Proceedings of the 2nd READI Workshop @ LREC2022, pages 54–61
Marseille, 24 June 2022

Open corpora and toolkit for assessing text readability in French

Nicolas Hernandez, Nabil Oulbaz, Tristan Faine
LS2N, Nantes Université
France
nicolas.hernandez@ls2n.fr, {nabil.oulbaz,tristan.faine}@etu.univ-nantes.fr

Abstract

Measuring the linguistic complexity or assessing the readability of written productions has been the concern of several researchers in pedagogy and (foreign) language teaching for decades. The children's language development and the second language (L2) learning are in focus with tasks such as age or reader's level recommendation, or text simplification. Despite the interest for the topic, open datasets and toolkits for processing French are scarce. In this paper, we present: (1) three new open corpora for supporting research on readability assessment in French, (2) a dataset analysis with traditional formulas and an unsupervised measure, (3) a toolkit dedicated for French processing which includes the implementation of statistical formulas, a pseudo-perplexity measure, and state-of-the-art classifiers based on MLP, SVM, fastText and fine-tuned CamemBERT for predicting readability levels, and (4) an evaluation of the toolkit on the three data sets.

Keywords: open-source, free, corpus, toolkit, readability assessment, French

1. Introduction

Text readability refers to the difficulty in understanding a given text. The difficulty depends on the reader's language ability and knowledge background as well as the linguistic complexity of the written object. Measuring the linguistic complexity or assessing the readability of spoken or written productions has been the concern of several researchers in pedagogy and (foreign) language teaching for decades. Children's language development (Blandin et al., 2020) or second language (L2) learning (Yancey et al., 2021) are mainly in focus with tasks such as age or reader's level recommendation (Rahman et al., 2020; Pintard and François, 2020), or text simplification (Javourey-Drevet et al., 2022).

Works on readability assessment can be classified into three approaches: (1) the statistical formulas, (2) the language model (LM)-based measures, and (3) the supervised approaches. The latter can be categorised further into two types: (3a) the (linguistic) feature-based and (3b) the deep learning-based approaches.

The formulas (1) are often called traditional because they correspond to early works in the field (Gunning, 1971; Smith and Senter, 1967; Kincaid et al., 1975; Mc Laughlin, 1969). Despite the fact they do not capture all the linguistic complexity of the discourse, they have the advantage to be easily implementable. The LM-based approaches (2) benefit from being unsupervised. With the advent of deep learning in especially Natural Language Processing (NLP), the LMs switch from statistical to neural ones (Martinc et al., 2021). They can be considered as formulas' evolution. The feature-based approaches (3a) were the standard approaches before deep learning became the new reference of doing machine learning (Balakrishna, 2015; Wilkens et al., 2022; Crossley et al., 2022). In practice, they remain quite competitive for readability tasks with end-users because they offer explicability and concrete (linguistic) objects that humans can discuss and understand. Deep neural architectures have been proposed to support the prediction of readability classes (Azpiazu and Pera, 2019b; Deutsch et al., 2020; Rahman et al., 2020; Martinc et al., 2021; Yancey et al., 2021). Works at the edge attempt to combine the advantage of a feature-based approach with a deep learning one (Deutsch et al., 2020; Qiu et al., 2021).

Despite the interest for the field, resources for processing French are scarce, while open datasets and toolkits exist in other languages. Free implementations of the readability formulas exist for processing English[1]. Linguistic feature-based approaches are also available as open source libraries for computing readability metrics in English[2] (Balakrishna, 2015) and in Portuguese.[3] The implementation of (Martinc et al., 2021)'s neural approaches have been proposed for German readability assessment[4] while Deutsch et al. (2020) and Qiu et al. (2021) released their code with the paper respectively for processing English and Chinese. The study of English is also supported by the availability of several corpora (Vajjala and Meurers, 2012; Vajjala and Lučić, 2018). Recently Crossley et al. (2022) initiated the creation of an open corpus in English.

