Towards Sentence-level Text Readability Assessment for French

Duy Van Ngo
CNRS - LORIA
Vandœuvre-Lès-Nancy, France
van-duy.ngo@loria.fr

Yannick Parmentier[1,2]
(1) Université de Lorraine - LORIA
Vandœuvre-Lès-Nancy, France
(2) Université d'Orléans - LIFO
yannick.parmentier@loria.fr

Abstract

In this paper, we report on some experiments aimed at exploring the relation between document-level and sentence-level readability assessment for French. These were run on an open-source tailored corpus, which was automatically created by aggregating various sources from children's literature. On top of providing the research community with a freely available corpus, we report on sentence readability scores obtained when applying both classical approaches (aka readability formulas) and state-of-the-art deep learning techniques (e.g. fine-tuning of large language models). Results show a relatively strong correlation between document-level and sentence-level readability, suggesting ways to reduce the cost of building annotated sentence-level readability datasets.

1 Introduction

Text readability assessment can be defined as the ability to automatically estimate the difficulty for someone to understand a given text. While it was primarily designed for selecting materials for textbooks (Dale and Chall, 1948) and based on statistical formulas modelling lexical and syntactic complexity, it has proved useful in many other contexts, such as evaluation of text simplification systems (Štajner and Saggion, 2013; Alva-Manchego et al., 2019), and has been extended to modern neural architectures (Martinc et al., 2021).

Depending on the context, estimating readability can take different forms (and rely on different scales)[1]. When aiming at assigning a textbook to pupils, a common scale corresponds to pupils' age. When aiming at assigning learning materials to second-language learners, a common scale corresponds to the Common European Framework of Reference (CEFR) for Languages (Council of Europe, 2002). In some contexts, assessing readability amounts to classifying a given text as simple or complex (e.g. when learning to assess readability on binary corpora such as the Geo-Geolino corpus for German (Hancke et al., 2012)).

Linear regression models developed for English (such as that of Flesch (1948)) were capable of capturing some degree of lexical and syntactic complexity. Some of these were later adapted to other languages such as French (Kandel and Moles, 1958). Still, some studies (e.g. (Richaudeau and Staats, 1981)) showed that they do not always correlate with field data.

Various attempts at using machine learning techniques for (mainly English) text readability assessment have been carried out since the seminal work of Si and Callan (2001), who combined statistical language models with surface linguistic features extracted from large datasets. One may cite in particular the work of Filighera et al. (2019) in deep machine learning, where authors developed specific word embeddings and neural architecture.

Readability assessment for French was revisited by François and Fairon (2012), who explored various statistical algorithms and experimented with several linguistic features. Recent advances in this domain include work by Blandin et al. (2020) who considered psycholinguistic features (e.g. emotional impact of texts), and by Martinc et al. (2021) who fine-tuned a pretrained BERT model for CEFR classification for French (Yancey et al., 2021).

As pointed out by Hernandez et al. (2022), a common bottleneck in machine learning-based French text readability assessment lies in the scarcity of useful (e.g. labelled) resources. We build upon their work to provide researchers with a tailored and open-source corpus while studying the relation between sentence-level and document-level readability assessment.

[1]Following the terminology used in machine learning / classification, we will refer to values of these scales as *classes*.

Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability associated with RANLP-2023, pages 78–84,
held in Varna Bulgaria, Sept 7, 2023.
https://doi.org/10.26615/978-954-452-086-1_008

The main contribution of this work is thus the compilation of an average size (1,228 documents) freely available corpus for French as a first language readability assessment, which has been preprocessed to remove noisy data and used to perform sentence-level automatic readability assessment with state-of-the-art BERT architectures, giving results in line with those obtained at the document level (Hernandez et al., 2022).[2]

2 Existing Models and Datasets

In this section, we briefly discuss the existing readability models and datasets related to our work. These models belong to two main categories: readability formulas and (deep and non-deep) machine learning-based approaches.

2.1 Readability Formulas

There have been plenty of approaches to measure the reading difficulty of a text such as Flesch Reading Ease (FRE) (Flesch, 1948), Kincaid Grade Level (KGL) (Kincaid et al., 1975), and Gunning Fog Index (GFI) (Gunning, 1969), to mention a few. The mutual simplicity-centric characteristic of these methods comes directly from the authors as they called the formulas "yardsticks" (Flesch, 1948; Gunning, 1969). Being rather arithmetic, these methods hold on to constants and self-defined coefficients in the effort of fitting the outcome in a fixed range of values while making use of similar variables (e.g. number of syllables).

