INVESTIGATING READABILITY OF FRENCH AS A FOREIGN LANGUAGE WITH DEEP LEARNING AND COGNITIVE AND PEDAGOGICAL FEATURES

KEVIN YANCEY ALICE PINTARD THOMAS FRANÇOIS

ABSTRACT: In this paper, we report three experiments evaluating a variety of feature sets and models intended to develop a new readability model for French as a Foreign Language (FFL) materials. The purpose of this model is to predict the levels of texts on a scale widely used in foreign language teaching, namely the six proficiency levels defined in 2001 in the Common European Framework of Reference for Languages (CEFR). Our new model has two main advantages. Firstly, it is the first readability model for FFL based on a deep learning model, which shows substantially better accuracy and generalizability than the previous state-of-the-art work. Secondly, it will be freely available through a website for the community of FFL teachers to use. We also investigated new cognitive and pedagogical features, but they failed to outperform existing features sets. Our best performing readability model, obtained by fine-tuning BERT, outperforms the state-of-the-art model by a gain of 8 percentage points in accuracy and adjacent accuracy.

KEYWORDS: Readability model, French as a foreign language, Beacco referentials, deep learning, cognitive and pedagogical features.

1. INTRODUCTION

In our digital societies, the growing prevalence of written texts both in professional and personal settings is well-documented (Belfiore *et al.* 2004). As a result, the ability to read efficiently has become a social necessity. Indeed, not only does it open access to different knowledge areas and to alterity, but it strongly correlates with socioeconomic status as well (Willms 2003). In second language (L2) acquisition, reading became a pivotal skill too and one of the main stakes of L2 curriculum is to support learners in their reading practice. Extensive reading especially, an approach in which learners process numerous relatively easy materials, has been shown to foster reading fluency, vocabulary size, and reading comprehension skills (Grabe 2014). In a further meta-review, Jeon & Day (2016) measured an overall Cohen's *d* effect of 0.57 of extensive reading over intensive or traditional reading settings.

However, extensive reading requires accessible texts, i.e. texts fitting the learner's reading proficiency level, in order to be efficient (Day & Bamford 2002). Finding such texts is a complex task, since many parameters can affect the text's level of difficulty for an L2 reader. For instance, various scholars investigated the relationship between learner's vocabulary size and reading comprehension. Hu & Nation (2000) suggested on this matter that learners must know at least 98% of the words in a fiction text for unassisted understanding. Later, Laufer & Ravenhorst-Kalovski (2010) established that the knowledge of 8,000 word families allows a learner to understand 98% of words in most texts. Although useful, such knowledge does not directly help L2 teachers select concrete materials readable by their students. Scientists in education have been developing readability formulas for this purpose, to help matching texts to the reading proficiency of a group of learners. This area of research on readability dates back to the 1920s with the seminal work of Lively & Pressey (1923), and has since expanded considerably.

Throughout the 20th century, numerous formulas have been developed, including the famous ones by Flesch (1948) and Dale & Chall (1948). In a survey, DuBay (2004: 2) reports that, by the 1980s, more than 200 readability formulas had already been published. A vast majority of these formulas share the same methodological approach. First, text excerpts are collected and their reading difficulty is measured with various techniques (for more details, see François (2014)). This measure is the dependent variable that the model aims to predict. Predictions are thus based on several linguistic variables (e.g. word length, sentence length, etc.) computed on the excerpts and included in the model as the independent variables (or predictors). Finally, a multiple linear regression model (MLR) is trained on the excerpts. MLR has been used by the large majority of scholars until the work of Si & Callan (2001), who applied machine learning and natural language processing (NLP) techniques to the field of readability. Their work gave rise to a large number of studies (see Section 2) relying on non-linear models – such as logistic regression, support vector model (SVM), or, more recently, deep learning – and on NLP-enabled engineered features (e.g. N-grams, tree-based syntactic features, etc.). In most of these studies, linguistic features were manually developed by experts, relying on findings about the reading process. Such feature design is called feature engineering and, in the case of readability, it has the shortcomings of performing a coarse aggregation of the word characteristics at the text level.

In this paper, we hypothesize that distributed representations such as word embeddings and deep learning models should achieve a better aggregation of linguistic information available at the different levels of text. To this aim, we compare the performance of engineered-based readability SVM models

with several state-of-the-art deep learning architectures: a hierarchical attention model (HAN) and a Fine-Tuned BERT model. To our knowledge, it is the first approach of readability for L2 French using deep learning. We also investigate new engineered features that have not been used for L2 French readability previously. At the cognitive level, we identified three types of features that have been shown to be correlated with the reading difficulty of texts (concrete/abstract words, OLD20, and average number of commonly-known senses), but have not been used in French readability models so far. We also developed new pedagogical variables, based on the official Reference Level Descriptors (RLD) for French (Beacco *et al.* 2008). Using the official RLD vocabulary lists should allow our models to better capture the different learner's interlanguage stages. To our knowledge, the only work to leverage RLD-based variables is due to Xia *et al.* (2016). It was dedicated to L2 English and resorted on the English Vocabulary Profile.

The paper includes three experiments and is structured as follows: section 2 offers a brief summary of previous studies in readability and further motivates our main assumptions. Section 3 describes the methodology for our experiments, first introducing the data used to train and evaluate our models (Section 3.1), then detailing the features entered in the models (Section 3.2), before describing the architecture of the three machine learning models we compared (Section 3.3). Section 4 describes the three sets of experiments and reports the results. The first set assesses the contribution of the different types of features; the second compares the model architecture and the third one is dedicated to the discussion of generalization power of our models, an issue that is often overlooked in readability. Finally, we discuss the results in Section 4.4, before concluding.

2. PREVIOUS WORK IN READABILITY

2.1 Advances in the field of readability

As explained above, readability is a field that has been extensively investigated by scholars during the last century. Work from the 20th century has been thoroughly described in previous syntheses (Chall & Dale 1995; DuBay 2004; François 2011; Benjamin 2012; Vajjala 2021). To briefly summarize this large body of work, it is enough to say that, after a first wave of pioneering research in the 1930s – the most exhaustive of which is Gray & Leary (1935) – it began the classical period that spans across the 1940s and 1950s. During this period, readability formulas were kept as simple as possible and included two or three predictors only. Some of the most famous formulas date back to that period,

such as those of Flesch (1948), Dale & Chall (1948) or Gunning (1952). The classic period ended when cloze tests replaced reading comprehension tests as a measure of the reading difficulty of texts (Coleman 1965; Bormuth 1966). The first computerized formulas also appeared at that time (Smith & Senter 1967), which eventually led to an even more superficial view of text difficulty, due to computer limitations. As a result, and also under the influence of advances in cognitive psychology and linguistics, increasingly sharp criticism was directed at readability formulas (Kintsch & Vipond 1979; Kemper 1983), leading the field into a sort of dormancy at the end of the 20th century.

Interest in the field of readability intensified again at the dawn of the 21st century, when techniques from machine learning and NLP helped deal with some of these criticisms. NLP techniques made it possible to take more complex linguistic phenomena into account. For instance, Collins-Thompson & Callan (2005) showed that including word distributions across grade levels within a multinomial Naïve Bayes classifier outperforms classic formulas. Schwarm & Ostendorf (2005) used a syntactic parser to extract several parser-based features, whereas Pitler & Nenkova (2008) investigated various semantic and discourse features, such as lexical chains and discourse relations. A more detailed overview of recent readability models can be found in the surveys by Collins-Thompson (2014) and François (2015).

One aspect that should be stressed is that all these previous approaches relied on engineered features, i.e. features engineered by experts in line with psycholinguistic theories describing what makes a text difficult to read. On the one hand, the use of engineered features has the advantage of producing more interpretable models and fosters the use of variables whose effect on reading is theoretically or empirically proven. On the other hand, engineered features are generally computed at the text level and need to aggregate the information from lower levels such as words and sentences. The common aggregation method in the field is the mean (Flesch 1948; Gunning 1952; Chall & Dale 1995; Schwarm & Ostendorf 2005; Vajjala & Meurers 2012). More recently, Chen & Meurers (2018: 2) compared several ways to aggregate word frequency and concluded: "High quality readability assessment depends on [...] a method for aggregating lexical frequency information that represents the distribution of word frequencies in a text more richly than using a single mean".

