The iRead4Skills Intelligent Complexity Analyzer

Wafa Aissa[1,*], Raquel Amaro[2], David Antunes[3], Thibault Bañeras-Roux[1], Jorge Baptista[3,6], Alejandro Catala[5], Luís Correia[4], Thomas François[1], Marcos Garcia[5], Mario Izquierdo-Álvarez[5], Nuno Mamede[3,7], Vasco Martins[4], Miguel Neves[4], Eugénio Ribeiro[3,8], Sandra Rodriguez Rey[5] and **Elodie Vanzeveren[1]**
[1]UCLouvain, [2]NOVA Lisboa, [3]INESC-ID Lisboa, [4]MindShaker,
[5]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela
[6]Faculdade de Ciências Humanas e Sociais, Universidade do Algarve
[7]Instituto Superior Técnico, Universidade de Lisboa
[8]Instituto Universitário de Lisboa (ISCTE-IUL)
Correspondence: raquelamaro@fcsh.unl.pt, marcos.garcia.gonzalez@usc.gal, jbaptis@ualg.pt, thomas.francois@uclouvain.be

Abstract

We present the iRead4Skills Intelligent Complexity Analyzer[1], an open-access platform specifically designed to assist educators and content developers in addressing the needs of low-literacy adults by analyzing and diagnosing text complexity. This multilingual system integrates a range of Natural Language Processing (NLP) components to assess input texts along multiple levels of granularity and linguistic dimensions in Portuguese, Spanish, and French. It assigns four tailored difficulty levels using state-of-the-art models, and introduces four diagnostic yardsticks—textual structure, lexicon, syntax, and semantics—offering users actionable feedback on specific dimensions of textual complexity. Each component of the system is supported by experiments comparing alternative models on manually annotated data.

1 Introduction

Reading skills are foundational to career advancement, lifelong learning, and informed participation in society. Both UNESCO[2] and OECD (2013a) define literacy as the process of engaging and understanding written texts in order to achieve one's goals and develop one's potential. Along the same lines, high literacy levels have been proven to be beneficial to other individual skills, such as problem-solving, digital literacy, and communication, showcasing its transversal aspect (OECD, 2013b). However, a minority of adults continue to experience difficulties with basic literacy skills—a minority that is nonetheless worth close attention due to the profound impact low-literacy can have on individuals' everyday lives. In fact, 20% of EU adult population exhibits low-literacy and numeracy skills (European Association for the Education of Adults, 2021). These challenges affect access to employment, healthcare, and civic participation, ultimately limiting individuals' full engagement in society (European Commission, 2021).

Reflecting the view that literacy is a continuous, lifelong process, Adult Learning (AL) programs include the improvement of literacy through vocational training (ANQEP, 2021). The training and development of reading skills rely on adequate written materials, which, unlike those available for children, are not easily accessible for adults. Consequently, selecting or designing texts suitable for low-literacy adults represents a major challenge and a highly time-consuming task for AL teachers across all fields (from the sciences to history). This underscores the need for tools that can assess and adapt text readability. With this in mind, the iRead4Skills project[3] aims to contribute to the solution of this problem by devising an automatic system that evaluates texts' complexity and suggests appropriate readings according to the user's reading skills. To do this, it is important to understand what text complexity entails for low-literacy adult native speakers and how it can be identified in written materials.

Building on a foundational assessment combining text complexity analysis, ethical governance, multilingual collaboration, and learner needs identification, the project defined four structured levels of text complexity (Very Easy, Easy, Plain +Complex) characterized by a large set of descriptors which were validated by a panel of experts. Language experts then collected and annotated multilingual datasets in Portuguese (PT), Spanish (SP), and French (FR) to train models for automatic

[*]Authors listed in alphabetical order.
[1]Accessible at demo02.iread4skills.com
[2]https://uis.unesco.org/node/3079547
[3]https://iread4skills.com/

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 73–84
November 4-9, 2025

readability assessment. These resources support the development of an Intelligent Complexity Analyzer (ICA) that integrates psycholinguistic insights with NLP techniques to predict text difficulty and provide interpretable feedback on readability.

This paper presents a user-friendly web interface for an ICA and introduces a novel classification of four diagnostic yardsticks (textual structure, lexicon, syntax, and semantics) supported by experimental analysis on newly annotated data. Designed to promote equitable access to written content, our approach embeds sustainability, dissemination, and impact evaluation to ensure lasting benefits for low-literacy adult education and digital inclusion.

Apart from this introduction, Section 2 discusses related work, Section 3 outlines the current state of the platform and its functional components, Section 4 reports on the evaluation, and Section 5 concludes with a discussion of future directions.

2 Related Work

Our study falls within the scope of readability and text complexity analysis research.

Readability: Readability research has traditionally focused on general adult readers or schoolchildren, with early formulas such as Flesch (Flesch, 1948) and Gunning Fog (Gunning, 1952) still widely used today. While some studies have applied these metrics to specialized domains like medical texts (Mcinnes and Haglund, 2011) or contracts (Arbel, 2024), they remain poorly adapted to the needs of low-literate adults. Only a few efforts have developed readability models for this population, including work in Portuguese (Aluisio et al., 2010), Italian (Dell'Orletta et al., 2011), Spanish (Saggion et al., 2015), and German (Weiss et al., 2018), with no such models yet available for French. Recent advances in readability modeling have leveraged deep learning (Nadeem and Ostendorf, 2018; Azpiazu and Pera, 2019; Martinc et al., 2021) and hybrid approaches combining linguistic features with neural architectures (Lee et al., 2021; Liu and Lee, 2023; Wilkens et al., 2024). Recently, the use of generative Large Language Models (LLMs) has also been explored (Jamet et al., 2024; Ribeiro et al., 2024, 2025; Aissa et al., 2025). In this work, we adapt readability assessment models to texts used in low-literacy adult education and evaluate their effectiveness in this context.