In terms of toolkit for processing French, the CENTAL Lab. offers AMesure,[5] an on-line demonstration application to analyse lexical, syntactic and textual difficulties of French administrative texts and rate the readability with a scale from 1 to 5 (François et al., 2018). Recently, the CENTAL has deployed another

[1] https://github.com/cdimascio/py-readability-metrics
[2] https://bitbucket.org/nishkalavallabhi/complexity-features
[3] https://github.com/vwoloszyn/pylinguistics
[4] https://github.com/kinimod23/GRANT
[5] https://cental.uclouvain.be/amesure

web service called FABRA[6] to assess reading difficulty in French. The toolkit is based on the aggregation of several linguistic features (Wilkens et al., 2022). Based on fine-tuning BERT on texts from French as a Foreign Language (FFL) course material following the Common European Framework of Reference for Languages (CEFR), (Yancey et al., 2021) will offer a web interface[7] for readability evaluation. Without discussing the performance of these deployed analysers, the quality of a toolkit as a service will depend on both the bandwidth availability and the power of the server. In addition, it will act as a blackbox and will not allow modification. Although there are nice projects funded by the National French Agency such as texttokids[8], there are little corpora freely available yet. We can mention the works of (Gala et al., 2020) and (Azpiazu and Pera, 2019a) who make available French corpora with aligned original and simplified texts. Our contributions are:

1. (1) three open corpora for supporting research on readability assessment in French,
2. (2) a dataset analysis with traditional formulas and an unsupervised measure,
3. (3) a toolkit dedicated for French processing which includes the implementation of statistical formulas, a pseudo-perplexity measure, and state-of-the-art classifiers based on multi-layer perceptron (MLP), Support Vector Machine (SVM), fastText and fine-tuned BERT for predicting readability levels,
4. and (4) an evaluation of the toolkit on the three data sets.

The library and corpora will be made available under open license in a repository later on.

The rest of the paper is structured as follows: Section 2 introduces the related work on readability measures and prediction techniques. We also say a few words on the grades system in France. Section 3 presents the corpora we collected for supporting readability studies and recommendation or prediction tasks. Section 4 presents a thorough analysis of our corpora as well as the report of the results of state-of-the art prediction systems.

2. Related Work

The readability assessment issue has been addressed by several researchers trying to find pertinent factors to take into account in order to automate this task. Martinc et al. (2021) offer a consolidated review of the major approaches.

[6]https://cental.uclouvain.be/fabra
[7]https://cental.uclouvain.be/amesure
[8]https://texttokids.irisa.fr/project

2.1. Traditional formulas

Readability measures mentioned in this section refer to methods based on mathematical functions linking text structural characteristics to a simple value of readability as perceived by humans. The structural characteristics are statistical measures on each text such as total words, total sentences, number of long words and number of syllables.

The Gunning fog index (GFI) formula (Gunning, 1971) takes into consideration the total number of words and sentences and the number of long words (long words are defined as words longer than 7 characters). GFI value and readability are negatively correlated meaning that a high GFI value indicates a higher readability measure. The Automated readability index (ARI) formula (Smith and Senter, 1967) corresponds to the number of study years needed to understand a text. It uses as features, similar to GFI, the total number of words and sentences in a text with the addition of the total number of characters. The Flesch reading ease (FRE) formula (Kincaid et al., 1975) brings an addition to the already mentioned formulas. It uses total number of syllables in a text to compute a score that increases with more readable documents. The Flesch-Kincaid grade level (FKGL) (Kincaid et al., 1975) is a similar formula to FRE, it corresponds to the number of years of education needed to understand a certain text. The Simple Measure of Gobbledygook (SMOG) formula (Mc Laughlin, 1969) similar to FKGL and ARI returns the number of years of education required to understand a text. It uses the number of polysyllables - the number of words containing three or more syllables in a text. Flesch's reading ease has been adapted to French language by (Kandel and Moles, 1958). They made changes to the coefficients of FRE to take into account the length difference between French and English Words. Their formula is named Reading Ease Level (REL).

2.2. Language model-based measures

Perplexity (ppl) is a common intrinsic metric for evaluating language models. It is defined as the exponential average negative log-likelihood of a sequence. For masked language models like BERT (Devlin et al., 2018), Salazar et al. (2020) proposed an adaptation called the pseudo-perplexity (pppl). The lower the score is the better the language model is able to "predict" a given text.

