe that the models trained in multiple languages demonstrate competitiveness with those trained exclusively on the target language. Remarkably, the top-performing model for Dutch results from training on the three languages, albeit with a higher degree of variability. This finding indicates the presence of cross-language aspects influencing text readability, and data from different languages can contribute to training more robust models. This observation is further supported by the adjacent accuracy results, which show that models trained on data from multiple languages consistently achieve the highest performance across all languages. An intriguing outcome pertains to the adjacent accuracy for German, which, despite having the highest accuracy, exhibits the lowest adjacent accuracy. This suggests certain challenges in capturing aspects of German texts' difficulty or potential issues with the annotation process, as the errors tend to be more severe.

Finally, as displayed in Table 3, when considering only three proficiency levels, the overall accuracy increases, exceeding 80% for every language, as anticipated. Notably, the model trained on data from both English and Dutch achieves the highest performance for both languages, while the model trained solely on German data remains the top performer for German. These results align with the high adjacent accuracy observed when considering the six CEFR levels for the models trained on English and Dutch data. Regarding adjacent accuracy, as mentioned in Section 3.3, the results for the three-level scenario are omitted, as the metric lacks informativeness in this context. In fact, every model trained on data from the target language achieves 100% adjacent accuracy, while the remaining models achieve at least 95%.

Table 4 Accuracy (Acc.) and adjacent accuracy (Adj. Acc.) results (%) using two-step training for six CEFR levels.

	English		Dutch		German	
	Acc.	**Adj. Acc.**	**Acc.**	**Adj. Acc.**	**Acc.**	**Adj. Acc.**
EN	62.89±3.01	97.11±1.39	64.58±0.83	99.17±0.00	65.03±2.26	86.93±2.26
NL	64.67±2.67	96.89±0.38	64.58±1.10	99.17±0.42	65.69±0.98	86.60±2.26
DE	61.33±2.40	95.78±2.34	65.97±1.34	99.31±0.24	67.32±1.50	87.91±4.42
EN+NL	63.56±3.67	97.11±1.02	65.83±1.10	**99.72±0.24**	62.75±4.27	**90.52±3.00**
EN+DE	**66.44±1.54**	**97.78±0.77**	**66.39±0.96**	99.58±0.42	**68.95±2.04**	89.87±3.00
NL+DE	64.67±0.67	96.22±1.02	65.28±0.48	**99.72±0.48**	66.99±2.04	87.58±1.13
All	62.00±3.06	97.56±0.77	65.28±2.51	99.31±0.24	66.99±2.83	88.89±1.50

4.2 Two-Step Training

Table 4 displays the accuracy and adjacent accuracy results achieved through two-step training, considering the six CEFR levels. All models demonstrate competitive performance with the model trained solely on the target language in terms of accuracy. However, when analyzing the results for Dutch and German, we observe distinct behavior. For Dutch, any pre-training using data from another language leads to improved performance. Fine-tuning the model trained on English data yields the same average accuracy but with slightly improved stability. This phenomenon can be attributed to Dutch's intermediary position between English and German. Consequently, it can leverage the generic difficulty information captured in the examples of both languages. Moreover, Dutch benefits from having the highest number of examples, allowing the second fine-tuning step to overturn biases related to specific aspects of the other languages. In contrast, German, arguably the most grammatically complex of the three languages and having the fewest examples, demonstrates improved performance only when fine-tuned with the model trained on English and German data. The limited contribution of data from other languages is insufficient to address German's more complex grammatical aspects, making it challenging to overturn previous biases. Overall, the optimal performance for every language is achieved by fine-tuning the model trained on English and German data. This suggests that, due to the higher number of Dutch examples, the background models trained on Dutch data fail to appropriately attribute significance to specific aspects of the other languages.