Text complexity analysis platforms:

A number of open access tools have been developed for assessing text complexity, though they are not tailored specifically to the needs of low-literacy adult readers (Table 1). These tools vary not only in the granularity of the readability analysis they provide to users—ranging from the document level to the segment level (sentence or paragraph) and the word level (single words or multi-word expressions)—but also in user accessibility, providing diagnoses through a graphical user interface (GUI) or via code, and in language support, which influences their suitability for multilingual contexts. In Portuguese, the CLAVIS system (Curto et al., 2015) utilizes NLP tools to extract 52 linguistic features from texts, which are then classified according to their readability levels. ALT (Moreno et al., 2022) is a web-based readability analysis tool that adapts traditional formulas to the Portuguese language and provides users with a composite readability score, lexical statistics, and visual feedback. LX-Proficiency also provides a Flesch analysis and proficiency classification for European Portuguese (Branco et al., 2014). The Coh-Metrix tool (Graesser et al., 2004) has been adapted to both Brazilian Portuguese (Scarton and Aluísio, 2010), and Spanish (Quispesaravia et al., 2016), providing cohesion and readability indices, and achieving strong classification performance between simple and complex texts. Besides, MultiAzterTest (Bengoetxea and Gonzalez-Dios, 2021) offers a multilingual approach—covering Spanish, English, and Basque—based on more than 125 linguistic features. In French, the closest system is AMesure (François et al., 2020), which focuses on standard readers of administrative texts and relies primarily on pre-deep learning techniques. More broadly, a recent system by Trokhymovych et al. (2024) introduces a multilingual readability model for Wikipedia articles in 14 languages. While their system achieves competitive results across

Platform	Granularity	GUI	PT	SP	FR
CLAVIS	D	✓	✓	✗	✗
ALT	D, S, W	✓	✓	✗	✗
LX-Proficiency	D, S, W	✓	✓	✗	✗
Coh-Metrix-Port	D, S, W	✓	✓	✗	✗
Coh-Metrix-Spa	D, S, W	✗	✗	✓	✗
MultiAzterTest	D, S, W	✓	✗	✓	✗
AMesure	D, S, W	✓	✗	✗	✓
Wikipedia system	D	✗	✓	✓	✓
iRead4Skills (ours)	D, S, W	✓	✓	✓	✓

Table 1: Overview of open-access complexity analysis platforms. D: document-level, S: Segment-level, W: word-level.

languages, its applicability to low-literacy adults is limited: it was trained exclusively on Wikipedia texts rather than on materials targeting low-literate readers, and it outputs a single difficulty score that cannot be directly mapped to our multi-level difficulty scale.

More generally, while existing platforms offer valuable insights into text complexity, they are typically designed for standard readers or language learners and do not address the specific needs of low-literacy adults, most rely either on traditional readability formulas or rich linguistic metrics. Our work contributes to this emerging area by providing an ICA tool specifically tailored to low-literacy adults, integrating linguistic features and NLP within an accessible, multilingual interface.

3 The iRead4Skills Platform

This section presents the main features of our platform, whose interface is shown in Figure 1.

3.1 Platform overview

The iReadSkills platform is a web-based tool designed to support the evaluation of text readability for adult learners with low-literacy skills. It enables educators and content developers to assess and visualize the linguistic complexity of texts in PT, SP, and FR. Users can input any text of their choice for analysis, making the platform adaptable to diverse educational contexts. It combines a Graphical User Interface (GUI) with NLP-based models, providing both an overall readability score and interpretable feedback on syntactic, lexical, semantic, and structural difficulty (top and bottom of the right-hand column in Figure 1, respectively). In addition, it offers word-level complexity annotations to help identify specific sources of reading difficulty in the text. Unlike general-purpose readability tools, iReadSkills is tailored to the needs of adult education, with an emphasis on accessibility, transparency, and language-sensitive analysis.

3.2 Functional Components

We present the interface components, focusing on the complexity classifier and yardstick evaluation, along with a brief description of the annotation module and highlighting of difficult phenomena.

3.2.1 Readability classification

The text difficulty classification component implements readability models designed to assess input texts in reference to the difficulty scale defined in the project (Monteiro et al., 2023):

Very Easy: Texts that are fully or almost fully understood by everyone, including people with very low schooling (i.e., that did not finish the primary school, ca. 6th year) and almost no reading experience. It roughly corresponds to the A1 level in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001).

Easy: Texts that are fully or almost fully understood by people with low schooling (i.e., that completed the primary school but no more than the 9th year) and have poor reading experience. It roughly corresponds to CEFR A2 level.

Plain: Texts that are understood the first time they are read by people that completed the 9th year and have a functional-to-average reading experience. It roughly corresponds to CEFR B1 level.

+Complex: Texts that exceed the characteristics of the previous categories. They often present significant challenges to people with low literacy.