Recently, two advances in the field of NLP have allowed for more complex ways of aggregating information: the use of distributed representations of texts (e.g. embeddings) and deep learning models that can automatically organise distributed representations into higher-level features (Lee *et al.* 2009). In readability, Cha *et al.* (2017) showed, on the Common Core Standards cor-

pus, that bag-of-words indeed performed poorly, whereas using features from semantic spaces (e.g. word embeddings or Brown clustering) within a SVM model outperforms a model based on engineered features and mean aggregation. Whigham *et al.* (2018) built on this idea, adding a transitional network over the clustered word2vec vectors to better capture structural characteristics of texts. Nadeem & Ostendorf (2018) compared two deep learning models: the Gated Recurrent Unit (GRU) and hierarchical recurrent neural networks (RNN) with various attention mechanisms. Interestingly, none of their neural networks was able to beat their baseline engineered model by Vajjala & Meurers (2014). Based on a large set of experiments, Martinc *et al.* (2021) report that "neural classifiers can either outperform or at least compare with approaches leveraging extensive feature engineering". Finally, we should also mention the work by Filighera *et al.* (2019), who compared various generations of language models, from word2vec and GloVe to ELMO (Peters *et al.* 2018) and BERT (Devlin *et al.* 2019).

One reason that could explain why engineered-based models remain competitive for readability would be that, despite the popularity of word embeddings, they might not be able to encode all properties of a text that matter for readability prediction. Jiang *et al.* (2018) first investigated this issue with their knowledge-enriched word embedding approach that, roughly said, integrates age-of-acquisition, word frequencies, and word length within a knowledge graph and combines it with word embedding. Similarly, Le *et al.* (2018) combined word embedding representing word meanings with frequency embeddings (based on word frequencies) and found that the latter indeed help improve model's performance. This is the reason why, in this paper, we will investigate both models based on distributed representations such as BERT and models combining engineered features.

2.2 L2 readability

Our work also focuses on L2 French readability, a domain that has not yet been investigated in light of the two new advances described above: distributed representations and deep learning.

In general, L2 readability has been overlooked, with most formulas being based on L1 data applied to L2 contexts (François 2011). This situation is due to the belief, until the 1970s, that reading processes were deemed universal regardless of the readers background. We now know it is not the case and Koda (2005: 7) highlights three main differences between L1 and L2 readers as follows: (1) L2 readers must deal with interferences from their L1; (2) they are generally older and therefore have a better conceptual and world knowledge;

(3) they cannot rely on linguistic knowledge acquired orally as native speakers do. Such differences are likely to require different readability models to be accounted for. Tharp (1939) might be the first one to have criticized the use of L1 readability formulas for L2 learners, partly because of their inability to capture the L1-L2 lexical interferences (i.e. cognates). However, using L1 readability formulas for a L2 audience was not really challenged before the work of Hamsik and Brown in the 1980s (Greenfield 2004: 8-10). But if Brown's experiment did suggest L1 formulas were inappropriate for L2 learners, Greenfield (2004) reached the opposite conclusion as difficulties encountered by native speakers and ESL learners on his corpus texts were relatively similar.

One limitation of Greenfield (2004)'s study is that it does not use any features specific to the L2 context, such as the pedagogical progression defined in L2 curriculum, cognates or other types of interferences. With the rise of AI readability, more scholars have designed readability models for L2 readers and more features were considered. Heilman *et al.* (2007) proposed a model for L2 English that takes into account the specific pedagogical progression of 22 syntactic structures. Weir & Anagnostou (2008) and François & Watrin (2011) focused their efforts on multi-word expressions, known to be difficult to acquire for L2 learners. On another level, Crossley *et al.* (2008) investigated the effect of cognitive variables for L2 readability, with one measure of textual cohesion in particular. Beinborn *et al.* (2014) further explored lexical readability by capturing cognates within a readability formula and confirming their effect. For Swedish, Pilán *et al.* (2016) combined 61 variables (length-based, lexical, morphological, syntactic and semantic) and were able to outperform a classic formula for Swedish, LIX. Finally, Xia *et al.* (2016) proposed a generalization method to train L2 readability models on L1 data. To our knowledge, they were also the first to use features based on the official CEFR reference level descriptors, namely the English vocabulary profile.

As regards French readability, a limited amount of studies has been published. The most popular L1 formula might be the adaptation of Flesch formula to French by Kandel & Moles (1958). Even though more recent models have been proposed (Henry 1975; Mesnager 1989; Daoust *et al.* 1996; Dascalu 2014; Blandin *et al.* 2020), none of them are available to the public. For French L2, the situation is the same. Cornaire (1988) showed that L1 Henry's formulas could be applied to L2 readers. Uitdenbogerd (2005) developed a formula too, including cognates along with sentence length, but had little data to train it. More recently, François & Fairon (2012) proposed the first AI readability for FFL, leveraging NLP tools to capture richer linguistic information in the texts, such as tense distribution, multi-word expressions, N-grams frequencies, etc. However, none of these formulas are freely accessible to FFL teachers or

researchers, nor do they try to capture the different stages of FFL learners' interlanguage, for instance using the CEFR reference level descriptors as done by Xia *et al.* (2016).

In this paper, we will propose a new readability formula for FFL, based on the later advances in NLP such as distributed representations and deep learning, but we will also assess the effect of two types of variables that were overlooked in FFL readability, namely cognitive variables and pedagogical variables.

3. METHODOLOGY

3.1 The Data

Training a readability model requires a corpus in which the reading difficulty of each text has been evaluated according to a reference scale. For FFL purposes, an obvious choice was the beginner/intermediate/advanced continuum, redefined in the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) as the six following levels: A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). This scale has become the reference for foreign language teaching and testing in Europe, and all pedagogical materials published after 2001 should be evaluated using this scale. This characteristic can be leveraged to build CEFR-annotated corpus, as done earlier by François & Fairon (2012). Reapplying this methodology, we expanded the original corpus to create FLE-CORP, a larger and more diverse collection of texts found in FFL pedagogical material. For validation purposes, we also got access to a corpus of simplified readers used to build the FLELex resource (François *et al.* 2014), which is briefly introduced at the end of this section.

The FLE-CORP *Corpus*

The first step to build this corpus was to collect new material while applying the same selection criteria as in François & Fairon (2012): (1) textbooks have to be published after 2001 and have a CEFR level and (2) they must be intended for adults or teenagers learning FFL for general purposes. The original corpus of François & Fairon (2012) (called FF-2012), which we were able to reproduce, is composed of 1793 texts taken from 27 textbooks published between 2001 and 2008. For this study, 20 new textbooks ranging from levels A1 to C2 were collected, from which we similarly extracted texts related to a reading comprehension task only. Selected texts were scanned with optical character recognition tools and manually revised. We then assigned to

each text the CEFR level of the book it came from, which had therefore been given by experts.[1] At the end of this process, we were able to expand *FF-2012* by 2,769 texts, totaling 4,562 texts distributed across the 6 CEFR levels and among 8 text types. These types are *narrative*, *informative*, *text* (other), *dialogue*, *mail/e-mail*, *sentence* (short unauthentic series of sentences to fit in a specific lesson, usually found in A levels textbooks), *gap filling exercise*, and a *varia* category including recipe, poetry, song, advertisement, etc. The whole corpus presents a major weakness: it is heavily unbalanced towards the beginner levels, which also include more shorter texts. We thus carried out two operations in order to correct this unbalanced distribution: merging C1 and C2 levels and undersampling to balance the categories.