In terms of adjacent accuracy, for English and Dutch, the patterns mirror those observed in terms of accuracy, with the highest performance differing by less than half a percentage point from that seen in the single-step training scenario. This minor variation was expected since the performance in both cases is close to perfect. In contrast, for German, additional models surpass the performance of the one trained solely on German data. Notably, the highest-performing model achieves over 90% adjacent accuracy and is trained on English and Dutch data in the first step. This finding further supports the claim that some cross-language aspects influence text readability assessment.

Finally, in Table 5, we observe that, akin to when considering the six CEFR levels, when only three levels are considered, the overall performance also improves compared to the single-step training scenario. Notably, for German, the highest performance is achieved by fine-tuning the model trained on English and German data in the first step, similarly to what was observed in the six-level scenario. However, for English, the highest performance is now attained by fine-tuning the model trained on English and Dutch data in the first step,

Table 5 Accuracy (%) results using two-step training for three CEFR levels.

	English	Dutch	German
EN	79.11±1.92	85.28±0.64	80.07±2.26
NL	80.89±2.04	84.72±1.58	80.39±1.70
DE	80.22±3.79	84.58±1.67	81.04±5.74
EN+NL	**81.55±2.04**	85.97±0.64	81.05±3.00
EN+DE	80.45±2.04	**86.25±1.10**	**83.01±1.50**
NL+DE	80.00±2.40	85.00±1.25	78.76±3.44
All	80.22±4.29	85.56±1.68	81.37±1.70

which is consistent with the findings from the single-step training scenario. This observation suggests that, from a broader perspective, certain aspects contributing to text difficulty are comparable in both languages, and that the prediction for the English language benefits from the larger number of examples available in Dutch. Once again, the adjacent accuracy results are omitted because every model achieves 100% performance, rendering the metric uninformative in this context.

4.3 Overall Discussion

Based on the results presented in the preceding sections, it is evident that models trained in one language cannot be directly applied to predict the readability level of texts in other languages. Nevertheless, data from diverse languages can be utilized to create background models that are able to capture certain cross-language aspects of text readability, which can subsequently be adapted to specific languages. However, indiscriminately adding training data from different languages does not appear to be the most effective approach, though clear patterns of influence between related languages were not identified. Therefore, crafting specialized models for each language remains essential to achieve optimal performance on the task.

Nonetheless, this line of research presents several promising avenues to pursue, which can yield more robust conclusions and help address the identified issues. Specifically, the datasets used in this study are limited in both size and the number of languages considered. Expanding the data to encompass additional sources could potentially lead to the development of a more generic model of text readability, serving as a universal starting point for training language-specific models. Moreover, the choice of the base model, particularly the data used for its pre-training, may contribute to imbalances in performance across different languages. Therefore, it is crucial to explore alternative base models, especially considering the recent advancements in the field of large language models. Taking a more language-specific approach, involving proficient speakers of each language to examine the prediction errors of the models becomes essential. Such analysis might reveal specific patterns that the current models miss and highlight potential discrepancies in annotation across languages.

To conclude this discussion, a remark concerning variability deserves attention. Despite the apparent high variability observed for some models, it is essential to acknowledge that (except for some low-performance models) this variability stems from the differing difficulty levels of the test sets in each split. Most importantly, every claim made in the results' discussion is not solely based on average results but also on the consistency observed across multiple runs.

5 Conclusion

In this study, we conducted an investigation into automatic readability level classification from a cross-language perspective, focusing on the CEFR levels. Our primary objectives were twofold: first, to evaluate the potential of data from one language to enhance classification performance in different languages, and, second, to explore the effectiveness of utilizing scarce labeled resources more efficiently. To achieve these goals, we employed a pre-trained multilingual Transformer-based language model, fine-tuned it using annotated data from one language or a combination of languages, and then examined its generalization capacity even to unseen languages. Additionally, we performed further fine-tuning of the models using data from the target language to assess whether models trained on data in different languages were able to capture generic information related to text readability and could subsequently be further specialized to capture the specific characteristics of the target language. Through experiments covering the English, Dutch, and German languages, we arrived at insightful conclusions. It became apparent that direct generalization to unseen languages presents considerable challenges. However, when paired with data from the target language, multilingual data can be leveraged and proves beneficial to the task, as it allows the models to capture cross-language aspects of text readability. This leads to the development of more robust and improved-performing models overall.