Separate readability models are built for each target language (PT, SP, and FR), leveraging our collected datasets annotated for difficulty by experienced teacher annotators (see Section 4.1; additional details are in Appendix B.1). We explore multiple modeling strategies (see Section 4.2), including traditional Machine Learning (ML) approaches with engineered linguistic features, Deep Learning (DL) methods based on Transformer architectures, and hybrid models that integrate both feature-based and representation-based techniques.

3.2.2 Readability yardsticks

To provide users with a more fine-grained and actionable understanding of text complexity, our platform also evaluates texts along four key linguistic dimensions (see below). These yardsticks serve as diagnostic indicators, offering insight into specific aspects that contribute to the overall readability of a text. Leveraging the rich information returned by other components (see Section 3.2.3), we define the following four main yardsticks, which provide a quick and interpretable overview of the complexity of a given text:

Textual Structure: This yardstick captures surface-level properties and includes descriptive measures of the document, e.g., sentence, word length and word count.

Lexicon: Measures lexical complexity through word usage patterns, and is computed using features such as lexical complexity, lexical frequency,

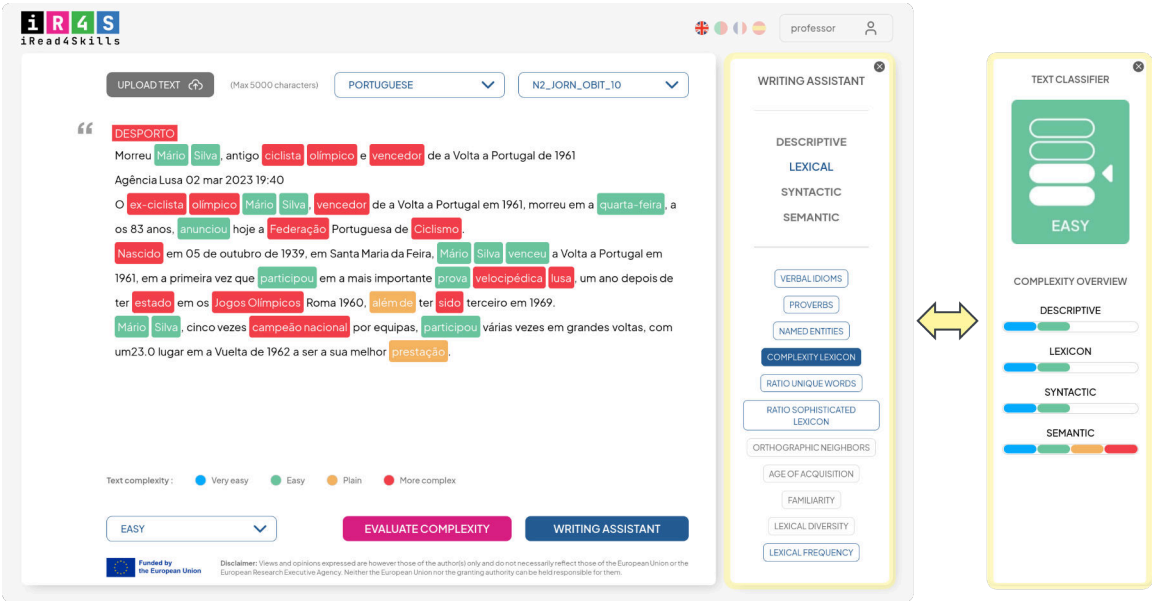

Figure 1: Interface of the iRead4Skills platform. The *Descriptive* yardstick shown is 'Textual Structure' (as defined in the paper). When using the writing assistant for in-text highlighting, the overall assessment view is replaced by the assistant's options. Both views are shown here together but correspond to different screens. The writing assistant will only be available to authorized users.

lexical diversity or age of acquisition.

Syntax: Reflects syntactic complexity, particularly the use of advanced constructions. It relies on parse depth as well as ratios of different syntactic phenomena such as subordination, coordination, auxiliary verbs, passive constructions, modifiers of noun and prepositional phrases.

Semantics: Accounts for semantic difficulty such as ambiguity or abstractness, and is measured using information from polysemy ratios and proportions of concrete or abstract words.

To compute a complexity score for each category, we compare two methods (threshold- and distribution-based) against a baseline (Section 4.3). The threshold-based approach uses predefined thresholds to assign complexity levels, with feature selection grounded in theoretical assumptions, while the distribution-based method models feature distributions using Gaussian Mixture Models (GMMs) (Reynolds, 2009) for each yardstick and complexity level.

3.2.3 Difficult phenomena annotation

This last component focuses both on identifying linguistic features that contribute to reading difficulties for low-literacy adults and highlighting specific phenomena relative to these features.

Annotator: To do this, we develop a heuristic annotator that leverages parse-based NLP models (Qi et al., 2020; Kitaev et al., 2019) to extract syntactic and semantic phenomena, alongside a curated set of readability features relevant to text complexity (see Table 5 in Appendix C). The annotation follows the BIO format (Ramshaw and Marcus, 1995), enabling structured labeling of multi-token and nested phenomena such as subordination and auxiliary verb chains. Scalar features, like sentence complexity or length, receive numerical annotations. Features are computed at multiple levels (document, sentence, and token) using metrics such as largest instance size, token coverage, and frequency relative to main verbs.