CEFR descriptors for the C1 and C2 levels are similar with regards to the type of texts to be understood, namely long and complex texts including literary works, specialized articles and manuals or long technical instructions. In fact, both levels usually share one textbook in which modules can be studied in any order (e.g. *Cosmopolite 5*, *Tendances C1/C2*) and texts are sometimes graded as C1/C2. The main difference between both levels lies in the reading context: while C2 learners are expected to read any kind of text unassisted, C1 learners can understand them "provided there are opportunities for rereading and they have access to reference tools" (Council of Europe 2001). In other words, linguistic characteristics of these texts, that readability formulas should model, do not matter as much as the condition in which learners are reading. As a result, we decided to merge C1 and C2 level into a single C level for the rest of the paper.

As mentioned above, the corpus is unbalanced, with more than a thousand texts at the A1, A2 and B1 levels and only 467 and 580 texts at B2 and C levels, respectively. We therefore decided to undersample it in order to get more balanced categories. Before undersampling, we carried out an outlier analysis, based on two principles. First, it is acknowledged that short texts are an issue for automatic readability assessment. Therefore, we rejected 514 very short texts made of less than 20 words in A1 or 30 words in A2 to C levels. Most of these texts were of the *sentence* genre, meaning that we mainly removed short reading exercises. Second, some documents in the *varia* genre, such as poems, songs, recipes, website pages or classified ad, have atypical characteristics (e.g. absence of punctuation, overuse of abbreviations, etc.) that are prone to disrupting readability models. To identify and root these texts out, we applied a standard outlier detection methodology. We first selected five variables, each one representative of a category of text features (lexical, syntactic,

[1] For the shortcomings of this approach, see François (2014).

discursive, cognitive and pedagogical). Then, for each distribution of scores corresponding to one CEFR level and one of the features, we transformed the distribution into z-scores and removed the texts for which at least one of the features was located beyond three standard deviations from the mean. Consequently, 97 texts were dismissed, most of which belonging to the *varia* type, as expected.

Finally, we performed a random sampling on levels A1 to B1 in order to reduce their number of texts to 580 and obtain a more balanced dataset. The resulting corpus, called FLE-CORP in the rest of the paper, includes 2,751 texts, for a total of 566,676 words, as described in Table 1.

TEXT TYPE	A1	A2	B1	B2	C	ALL LEVELS
text	98	130	254	100	281	863
narrative	21	39	28	28	55	171
informative	62	95	84	66	107	414
dialogue	105	51	25	15	0	196
mail/e-mail	43	31	25	24	12	135
sentence	106	64	26	56	42	294
blank sentence	22	16	12	14	2	66
varia	123	154	126	139	70	612
WORDS	61,159	82,598	113,588	122,722	186,609	566,676
words per text	105	142	196	278	328	206
TEXTS	580	580	580	442	569	2751

TABLE 1: FLE-CORP DESCRIPTION.

Validation Corpus

Due to the lack of available datasets, the performance of most readability formulas published so far has been estimated on the same corpus used to train them, using a cross-validation scheme for the most recent ones. Since such procedure is likely to overestimate the models generalization capabilities, we decided to use another corpus presenting slightly different characteristics than the FLE-CORP. This validation corpus, called GRADED READERS, has not been used for readability yet, but was collected as part of the FLELex project by François *et al.* (2014). It consists of 29 simplified readers, labelled according to the CEFR from level A1 to B2 and split according to their chapters. The resulting 317 texts are distributed as follows: 71 A1, 114 A2, 84 B1, and 48 B2. The total number of words amounts to 244,604. It should be mentioned that as texts were extracted from readers, most of them are narrative. In addition, the difficulty of simplified readers is generally much more controlled

than excerpts used in textbooks, since authors use pedagogical guides such as vocabulary lists to write them. We can therefore expect Graded Readers to be more homogeneous than FLE-Corp.

3.2 Text Features

In this section, we describe the five sets of features used in our experiments. Lexical, syntactic and discursive features are first described, then we introduce the cognitive and pedagogical feature sets. These three sets of features can be referred to as engineered features, as they are defined by researchers based on theoretical or empirical knowledge about their effect on reading (e.g. word length, conceptual density). As mentioned previously, this paper also investigates the use of distributed representations, which are introduced later in this section. Finally, we also used log-term frequency features, used in previous work, for comparison purposes.

Lexical, Syntactic and Discursive Features

As the main set of engineered features, we decided to replicate 43 features out of the 46 listed in (François & Miltsakaki 2012: 53), hereafter refer to as the "LSD" features. These features capture together lexical, syntactic, and discursive information about the text. They are briefly introduced in the Appendices section, to which the reader can also refer to find the explanation for the names of variables used in this section. Three variables out of the 46 original ones were not replicated : they are the ones based on multi-word expressions variables (NCPW, NAColl, and MeanNGProb-G). They were dropped from the model as there were not possible for us to reproduce and did not bring much information to the model anyway (François & Watrin 2011).

Cognitive Features

Regarding cognitive features, we relied on psycholinguistic studies to identify 3 types of variables that were never used for French L2 so far.

1. The proportion of abstract and concrete words

   Neuroscientists and psycholinguists have shown that abstract and concrete words are processed differently in the brain (Paivio 1991; Just *et al.* 2004) and that concreteness gives an advantage in recognition and recall tasks due to their higher degree of imageability (Jessen *et al.* 2000; Steacy & Compton 2019). Goriachun & Gala (2020) recently automatically developed a list of 7,898 abstract and concrete French nouns, ob-

tained through the collection of the nearest neighbors and syntactic co-occurrences of 61 seed concrete and abstract words in the *Les Voisins De Le Monde* lexical database. Based on this list, the largest available for French, we derived 3 variables: the proportion of abstract words, the proportion of concrete words and the text coverage of the list.

2. The average OLD20 and PLD20 per word

   The *Orthographic Levenshtein Distance 20* (OLD20) is a measure introduced by Yarkoni *et al.* (2008) to improve Coltheart's orthographical neighborhood index, which has been widely used in psycholinguistic studies to reflect the interference produced by a high number of orthographic neighbors in visual word recognition tasks and first applied to readability by François & Fairon (2012). In order to go beyond Coltheart's definition of orthographic neighbors (i.e. words that can be produced by changing a single letter in a word of the same length), Yarkoni *et al.* (2008) proposed, for a given word, to compute its Levenshtein distance to all the other words in the lexicon, no matter their length. The OLD20 is then obtained by calculating the average Levenshtein distance of the 20 closest words found in the lexicon. The authors showed that their new measure explains more variance than the number of orthographic neighbors. PLD20 stands for *Phonological Levenshtein distance* and adapts OLD20 at the phonemic level. For French, New & Pallier (2019) applied the same process on the 125,653 orthographically different entries of the *Lexique 3.6* database. We used their work to compute the average OLD20 and PLD20 per word in each text of our corpus.

3. The average number of commonly known senses per word

   Polysemy is an important factor in reading comprehension as it requires the reader to select the appropriate meaning of a word according to its context. There is a consensus in scientific literature about semantic plurality inhibiting word recognition in context (Hino *et al.* 2002) but the question of how to parametrize this effect is still open. Indeed, simply counting the number of meanings per word in a dictionary has shown its inadequacy, since a common reader does not usually know all existing meanings of the words (Gernsbacher 1984). Measuring lexical polysemy as the number of commonly known senses per word better reflects the psychological reality of this factor and several methods have been tested to do so (Millis & Button 1989). An automatic approach explored by François *et al.* (2016) for French language is to derive information from the large semantic network *BabelNet* (Navigli & Ponzetto 2012),

in order to detect all possible meanings of each word, before merging the similar ones together. In this process, highly specialized senses are left out while a reduced number of distinct and commonly known meanings comes forth,[2] resulting in this case in a list of 23,242 French words automatically annotated with their level of commonly known polysemy. For our study, we used this lexical resource as a variable representing the average number of commonly known senses per word.