Martinc et al. (2021) also proposed a ranked sentence readability score (RSRS) which exploits language models to estimate a readability score for each word in a specific context.

2.3. Supervised approaches

Many traditional machine learning algorithms were experimented for the readability prediction task (Schwarm and Ostendorf, 2005; Vajjala and Meurers,

2012). These methods used various kind of features: traditional formulas scores, discourse cohesion measures, lexico-semantic features, syntactic and language model measures. The literature reveals that Support Vector Machine (SVM) classifier was giving the best results for (Martinc et al., 2021).
Feature-based approaches are language and genre-dependent. With the success encountered by Deep Learning methods for tackling numerous NLP tasks, end-to-end neural architectures were also proposed for difficulty estimation or readability classification.
Filighera et al. (2019) designed architectures comprising three global layers: an input layer made of contextual and non-contextual word embeddings (word2vec (Mikolov et al., 2013), BERT (Devlin et al., 2018), ...), an intermediate layer dedicated to the building of a text representation (thanks to Bi-LSTM or CNN layers), than a final dense layer to perform the prediction. Martinc et al. (2021) proposed a classifier by fine-tuning a pre-trained BERT model on a specific readability corpus. This latter approach correspond to the state-of-the-art performances. This approach gave the best results in Yancey et al. (2021) in a CEFR classification task of French as a foreign language.

2.4. Ages, grades, readability levels...

Age	Cat.	LC	FR grade	CEFR	US grade
<6	Pre.	lc1	PS, MS, GS		Kinder.
6-9	Prim.	lc2	CP, CE1, CE2	A1	1-3
9-12	Prim., Sec.	lc3	CM1, CM2, 6e	A1-A2	4-6
12-15	Sec.	lc4	5e, 4e, 3e	A2-B1	7-9
15-18	High		2nd, 1st, terminal	B1-B2	9-12

Table 1: Alignment of age, grades in French (FR) and in US, French learning cycle (LC), category (Cat.) such as Preschool (Pre.), Primary (Prim.), Secondary (Sec.) and High School, Kindergarten, and the Common European Framework of Reference for Languages (CEFR).

Since 2014, the French primary school (*primaire*) has been split into four learning cycles[9]. To erase any maturity differences, the learner has 3 years to acquire the required skills before the next stage: cycle 1 "first learning" (under 6, PS-GS), cycle 2 "fundamental learning" (6-8, CP-CE2), cycle 3 "consolidation" (9-11, CM1-6e) and cycle 4 "enhancement" (12-14, 5e-3e). At the primary school, the reading levels follows this development.
In order to provide a basis for recognising language qualifications, the Council of Europe proposed to "organise language proficiency in six levels, which can be regrouped into three broad levels: Basic User (beginner A1, intermediate A2), Independent User (B1, B2) and Proficient User (C1, C2)" called the The Common European Framework of Reference for Languages (CEFR).[10]A1 corresponds to beginner at primary school, A2 to intermediate at secondary school, B1 to newly independent at the end of the compulsory education (*collège*), B2 to advanced at high school (*baccalauréat*), C1 to autonomous learner, C2 to master.
Table 1 attempts to provide an overview of the alignment between the ages, grades and the education syllabus.

3. Datasets

Our datasets result from the compilation of various sources releasing children's and young adult's books under open licences (mainly in CC BY). These include the following projects: *littérature de jeunesse libre*, *StoryWeaver*, *Bibebook*, *Je Lis Libre*, WikiSource and Gutenberg. Some of these sources are collecting and packaging books coming from other sources. For more convenience, we will refer here to three distinct packages: *littérature de jeunesse libre* (*ljl*), *Bibebook* (*bb*) and *Je Lis Libre* (*jll*). Books belong to the literary genre (children story, adventure novel, poetry, theatre play...).
The *littérature de jeunesse libre (ljl)*[11] corpus compiles children's books acquired from the StoryWeaver platform which defines four reading levels:[12] (lv1) beginning to read (easy words with repetition, short sentences, up to 250 words), (lv2) learning to read (simple concepts, from 250 to 600 words), (lv3) reading independently (popular topics with well sketched-out characters, 600 to 1500), (lv4) reading proficiently (rich vocabulary, word play, more than 1500 words). In our interpretation, we consider lv1 and lv2 covering the second learning cycle (lc2), and lv3 and lv4 covering the third one (lc3). Books are mainly children stories translated from Hindi or African literature. The 746 books were written by 460 distinct authors.
With the *bibebook (bb)* project, the Association de Promotion de l'Ecriture et de la Lecture (APEL) aims at promoting writing and reading activities for young adults. The corpus references books[13] that are in the public domain (i.e. with authors who died more than 70 years ago), and which are known as classic masterpieces that young adults read in French secondary