Highlighting: Following the text analysis, this module highlights linguistic phenomena identified as difficult based on the user's proficiency level. We computed statistical thresholds for each feature and complexity level using the annotated corpus (see Section 4.1) and an Interquartile Range (IQR)-based threshold ($Q3 + 1.5 \times (Q3 - Q1)$) to detect complex instances. Thresholds were applied per feature and complexity level, using upper bounds for most features (e.g., parse depth, word syllables) and lower bounds where lower values indicate complexity (e.g., familiarity, $Q1 - 1.5 \times (Q3 - Q1)$). Highlighting is triggered at document, sentence, or token level depending on the feature type. We are currently assessing, in collaboration with teachers, the most effective way to highlight difficult phenomena in text.

Language	Method	Accuracy	Adj. Acc.	Macro F_1
Portuguese	Feature-based (Random Forest)	53.89	91.36	52.17
	Fine-tuning (Albertina 100M)	51.87	92.99	50.79
	Hybrid (Prob. Distrib.)	**55.30**	**93.93**	**53.80**
Spanish	Feature-based (Decision Tree)	39.88	83.70	32.91
	Fine-tuning (BETO)	**47.47**	88.80	39.89
	Hybrid (Prob. Distrib.)	46.62	**89.88**	**43.24**
French	Feature-based (Random Forest)	62.77	98.05	47.78
	Fine-tuning (CamemBERT)	64.04	98.71	**60.36**
	Hybrid (Prob. Distrib.)	**67.32**	**99.14**	56.26

Table 2: Results of the top models in each category for the classification task across the three languages.

4 System Evaluation

We describe the experiments of classifying textual complexity and each yardstick, reporting the best results on each case.

4.1 Datasets

We evaluate the system using the iRead4Skills Dataset (Pintard et al., 2024), which includes written texts tailored to low-literacy adult readers, collected from a wide range of genres and text types. These documents were carefully selected to reflect the diversity of real-world reading materials and a subset of it was classified and validated by experts using the project's levels of reading complexity. Further details on the annotation process and dataset characteristics are in Appendix B.1.

Additionally, and as a contribution of this paper, we selected 60 documents in each target language, which were annotated by experts with the corresponding yardstick levels to evaluate this fine-grained level of readability classification. Appendix B.2 describes the annotation protocol.

4.2 Text Difficulty classification

As mentioned, we compared ML algorithms, Transformer-based DL models, and hybrid approaches for readability classification, with the best-performing models integrated into our tool.

Classical Machine Learning: To train ML models we use features from four large categories based on Wilkens et al. (2022): descriptive (length, counts and structure), lexical, syntactic, and discourse features; and also word-embeddings. We aggregate paragraph-, sentence-, and word-level features using statistical metrics like mean, max, and length. In addition to the complete set of features, we also experiment with feature selection methods to identify the most impactful features. Appendix A.1 describes the feature selection methods and the ML algorithms evaluated for this task.

Deep Learning models: We adopted the standard approach for fine-tuning transformer models[4] on our datasets. The list of models evaluated for each language can ben found in Appendix A.1.

Hybrid approaches: We implemented the soft-label architecture from Lee et al. (2021). We used our best performing fine-tuned DL models to predict classification probabilities for each difficulty class based on the input text. These probabilities were subsequently concatenated with the machine learning features to train hybrid models.

Results: To measure performance, we used common metrics in readability classification, namely Accuracy, Adjacent Accuracy, and Macro F_1. Accuracy reflects exact correctness, Adjacent Accuracy captures near-miss predictions between adjacent levels, and Macro F1 provides balanced evaluation across all classes. Table 2 presents the top-performing models across categories and languages. Overall, the hybrid approach achieves the highest performance. An exception occurs in French, where the fine-tuned model alone yields a higher F_1 score, albeit with a slightly lower precision.

Concerning generative LLMs, which are increasingly being used for readability assessment due to their flexibility and broad language capabilities (Jamet et al., 2024; Ribeiro et al., 2024), our experiments on the French and Portuguese datasets specifically designed for low-literacy adults revealed that, although LLMs show promising results, they still struggle to generalize effectively to

[4]We have also evaluated variants, such as partial fine-tuning, with similar or lower results.

texts outside their pretraining distribution and to adjust to readability scales tailored for this population. As demonstrated in our recent work (Aissa et al., 2025; Ribeiro et al., 2025), few-shot prompting improves LLM performance compared to zero-shot prompting, especially through a careful selection of examples. However, models specifically trained and fine-tuned for this readability task still outperform LLMs. For French, the top-performing LLM, DeepSeek-70B (DeepSeek-AI, 2025), achieved a macro-F1 score of 48.95%, which is below the 60.36% macro-F1 score obtained by our fine-tuned CamemBERT model. Similarly, for Portuguese, the top-performing LLM, GPT-4o mini (OpenAI et al., 2024), achieved 45.64% accuracy, 92.99% adjacent accuracy, and 45.44% macro-F1 score, all of which are below the results achieved by the hybrid model. Nonetheless, LLM capabilities for this task remain underexplored, offering an interesting avenue for future research.

4.3 Yardsticks evaluation

The targeted yardstick framework enables a more fine-grained understanding of the linguistic factors contributing to textual complexity. However, a key limitation to this analysis is that standard datasets are annotated only at the document level, and obtaining large-scale fine-grained annotations at each dimension would require significant human effort.

