Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1217–1233
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

FABRA: French Aggregator-Based Readability Assessment toolkit

Rodrigo Wilkens*, **David Alfter***, **Xiaoou Wang***, **Alice Pintard***,
Anaïs Tack†, **Kevin Yancey**◇, **Thomas François***
*Cental, IL&C, UCLouvain, †Stanford University, ◇Duolingo
{rodrigo.wilkens, david.alfter, xiaoou.wang, alice.pintard}@uclouvain.be,
atack@stanford.edu, kyancey@duolingo.com, thomas.francois@uclouvain.be

Abstract

In this paper, we present the FABRA readability toolkit based on the aggregation of a large number of readability predictor variables. The toolkit is implemented as a service-oriented architecture, which obviates the need for installation, and simplifies its integration into other projects. We also perform a set of experiments to show which features are most predictive on two different corpora, and how the use of aggregators improves performance over standard feature-based readability prediction. Our experiments show that, for the explored corpora, the most important predictors for native texts are measures of lexical diversity and dependency counts while the most important ones for foreign texts are syntactic variables illustrating language development, as well as features linked to lexical gradation in FFL textbooks. FABRA has the potential to support new research on readability assessment for French.

Keywords: readability, French, readability score encoding, automatic readability assessment

1. Introduction

International surveys on reading abilities such as PISA (Schleicher, 2019) regularly remind us that about 20% of the 15-year-old students can be considered as poor readers. Reading deficiencies can have severe consequences on people's lives, such as the inability to access crucial information (e.g., medical data) (Friedman and Hoffman-Goetz, 2006), to process administrative applications (Kimble, 1992), or to find a job. This challenge has been addressed in a variety of ways, as the development of methods to assess reading difficulty of texts for a given audience within the field of readability. The readability research can be traced back to the 1920s', with the seminal work of Lively and Pressey (1923) that originated a tradition of statistical models aiming to predict the reading difficulty of texts, called readability formulas.[1] The evolution of readability formulas across the 20th century can be roughly summarized as follows: initially, readability formulas were computed by hand and were therefore designed as a trade-off between reliability and minimization of effort. The most famous ones combine two textual features, such as syllable count (Flesch, 1948), sentence length (Flesch, 1948; Dale and Chall, 1948), or proportion of easy words (Dale and Chall, 1948). Later, with the dawn of computers, the first automatized formulas appeared, such as the Automated Readability Index (Smith and Senter, 1967). In addition, readability formulas included more features (Bormuth, 1966; Coleman and Liau, 1975) that mostly remained surface features in the 70s, then evolved to capture more sophisticated characteristics of texts, such as coherence or inference load (Kintsch and Vipond, 1979). For the rest of the 20th, the statistical models used to design readability formulas hardly evolved, whereas innovations occurred both in how reading difficulty was measured (e.g. cloze test and reading time) and what text characteristics were being measured.

With the advent of the 21th century, a revolution took place in readability, as the use of Natural Language Processing (NLP) techniques enabled researchers to automatically capture complex textual features, and sophisticated Machine Learning (ML) algorithms allowed to better combine them. One of its main achievements is the large set of textual features that can now be automatically measured via NLP. Not only can NLP-enabled features help readability models to become more accurate (François and Miltsakaki, 2012), but they also have applications in other domains. For instance, textual features designed for readability studies can be applied to other contexts, such as, the automatic assessment of writing skills of foreign language learners (Crossley and McNamara, 2012). In addition, their rather good interpretability allows them to be included in tools that help writers simplify text by analyzing the reading difficulties of the text (François et al., 2020).