2.2 Machine Learning-based Approaches

Text readability assessment can be viewed as a classification problem (François and Fairon, 2012; Hancke et al., 2012; Vajjala and Lučić, 2018). While statistical models can be employed upon the extraction and quantification of linguistic features of a text for reading difficulty evaluation (François and Fairon, 2012), approaches using Large Language Models (LLMs) require less symbolisation of linguistic features and yet demonstrate dominant performance amongst the rest (Hernandez et al., 2022). LLMs, especially pretrained LLMs with BERT-like architecture have drawn attentions of users from a wide range of fields and practical usages. Having been pretrained on massive datasets, these LLMs with fine-tuning techniques achieved state-of-the-art results in many Natural Language Understanding tasks (Devlin et al., 2018) where the text readability assessment task manifests itself. The prospect of utilising fine-tuned pre-trained LLMs for text readability assessment is noticeable (Hou et al., 2022). Recent prominent encoders are proven to store linguistic features without explicit guidelines. Since the readability of a text is correlated with such features (including but not limited to lexical, syntactic, and semantic features), features contained in document embeddings are definitely valuable in the understanding of text complexity.

2.3 Datasets

Despite the fact that labelled datasets are crucial for classification tasks, the French language has experienced a shortage of such datasets for readability assessment under the L1 learner-centric theme.

Still, some French corpora dedicated to this task do exist. One may cite the corpus collected by Daoust et al. (1996) within the SATO-CALIBRAGE project aiming at assisting teachers in the selection and creation of adapted learning materials, and which consits of 679 texts from textbooks from primary and secondary schools in Quebec. François et al. (2014) developed AMESURE, a collection of 105 administrative texts automatically annotated into 5 readability classes. More recently, Wilkens et al. (2022) compiled FLM-CORP, a carefully curated corpora gathering 334 texts from Belgian textbooks of French literature, history, and sciences. Unfortunately, these corpora are not openly accessible due to copyright constraints.

Regarding open corpora supporting French, Hernandez et al. (2022) created three open corpora by collecting free books on the internet from the following sources: Je Lis Libre[3] (JLL), Litterature de Jeunesse Libre[4] (LJL), and Bibebook[5] (BB). Each of these corpora uses specific readability scales (having 3 to 4 classes). Altogether these corpora contain 998 texts. Classifications conducted using the corpora provision promising results, showing that document-level text readability assessment can be achieved using a fine-tuned BERT model with a macro F1 score from 69% on LJL to 92% on JLL, depending on the characteristics of each corpus.

[2]This work was financially supported by the French Scientific Research Center (CNRS) within the GramEx project.

[3]http://www.crdp-strasbourg.fr/je_lis_libre/

[4]https://litterature-jeunesse-libre.fr/bbs/

[5]http://www.bibebook.com/

3 Experimental Framework

In this work, we are studying sentence-level readability assessment for French. The readability scale we are using comes from the context of this work, namely the implementation of a computer assisted language learning environment for French L1 learners. We consider the five following levels (classes):

0 emergent readers
1 short and easy texts
2 long and easy texts
3 lower-intermediate texts
4 upper-intermediate texts

These levels are loosely related to the French primary school curriculum, and match the categories available in the online resource used to create our corpus (namely StoryWeaver, see below).

3.1 Corpus Construction

The target corpus is designated to be the consolidation of contents from French books available on the StoryWeaver[6] website under a creative commons licence. The website lists 1257 children stories in French language that belong to 5 readability levels, categorised as described in Table 1 (and which, as mentioned above, match our readability classes).

Level	Word count	Other descriptions
0	< 50	Familiar words, word rep.
1	$50 - 250$	Easy words, word rep.
2	$250 - 600$	Simple concepts
3	$600 - 1500$	Longer sentences
4	> 1500	Long & nuanced stories

Table 1: StoryWeaver level description

To back up the claim that simpler texts tend to be more repetitive, we computed repetition rates for each of these levels. Results are given in Table 2 below (level 4's lower repetition rate comes from its relatively small number of tokens).