To overcome this challenge, we hypothesize that the correlation between each yardstick's complexity level and the corresponding features can be partially inferred from document-level annotations. Specifically, we partition the overall feature space into subspaces corresponding to each yardstick, as detailed in Section 3.2.2, and analyze each dimension separately assuming each yardstick inherits the document-level complexity annotation. Under this assumption, our aim is to identify correlations between feature distributions and complexity within each yardstick, thereby enabling dimension-specific analyses without the need for additional annotation. To validate this assumption, we rely on the documents annotated at the yardstick level (see Section 4.1 and Appendix B.2). This data allows us to assess whether the models (trained on document-level labels) can effectively capture variability in complexity across dimensions, particularly in cases where the global annotation does not fully reflect the complexity of individual yardsticks.

Methods: To assign a complexity level to each yardstick, we compare threshold- and GMM-based modeling approaches. This comparison allows us to evaluate the effectiveness of simple statistical thresholds against a probabilistic clustering technique for accurately classifying yardstick levels. Appendix A.2 describes each of the methods. As a baseline, we adopt a model in which each yardstick simply inherits the document-level complexity label, predicted by the best-performing document-level classifier available for each target language. This configuration approximates the best-case scenario for global complexity estimation in a real-world setting. If our yardstick-specific models outperform this baseline, it would indicate that the same annotated data can be used to extract more informative, dimension-specific insights, thus representing an information gain in the analysis of textual complexity.

Results: To compare the above methods, performance is evaluated using accuracy and Macro F_1; and QWK, which accounts for the ordinal nature of complexity levels and quantifies agreement between human annotations and model predictions. Results of the yardstick's classification are reported in Table 3.

For Portuguese, the threshold-based approach outperforms the baseline in predicting structure and semantic complexities, as indicated by higher accuracy, QWK scores and Macro F_1. When it comes to lexical complexity, the metrics indicate that our approach gets more labels correct, but its mistakes are more severe (e.g., predicting far-off classes). Lastly, it underperforms in syntactic complexity, likely due to difficulties in capturing the specific cues required for this dimension—an issue that corresponds to the low inter-annotator agreement (Krippendorff's α), suggesting that syntax is inherently more challenging to assess and model. Overall, performance improvements tend to occur in dimensions where annotators showed higher agreement, indicating a correlation between annotation reliability and model learnability. For Spanish, GMM-based methods consistently outperform the baseline across all dimensions, with particularly substantial gains in structure and more modest improvements in syntax, supporting the advantage of modeling each yardstick in its dedicated feature subspace. Similarly, in French, GMMs yield better results than the baseline in structure and lexical dimensions, while syntax shows mixed outcomes: the GMMs achieved higher QWK, indicating better ordinal consistency, but lower accuracy and F_1. In the semantic dimension, GMMs underperform,

	Method	Structure			Lexical			Syntax			Semantics		
		Acc	QWK	F_1	Acc	QWK	F_1	Acc	QWK	F_1	Acc	QWK	F_1
PT	baseline	0.350	0.445	0.299	0.383	**0.359**	**0.319**	**0.367**	**0.284**	**0.281**	0.417	0.408	0.382
	best (TR-based)	**0.633**	**0.702**	**0.564**	**0.467**	0.241	0.199	0.167	0.180	0.186	**0.433**	**0.424**	**0.388**
SP	baseline	0.350	0.243	0.316	0.350	0.421	0.342	0.366	0.262	0.260	-	-	-
	best (GMMs)	**0.433**	**0.616**	**0.442**	**0.416**	**0.545**	**0.407**	**0.400**	**0.337**	**0.350**	-	-	-
FR	baseline	0.467	0.530	0.363	0.400	0.568	0.363	**0.567**	0.655	**0.523**	**0.450**	**0.581**	**0.392**
	best (GMMs)	**0.550**	**0.646**	**0.476**	**0.567**	**0.661**	**0.458**	0.450	**0.686**	0.446	0.400	0.396	0.261

Table 3: Results per language and yardstick. Evaluation metrics include Accuracy (Acc), Quadratic Weighted Kappa (QWK), and Macro F_1. Semantic yardstick is omitted for Spanish due to temporary lack of external resources.

likely due to the yardstick's reliance solely on the concreteness ratio, which appears insufficient to capture semantic complexity.

5 Conclusions and future work

This paper presented the iRead4Skills ICA, a new multilingual tool specifically designed to assist educators and content developers in addressing the needs of low-literacy adults, by analyzing and diagnosing text complexity. The components build upon previous empirical findings from readability and text complexity analysis combined with current NLP techniques, and offers diagnosis through a text complexity analyzer. Crucially, the tool introduces four novel diagnostic yardsticks (textual structure, lexicon, syntax, and semantics) that provide actionable feedback on specific dimensions of textual complexity. Both the classification of texts and yardsticks were developed and validated through experimentation with annotated data. The platform is model-agnostic, allowing seamless updates with improved models over time. Currently, we are enhancing these models and conducting evaluations with teachers to identify the most effective and useful methods for highlighting difficult phenomena, aiming to support reading skill development and ensure lasting benefits for low-literacy adult education. Concretely, two main use cases will be examined: (1) evaluating complexity assessment and difficult phenomena annotation to help teachers select and design materials suitable for low-literate adults; and (2) evaluating the usefulness of readability predictions for directly recommending appropriate readings to low-literate adults. Moreover, further research is currently being conducted to extend the system beyond complexity analysis and provide writing assistance. For this, LLMs appear to be a natural choice, both for offering simplified text alternatives and for providing more general but targeted suggestions. Exploring these directions will be the focus of our next development stage. Future updates and the most recent version of the tool can be accessed at: https://iread4skills.com/.