[9] Loi d'orientation sur l'éducation de 1989, modifiée en 2014 par un décret de 2013 https://www.education.gouv.fr/bo/13/Hebdo32/MENE1318869D.htm?cid_bo=73449

[10] https://www.coe.int/en/web/common-european-framework-reference-languages

[11] litterature-jeunesse-libre.fr/bbs/

[12] storyweaver.org.in/reading_levels

[13] www.bibebook.com/visual-search?f%5B0%5D=field_genre%3A1267

school (such as La Fontaine's tales, Molière's plays, Vernes's adventure novels, Zola's novels, Racine's plays). Books are organised in three levels of difficulty: easy reading (age 10-12), intermediate reading (12-15), and advanced reading (15-18). The 208 books are written by 72 distinct authors.

The *je lis libre*[14] project is a small database which refers to a subset of books present in *bibebook* database. The organisation is different and follows the reading recommendation from the Ministry of Education for a given secondary school grade: grades from 6 to 3 (3 being higher than 6 in the French education system).

To collect the books, we scrapped each website (while respecting the `robots.txt` restrictions) to get the pdf or epub files of each document, and used common tools, such as the pdftotext python library [15] to convert them into text format. Thanks to adhoc filters or manual operations, we were able to clean them as much as possible by removing meta-data descriptions (header and footer).

Dataset statistics are presented in Table 2. Sentence splitting and word tokenization were performed thanks to the NLP `spaCy` library and its `fr_core_news_sm`[16] model.

When looking at the number of tokens or the number of documents for each readability class, we clearly see that the corpora are unbalanced. We can also note that the corpora are small in terms of number of documents while being big in terms of number of sentences and tokens. We do not report here the average number of tokens per document but we can easily infer from the Table that the document size in the *ljl* corpus goes from 150 to 1,500 words approximately, and to tens of thousands of words in the *bb* and *jll* corpora.

The vocabulary size for *ljl* corpus is 23,123 words, 36,011 for *jll* and 38,503 for *bb*. The latter two are somewhat comparable, however the *ljl* corpus is lacking diversity in its words.

4. Datasets analysis and class prediction

In this section, we report:

- First the readability analysis of our corpora thanks to the traditional formulas and the pseudo-perplexity measure (cf. Section 4.1) ;
- Then we evaluate baseline approaches over the corpora and provide preliminary results for the class prediction task (cf. Section 4.2).

In both studies, we did not use the raw versions of the corpora. For each corpus, due to the imbalance between the classes, the size of the documents and the small number of documents we have at our disposal for each class, we decided to artificially generate new documents (d') from the big ones. New documents were generated to be between 140 and 200 words, with all beginning and ending not starting or ending in the middle of sentences. In (Crossley et al., 2022), the authors did the same to build up their corpus. The distinction is that our generation is automatic and consequently our generated documents may not correspond to an idea unit. For the *ljl* corpus, the strategy was to split the big documents into smaller pieces while for *bb* and *jll*, which comes with much larger documents, the strategy was to select text excerpts. We could not get smaller pieces with the *ljl* corpus. For the *bb* and *jll* corpora, we generated documents to obtain about 1k of documents per class. The number of generated documents remains proportional to the number of actual documents.

Last column of Table 2 indicates the number of generated documents.