Level	#uniq. lemmas	#tokens	Rep. rate(%)
0	852	4,076	20.90
1	4,919	63,255	7.78
2	8,776	157,372	5.58
3	10,318	192,137	5.37
4	9,014	128,440	7.02

Table 2: Repetition rate of unique lemmas

[6] https://storyweaver.org.in/en/

Data Retrieval The books are filtered by readability levels before their ID and level are extracted and stored in a PolaRS[7] dataframe. Afterwards, the URL to each story is constructed by concatenating the path with its ID. Thanks to Selenium[8] on Python, each story with its basic information, including title, author, level, and translator (optional, only applicable if the story is not originally written in French) are automatically scraped from the HTML documents and stored along with a local path to the downloaded PDF file.

Pre-processing After the removal of duplicated stories, there are a total number of 1256 stories of five readability levels downloaded. The PyPDF2 library is used to extract texts from the PDF files. To minimise the presence of unwanted texts such as authors' name, acknowledgements, credits, etc., the cover page of every book is ignored along with the last four pages since these pages do not contribute to the individual content of the book. Besides, to ensure the lowest rate of noises possible for the corpus, page numbers are excluded, along with sentences whose length is smaller than 4 tokens or greater than 28 tokens. Table 3 outlines key properties of the corpus as a result of the preprocessing step. We used the SpaCy library and its `fr_core_news_md` pipeline[9] for sentence segmentation and tokenization.

Level	#documents	#sentences	#tokens
All	1,228	52,168	545,280
0	84	700	4,076
1	424	7,903	63,255
2	421	16,672	157,372
3	215	16,748	192,137
4	84	10,145	128,440

Table 3: Corpus level-based x-counts

The resulting corpus, named FSW (for French StoryWeaver), is freely available under a Creative Commons CCBY4.0 license.[10]

3.2 Readability Assessment

We applied both traditional readability formulas and deep learning classification models to our tai-

[7] https://pola-rs.github.io/polars/polars/index.html
[8] https://www.selenium.dev/
[9] https://spacy.io/models/fr/
[10] https://gitlab.inria.fr/vngo/fsw-corpus

lored corpus as described below.[11]

Traditional Metrics Though French is not the target language for the KGL and FRE metrics, the scores do depict the complexity with regards to the average words per sentence and syllables per word. For the statistics concerning the mean KGL scores, FRE scores, token counts, and syllable counts, see Table 4 below.

Level	KGL	FRE	tokcount	sylcount
0	-3.23	127.69	48.52	48.19
1	-0.50	113.60	149.19	152.54
2	1.05	105.25	373.80	404.34
3	2.19	100.46	893.66	995.62
4	2.76	97.50	1529.05	1772.29

Table 4: Basic metrics of each readability class

These results confirm that there exists a strong correlation between each pre-existing text readability level and the KGL and FRE metrics.

Fine-tuned CamemBERT for Classification To examine the distinctiveness of documents from different readability levels from a LLM perspective, we consider fine-tuning and evaluating CamemBERT models (Martin et al., 2020) with the corpus we obtained. We conduct two experiments using the `camembert-base` model[12], attempting to decipher the correlation between document-level and sentence-level readability (keeping in mind that the distinctiveness of classes is a key factor). Due to the insignificant volume of data compared to other classes, the documents with level 0 are ignored. If not explicitly mentioned, we fine-tune pretrained CamemBERT models with 5 epochs and the batch size of 64 using the `grele` cluster of Grid5000[13]. The fine-tuning process on this cluster with a single GTX 1080Ti GPU takes approximately 30 minutes.

Randomly Split Datasets In this first experiment, we examine the performance of a fine-tuned CamemBERT model for classification. We use SpaCy to collect the sentences from each document. These sentences are assigned the level from the document they are originally from. The dataset made of sentences labelled with their level is then randomly split into two subsets: train and test sets. We finetune the CamemBERT model on the train set and evaluate it on the balanced test set. The classification result is illustrated in Figure 1.

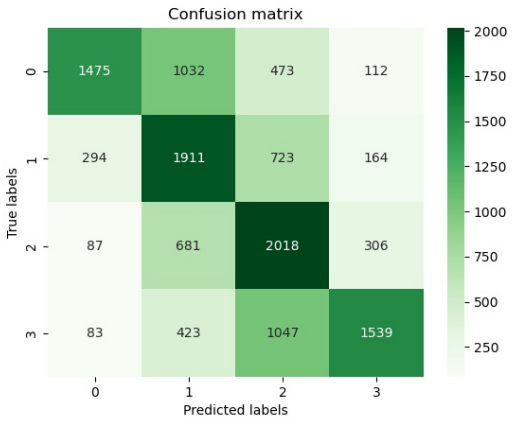

Figure 1: Classification results on randomly split test set

The fine-tuned model performs relatively well on the test set, and is able to classify most of the sentences to the readability level of the documents where they are extracted from. Indeed F1 scores range from 53% to 59% depending on the level (recall that document-level readability assessment on LJL, that is, a corpus of children's book comparable to ours, made by Hernandez et al. (2022) reached 69%). Table 5 shows the details of our classification attempt.