For English, various toolkits have been proposed to help researchers automatically compute textual characteristics for text readability (Graesser et al., 2004; Lu, 2010; Chen and Meurers, 2016). However, for French, no such tool exists, which limits the design of new readability formulas as well as the use of readability features to support clear writing. In this paper, we want to support research on readability for the French language by contributing the *French Aggregator-Based Readability Assessment* (FABRA), a readability architecture that automatically extracts and calculates 435 language variables relevant for readability predictions that can be divided into the following groups: length-based, lexical, syntactic, and discourse. Besides being the first readability tool for French, it is also distinctive among all readability toolkits for the fact that

[1] Readability cannot be confused with Text Simplification that aims to modify a text, making it simpler (Saggion, 2017).

it does not only allow to calculate averages over language variables, as is usual, but as much as 18 statistical aggregators. As a result, our toolkit is able to compute 5892 readability scores. Our main contributions are (1) a toolkit that automatically extracts various features, along with 18 different statistical aggregators, (2) a study of a richer description of the language variables, and (3) an evaluation of the toolkit on two data sets. Concerning future uses of FABRA, we see at least two use cases: feature engineering for ML and linguistic studies exploring language variables patterns.

The rest of the paper is structured as follows: after a summary of related works (Section 2), we describe FABRA's architecture and specify the implemented feature types (Section 3). We then report on various experiments aimed at assessing the interest of the toolkit for readability. Section 4 introduces the methodology of our experiments as well as two corpora designed for French native learners (L1) and French as a foreign language (L2). In Section 5, we first assess the importance of the various features on the two corpora, stressing the contribution of the aggregators, then we create a readability model based on the best feature of each subtypes. Finally, the results are discussed in Section 6.

2. Automatic readability assessment

Readability studies started in the early 20th century and have led to hundreds of different formulas. As the main contributions to this field have been described in previous surveys (Chall and Dale, 1995; DuBay, 2004; François, 2011; Benjamin, 2012; Collins-Thompson, 2014; Martinc et al., 2021; Vajjala, 2021), we focus on some recent work for English and French.

Readability has long been modeled using a feature-based approach (Vajjala, 2021). Features offer a certain level of interpretability that other methods such as deep learning are lacking. For instance, Collins-Thompson and Callan (2005) showed that including word distributions across grade levels within a multinomial Naïve Bayes classifier outperforms classic readability formulas such as (Flesch, 1948). Schwarm and Ostendorf (2005) explored several syntactic features based on parsing trees, whereas Pitler and Nenkova (2008) designed various semantic and discourse features, for instance capturing properties of lexical chains and discourse relations. Most recent works rely on distributed representations of texts (e.g. embeddings) (Cha et al., 2017; Filighera et al., 2019) and deep learning models (Nadeem and Ostendorf, 2018; Martinc et al., 2021).

As regards French, few studies have been published. The first formula for French was designed by Kandel and Moles (1958) as an adaptation of Flesch's formula. Henry (1975) later dedicated his Ph.D. thesis to the development of a formula for native readers, which has been adapted to L2 French readers by Cornaire (1988). The first application of some NLP techniques for French readability can be traced back to the tool Daoust et al. (1996), although it is not until François and Fairon (2012) that full-fledged AI-readability was applied to French. More recently, Dascalu (2014) proposed a tool inspired by Coh-Metrix (Graesser et al., 2004). Finally, deep learning techniques were applied for both L1 French readers (Blandin et al., 2020) and L2 readers (Yancey et al., 2021). In their experiment, Yancey et al. (2021) showed that fine-tuning BERT on about 500 texts per level clearly outperformed a readability formula based on features, but it comes with a total lack of interpretability of the model.

This brief overview of the readability field reminds us that several significant breakthroughs were a direct consequence of our ability to automate more features and to extract more sophisticated ones. Surprisingly, there is only a limited amount of readability tools that can help researchers to automatically capture features in their corpus. Such tools are available for a small set of languages (e.g., Italian (Dell'Orletta et al., 2011; Tonelli et al., 2012; Okinina et al., 2020), Greek (Mikros and Voskaki, 2021), Portuguese (Scarton and Aluısio, 2010), Arabic (Al-Twairesh et al., 2016), German (Chen and Meurers, 2016), and Japanese (Sato et al., 2008)), but most works target English.