Pedagogical features

We also designed 5 variables based on the pedagogical knowledge provided by the official Reference Level Descriptors (RLD) for French (Beacco *et al.* 2008). This pedagogical resource is based on the CEFR framework and presents a selection of linguistic forms (e.g. vocabulary, syntactic structures, phonemes, etc.) to be mastered at each level. The classification of linguistic forms to a given level was performed based on criteria selected by the authors for their relevance and objectivity: essentially the official descriptors from the CEFR, collective knowledge and experience of experts, teachers and examiners, and examples of learner productions deemed to be at a particular level (Beacco *et al.* 2008). We used the RLD chapters 4 and 6, relative to lexicon, since they prove to be a valuable source of information when trying to infer the CEFR level of words from their frequency in FFL material (Pintard & François 2020). Those chapters share the same architecture across levels A1 to B2,[3] organizing words within semantic categories, specifying the part-of-speech (POS) categories and sometimes providing a context. Polysemous words can have up to 8 entries across the four levels (e.g. "être", *to be*). However, as these semantic categories diverge from sense inventories used in word sense disambiguation systems, we decided to ignore the information on semantic category from the French RLDs. When a form is linked to more than one CEFR levels, we kept the lowest one. This method led us to drop about 2,968 entries, going from 8,486 to 5,518 entries. We then defined 5 pedagogical features from the RLDs, namely the proportion of words from the text assigned to A1, A2, B1 and B2 levels by Beacco and his collaborators, and the proportion of words not covered by the RLDs, supposedly being the most difficult ones.

[2] The method used to reduce the number of meanings per entries is refered to as NASARI: *a Novel Approach to a Semantically–Aware Representation of Items* (Camacho-Collado *et al.* 2015).

[3] The RLD for C levels does not contain information relative to lexicon like the other ones.

Features based on Distributed Representations

As a fourth source of information, we also selected features based on distributed representations contributed by Deep Learning. These representations come in the form of densely-encoded embedding vectors that capture semantic and syntactic information from the text. These are learned by training Deep Learning models on large corpora. In this paper, we use two kinds of embeddings that are pervasive in modern NLP:

fastText – fastText provides state-of-the-art word embeddings (Bojanowski *et al.* 2016). One of its key innovations was to incorporate sub-word information to facilitate representation sharing across words, enabling it to efficiently learn useful representations, even for very rare words.

BERT – Pre-Training for Deep Bidirectional Transformers (BERT) is a state-of-the-art language model which represents tokens as contextual word embeddings (Devlin *et al.* 2019). Because the embeddings are contextual, the special end token captures semantic and syntactic information of the entire passage, and thus can be used as an embedding to represent that passage. In our case, we extract embeddings for sentences, and then take the mean across sentences to produce passage-level embeddings.[4] We use a pre-trained BERT model named CamemBERT (Martin *et al.* 2019) specially trained for French.

Log Term Frequency

For a baseline model, we also use log-term frequency, i.e. a set of features representing the log of the frequency with which each term occurs in the passage. This is similar to unigram features used in previous works (Vajjala & Lučić 2018). These features are motivated by the fact that vocabulary is very important to readability (Hu & Nation 2000), and that these kinds of features are very easy to implement.

3.3 Models

In our experiments, we evaluate the three models listed below:

Linear Support Vector Machine (SVM) – Support Vector Machines (SVMs) are a commonly used classifier that work by attempting to maximize the margin between the classification boundary and the nearest labeled data

[4] We found that this produced slightly better results than extracting one embedding for the entire passage, and also avoids the need to truncate long passages.

points. We use a linear kernel with L2 regularization, optimized via a hyperparameter sweep over the values $\sqrt{0.125}$, 0.25, $\sqrt{0.25}$, 0.5, $\sqrt{0.5}$, 1, $\sqrt{2}$, 2, $\sqrt{8}$, 4, $\sqrt{32}$, 8, $\sqrt{128}$, and 16.

Hierarchical Attention Network (HAN) – Hierarchical Attention Networks are a Deep Learning Architecture used for document classification, originally proposed by Yang *et al.* (2016) and was first evaluated for its applications to reabability by Nadeem & Ostendorf (2018). These networks aggregate sequences of embeddings using two layers: an encoding layer using a recurring network structures such as RNN, LSTM, GRU, or bi-directional GRU (as in our case), and an attention layer that aggregates the encoded elements into a fixed-length embedding. In this way, HANs aggregate sequences of token embeddings into sentence embeddings, and sequences of sentence embeddings into document embeddings. The document embedding is then fed into a linear classifier to make the prediction. For our purposes, an advantage of a HAN model is that it can incorporate arbitrary features designed for readability, while, unlike the SVM model, also modeling non-linear relationships and interactions between features.

Fine-Tuned BERT – BERT is a state-of-the-art language model that is frequently used to encode passages into embeddings for downstream tasks, as we described in subsection 3.2. However, a pre-trained BERT model can also be fine-tuned by training it directly on labeled documents. This is a form of transfer learning and enables more tailored feature extraction than can be achieved by just using embeddings extracted from the pre-trained model alone. Since the model's weights are initialized via pre-training on large corpora, it can often be trained for a particular classification task with very small corpora despite its numerous parameters. As before, we use CamemBERT (Martin *et al.* 2019) as implemented by Wolf *et al.* (2020).[5]

3.4 Metrics

To evaluate the performance of our different models, we applied three metrics, commonly used in readability:

Accuracy (Acc.) – The proportion of passages that were correctly classified by the model.

[5] Note: BERT models are limited to 512 tokens, and so passages longer than this are truncated.

Adjacent Accuracy (Adj. Acc.) – The proportion of passages that were correctly classified plus those classified within 1 class of the correct class.

Pearson Correlation (R) – The Pearson correlation between (i) weighted average computed from the predicted level probabilities of each document and (ii) the true class.

4. EXPERIMENTS

In the following subsections, we present three experiments. In subsection 4.1, we explore the utility of various features for FFL readability. In subsection 4.2, we explore the use of Deep Learning models instead of linear SVMs. In subsection 4.3 we run cross-corpus experiments to evaluate how well these models will generalize to texts drawn from sources other than those the model was trained on.

4.1 Experiment 1: Features

In this experiment, we explore the utility of the different types of features we implemented, including cognitive, pedagogical, and deep learning features. All features are evaluated with SVM models (see subsection 3.3), using 10-fold cross validation on the FLE-CORP corpus. Results are shown in Table 2.

For this experiment, we had three baselines. The first is the majority class baseline, which simply shows what can be trivially achieved by predicting all texts to be assigned to the most common class. It serves as a reference for the difficulty of the classification task. The second baseline model includes the 43 lexical, syntactic, and discursive features described in section 3.2 (LSD) and reproduces the model by François & Fairon (2012). It achieves 50 % accuracy and 82 % adjacent accuracy, which is in line with the results published by François & Fairon (2012) – 49 % accuracy and 80 % adjacent accuracy – on a much smaller corpus. Our third baseline uses only the log term-frequency feature set described in section 3.2. This baseline appears as a particularly strong one, as it increases accuracy by 3 percentage points and adjacent accuracy by 1 point compared to the LSD baseline and ends up being the second most efficient set of features on the FLE-CORP.

From here, we explore whether adding cognitive and/or pedagogical features improve the LSD baseline model's accuracy, shown in the second group of rows in Table 2. We evaluate each feature type separately, and in combination with other types of features. Evaluating each individually, we find that both are strong predictors of readability, achieving an accuracy of 35 %

features	FLE-CORP		
	Acc.	Adj. Acc.	R
Majority Class	0.21	0.42	-
Lex., Syn., Disc. (LSD)	0.50	0.82	0.71
Log Term-Frequency	0.53	0.83	0.71
Cognitive (Cog.)	0.35	0.67	0.44
Pedagogical (Ped.)	0.40	0.69	0.52
Cog. + Ped.	0.42	0.74	0.57
LSD + Cog.	0.50	0.82	0.70
LSD + Ped.	0.50	0.83	0.71
LSD + Ped. + Cog.	0.51	0.82	0.70
fastText	0.50	0.82	0.70
BERT	0.51	0.84	0.71
LSD + Ped. + Cog. + BERT	**0.56**	**0.87**	**0.77**

TABLE 2: FEATURE EXPERIMENT RESULTS.

and 40 %, respectively. However, when added with LSD features, they fail to substantially improve that baseline's metrics.