Level	Precision	Recall	F1 Score
1	76.07	47.70	58.64
2	47.22	61.80	53.54
3	47.36	65.27	54.89
4	72.56	49.77	59.04

Table 5: Classification scores on randomly split test set

When compared with the application of neural transformer-based models to document-level text readability assessment in English using datasets such as WeeBit, whose performance reaches an F1 score of 85% (Martinc et al., 2021),[14] these results may seem somehow limited, suggesting that our corpus is still relatively noisy. Another reason for our scores may come from the model itself. Indeed Martinc et al. (2021) used a model which was

[11] For a more exhaustive evaluation of comparable corpora against non-deep machine learning-based approaches such as SVM, see (Hernandez et al., 2022).

[12] https://huggingface.co/camembert-base

[13] https://www.grid5000.fr/

[14] Even better performances (99% classification accuracy) have been obtained for English by mixing handcrafted linguistic features with transformer-based models (Lee et al., 2021). We could not experiment with these hybrid models as they do not support French.

pretrained on documents of a somehow homogeneous type (books and wikipedia articles) while we used CamemBERT whose pretraining relied on much diverse documents (coming from Common Crawl). Furthermore our corpus is mainly made of children's books, whose content may be less close to the pretraining data. To put these results into perspective, one can note that Martinc et al. (2021) also applied BERT-like architectures on Slovenian school books, and obtained a F1 score of 41%. Eventually, the size of the model's input data (document-level vs sentence-level assessment) may also impact its performance.

Disjoint Datasets This experiment is conducted to test the generalisability of the model and eliminate the possible cross contamination that may lead to the model trying to identify documents using the given sentences rather than the readability level of the sentence itself. We split the dataset into two subsets, train and test sets, in which the documents in each set are disjoint. In other words, all the sentences in the train set belong to none of the documents in the test set. We fine-tune the CamemBERT model on the train set and evaluate its performance on the balanced test set. The classification result with regards to the confusion matrix is displayed in Figure 2.

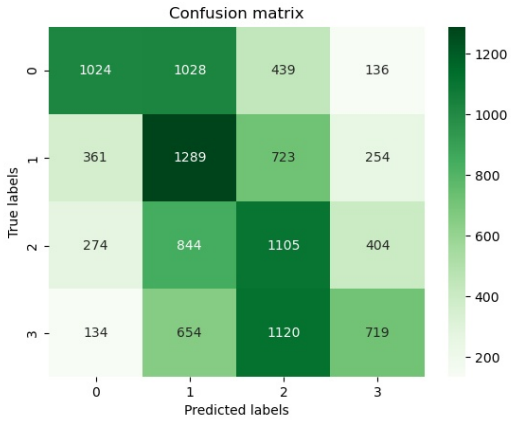

Figure 2: Classification results on disjoint test set

Furthermore, for the detailed classification results with regards to the precision, recall, and F1 scores, see Table 6.

Despite the lower scores compared to the model fine-tuned on the randomly split dataset, the model in this experiment reveals interesting common phenomena, such as significantly higher precision scores for level 1 and level 4 while maintaining relatively high recall scores for the two middle levels. Besides, there exists a higher confusion rate between sentences of adjacent classes than that of distant classes.[15] Except the prediction that labels many level 4 sentences as level 3, the model does portray a distinctiveness between levels, and maintains a consistent reduction of confusion rate as the differences between levels increase.

Level	Precision	Recall	F1 Score
1	57.11	38.98	46.33
2	33.79	49.07	40.02
3	32.62	42.06	36.75
4	47.52	27.37	34.73

Table 6: Classification scores on disjoint test set

About Sentence-level Readability Assessment In these experiments, we took the document's level as a reliable level for the sentences contained in the said document. This may seem unreasonable as texts cannot be expected to contain only sentences of a given level. Recall that we performed some preprocessing on the input data to remove very short and very long sentences, and that our input data belong to a specific domain, namely children's stories. We think that in this context, the impact of this sentence labelling is weaker than in a general setting (i.e., when more diverse texts are used). Furthermore, we aim at studying the performance of readability assessment under such a heuristic. Our results tend to show than it remains reasonable considering the prediction performances on adjacent classes (i.e., when allowing for "minor" errors).