Limitations

First, it is worth mentioning that some components of the platform are still in early stages of development; however, full functionality is expected in the coming months. While global complexity assessment is already quite mature, the yardsticks still require refinement, although they are already adding value and outperforming the baseline in most cases. Secondly, the user interface is being iteratively improved through user testing, with further changes expected to enhance usability and user experience. We are exploring the most effective way to highlight text annotations that indicate difficult phenomena to the user, so this aspect is expected to be improved. Thus, although initial experiments have shown promising results, more extensive testing is currently underway to validate and enhance the performance of the different modules. In this regard, in certain cases the data supporting the experiments is limited or it has relatively few annotations, which may affect model robustness. Finally, as commonly found in readability research, inter-annotator agreement sometimes remains low, reflecting the inherent challenges of subjective text complexity assessments.

Ethical considerations

The research activities necessary to build the machine learning models and develop the iRead4Skills webtool presented in this demo paper have followed well-established ethics high standards in

the academia. Ethics approval has been obtained from the University ethics committee, based on conformance of the protocols and user-centered research methods involved when involving stakeholders/users for either requirements elicitation and design feedback. The project started by establishing solid coordination and governance mechanisms to ensure ethical, legal, and data protection compliance, alongside effective collaboration among academic, technical, and field partners working across three target languages: Portuguese, Spanish, and French. A comprehensive needs analysis was then conducted, combining a review of existing research with newly designed surveys to identify the reading challenges faced by adult learners in vocational and lifelong education contexts.

Given the ethics guidelines for trustworthy Artificial Intelligence (AI) by the High-Level Expert Group on AI for the European Commission, some important rationale and decisions are set behind the construction of our models and tool design. Specifically, given the unique target users (low-literacy adults and their teachers), our corpora was specifically compiled and curated by linguists and experienced teachers who work with such target population. Texts are made available for research purposes. We rely on open source frameworks when necessary to ensure reliability and transparency. The interfaces are intended as supporting tools, allowing users to inspect and interpret the outputs, supporting human agency and decision making. Non deep learning models are also considered given the lower carbon footprint (as empirically measured in the constructions of our models in our internal reports). Thus, the administrator can choose the models available for each language. Finally, once the ML models are fully ethically tested in context, they will be released openly along with a statement on the scope and terms of use to ensure other developers/researchers understand their intended use, limitations and how to use them responsibly.

Acknowledgments

This work was supported by the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837), by the Galician Government (ED431F 2021/01, ED431G 2023/04, and ED431B 2025/16), by a Ramón y Cajal grant (RYC2019-028473-I), and by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT) (References: UIDB/50021/2020, DOI: 10.54499/UIDB/50021/2020 and UID/03213).

References

Wafa Aissa, Thibault Bañeras-Roux, Elodie Vanzeveren, Lingyun Gao, Rodrigo Wilkens, and Thomas François. 2025. Assessing french readability for adults with low literacy: A global and local perspective. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

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.

Raquel Amaro and Thomas François. 2023. iRead4Skills - Annotation Schema. Published October 30, 2023.

Raquel Amaro, Ricardo Monteiro, Thomas François, and Justine Nagant de Deuxchaisnes. 2024. iRead4Skills - Data Set 2: Annotated Corpora Report. Published November 30, 2024.

ANQEP. 2021. Referencial de competências-chave de educação e formação de adultos – nível básico. Technical report, Agência Nacional para a Qualificação e o Ensino Profissional, I.P. (ANQEP), Portugal.

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. CamemBERT 2.0: A Smarter French Language Model Aged to Perfection. *arXiv preprint arXiv:2411.08868*.

Yonathan A Arbel. 2024. The readability of contracts: Big data analysis. *Journal of Empirical Legal Studies*, 21(4):927–978.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Kepa Bengoetxea and Itziar Gonzalez-Dios. 2021. Multiaztertest: a multilingual analyzer on multiple levels of language for readability assessment.

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.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *Proc. PML4DC at ICLR 2020*.

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

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.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Ding and Hanchuan Peng. 2003. Minimum redundancy feature selection from microarray gene expression data. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pages 523–528.

European Association for the Education of Adults. 2021. Basic skills development in selected European countries: State of play report. Report, March 2021. `https://eaea.org/wp-content/uploads/2021/03/BLUESS-state-of-play_report.pdf`.

European Commission. 2021. Upskilling pathways – new opportunities for adults. Access link.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Thomas François, Adeline Müller, Eva Rolin, and Magali Norré. 2020. Amesure: a web platform to assist the clear writing of administrative texts. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 1–7.

Arthur Graesser, Danielle McNamara, Max Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 36:193–202.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

Henri Jamet, Maxime Manderlier, Yash Raj Shrestha, and Michalis Vlachos. 2024. Evaluation and simplification of text difficulty using LLMs in the context of recommending texts in French to facilitate language learning. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 987–992.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fengkai Liu and John SY Lee. 2023. Hybrid models for sentence readability assessment. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 448–454.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Nicholas Mcinnes and Bo JA Haglund. 2011. Readability of online health information: implications for health literacy. *Informatics for health and social care*, 36(4):173–189.