Very likely, the first tool for readability was Coh-Metrix (Graesser et al., 2004; McNamara et al., 2014), which automatically extracts various text descriptors using NLP and makes the results available through a web interface. Several researchers used this tool as a common ground for feature extraction[2]. More recently, Chen and Meurers (2016) released a web-based tool to compute 154 readability features for English and German, the Common Text Analysis Platform (CTAP). It was later extended to Italian (Okinina et al., 2020), for which about 400 features can be computed. A team from Georgia State University also developed a suite of tools for readability, that are specialized for a given linguistic level. It includes TAALES for the assessment of lexical sophistication (Kyle and Crossley, 2015), TAACO for text coherence (Crossley et al., 2016), and TAASSC that can assess syntactic complexity and sophistication (Kyle, 2016). Although such toolkits to compute readability metrics are now available for various languages, none are yet available for French, which is why this paper aims to fill this gap.

3. FABRA

In order to make our tool accessible to the largest audience possible, we use a service-oriented architecture. Users can interact with the tool through a web interface to annotate a few paragraphs, whereas more advanced users can take advantage of the architecture by skipping the GUI and use a restful API. The toolkit architecture can be divided into controller and annotators. The controller is the high-level service used as the main entry to the tool. Its principal function is to connect with the other services running the entire pipeline and aggregate

[2] This tool has also been extended to other languages (e.g., Scarton and Aluısio (2010) for Portuguese).

the other services' results. The second part of the tool is a set of annotators – specialized services – that provide the readability variables. The readability annotation services focus on identifying different levels of information by combining dictionaries and NLP tools. At the highest level, the provided annotation may be categorized in the following groups: length-based, lexical, syntactic, and discourse. Each one of these groups is further subdivided into feature families. In the remainder of this section, we describe the variables (Sections 3.1, 3.2, 3.3 and 3.4) and the aggregation methods (Section 3.5) proposed to improve the variables description.[3] The next four subsections present different families of language variables in bold and, in italic, the number of language variables (l.v.).

3.1. Length-based variables

Length has been a recurring proxy for readability since early works (Björnsson, 1968; Smith, 1961; Flesch, 1948). The variables in this group are the most studied and, despite their weaknesses, are still influential.

Word Length *(3 l.v.)* We implemented word length measured with characters (entire word and stem) and syllables, excluding punctuation. For stemming, we use the French SnowBall Stemmer (Bird and Loper, 2004). For syllables, we use the pre-syllabified Lexique3 list (New, 2006) extended with *espeak*, a Linux Text-To-Speech (TTS) program.

Sentence Length *(1 l.v.)* The number of tokens per sentence, excluding punctuation.

3.2. Lexical variables

Lexical information is widely used in readability scores. These variables aim to encode external knowledge in readability models.

Graded Lexical *(10 l.v.)* Graded lexical resources are commonly used for foreign language acquisition. They summarize curricula and pedagogical information and help readability models to encode readers' expected knowledge. In our toolkit, we use Reference Level Descriptors (RLD) (Beacco et al., 2008) and FLELex (François et al., 2014) as graded resources, both targeting Common European Framework of Reference (CEFR) levels. The first is a collection of books compiled from FFL experts knowledge and examples of learner productions. The second, FLELex, was built from a textbook corpus and indicates, for each word, the distribution of (normalized) frequency across the levels of the CEFR, as well as its cumulative frequency. In terms of annotation, we identify the word usage in each level of the RLD, and, for FLELex, we calculate the average frequency per CEFR level by looking up the frequency distribution over CEFR levels, summing the frequencies level-wise and normalizing by the number of words in the sentence.

Orthographic Neighbours *(8 l.v.)* As words are based on a small set of letters and syllables, the lexical retrieval during reading requires selecting the target representation from alternative lexical candidates (Andrews, 1997; Balota et al., 2007; Coltheart, 1977). Andrews (1997) found a relation between lexical retrieval and the number of lexical neighborhoods and the frequency of the neighborhoods. Following these works, we identify the number of neighbors, and their average and cumulative frequency in a reference corpus (New, 2006). In addition, we also use a variation for each of them that considers neighbors with higher frequency. Finally, we also used New and Pallier (2019) work on OLD20 and PLD20 for French. These measures are obtained by calculating each word's average Orthographic and Phonologic Levenshtein distance from the 20 closest words found in Lexique3.