4 Conclusion

In this paper, we presented a freely available corpus for French sentence-level text readability, which was automatically extracted from online resources and evaluated against state-of-the art deep-learning techniques. Results show some correlation between document-level and sentence-level readability assessment, which suggests that extending training corpora could be done by considering labelled documents, thus saving annotation costs.

Acknowledgments

We are grateful to Claire Gardent and anonymous reviewers for their valuable comments on this work.

[15] Considering adjacent levels is also done by François and Fairon (2012) to distinguish minor errors from more serious ones.

Lay Summary

”Is it possible to automatically assign a given text a readability score, which would reflect the difficulty for someone to understand this text ?” is a question which has been discussed by researchers from various fields including linguistics, science of education, or computer science for decades. Being able to compute such scores could for instance help teachers to select learning materials depending on their target audience. First attempts at computing such scores were based on so-called readability formulas, where readability was a function of various linguistic properties (e.g. sentence length).

More recent work applied techniques borrowed from the field of machine learning to this task, reaching state-of-the-art results. Such approaches require labelled data, that is, texts whose content has been labelled with a readability score by a human annotator. Freely available such labelled data is still relatively rare for other languages than English, especially French.

With the growing availability of texts (and computing power), new techniques of machine learning called deep learning (or sometimes simply AI) arose. Such techniques use so-called deep neural networks, which correspond to very large parameterized networks capable of learning implicit patterns from input data. These techniques were in particular used to create large language models (LLMs) which are trained on extremely large datasets and can be adapted to specific tasks via an additional training phase called fine-tuning.

In this work, our objective is (1) to create an average size open dataset for French, which would associate sentences with a readability score, and (2) to study how well would a LLM fine-tuned with this dataset would perform.

While a similar study has already been done at the document level reaching relatively good results (80% in terms of average accuracy), here we focus on the sentence level. We aim at finding whether assigning sentences with their document readability level (e.g. in case of lacking sentence-labelled data) would still be a viable option. The experiments we ran tend to show that such an assignment does not prevent the fine-tuned LLM from performing well, in so far as the LLM makes relatively few strong errors (i.e., it rarely computes readability scores which are not close to the target scores).

References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. 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): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Alexis Blandin, Gwénolé Lecorvé, Delphine Battistelli, and Aline Étienne. 2020. Recommandation d'âge pour des textes. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : Traitement Automatique des Langues Naturelles (Articles courts)*, pages 164–171, Nancy, France. Association pour le Traitement Automatique des Langues. Age recommendation for texts.

Council of Europe. 2002. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment : Case Studies*. Language Learning. Council of Europe Publishing/Éditions du Conseil de l'Europe.

Edgar Dale and Jeanne Sternlicht Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27(1):11–28.

François Daoust, Léo Laroche, and Lise Ouellet. 1996. Sato-calibrage : présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, 25(1):205–234.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings*, page 335–348, Berlin, Heidelberg. Springer-Verlag.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Thomas François, Laetitia Brouwers, Hubert Naets, and Cédrick Fairon. 2014. AMesure: a readability formula for administrative texts (AMESURE: une plateforme de lisibilité pour les textes administratifs) [in French]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 467–472, Marseille, France. Association pour le Traitement Automatique des Langues.

Thomas François and Cédrick Fairon. 2012. An “ai readability” formula for french as a foreign language.

In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 466–477.

Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

Nicolas Hernandez, Nabil Oulbaz, and Tristan Faine. 2022. Open corpora and toolkit for assessing text readability in french. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 54–61.

Shudi Hou, Simin Rao, Yu Xia, and Sujian Li. 2022. Promoting pre-trained lm with linguistic features on automatic readability assessment. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 430–436.

Liliane Kandel and Abraham Moles. 1958. Application de l'indice de flesch à la langue française. *Cahiers Etudes de Radio-Télévision*, 19(1958):253–274.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

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.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. 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.

François Richaudeau and Donna M. Staats. 1981. Some French Work on Prose Readability and Syntax. *Journal of Reading*, 24(6):503–508.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 574–576, New York, NY, USA. Association for Computing Machinery.

Sanja Štajner and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382, Nagoya, Japan. Asian Federation of Natural Language Processing.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

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.

Kevin Yancey, Alice Pintard, and Thomas Francois. 2021. Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e linguaggio, Rivista semestrale*, 2021(2):229–258.