Ricardo Monteiro, Raquel Amaro, Susana Correia, Alice Pintard, Roser Gauchola, Michell Moutinho, and Xavier Blanco Escoda. 2023. iRead4Skills Complexity Levels. Project Deliverable D3.1, iRead4Skills.

Gleice Carvalho de Lima Moreno, Marco PM de Souza, Nelson Hein, and Adriana Kroenke Hein. 2022. ALT: um software para análise de legibilidade de textos em Língua Portuguesa. *arXiv preprint arXiv:2203.12135*.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 45–55.

OECD. 2013a. *The Survey of Adult Skills Reader's Companion*. OECD Publishing, Paris.

OECD. 2013b. *Technical Report of the Survey of Adult Skills (PIAAC)*. OECD Publishing, Paris.

OpenAI et al. 2024. GPT-4o System Card. *Computing Research Repository*, arXiv:2410.21276.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in Python. *the Journal of Machine Learning Research*, 12:2825–2830.

Alice Pintard, Thomas François, Justine Nagant de Deuxchaisnes, Sílvia Barbosa, Maria Leonor Reis, Michell Moutinho, Ricardo Monteiro, Raquel Amaro, Susana Correia, Sandra Rodríguez Rey, Marcos Garcia González, Keran Mu, and Xavier Blanco Escoda. 2024. iRead4Skills Dataset 1: Corpora by Complexity Level for FR, PT and SP. Project Deliverable D3.2, iRead4Skills.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Andre Quispesaravia, Walter Perez, Marco Sobrevilla Cabezudo, and Fernando Alva-Manchego. 2016. Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4694–4698, Portorož, Slovenia. European Language Resources Association (ELRA).

Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Proceedings of the Workshop on Very Large Corpora (VLC)*, pages 82–94.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Douglas Reynolds. 2009. *Gaussian Mixture Models*, pages 659–663. Springer US, Boston, MA.

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

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

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-*. In *Proceedings of the Portuguese Conference on Artificial Intelligence (EPIA)*, page 441–453.

Sandra Rodríguez Rey, André Bernárdez Braña, and Marcos Garcia. 2025. Exploring Linguistic Features in a New Readability Corpus for Spanish. *Procesamiento del Lenguaje Natural*, 74(0):221–239.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Trans. Access. Comput.*, 6(4).

Rodrigo Santos, João Rodrigues, Luís Gomes, João Ricardo Silva, António Branco, Henrique Lopes Cardoso, Tomás Freitas Osório, and Bernardo Leite. 2024. Fostering the Ecosystem of Open Neural Encoders for Portuguese with Albertina PT* Family. In *Proceedings of the Annual Meeting of the Special Interest Group on Under-resourced Languages (SIGUL)*, pages 105–114.

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.

Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. An Open Multilingual System for Scoring Readability of Wikipedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6296–6311, Bangkok, Thailand. Association for Computational Linguistics.

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.

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.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

A Technical details

A.1 Text classification

Experimental Data For Portuguese and Spanish we selected the best performing model based on test set evaluation with model selection guided by performance on validation data. For French, due to the considerably smaller size of the annotated dataset (see Table 4 in Section B.1), experiments were conducted using stratified 5-fold cross-validation with a split of 70% for training, 10% for validation, and 20% for testing.

Feature selection methods: To train ML models for readability classification we experimented, in addition to the complete set of features, with feature selection methods to identify the most impactful features, aiming to improve model accuracy and generalization for readability prediction. Specifically, we explored (i) the *minimum Redundancy Maximum Relevance* feature selection algorithm (Ding and Peng, 2003), (ii) Spearman ρ correlation between the features and the levels of the training data, (iii) k-Best, by selecting the most relevant k features according to the ANOVA F value, and (iv) the Recursive Feature Elimination approach.

ML algorithms: Regarding algorithms, we evaluate a selection of algorithms from various families, including Random Forest, SVM, Decision Trees, kNN, and Gradient Boosting. We tuned hyperparameters using grid search in each case.

Transformer models: For French, we employed variants of CamemBERT, including CamemBERT-base[5] (Martin et al., 2020), CamemBERT-v2-base[6] (Antoun et al., 2024), and a Sentence-BERT[7] (Reimers and Gurevych, 2019) model based on CamemBERT-base.

For Portuguese, we relied on the Albertina PT-PT[8] family of foundation models (Rodrigues et al., 2023; Santos et al., 2024).

For Spanish, we used a few variants of the BERT model, notably multilingual BERT[9] (Devlin et al., 2019), BERT for Spanish[10] (Cañete et al., 2020) and RoBERTa by MarIA[11] (Fandiño et al., 2022).

A.2 Yardstick analysis

Threshold-based: We use the statistical thresholds computed for different complexity levels. For each yardstick, we select relevant linguistic features and compare their values to predefined thresholds to estimate the complexity level. These levels are converted into numerical scores and aggregated to yield a final overall score for the yardstick. We identified the most informative features and aggregation methods through experiments with various theoretically motivated combinations, evaluated against a baseline. This approach balances linguistic theory with data-driven insights from annotated corpora.