Lexical Norms *(4 l.v.)* Psycholinguistic norms influence the reading comprehension of young readers (Crossley et al., 2017; Beinborn et al., 2014), and their scores have been associated with writing quality and development (Sadoski et al., 1995; Crossley et al., 2019; Crossley, 2020). These norms are deeply language- and culture-dependent and take considerable amounts of time to collect, so we resort to lists previously available. For lexical coverage, we combine lists when compiled following similar methodologies.[4] A complementary measure is word polysemy (Beinborn et al., 2014). One can see this measure as a source of confusion (i.e., the wrong sense is retrieved) or a fixation factor (i.e., readers may be more familiar with the word form). Considering the limited resources available, 4 psycholinguistic norms are used: age of acquisition (Ferrand et al., 2008; Alario and Ferrand, 1999), familiarity (Desrochers and Thompson, 2009; Ferrand et al., 2008; Bonin et al., 2003; Desrochers and Bergeron, 2000), concreteness (Bonin et al., 2018; Bonin et al., 2011; Desrochers and Thompson, 2009; Bonin et al., 2003; Desrochers and Bergeron, 2000), and polysemy (Sagot and Fišer, 2008).

Lexical Diversity *(114 l.v.)* Hapax legomena are words that only occur once in a document, and they generally make up about 40-60% of a text (Kornai, 2007). Some works related their number to readability, as readability decreases when a bigger proportion of hapax legomena is present (Islam et al., 2012). For lexical diversity, we count the number of hapax legomena per text, both on the token and lemma levels. A widely used measure of lexical diversity is the type-token-ratio (TTR) (Washburne and Morphett, 1938; Henry, 1975; Kemper et al., 1993), i.e., the number of types (unique word forms) divided by the number of tokens (word forms). However, TTR and its variants are critiqued for being dependent on text length (Patty and Painter, 1931; Heaps, 1978; Hess et al., 1986; Arnaud and Béjoint, 1992). We therefore also

[3] Variables that require tokenization, lemmatization, POS tagging, morphology information or dependency relation use the information provided from Stanza parser (Qi et al., 2020).

[4] The last proposed norm is used when lists overlap.

implemented alternative diversity scores: Moving Average TTR (MATTR; Covington and McFall (2010)) with a window size of 100 words; Corrected TTR (CTTR; Carroll (1964)); Root TTR (RTTR; Guiraud (1959)); Bilogarithmic TTR (LogTTR; Herdan (1960); Herdan (1966)); SquaredTTR (Chaudron and Parker, 1990); and UberIndex (Arnaud and Béjoint, 1992). We doubled these scores by distinguishing lemma and surface forms. Moreover, we also differentiate all words, content words, adjective, adverb, adjective and adverb, nouns and pronouns, and verb, which has the occurrence of content words as a denominator during the division, and weighted verb, which has the verbs as a denominator.

Lexical Frequencies *(28 l.v.)* Lexical frequency is a strong predictor of lexical complexity and readability (Rayner and Duffy, 1986). In this work, we use the following lists aiming to capture different language registers: Lexique3 as an approximation for general language, CHILDES for (productive) child language, and FLELex for language aimed at learners of French as a second language. For FLELex, we use the total frequency of each word, as opposed to the frequency distribution over CEFR levels described in *Graded Lexical*. Lexical frequency is calculated for each of the following categories: all words, content words, functional words, nouns (excluding proper nouns), verbs and adjectives. Lexical frequency is calculated on a word form basis for CHILDES list and the word form list from Lexique3, and on a lemma basis for FLELex and the lemma list from Lexique3.