R_{class}	$\#d$	$\#s$	$\#t$	$\#d'$
littérature de jeunesse libre (ljl)				
lv1	240	4,880	38,976	240
lv2	314	13,049	128,019	628
lv3	134	10,354	124,901	670
lv4	58	7,743	101,165	522
Bibebook (bb)				
easy	52	285,339	4,391,733	988
interm.	91	54,465	857,645	1,729
advan.	65	507,049	8,099,112	1,253
Je Lis Libre (jll)				
6e	13	57,399	1,349,523	1,285
5e	12	50,664	960,218	1,187
4e	10	87,234	1,616,076	989
3e	9	33,414	475,616	890

Table 2: Dataset statistics with readability class (R_{class}), number of documents (#d), of sentences (#s), of tokens (#t), and the number of artificial documents (#d'). The readability classes follow an increasing order: $lv1 < lv2 < lv3 < lv4$, $easy < interm. < advan$ and $6e < 5e < 4e < 3e$.

4.1. Dataset analysis

Table 3 reports the scores given by the traditional formulas and the pseudo-perplexity measure presented respectively in Section 2.1 and 2.2. The scores were averaged over all the documents of a given class. The *pppl* measure was computed by using the generative GPT model `gpt-fr-cased-small`.[17] For each measure, we calculated the Pearson coefficient ($p - score$) in order to estimate the linear correlation between these values and the levels labeled in each corpus.

Regarding the *ljl* corpus, the computed scores of each measure match the classes: The higher a readability class is, the higher the scores are. This is translated into a positive Pearson correlation score except for the

[14] `www.crdp-strasbourg.fr/je_lis_libre`
[15] `https://github.com/jalan/pdftotext`
[16] `https://spacy.io/models/fr`
[17] Sourced by `https://huggingface.co/asi`

R_{class}	GFI	ARI	FRE	FKGL	SMOG	REL	PPPL
littérature de jeunesse libre (ljl)							
lv1	44.61	14.12	78.6	4.28	15.97	94.38	54.59
lv2	66.88	19.8	67.61	6.32	18.65	84.55	57.79
lv3	91.21	25.66	59.04	8.06	21.11	76.81	63.80
lv4	105.52	27.87	54.92	8.81	22.15	73.81	62.87
p-score	0.48	0.49	-0.40	0.45	0.49	-0.40	0.04
Bibebook (bb)							
easy	122.6	35.56	57.04	9.42	23.85	74.49	152.33
interm.	128.93	36.71	56.04	9.67	24.06	73.56	414.00
advan.	122.6	36.26	58.03	9.38	23.95	75.30	161.62
p-score	-0.003	0.012	0.021	-0.006	0.005	0.019	-0.007
Je Lis Libre (jll)							
6e	119.82	46.38	77.38	7.96	23.74	91.45	177.68
5e	132.39	40.75	60.49	9.53	24.38	77.17	114.06
4e	102.42	36.12	81.63	6.27	21.69	95.73	172.71
3e	104.06	34.36	79.84	6.24	21.12	94.32	169.45
p-score	-0.11	-0.19	0.12	-0.17	-0.19	0.13	0.02

Table 3: Traditional formulas and pseudo-perplexity scores for all the readability class (R_{class}) of each corpus. The Pearson coefficient shows the correlation between the scores and the classes.

FRE measure since lower scores indicate that a text is less readable (negative $p-score$). We observe also that despite a positive increment, the lv3 and lv4 classes are closer than each of the other class pairs. This can indicate some difficulties to differentiate between them.

Looking at the *bb* and *jll* corpora, there is no significant correlation between the scores and their respective classes. We note, however, that for both corpora, the measures depict a peak in difficulty for the intermediate classes (namely the "intermediate" class in *bb* and the "5e" class in *jll*. In addition, the small deviation between the scores of the "4e" and the "3e" classes in the *jll* corpus seems to indicate there is no clear difference between the classes.

Concerning the pseudo-perplexity scores, the Pearson coefficient does not detect any correlation with the readability classes. But the *pppl* seems to confirm the closeness in the language of the lv3 and lv4 classes of the *ljl* corpus. It also confirms that the intermediate classes of the *bb* and *jll* corpora seem to follow an unexpected behaviour.