Gaussian Mixture Models (GMMs): We adopt a probabilistic modeling approach by fitting a GMM using the Scikit-learn[12] library (Pedregosa et al., 2011) for each complexity level across the different yardsticks. Each model is trained under the assumption that the global document-level label can be projected onto all dimensions. Additionally, to accommodate documents of varying lengths, features at both the sentence and the token level are aggregated at the document level by using statistical descriptors such as mean, standard deviation, skewness, etc., resulting in fixed-size feature vectors suitable for modeling.

B Data and Annotation Protocol

B.1 Datasets

Table 4 presents the size of each dataset by language. For Portuguese and Spanish (Rodríguez Rey et al., 2025), all documents (2,933 and 2,563 respectively) were annotated by linguists, with a

[5] https://huggingface.co/almanach/camembert-base

[6] https://huggingface.co/almanach/camembertv2-base

[7] https://huggingface.co/dangvantuan/sentence-camembert-large

[8] https://huggingface.co/collections/PORTULAN/albertina-66a39cf7e2460605f3f1a9c2

[9] https://huggingface.co/google-bert/bert-base-multilingual-cased

[10] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[11] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

[12] https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

subset (428 and 406) further validated by additional language experts and used as the test set (for Spanish, only documents with majority votes were selected). For French, only the validated subset (461 documents) was annotated. For detailed information on the annotation process, including the annotation guidelines and decisions used for classifying text complexity, we refer to the iRead4Skills annotation schema (Amaro and François, 2023). The annotated corpora report (Amaro et al., 2024) provides concrete examples of labeled texts, inter-annotator agreement statistics, and coverage across multiple languages, illustrating the application of the schema in practice.

	Corpus	Validated	Train	Dev	Test
PT	2933	428	1986	519	428
SP	2563	406	1765	442	356
FR	2200	461	322	46	93

Table 4: Dataset splits by language: number of texts per corpus and dataset.

B.2 Yardsticks Annotation Protocol

All three languages used comparable descriptors for each complexity level as defined in the iRead4Skills Complexity Levels (Monteiro et al., 2023).

Portuguese: We selected 60 texts: 40 chosen at random (10 from each level) and 20 (5) selected based on the highest predicted probability of belonging to the target level. Three professional linguists, each with extensive experience in language proficiency assessment and language teaching, were then asked to read the texts carefully and assign a difficulty level to each yardstick according to descriptors developed within the project. Krippendorff's alpha (ordinal) was used to calculate inter-annotator agreement for each yardstick. The agreement scores were as follows: Structure: 0.650, Lexical: 0.495, Syntax: 0.320, Semantics: 0.397.

Spanish: We randomly selected 60 texts (15 for each level), and asked a professional linguist to read it carefully, assess it, and assign a level for each yardstick and text, based on the detailed descriptors of the project. Overall, the annotation process was smooth, with the annotator encountering no significant difficulties in assigning yardstick levels. For future work, we plan to engage additional annotators to facilitate the assessment of inter-annotator agreement.

French: We randomly selected 60 texts (15 per level) and hired two linguistics students, who were compensated, to perform the annotation. Prior to the annotation process, annotators were briefed on the project objectives and task specifications, and were provided with an annotation guide. A follow-up meeting was held after completion of the first set to address questions and identify potential difficulties; no major issues were reported. Additionally, a member of the project team served as a third annotator. The time required to annotate each set ranged from fifteen minutes to one hour. We used Krippendorff's alpha (ordinal) to assess inter-annotator agreement. The agreement scores were as follows: Structure: 0.422, Lexical: 0.398, Syntax: 0.497, and Semantics: 0.420.

C Annotated features

Table 5 displays the features implemented for each language. Features are based on those described by Wilkens et al. (2022).

feature	**PT**	**SP**	**FR**
Auxiliary verbs	✓	✓	✓
Passive construction	✓	✓	✓
Subordination	✓	✓	✓
Coordination	✓	✓	✓
Clitic pronouns	✓	✓	✓
NP/PP modifiers	✓	✓	✓
Depth of parse tree	✓	✓	✓
Support-verb constructions	✓	✗	✗
Causative operator-verb	✓	✗	✗
Vocative	✓	✗	✓
Echo complements	✓	✗	✗
Verbal idioms	✓	✗	✗
Proverbs	✓	✗	✗
Named Entities	✓	✓	✓
Complexity	✓	✓	✓
Number of sentences	✓	✓	✓
Sentence length	✓	✓	✓
Word length	✓	✓	✓
Word length (in syllables)	✓	✗	✓
Ratio Hapax	✓	✓	✓
Ratio of surface forms [P0-P75]	✓	✓	✗
Ratio of lemmas [P0-P75]	✓	✓	✗
Ratio Sophisticated Words	✓	✗	✓
Orthographic Neighbors (Nb.)	✓	✗	✓
Orthographic Neighbors (Cum. Freq.)	✓	✗	✓
Age of acquisition	✓	✓	✓
Word familiarity	✓	✓	✓
Ratio of abstract words	✓	✗	✗
Ratio of concrete words	✓	✗	✓
Ratio of Polysemous Words	✓	✗	✗
Nb. of words before verb	✓	✓	✓
Nb. of words after verb	✓	✓	✓
Dialogue quote	✓	✗	✗
Lexical diversity (MATTR)	✗	✗	✓
Lexical frequency	✗	✓	✓

Table 5: Summary of the implemented features per language.