Lexical Sophistication *(51 l.v.)*
Lexical sophistication is calculated as the ratio between the number of *sophisticated* words of a "class" and the total number of words of the same "class" (Linnarud, 1986; Hyltenstam, 1988), with sophisticated words being words not found in the Gougenheim list, and "class" being either "all POS tags", "lexical POS tags" or "verbs". In addition, lexical frequency is a strong predictor of word complexity (Ryder and Slater, 1988), and thus readability. Instead of directly working with frequencies, one can subdivide the frequency spectrum into *frequency bands* (François, 2011; Chen and Meurers, 2018), i.e., contiguous slices containing approximately the same number of words. We use four word lists, namely (1) Gougenheim (Gougenheim et al., 1964), a list of easy French words (8.775 tokens), (2) CHILDES list, a list derived from the French part of the CHILDES corpus (MacWhinney, 2000) of child language (11.479 tokens), (3) Lexique3 (New, 2006), a list derived from subtitles (142.694 tokens), and (4) FLELex (François et al., 2014) [5] (14.236 tokens). For each list, we retain the top 9.000 most frequent words, and subdivide them into 9 equal slices[6]. Those lists are further subdivided into a lemma (Lexique3 and FLELex) and surface (Lexique3, CHILDES and Gougenheim) form lists. Each of the slices is named K_n with n being the i^{th} slice. We computed the proportion of words in each of the slices per sentence.

3.3. Syntactic variables

The differences in syntactic usage may also indicate language mastery (Bates et al., 1994; Tardif et al., 1997). In this work, we use the Stanza parser (Qi et al., 2020) to extract the syntactic information.

Part-of-speech Tags *(17 l.v.)* Certain part-of-speeches have been found to be good indicators of reading difficulty (Bormuth, 1966). For POS tags, we count the number of each of the 17 Universal POS tags per sentence (e.g., NOUN, VERB and ADJ).

Morphology Features *(34 l.v.)* We count the morphological features combinations for each morphology tags per sentence (e.g., pronoun and plural form).

Verb Tenses *(24 l.v.)* Certain verb tenses or moods make texts more complex to understand than others. Although it is difficult to determine precisely which ones are most likely to cause difficulty, studies point to their importance (Carreiras et al., 1997; Truitt and Zwaan, 1997; François, 2009; Gillie, 1957).

Dependency Relations *(37 l.v.)* In addition to syntactic information at the word level, we count the types of dependency relations between words as the number of each of the Universal Dependency tags per sentence (e.g., nominal subject and indirect object).

Language Development *(25 l.v.)* The syntactic structure also contains evidence of language development (Nenkova et al., 2009). Therefore, we use two well-known marks of syntactic development: T-units (Hunt, 1965)[7] and Yngve index (Yngve, 1960; Frazier, 1985). Motivated by Yngve's work, we also include other constituent-level measures, aiming at a richer representation. For this, we use the constituency parser (Kitaev and Klein, 2018) to extract occurrences of different types of phrases, the depth of a phrase in the sentence, the sentence depth, and the total number of phrases in a sentence. These measures have already been explored in other works, such as Vajjala and Meurers (2013). In addition, we also count the number of words before and after the main verb, following Graesser et al. (2011). We also measure the syntactic similarity between sentences (Baayen et al., 1993) by comparing both the similarity between one sentence and the next one as well as the similarity between a sentence and all other sentences in the document.

3.4. Discourse variables

Discursive elements, such as anaphoric relations and textual coherence, are also associated with a text's

[5] FLELex is described in the *Graded Lexical* variables.

[6] It should be mentioned that, for the Gougenheim list, the first slice contains 1063 words for historical reasons and 775 words in the last one.

[7] We highlight that the T-units measure should be taken with caution for L2 studies (Biber et al., 2011; Crossley and McNamara, 2014).

readability. In this sense, discourse variables model relationships between elements across sentences.

Dialogue *(3 l.v.)* Following Henry (1975) who argues that texts containing dialogue might be simpler, we identify the usage of exclamation and question marks considering all sentence stops, considering all sentence stops and colons, and the presence of dialogue quotes.

Referential Expressions *(6 l.v.)* Bormuth (1969) discusses the need for anaphora resolution, remarking on the relation between reading difficulty and the density of anaphoric elements as well as their distance. We consider that coreference resolution will put a significant overhead on the tool. Therefore, we simplify Bormuth's original measures and count the proportion of pronouns between all nouns and all words, the ratio of possessives, the ratio of personal pronouns, and the proportion of definite article considering all words and considering only nouns.