While in primary school the guideline is to pursue the children's development and to increase iteratively the linguistic complexity of the text, it seems that the reading recommendations in secondary school does not follow the same objective. Indeed the pedagogical choices are often to follow an historical progression, from old written texts to more contemporary ones.

Further observations of the corpus are necessary to clarify these numbers.

4.2. Readability class prediction

The current section reports the results obtained with four baselines over the three corpora for a class prediction task. The baselines differ from the text representation and the learning and classification algorithm. Two baselines are feature-based approaches and rely directly on words. One is based on non-contextual subword embeddings; it is fastText (Joulin et al., 2016). And the last one is based on contextual embeddings; it is BERT (Devlin et al., 2018).

4.2.1. Classifiers

In practice, thanks to the `scikit-learn`[18] library, we experimented several traditional machine learning algorithms (SVM, Random Forest, Logistic regression, multinomial Naive Bayes and multi-layer perceptron (MLP)) with normalised (or not) bag-of-words and TF-IDF text representations. We report only the very best of these approaches, namely the SVM and the MLP classifiers with a TF-IDF representation without any text normalisation.

FastText is a word embedding method that is an extension of the word2vec model (Mikolov et al., 2013). Instead of learning vectors for words directly, fastText represents each word as sub-word character n-grams. This offers more robustness to deal with previously unseen words. A document vector is obtained by averaging the subword embeddings. For the classification task, a multinomial logistic regression is used, where the document vector corresponds to the features.

Unlike word2vec-like models, BERT provides contextual embeddings to represent the meaning of words in context. BERT benefits from a bidirectional architecture based on Transformers and their attention mechanism. BERT can easily be used for classification task by adding a supplement dense layer. Training BERT for a classification task results in fine-tuning a pretrained BERT model with an additional layer for the

[18] `https://scikit-learn.org`

(ljl)	lv1			lv2			lv3			lv4				
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc.	Macro F1
MLP	0.42	0.46	0.44	0.47	0.62	0.53	0.47	0.47	0.47	0.55	0.31	0.40	0.48	0.47
SVM	0.41	0.52	0.46	0.47	0.55	0.51	0.48	0.48	0.48	0.52	0.36	0.42	0.47	0.47
fastText	0.49	0.46	0.47	0.59	0.7	0.64	0.71	0.79	**0.75**	0.94	0.62	0.75	0.68	0.65
CamemBERT	0.77	0.46	**0.57**	0.69	0.72	**0.71**	0.7	0.75	0.72	0.74	0.78	**0.76**	**0.71**	**0.69**

(bb)	easy			intermediate			advanced				
	P	R	F1	P	R	F1	P	R	F1	Acc.	Macro F1
MLP	0.44	0.33	0.37	0.52	0.61	0.56	0.51	0.48	0.50	0.50	0.48
SVM	0.44	0.38	0.40	0.53	0.62	0.57	0.54	0.47	0.51	0.51	0.49
fastText	0.75	0.73	**0.74**	0.77	0.78	0.78	0.78	0.78	0.78	0.77	0.76
CamemBERT	0.71	0.71	0.71	0.83	0.8	**0.81**	0.8	0.84	**0.82**	**0.79**	**0.78**

(jll)	6e			5e			4e			3e				
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc.	Macro F1
MLP	0.63	0.79	0.70	0.80	0.63	0.70	0.42	0.37	0.39	0.55	0.57	0.56	0.50	0.59
SVM	0.58	0.80	0.67	0.76	0.61	0.68	0.41	0.32	0.36	0.51	0.48	0.50	0.57	0.55
fastText	0.93	0.89	0.88	0.96	0.9	**0.96**	0.81	0.82	0.8	0.97	0.95	0.81	0.84	0.77
CamemBERT	0.96	0.95	**0.96**	0.92	0.93	0.93	0.87	0.88	**0.87**	0.92	0.91	**0.92**	**0.92**	**0.92**

Table 4: Results on 'littérature de jeunesse libre' (ljl), 'Bibebook' (bb) and 'Je Lis Libre' (jll) corpora for the class prediction task. Best Accuracy, F1-score and Macro average F1-score values are in bold.

task. For our experiments, we used CamemBERT, a state-of-the-art language model for French (Martin et al., 2020). The implementations of the fastText and BERT classifiers were supported by the *ktrain* library (Maiya, 2020).