In conclusion, our study has yielded compelling findings that open up a new avenue for training automatic readability assessment models. However, several limitations warrant consideration for future research. Chiefly, the datasets used in the study are constrained in size and encompass only a limited number of languages of the same family. Consequently, further investigations should evaluate the potential of incorporating additional data to establish more robust conclusions concerning cross-language interactions and to develop a more generic model of text readability, which can subsequently be fine-tuned for specific languages. For instance, it would be interesting to explore the 13 languages currently covered by the UniversalCEFR initiative and assess whether the patterns observed in this study are replicated in other families, as well as in cross-family settings.

While our study prioritized exploring the cross-language aspects of automatic readability level classification rather than aiming for the highest performance possible, the findings lay the groundwork for future research, opening for comparison with higher-performing base models. It is essential to acknowledge that multilingual models may exhibit reduced performance compared to their single-language counterparts. Consequently, a critical consideration is whether the performance gains achieved by leveraging data from different languages can sufficiently compensate for the lower base performance of multilingual models. A potential path to address this concern is to leverage data from different languages by translating it into the target language instead of training a multilingual model directly on it. However, text complexity aspects can be lost or increased in translation. As such, future investigations should take these factors into account for a more comprehensive assessment of model performance and a more accurate estimation of the impact of multilingual data on the task of readability level classification.

Another important aspect to consider is that the current top performing approaches to text readability assessment rely on hybrid models that combine model fine-tuning with the use of handcrafted features. While this has not been thoroughly explored, the importance of specific features for readability assessment is expected to show higher variability across languages. Thus, the use of multilingual data in the context of hybrid models poses additional challenges that require further research.

References

1 Wafa Aissa, Raquel Amaro, David Antunes, Thibault Bañeras-Roux, Jorge Baptista, Alejandro Catala, Luís Correia, Thomas François, Marcos Garcia, Mario Izquierdo-Álvarez, Nuno Mamede, Vasco Martins, Miguel Neves, Eugénio Ribeiro, Sandra Rodriguez, and Elodie Vanzereven. The iRead4Skills Intelligent Complexity Analyzer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 73–84, 2025. doi:10.18653/v1/2025.emnlp-demos.6.

2 Wafa Aissa, Thibault Bañeras-Roux, Elodie Vanzeveren, Lingyun Gao, Rodrigo Wilkens, and Thomas François. Assessing French Readability for Adults with Low Literacy: A Global and Local Perspective. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 20528–20550, 2025. doi:10.18653/v1/2025.emnlp-main.1036.

3 Soroosh Akef, Amália Mendes, Detmar Meurers, and Patrick Rebuschat. Investigating the Generalizability of Portuguese Readability Assessment Models Trained Using Linguistic Complexity Features. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 332–341, 2024. URL: https://aclanthology.org/2024.propor-1.34/.

4 Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. Readability Assessment for Text Simplification. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, 2010. URL: https://aclanthology.org/W10-1001.

5 António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 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, 2014. doi:10.1007/978-3-319-09761-9_29.

6 Mark Breuker. CEFR Labelling and Assessment Services. In Georg Rehm, editor, *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 277–282. Springer International Publishing, 2023. doi:10.1007/978-3-031-17258-8_16.

7 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

8 Miriam Cha, Youngjune Gwon, and HT Kung. 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, 2017. doi:10.1145/3132847.3133104.

9 Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001. URL: https://rm.coe.int/1680459f97.

10 Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas. *Discourse Processes*, 54(5-6):340–359, 2017. doi:10.1080/0163853x.2017.1296264.

11 Pedro Curto, Nuno Mamede, and Jorge Baptista. Automatic Text Difficulty Classifier. In *Proceedings of the International Conference on Computer Supported Education (CSEDU)*, volume 1, pages 36–44, 2015. doi:10.5220/0005428300360044.