Finally, we explore whether using Deep Learning distributed representations as features improves performance. Surprisingly, both word-level fastText embeddings and contextual BERT embeddings achieve similar results as the LSD baseline. When combined with the previously evaluated engineered features, they achieve a modest lift in all metrics, achieving 56 % accuracy and 87 % adjacent accuracy .

4.2 Experiment 2: Model Architecture

In the previous experiment, we explored the contribution of various types of features within the context of linear SVM models that are commonly used in readability formula. In this section, we explore whether moving to Deep Learning gives additional gains over what can be achieved with linear SVM models. Specifically, we evaluate the two deep learning models described in Section 3.3. We compare these to two previous baselines and the best performing SVM: the one combining LSD, pedagogical, cognitive, and BERT features. We also compare to the results of the SVM model using the fastText embedding, for a more direct comparison to the HAN model.

The results are shown in Table 3. First, comparing the fastText SVM to the HAN model, we find that the architecture of the Hierarchical Attention Model (HAN) achieves modestly better results than the linear SVM using the same embedding features (i.e., fastText, SVM). However, we see in the last line

Features	FLE-CORP		
	Acc.	Adj. Acc.	R
LSD SVM	0.50	0.82	0.71
Log Term-Frequency SVM	0.53	0.83	0.71
fastText SVM	0.50	0.82	0.70
LSD + Ped. + Cog. + BERT	0.56	0.87	0.77
Hierarchical Attention (HAN)	0.53	0.85	0.75
Fine-Tuned BERT	**0.58**	**0.90**	**0.82**

TABLE 3: MODEL EXPERIMENT RESULTS.

that Fine-Tuned BERT does better than all SVM baselines, and dramatically better than the previously state-of-the-art LSD baseline, raising accuracy by 8 percentage points from 50 % to 58 %, accompanied by similar increases in the other metrics. Interestingly, Fine-Tuned BERT also outperforms BERT in accuracy by 7 percentage points, confirming that tuning it even on a relatively small number of texts can help a lot for readability.

4.3 Experiment 3: Generalization

In this final experiment, we attempt to evaluate how well the models generalize to texts sampled from sources other than those used in the training corpora, but on which the model could be validly applied. Textbooks are a convenient source of CEFR-labeled passages, but their content is also determined by the pedagogical progression and by commercial constraints specific to each publisher or each textbook series. It is therefore possible that the models we have evaluated may overfit to these patterns, and thus will not generalize well to texts sampled from other sources. To determine each model's generalizeability to other sources of texts, we evaluate them using two new setups:

Series-Split – The FLE-CORP includes as much as 47 textbooks, but they belong to a limited amount of series, each of which is likely to be characterized by specific features. In this setup, we evaluated the models on FLE-CORP using cross validation as before, but with the folds sampled in a way that ensures that passages from the same series never appear in *both* the train and test dataset. In effect, we train on a set of textbooks series, and evaluate on a different set of textbook series. This ensures that the model's metrics cannot be inflated by relying on statistical cues that are only applicable to a given textbook series.

Graded Readers – In this setup, we do a cross corpus experiment where the model is trained on the full FLE-CORP corpus, and is then evaluated on the GRADED READERS corpus described at subsection 3.1. This is perhaps the most robust of our evaluations, as the train and test passages come from very different kinds of books.

The results are shown in Table 4. We see that not all the models that performed well in the previous evaluations performed well in this evaluation. In particular, the Log Term-Frequency model, that was our second best model in the first experiment, ends up among the worst models, indicating that it is prone to overfitting on the specific vocabulary of the training corpus. Interestingly, we can also notice that, unlike in experiment 1, adding BERT features to the engineered features failed to improve the performance of the models using engineered features alone. Fine-Tuned BERT, on the other hand, continues to perform considerably better than all other models.

	SERIES-SPLIT			GRADED READERS		
Features	Acc.	Adj. Acc.	R	Acc.	Adj. Acc.	R
Most Common Class	0.03	0.20	-0.47			
Lex., Syn., Disc. (LSD)	0.44	0.80	0.66	0.42	0.82	0.56
Log Term-Frequency	0.39	0.78	0.59	0.28	0.68	0.32
LSD + Cog.	0.44	0.79	0.66	0.41	0.81	0.55
LSD + Ped.	0.45	0.80	0.67	0.44	0.81	0.54
LSD + Ped. + Cog.	0.44	0.78	0.64	0.42	0.83	0.55
fastText	0.40	0.76	0.60	0.41	0.83	0.46
BERT	0.38	0.79	0.61	0.42	0.83	0.37
LSD + Ped. + Cog. + BERT	0.43	0.81	0.6	0.43	0.82	0.42
Hierarchical Attention (HAN)	0.44	0.79	0.66	0.32	0.86	0.52
Fine-Tuned BERT	**0.48**	**0.86**	**0.74**	**0.50**	**0.94**	**0.59**

TABLE 4: MODEL EXPERIMENT RESULTS.

4.4 Analysis and Discussion

A first important result is that fine-tuning BERT appears to improve substantially on François & Fairon (2012), significantly beating the baseline in all three experimental setups we evaluated. Transfer learning seems to be its greatest advantage: despite its much larger number of parameters and small training corpus size, it was able to do better than using pre-trained BERT embeddings as features and the Deep Learning HAN model. A follow-up analysis about the

effect of text genres on the quality of predictions, detailed in Appendix III, revealed that fine-tuning BERT was also the most versatile model of all the ones tested in this paper. However, there may still be reasons to use the feature-engineering-based models. In particular, an advantage of linear models on engineered features is that they are easily interpretable, and this characteristic could be used to provide feedback on how to adjust the text to better suit a particular CEFR level.

We also showed that both cognitive and pedagogical features had strong predictive capability for this task. However, none of them were able to increase performance when combined with LSD features. We believe that this is due to the overlap in information with variables already in the LSD model. For instance, the feature *PA-Alterego*, which corresponds to the proportion of absent words from a list of A1 words found in the *Alter Ego* textbook, correlates as much as 0.66 with one of the pedagogical features, and between 0.25 and 0.44 with the others (more detailed information is available in Appendix II. For English, Xia *et al.* (2016) seemed to reach a similar conclusion, as their pedagogical features based on EVP were included in the lexical category that did not perform better than the traditional features. However, the pedagogical set, with only 5 features, remains a very efficient way to achieve decent results. It is worth stressing that they achieved better performance than the previously best-performing feature group (i.e., lexical features) from François & Fairon (2012), while having far fewer features (5 vs 16).

We also discovered that certain features appeared to do very well when assessed in a cross-validation setting, but failed to generalize adequately when evaluated on texts extracted from sources slightly different from the training one. In particular, the Log Term-Frequency features generalized very poorly. This may partly have to do with the large number of features involved in all of these models. It may also be that these features encode writing constraints applied to texts to keep them simple for L2 learners (e.g., internal word lists used by textbook publishers), which do not generalize to other sources. This is an important discovery, as readability models are often trained and evaluated on corpora extracted from a limited set of sources which are split between train and test or through cross validation, as we did in Experiment 1. This suggests that we need to be careful when evaluating L2 readability models, as high accuracy scores can be misleading if they are not tested across heterogeneous corpora, as already showed by Nelson *et al.* (2012) in another context.

5. CONCLUSION

This paper has reported a set of three experiments about French L2 readability in which, for the first time, the effect of distributed representations and deep learning models was evaluated for this task. In addition, we introduced several new features for French or French L2: on the one hand, a set of cognitive features aimed at capturing polysemy, the abstractness level of words, and the perceptual ambiguity during word recognition caused by close orthographic neighbours. On the other hand, we defined a set of 5 pedagogical features encoding learners' progression in lexicon acquisition according to the CEFR levels, as described in the French RLD published by Beacco and his colleagues.