The evaluation of the algorithms is based on the precision, recall, F1-score, accuracy, macro average F1-score metrics. The reported results for MLP and SVM were obtained by cross validation by splitting each dataset into five folds. For fastText and CamemBERT, the scores were obtained by averaging the scores over five runs, eachone with a randomly selected dataset with 90% for training and 10% for validating. Optimal learning rate (lr) and number of epochs hyperparameters were set up by utilizing the following learning rate schedules: the triangular policy (Smith, 2015), the 1cycle policy (Smith, 2018), and SGDR Warm Restart (Loshchilov and Hutter, 2016). We began training with a maximum value for lr. This was set to 0.0001 for fastText and $2e^{-5}$ for CamemBERT.

4.2.2. Results

Table 4 presents the results respectively for the corpora *ljl*, *bb* and *jll*. The best models are fastText and CamemBERT. Both are competing with each other over the three corpora but CamemBERT slightly outperforms fastText. FastText remains competitive probably by taking advantage of of a vocabulary made of subwords. MLP and SVM achieve similar performance; SVM being better on the *ljl* and *bb* corpora.

For all the models we note that results are higher in the *jll* corpus than in the *bb* corpus. This may come from the fact that the task may be harder for the *bb* corpus since there is a larger number of documents and fewer number of classes to differentiate the documents. The lowest performance scores were obtained for the *ljl* corpus, but this may due to the size of the corpus which remains relatively small.

The difference of performance between the classes of a same corpus seem to match the imbalance in number of instances between the classes. This suggests that future experiments should benefit from taking into consideration class weights. In general, the results are not bad but there is room for improvement in particular on the prediction task on a very small corpus (i.e. the *ljl* corpus).

Despite the fact that the corpus and the number of classes were different, the results are consistent with the results of Yancey et al. (2021) who observed that best results were obtained with a fine-tuned CamemBERT model.

5. Conclusion

Supporting primary and secondary education and developing effective learning environments are part of the Unesco's open science recommendations and its Sustainable Development Goal 4 (SDG4).[19] What is noticeable about the modern age is the efforts for researchers to enable other peers to access to the data and tools they develop (Crossley et al., 2022; Wilkens et al., 2022). With this paper, we aim at contributing to the efforts. Our material contributions are three corpora and a library for assessing readability in French available

[19] `https://unesdoc.unesco.org/ark:/48223/pf0000259784`

under open licences[20].

There are prospects for improving and extending the current work. One major direction will be to deepen the data analysis and the assessment of the data quality. Indeed, the low correlation coefficients question the quality of the *bb* and *jll* corpora. We plan to use the distribution of the current measures to filter out the outliers and observe whether the correlation scores improve. These measures attempt to capture the lexical complexity as well the syntax complexity (with the pppl). In order to verify the reliability of these measures to distinguish the different classes, we will compute correlations with additional lexical complexity measures (for instance by computing the distribution of the Dubois-Buyse school lexicon (Ters et al., 1977) over the classes of each corpus) as well as complementary measures designed for capturing the semantic complexity and the discourse cohesion of the texts. One appealing aspect with such linguistic features is that they can support the implementation of readability measures which allow to build self-explanable systems. Eventually we will also manually annotate a sample of the corpus to confirm there is no issues in the way the texts have been categorised. The study of the classification errors may also allow to understand how to improve our datasets. Since the process of building documents is partially artificial, it is important to ensure that classifiers actually learn to distinguish between readability levels and not from hidden variables (such as authors, topics...). Attention will be paid to other datasets configurations to verify the independence of the classifiers to the variables.

Last, we plan to extend the corpora. Since the data annotated by Crossley et al. (2022) is available in numerous languages, we can study the possibility of transferring to French their manual annotation. New genres such as encyclopaedic textbooks[21] will be considered, this could allow us to compare texts written by children and texts written by adults for children.