12 Jacob Devlin, Ming-Wei Chang, Lee Kenton, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, volume 1, pages 4171–4186, 2019. doi:10.18653/v1/N19-1423.

13 William H. DuBay. *The Principles of Readability*. Impact Information, 2004. URL: https://files.eric.ed.gov/fulltext/ED490073.pdf.

14 Anna Filighera, Tim Steuer, and Christoph Rensing. Automatic Text Difficulty Estimation Using Embeddings and Neural Networks. In *Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL)*, pages 335–348, 2019. doi:10.1007/978-3-030-29736-7_25.

15 Luciana Forti, Giuliana Grego Bolli, Filippo Santarelli, Valentino Santucci, and Stefania Spina. 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, 2020. URL: https://aclanthology.org/2020.lrec-1.890.

16 Thomas François and Cédrick Fairon. 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, 2012. URL: https://aclanthology.org/D12-1043.

17 Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. AMesure: A Web Platform to Assist the Clear Writing of Administrative Texts. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (AACL-IJCNLP): System Demonstrations*, pages 1–7, 2020. doi:10.18653/V1/2020.AACL-DEMO.1.

18 Gemma Team. Gemma: Open Models based on Gemini Research and Technology. *Computing Research Repository*, arXiv:2403.08295, 2024. doi:10.48550/arXiv.2403.08295.

19 Nicolas Hernandez, Nabil Oulbaz, and Tristan Faine. Open Corpora and Toolkit for Assessing Text Readability in French. In *Proceedings of the Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 54–61, 2022. URL: https://aclanthology.org/2022.readi-1.8.

20 Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. UniversalCEFR: Enabling Open Multilingual Research on Language Proficiency Assessment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9703–9755, 2025. doi:10.18653/v1/2025.emnlp-main.491.

21 Simon Jönsson, Evelina Rennes, Johan Falkenjack, and Arne Jönsson. A Component Based Approach to Measuring Text Complexity. In *Proceedings of the Swedish Language Technology Conference (SLTC)*, pages 58–61, 2018. URL: https://sltc2018.blogs.dsv.su.se/files/2019/11/Final-compilation-of-abstracts-SLTC-2018-Nov-19.pdf.

22 Nikolay Karpov, Julia Baranova, and Fedor Vitugin. Single-sentence Readability Prediction in Russian. In *Proceedings of the International Conference on Analysis of Images, Social Networks and Texts (AIST)*, pages 91–100, 2014. doi:10.1007/978-3-319-12580-0_9.

23 J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 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, 1975. doi:10.21236/ada006655.

24 Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. NILC-Metrix: Assessing the Complexity of Written and Spoken Language in Brazilian Portuguese. *Computing Research Repository*, arXiv:2201.03445, 2021. doi:10.48550/arXiv.2201.03445.

25 Bruce W. Lee, Yoo Sung Jang, and Jason Lee. Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10669–10686, 2021. doi:10.18653/v1/2021.emnlp-main.834.

26 Fengkai Liu and John SY Lee. Hybrid models for sentence readability assessment. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 448–454, 2023. doi:10.18653/v1/2023.bea-1.37.

27 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository*, arXiv:1907.11692, 2019. doi:10.48550/arXiv.1907.11692.

28 Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179, 2021. doi:10.1162/coli_a_00398.

29 Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. EuroLLM: Multilingual Language Models for Europe. *Procedia Computer Science*, 255:53–62, 2025. doi:10.1016/j.procs.2025.02.260.

30 Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NeurIPS*, pages 3111–3119, 2013. URL: https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

31 Salar Mohtaj, Babak Naderi, and Sebastian Möller. 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, 2022. URL: https://aclanthology.org/2022.germeval-1.1.

32 Farah Nadeem and Mari Ostendorf. Estimating Linguistic Complexity for Science Texts. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, 2018. doi:10.18653/v1/W18-0505.

33 Kai North, Marcos Zampieri, and Matthew Shardlow. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42, 2023. doi:10.1145/3557885.