Our experiments revealed that neither our new sets of features, nor distributed representations such as fastText or BERT were able to outperform the baseline by François & Fairon (2012). The model that combined these engineering and embedding features, the one using Log-Term Frequency, and the HAN model were all able to beat this baseline in the cross-validation setting, but failed to keep this advantage on two other datasets: the SERIES-SPLIT and the GRADED READERS. However, fine-tuning BERT on our train data allows us to clearly surpass the previously state-of-the-art readability model by François & Fairon (2012) by 8 percentage points for the accuracy and 8 percentage points for adjacent accuracy. As regards the pedagogical set of features, we should mention that we only selected the two chapters of the French RLD dedicated to vocabulary. Future investigations should focus on the parameterization of the other chapters, dedicated to communication functions, syntactic structures, etc.

Our new readability model for L2 French based on fine-tuning BERT will be included in a web interface,[6] freely accessible to the general public. We hope that this accessibility will foster the use of readability tools among L2 teachers, textbook publishers, and for language assessment purposes.

REFERENCES

Beacco, J.C., S. Lepage, R. Porquier & P. Riba (2008). *Niveau A2 pour le français: Un référentiel*. Paris, France: Didier.

Beinborn, L., T. Zesch & I. Gurevych (2014). Readability for foreign language learning: The importance of cognates. *ITL Journal*, 165(2). 136–162.

Belfiore, M., T. Defoe, S. Folinsbee, J. Hunter & N. Jackson (2004). *Reading work: Literacies in the new workplace*. London & New York: Routledge.

[6] `https://cental.uclouvain.be/dmesure/`

Benjamin, R.G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1). 63–88.

Blandin, A., G. Lecorvé, D. Battistelli & A. Étienne (2020). Recommandation d'âge pour des textes. In *Proceedings of TALN 2020*. 164–171.

Bojanowski, P., E. Grave, A. Joulin & T. Mikolov (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Bormuth, J. (1966). Readability: A new approach. *Reading research quarterly*, 1(3). 79–132.

Camacho-Collado, J., M.T. Pilehvar & R. Navigli (2015). Nasari: a novel approach to a semantically-aware representation of items. In *NAACL*. 567–577.

Cha, M., Y. Gwon & H. Kung (2017). Language modeling by clustering with word embeddings for text readability assessment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2003–2006.

Chall, J. & E. Dale (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, MA: Brookline Books.

Chen, X. & D. Meurers (2018). Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3). 486–510.

Coleman, E. (1965). On understanding prose: some determiners of its complexity. *NSF Final Report GB-2604. Washington, DC: National Science Foundation*.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL Journal*, 165(2). 97–135.

Collins-Thompson, K. & J. Callan (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13). 1448–1462.

Cornaire, C. (1988). La lisibilité : essai d'application de la formule courte d'Henry au français langue étrangère. *Canadian Modern Language Review*, 44(2). 261–273.

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

Crossley, S., J. Greenfield & D. McNamara (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3). 475–493.

Dale, E. & J. Chall (1948). A formula for predicting readability. *Educational research bulletin*, 27(1). 11–28.

Daoust, F., L. Laroche & L. 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.

Dascalu, M. (2014). Readerbench (2)-individual assessment through reading strategies and textual complexity. In *Analyzing Discourse and Text Complexity for Learning and Collaborating*, 161–188. Dordrecht, NL: Springer.

Day, R. & J. Bamford (2002). Top ten principles for teaching extensive reading1. *Reading in a Foreign Language*, 14(2).

Dell'Orletta, F., S. Montemagni & G. Venturi (2014). Assessing document and sentence readability in less resourced languages and across textual genres. *ITL-International Journal of Applied Linguistics*, 165(2). 163–193.

Devlin, J., M.W. Chang, K. Lee & K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL 2019*. 4171–4186.

DuBay, W. (2004). *The principles of readability*. Impact Information. Disponible sur http://www.nald.ca/library/research/readab/readab.pdf.

Filighera, A., T. Steuer & C. Rensing (2019). Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*. Dordrecht, NL: Springer, 335–348.

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

François, T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Ph.D. thesis, Université Catholique de Louvain. Thesis Supervisors: Cédrick Fairon and Anne Catherine Simon.

François, T. (2015). When readability meets computational linguistics: a new paradigm in readability. *Rev. française de linguistique appliquée*, 20(2). 79–97.

François, T. & C. Fairon (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of EMNLP 2012*. 466–477.

François, T. & E. Miltsakaki (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of PITR 2012)*.

François, T. & P. Watrin (2011). On the contribution of MWE-based features to a readability formula for French as a foreign language. In *Proceedings of RANLP 2011*.

François, T., N. Gala, P. Watrin & C. Fairon (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of LREC 2014*. 3766–3773.

François, T. (2014). An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the NLP4CALL workshop 2014, NEALT Proceedings Series Vol. 22, Linköping Electronic 107*. 13–32.

François, T., M. Billami, N. Gala & D. Bernhard (2016). Bleu, contusion, ecchymose: triautomatique de synonymes en fonction de leur difficulté de lecture et compréhension. In *JEP-TALN-RECITAL*. Paris, France, 15–28.

Gernsbacher, M. (1984). Resolving 20 years of inconsistent interactions between lexi-cal familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology : General*, 113.

Goriachun, D. & N. Gala (2020). Identifying abstract and concrete words in French to better address reading difficulties. In E.L.R. Association (ed.) *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*. Marseille, France, 33–40.

Gougenheim, G., R. Michéa, P. Rivenc & A. Sauvageot (1964). *L'élaboration du français fondamental (1er degré)*. Paris, France: Didier.

Grabe, W. (2014). Key issues in l2 reading development. In *Proceedings of the 4th CELC Symposium for English Language Teachers-Selected Papers*. 8–18.

Gray, W. & B. Leary (1935). *What makes a book readable*. Chicago, IL: University of Chicago Press.

Greenfield, J. (2004). Readability formulas for EFL. *Japan Association for Language Teaching*, 26(1). 5–24.

Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill.

Heilman, M., K. Collins-Thompson, J. Callan & M. Eskenazi (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL 2007*. 460–467.

Henry, G. (1975). *Comment mesurer la lisibilité*. Bruxelles: Labor.

Hino, Y., S. Lupker & P. Pexman (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: Interactions between orthography,phonology, and semantics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28.

Hu, M. & P. Nation (2000). Unknown vocabulary density and reading comprehension. *Reading in a foreign language*, 13(1). 403–30.

Jeon, E.Y. & R.R. Day (2016). The effectiveness of er on reading proficiency: A meta-analysis. *Reading in a Foreign Language*, 28(2). 246–265.

Jessen, F., R. Heun, M. Erb, D.O. Granath, U. Klose, A. Papassotiropoulos & W. Grodd (2000). The concreteness effect: Evidence for dual coding and context availability. *Brain and Language*, 74. 103–112.

Jiang, Z., Q. Gu, Y. Yin & D. Chen (2018). Enriching word embeddings with domain knowledge for readability assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*. 366–378.

Just, M.A., S.D. Newman, T.A. Keller, A. McEleney & P.A. Carpenter (2004). Imagery in sentence comprehension: An fmri study. *NeuroImage*, 21. 112–124.

Kandel, L. & A. Moles (1958). Application de l'indice de Flesch à la langue française. *Cahiers Études de Radio-Télévision*, 19. 253–274.

Kemper, S. (1983). Measuring the inference load of a text. *Journal of Educational Psychology*, 75(3). 391–401.

Kintsch, W. & D. Vipond (1979). Reading comprehension and readability in educational practice and psychological theory. In L. Nilsson (ed.) *Perspectives on Memory Research*, 329–365. Hillsdale, NJ: Lawrence Erlbaum.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge. UK: Cambridge University Press.

Laufer, B. & G. Ravenhorst-Kalovski (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a foreign language*, 22(1). 15–30.

Le, D.T., C.T. Nguyen & X. Wang (2018). Joint learning of frequency and word embeddings for multilingual readability assessment. In *Proceedings of NLPTEA 2018*. 103–107.

Lee, H., R. Grosse, R. Ranganath & A.Y. Ng (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*. 609–616.