6. Acknowledgements

We are grateful to the the anonymous reviewers for their valuable comments which will help us to pursue this work in a more "high quality" direction. This work was partially supported by the French *Agence Nationale de la Recherche*, within its *Programme d'Investissements d'Avenir*, with grant ANR-16-IDEX-0007.

[20] `https://github.com/nicolashernandez/READI-LREC22`

[21] `https://fr.wikimini.org` (written by children) and `https://fr.vikidia.org`

7. Bibliographical References

Azpiazu, I. M. and Pera, M. S. (2019a). Is cross-lingual readability assessment possible? *In press*, 1(1):1–18.

Azpiazu, I. M. and Pera, M. S. (2019b). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Balakrishna, S. V. (2015). *Analyzing TextComplexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Ph.D. thesis, Universität Tübingen.

Blandin, A., Lecorvé, G., Battistelli, D., and Étienne, A. (2020). Recommandation d'âge pour des textes (age recommendation for texts). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 164–171, Nancy, France, 6. ATALA et AFCP.

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., and Malatinszky, A. (2022). A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 16 March.

Deutsch, T., Jasbi, M., and Shieber, S. (2020). Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Filighera, A., Steuer, T., and Rensing, C. (2019). Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning (EC-TEL)*, pages 335–348, Delft, The Netherlands. Springer.

François, T., Müller, A., Degryse, B., and Fairon, C. (2018). Amesure : une plateforme web d'assistance à la rédaction simple de textes administratifs. *Repères DoRiF*, 16 – Littératie et intelligibilité : points de vue sur la communication efficace en contexte plurilingue, novembre.

Gala, N., Tack, A., Javourey-Drevet, L., François, T., and Ziegler, J. C. (2020). Alector: A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Language Resources and Evaluation for Language Technologies (LREC)*, Marseille, France, May.

Gunning, R. (1971). *The Technique of Clear Writing*. McGraw-Hill.

Javourey-Drevet, L., Dufau, S., François, T., Gala, N., Ginestié, J., and Ziegler, J. C. (2022). Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of French. *Applied Psycholinguistics*, pages 1–28, January.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification.

Kandel, L. and Moles, A. (1958). Application de l'indice de flesch à la langue francaise. *Cahiers Etudes de Radio-Télévision*, pages 253–274.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Institute for Simulation and Training*.

Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts.

Maiya, A. S. (2020). ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., Éric de la Clergerie, Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.

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

Mc Laughlin, G. H. (1969). Smog grading-a new readability formula. *Journal of Reading*, 12(8):639–646.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Pintard, A. and François, T. (2020). Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 85–92, Marseille, France, May. European Language Resources Association.

Qiu, X., Chen, Y., Chen, H., Nie, J.-Y., Shen, Y., and Lu, D. (2021). Learning syntactic dense embedding with correlation graph for automatic readability assessment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3013–3025, Online, August. Association for Computational Linguistics.

Rahman, R., Lecorvé, G., Étienne, A., Battistelli, D., Béchet, N., and Chevelu, J. (2020). Mama/papa, is this text for me? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6296–6301, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Schwarm, S. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Smith, E. A. and Senter, R. (1967). Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories*, pages 1–14.

Smith, L. N. (2015). Cyclical learning rates for training neural networks.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay.

Ters, F., Mayer, G., and Reichenbach, D. (1977). L'échelle dubois-buyse d'orthographe usuelle française. *OCDL. 5e édition revue et corrigée*, 1.

Vajjala, S. and Lučić, I. (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, New Orleans, Louisiana, June. Association for Computational Linguistics.

Vajjala, S. and Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada, June. Association for Computational Linguistics.

Wilkens, R., Alfter, D., Wang, X., Pintard, A., Tack, A., Yancey, K., and François, T. (2022). Fabra: French aggregator-based readability assessment toolkit. In *In Proceedings of the thirteenth international conference on language resources and evaluation (LREC 2022), (submited)*.

Yancey, K., Pintard, A., and François, T. (2021). Investigating readability of french as a foreign language with deep learning and cognitive and pedagogical features. *Lingue e Linguaggio*, 2021(2):229–258.