34 Ildikó Pilán and Elena Volodina. 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, 2018. URL: https://aclanthology.org/W18-4606.

35 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9, 2019. URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

36 Sandra Rodríguez Rey, André Bernárdez Braña, and Marcos Garcia. Exploring Linguistic Features in a New Readability Corpus for Spanish. *Procesamiento del lenguaje natural*, 74:221–239, 2025. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6677.

37 Sandra Rodríguez Rey and Marcos Garcia. Clasificación Automática de Textos por Niveis de Lecturabilidade: Recursos e Modelos para o Galego. *Linguamática*, 17(2):33–56, 2025. doi:10.21814/lm.17.2.488.

38 Robert Reynolds. Insights from Russian Second Language Readability Classification: Complexity-Dependent Training Requirements, and Feature Evaluation of Multiple Categories. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, 2016. doi:10.18653/v1/W16-0534.

39 Eugénio Ribeiro, David Antunes, Nuno Mamede, and Jorge Baptista. Exploring Few-Shot Approaches to Automatic Text Complexity Assessment in European Portuguese. *Journal of the Brazilian Computer Society*, 31(1):690–710, 2025. doi:10.5753/jbcs.2025.5820.

40 Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. Automatic Text Readability Assessment in European Portuguese. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 97–107, 2024. URL: https://aclanthology.org/2024.propor-1.10/.

41 Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. Avaliação Automática do Nível de Complexidade de Textos em Português Europeu. *Linguamática*, 16(2):121–145, 2024. doi:10.21814/lm.16.2.449.

42 Eugénio Ribeiro, Nuno Mamede, and Jorge Baptista. Text Readability Assessment in European Portuguese: A Comparison of Classification and Regression Approaches. In *Proceedings of the International Conference on Computational Processing of Portuguese (PROPOR)*, pages 551–557, 2024. URL: https://aclanthology.org/2024.propor-1.59/.

43 Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *Computing Research Repository*, arXiv:1910.01108, 2019. doi:10.48550/arXiv.1910.01108.

44 Rodrigo Santos, João Rodrigues, António Branco, and Rui Vaz. 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, 2021. doi:10.1007/978-3-030-86230-5_56.

45 Valentino Santucci, Filippo Santarelli, Luciana Forti, and Stefania Spina. Automatic Classification of Text Complexity. *Applied Sciences*, 10(20):7285, 2020. doi:10.3390/app10207285.

46 Yao Ting Sung, Wei Chun Lin, Scott Benjamin Dyson, Kuo En Chang, and Yu Chia Chen. Leveling L2 Texts through Readability: Combining Multilevel Linguistic Features with the CEFR. *The Modern Language Journal*, 99(2):371–391, 2015. doi:10.1111/modl.12213.

47 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. LLaMA: Open and Efficient Foundation Language Models. *Computing Research Repository*, arXiv:2302.13971, 2023. doi:10.48550/arXiv.2302.13971.

48 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

49 Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin Yancey, and Thomas François. FABRA: French Aggregator-Based Readability Assessment Toolkit. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1217–1233, 2022. URL: https://aclanthology.org/2022.lrec-1.130.

50 Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. Exploring Hybrid Approaches to Readability: Experiments on the Complementarity between Linguistic Features and Transformers. In *Findings of the Association for Computational Linguistics: EACL*, pages 2316–2331, 2024. doi:10.18653/v1/2024.findings-eacl.153.

51 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. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *Computing Research Repository*, arXiv:1910.03771, 2019. doi:10.48550/arXiv.1910.03771.

52 Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. Text Readability Assessment for Second Language Learners. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, 2016. doi:10.18653/v1/W16-0502.

53 Kevin Yancey, Alice Pintard, and Thomas Francois. Investigating Readability of French as a Foreign Language with Deep Learning and Cognitive and Pedagogical Features. *Lingue e Linguaggio*, 20(2):229–258, 2021. doi:10.1418/102814.