Lee, H., P. Gambette, E. Maillé & C. Thuillier (2010). Densidées: calcul automatique de la densité des idées dans un corpus oral. In *Proceedings of RECITAL 2010*.

Lively, B. & S. Pressey (1923). A method for measuring the "vocabulary burden" of textbooks. *Educational Administration and Supervision*, 9. 389–398.

Martin, L., B. Muller, P.J.O. Suárez, Y. Dupont, L. Romary, É.V. de la Clergerie, D. Seddah & B. Sagot (2019). Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

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

Mesnager, J. (1989). Lisibilité des textes pour enfants: un nouvel outil? *Communication et Langages*, 79. 18–38.

Millis, M. & S. Button (1989). The effect of polysemy on lexical decision time :nowyou see it, now you don't. *Memory & Cognition*, 17.

Nadeem, F. & M. Ostendorf (2018). Estimating linguistic complexity for science texts. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*. 45–55.

Navigli, R. & S.P. Ponzetto (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193. 217–250.

Nelson, J., C. Perfetti, D. Liben & M. Liben (2012). Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Student Achievement Partners*.

New, B. & C. Pallier (2019). Manuel de lexique 3.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45. 255–287.

Peters, M.E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee & L. Zettlemoyer (2018). Deep contextualized word representations. In *Proceedings of NAACL 2018*.

Pilán, I., S. Vajjala & E. Volodina (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. 7(1). 143–159.

Pintard, A. & T. François (2020). Combining expert knowledge with frequency information to infer cefr levels for words. In *Proceedings of the workshop READI, LREC 2020*. 85—-92.

Pitler, E. & A. Nenkova (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP 2008*. 186–195.

Schwarm, S. & M. Ostendorf (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of ACL 2005*. 523–530.

Si, L. & J. Callan (2001). A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*. ACM New York, NY, USA, 574–576.

Smith, E. & R. Senter (1967). Automated Readability Index. Technical report, AMRL-TR-66-220, Aerospace Medical Research Laboratories, Wright-Patterson Airforce Base, OH.

Steacy, L. & D. Compton (2019). Examining the role of imageability and regularity in word reading accuracy and learning efficiency among first and second graders at risk for reading disabilities. *Journal of Experimental Child Psychology*, 178. 226–250.

Tharp, J. (1939). The Measurement of Vocabulary Difficulty. *Modern Language Journal*. 169–178.

Uitdenbogerd, S. (2005). Readability of French as a foreign language and its uses. In *Proceedings of the Australian Document Computing Symposium*. 19–25.

Vajjala, S. (2021). Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv:2105.00973*.

Vajjala, S. & I. 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*. New Orleans, Louisiana: Association for Computational Linguistics, 297–304.

Vajjala, S. & D. Meurers (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*. 163–173.

Vajjala, S. & D. Meurers (2014). Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2). 194–222.

Weir, G. & N. Anagnostou (2008). Collocation frequency as a readability factor. In *Proceedings of the 13th Conference of the Pan Pacific Association of Applied Linguistics*.

Whigham, P., M. Chugh & G. Dick (2018). Measuring language complexity using word embeddings. In *Australasian Joint Conference on Artificial Intelligence*. Springer, 843–854.

Willms, J. (2003). Literacy proficiency of youth: Evidence of converging socioeconomic gradients. *International Journal of Educational Research*, 39(3). 247–252.

Wolf, T., L. Debut & et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP2020: System Demonstrations*. 38–45.

Xia, M., E. Kochmar & T. Briscoe (2016). Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. 12–22.

Yang, Z., D. Yang, C. Dyer, X. He, A. Smola & E. Hovy (2016). Hierarchical attention networks for document classification. In *Proceedings of NAACL 2016*. San Diego, California: Association for Computational Linguistics, 1480–1489.

Yarkoni, T., D. Balota & M. Yap (2008). Moving beyond coltheart's n: A new measure of orthographic similarity. *Psychonomic bulletin & review*, 15. 971–979.

Kevin Yancey
Duolingo
5900 Penn Ave, Second Floor, Pittsburgh, PA 15206
USA
e-mail: kyancey@duolingo.com

Alice Pintard
UCLouvain
Place Montesquieu, 3, 1348 Louvain-la-Neuve
Belgium
e-mail: alice.pintard@uclouvain.be

Thomas François
UCLouvain
Place Montesquieu, 3, 1348 Louvain-la-Neuve
Belgium
e-mail: thomas.francois@uclouvain.be

APPENDIX I: FEATURES IN THE LSD BASELINE

FAMILY	TAG	DESCRIPTION OF VARIABLE
Lexical	PA-Alterego	Proportion of absent words from a list of easy words (*AlterEgo1*)
	Unigram	Probability of the text sequence based on unigrams
	X90FFFC	90^{th} percentile of inflected forms for content words only
	X75FFFC	75^{th} percentile of inflected forms for content words only
	PA-Goug2000	Prop. of absent words from 2000 first of Gougenheim *et al.* (1964)'s list
	MedianFFFC	Median of the frequencies of inflected content words
	PM8	Pourcentage of words longer than 8 characters
	NL90P	Length of the word corresponding to the 90^{th} percentile of word lengths
	NLM	Mean number of letters per word
	IQFFFC	Interquartile range of the frequencies of inflected content words
	MeanFFFC	Mean of the frequencies of inflected content words
	TTR	Type-token ratio based on lemma
	FCNeigh75	75^{th} percentile of the cumulated frequency of neighbors per word
	MedNeigh+Freq	Median number of more frequent neighbor for words
	Neigh+Freq90	90^{th} percentile of more frequent neighbor for words
Syntactic	NMP	Mean number of words per sentence
	NWS90	Length (in words) of the 90^{th} percentile sentence
	PL30	Percentage of sentences longer than 30 words
	PRE/PRO	Ratio of prepositions and pronouns
	GRAM/PRO	Ratio of grammatical words and pronouns
	ART/PRO	Ratio of articles and pronouns
	PRE/ALL	Proportions of prepositions in the text
	PRE/LEX	Ratio of prepositions and lexical words
	ART/LEX	Ratio of articles and lexical words
	PRE/GRAM	Ratio of prepositions and grammatical words
	NOM-NAM/ART	Ratio of nouns (common and proper) and gramm. words
	PPres	Presence of at least one present participle in the text
	PPres-C	Proportion of present participle among verbs
	PPasse	Presence of at least one past participle
	Infi	Presence of at least one infinive
	Impf	Presence of at least one imperfect
	Subp	Presence of at least one subjunctive present
	Futur	Presence of at least one future
	Cond	Presence of at least one conditional
	PasseSim	Presence of at least one simple past
	Imperatif	Presence of at least one imperative
	Subi	Presence of at least one subjunctive imperfect
Discourse	avLocalLsa-Lem	Average intersentential cohesion measured via LSA
	ConcDens	Estimate of the conceptual density with *Densidées* Lee *et al.* (2010)
	PP1P2	Percentage of P1 and P2 personal pronouns
	PP2	Percentage of P2 personal pronouns
	PPD	Percentage of personal pronouns of dialogue
	BINGUI	Presence of commas

APPENDIX II: FURTHER ANALYSIS OF COGNITIVE AND PEDAGOGICAL FEATURES

In Section 4.1, we showed that cognitive and pedagogical features strongly predicted CEFR levels in the FLE-CORP corpus, but that they fail to improve the LSD baseline. This indicates that the information provided by these features were already available in the baseline model. To explore this further, we computed the correlations between these features and the baseline features. Table 5 presents the correlation achieved when regressing the LSD features to each cognitive and pedagogical feature (*LSD Regression R*), as well as the most highly correlated individual features from that baseline. As a measure of importance of each feature, we also include the correlation that can be achieved when using each cognitive or pedagogical feature alone to classify the FLE-CORP corpus (*FLE-Corp R*). Each of these features is explained in section 3.2.

feature	FLE-CORP R	LSD Regression R	Most Correlated LSD Features
old20	0.34	0.87	NLM (0.79), PM8 (0.79)
pld20	0.37	0.87	NLM (0.83), PM8 (0.81)
old_pld_not_found	0.22	0.75	IQFFFC (-0.51), MeanFFFC (0.42)
p_abstract	0.24	0.25	NL90P (0.12), 75FFFC (-0.12)
p_concrete	0.20	0.55	PP1P2 (-0.26), ConcDens (-0.19)
p_concreteness_unknown	0.20	0.55	PP1P2 (0.27), PPD (0.20)
mean_senses_per_word	0.17	0.56	MeanFFFC (-0.26), ART_LEX (0.21)
senses_per_word_not_found	0.21	0.56	PRE_LEX (0.24), MedianFFFC (0.23)
p_a1_expressions	0.48	0.78	PA_Alterego1a (-0.66), PAGoug_2000 (-0.58)
p_a2_expressions	0.21	0.43	ART_LEX (0.31), NLM (0.21)
p_b1_expressions	0.38	0.56	PA_Alterego1a (0.43), NLM (0.42)
p_b2_expressions	0.37	0.76	PRE_GRAM (0.61), PRE_ALL (0.53)
not_found_expressions	0.31	0.74	PP1P2 (0.46), PPD (0.37)

TABLE 5: CORRELATIONS BETWEEN COGNITIVE AND PEDAGOGICAL FEATURES AND LSD BASELINE FEATURES.

We find that the most important cognitive and pedagogical features are strongly predicted by the LSD features, which explains why they failed to improve the LSD baseline when combined with it. In many cases, the most correlated features also make sense. For example, *PA-Alterego* and *PAGoug-2000* are both easy word lists, and both strongly predict *p_a1_expressions*. However, other results are more surprising, such as the fact that the information in old20 can be mostly predicted by the length of words.

APPENDIX III: MODEL PERFORMANCE BY GENRE

Nelson *et al.* (2012) and Dell'Orletta *et al.* (2014), among others, concluded that performance of readability models can be strongly influenced by the genre of the document being predicted. To investigate this in the context of our experiment, we took our cross-validation predictions from Experiments 1 and 2, and computed model performance metrics for each document category. The results are presented in Table 6.

	TEXT (863)			NARRATIVE (171)			INFORMATIVE (414)			VARIA (612)		
features	Acc.	Adj. Acc.	R	Acc.	Adj. Acc.	R	Acc.	Adj. Acc.	R	Acc.	Adj. Acc.	R
LSD	0.52	0.80	0.68	0.51	0.85	0.72	0.51	0.83	0.74	0.41	0.82	0.60
Log Term-Freq	0.49	0.79	0.62	0.56	0.81	0.67	0.55	0.87	0.75	0.48	0.84	0.66
Cog.	0.39	0.63	0.39	0.35	0.69	0.56	0.37	0.66	0.46	0.30	0.67	0.35
Ped.	0.40	0.65	0.48	0.41	0.71	0.52	0.42	0.72	0.56	0.30	0.65	0.35
Cog. Ped.	0.43	0.71	0.54	0.48	0.77	0.65	0.43	0.72	0.60	0.33	0.72	0.42
LSD Cog.	0.52	0.79	0.67	0.50	0.82	0.70	0.48	0.83	0.70	0.40	0.82	0.59
LSD Ped.	0.52	0.81	0.69	0.55	0.85	0.72	0.48	0.84	0.74	0.42	0.82	0.59
LSD Ped. Cog.	0.53	0.80	0.67	0.53	0.84	0.71	0.48	0.84	0.73	0.42	0.81	0.58
fastText	0.53	0.79	0.64	0.49	0.82	0.69	0.47	0.85	0.74	0.46	0.82	0.63
BERT	0.51	0.81	0.65	0.53	0.85	0.74	0.49	0.88	0.77	0.45	0.83	0.62
LSD P. C. BERT	0.56	0.84	0.71	0.57	0.83	0.74	0.57	0.91	0.82	0.51	0.87	0.71
HAN	0.55	0.80	0.68	0.54	0.88	0.77	0.54	0.89	0.80	0.44	0.87	0.70
Fine-Tuned BERT	0.59	0.90	0.79	0.59	0.91	0.82	0.58	0.92	0.83	0.51	0.90	0.75

	DIALOGUE (196)			SENTENCE (294)			BLANK-SENTENCE (66)			MAIL (135)		
features	Acc.	Adj. Acc.	R	Acc.	Adj. Acc.	R	Acc.	Adj. Acc.	R	Acc.	Adj. Acc.	R
LSD	0.65	0.96	0.73	0.51	0.77	0.68	0.43	0.77	0.53	0.5	0.86	0.72
Log Term-Freq	0.68	0.93	0.76	0.58	0.81	0.71	0.55	0.88	0.75	0.58	0.88	0.73
Cog.	0.44	0.81	0.21	0.35	0.68	0.39	0.26	0.65	0.11	0.30	0.66	0.43
Ped.	0.57	0.84	0.45	0.42	0.69	0.46	0.40	0.77	0.42	0.42	0.72	0.66
Cog. Ped.	0.62	0.88	0.50	0.40	0.70	0.47	0.37	0.78	0.35	0.44	0.72	0.61
LSD Cog.	0.63	0.97	0.71	0.54	0.79	0.68	0.38	0.74	0.36	0.55	0.86	0.74
LSD Ped.	0.65	0.96	0.73	0.53	0.78	0.71	0.40	0.77	0.51	0.51	0.84	0.69
LSD Ped. Cog.	0.67	0.95	0.71	0.53	0.78	0.66	0.43	0.78	0.48	0.55	0.85	0.70
fastText	0.62	0.92	0.61	0.50	0.84	0.69	0.41	0.72	0.45	0.51	0.87	0.71
BERT	0.67	0.93	0.67	0.53	0.79	0.64	0.52	0.83	0.67	0.54	0.89	0.76
LSD P. C. BERT	0.71	0.96	0.76	0.58	0.82	0.73	0.48	0.88	0.66	0.54	0.88	0.73
HAN	0.67	0.95	0.73	0.49	0.80	0.71	0.45	0.80	0.66	0.57	0.88	0.73
Fine-Tuned BERT	0.71	0.98	0.81	0.63	0.88	0.84	0.54	0.83	0.76	0.59	0.92	0.82

TABLE 6: FLE-CORP RESULTS BY GENRE.

We note that in general the models perform most poorly on *varia*. This may be due to the unusual nature of these texts, which includes things such as poems and advertisements. On the opposite end of the spectrum, the performance

	A1 (580)		A2 (580)		B1 (580)		B2 (442)		C (569)	
features	Acc.	Adj. Acc.	Acc.	Adj. Acc.	Acc.	Adj. Acc.	Acc.	Adj. Acc.	Acc.	Adj. Acc.
Cog.	0.55	0.73	0.22	0.80	0.29	0.47	0.0	0.72	0.63	0.63
Ped.	0.67	0.83	0.30	0.78	0.28	0.48	0.0	0.71	0.64	0.64
LSD P. C. BERT	0.73	0.95	0.52	0.91	0.49	0.83	0.29	0.82	0.72	0.81
Fine-Tuned BERT	0.83	0.99	0.57	0.98	0.56	0.91	0.18	0.83	0.67	0.80

TABLE 7: FLE-CORP RESULTS BY LEVEL

metrics are highest on *Dialogue* across all models. We surmise this is because, in our corpus, this category consists mainly of lower-level CEFR texts, which are often easier to distinguish. We can see this in Table 7, where we break down our performance metrics across CEFR levels for the entire corpus, and we see that the best metrics are in the A1 and A2 categories. Looking into this table further, we notice that all models seem to struggle to distinguish B2 texts from adjacent levels, and so tend to classify these as either B1 or C level texts, which are the more common classes in our corpus.

Most importantly, however, Table 6 shows that our conclusions from experiments 1–3 remain true regardless of which genre the readability models are evaluated on. In particular, Fine-Turned BERT matches or beats all models in all textual genres, which demonstrates the versatility of this model.