Content Overlap *(10 l.v.)* Word repetition is indicative of textual cohesiveness (Baayen et al., 1993). In this regard, we identify the proportion of words shared by adjacent and all sentences. For each of these cases, we consider the overlap of all words, only nouns, pronouns and nouns, and content words.

Text Likelihood *(49 l.v.)* N-grams models can be seen as coherence measures in the readability context (Schwarm and Ostendorf, 2005; Kate et al., 2010; Si and Callan, 2001). These variables are anchored in the comparison of a word-level distribution in a reference language model. In this work, we follow Pitler and Nenkova (2008) who showed that even a single unigram model is an efficient predictor. Our model was trained on the movie subtitles corpus behind Lexique3 (New et al., 2007), and we extract the variables both at surface and lemma levels: the probability of all words, only the content words, only the functional words, only nouns, only verbs and only adjectives. For each of them, we consider the sentence average and geometric average (which corresponds to a unigram model).

Text Coherence *(8 l.v.)*
Latent Semantic Analysis (LSA; Deerwester et al. (1990)) has been used to measure textual coherence (Foltz et al., 1998; Landauer et al., 1998) under the hypothesis that more coherent texts are easier to read. A complementary method consists of applying a Single Value Decomposition over a Positive Pointwise Mutual Information (PPMI) matrix (Bullinaria and Levy, 2007; Levy et al., 2015) to generate word embeddings indicating shared contexts between words. In order to obtain LSA and PPMI models, we used the frWaC (Baroni et al., 2009), a large corpus (1.6 billion words) that covers a great range of themes. For the LSA model, stopwords and punctuations were removed and only the 100,000 most frequent tokens/lemmas are used. The number of dimensions (topics) is set to 250 for the LSA training. For the calculation of the PPMI matrix, we did not remove stopwords (Bullinaria and Levy, 2007), the window size was set to 2 (Bullinaria and Levy, 2007; Levy et al., 2015) and the final number of dimensions is 500. For both metrics, our toolkit calculates the cosine similarity of all pairs of adjacent sentences as well as for each sentence with all the other sentences. Since the PPMI approach generates word-level vectors, we average all the word vectors of a sentence before using the two aforementioned coherence metrics.

3.5. Feature Aggregators

Annotation is performed at word or sentence level, thus each language variable is encoded as a list of descriptors. Following our goal of a more detailed description of the language variables, we aggregate each descriptor's list using the following 18 descriptive statistic scores: measures of range (i.e., sum, min, max and length), separation (i.e., median, first and third quantile, and eightieth and ninetieth percentiles), central tendency (average and mode), dispersion (i.e., variance, standard deviation, relative standard deviation (RSD), interquartile range (IQR) and Dolch[8]), and description of the curve (i.e., skewness and kurtosis). Therefore, the toolkit can output up to 321 × 18 + 114 = 5892 readability scores from the combination of aggregators and language variables[9]. We emphasize that some language phenomena may not be observed in a text. Thus the toolkit does not present their aggregations in the output, aiming to avoid misinterpretation. Also, some aggregations may be desirable even if the language phenomenon is rare in the text. Consequently, we only suppress aggregators when a single value is identified and let the user decide the minimum acceptable value of observations using the length aggregator.

4. Toolkit Evaluation Methodology

In order to illustrate the large range of research works made possible by using our toolkit and to provide some guidelines for its use as a standard readability assessment tool for French, we investigate various possibilities of leveraging FABRA's outputs on two corpora consisting of textbooks targeting natives and second language learners described in Section 4.1. First, we perform a variable importance analysis to extract the most useful variables by corpus and variable group. More importantly, we also explore the Spearman correlation among the variables of the same family. Finally, we train machine learning models to test the readability prediction per se, using the readability formula for French by Kandel and Moles (1958).

4.1. Corpus

Assessing our toolkit's performance requires corpora in which the reading difficulty of each text has been

[8] In this work we extend the Dolch measure (Daoust et al., 1996), defined as the 90th percentile subtracted from the median, to be applicable to other variables than sentence length.

[9] We highlight that the 114 variables from *Lexical Diversity* are self-aggregated. In other words, the information is a cardinal value and not a distribution.

evaluated according to a reference scale. A common way to build such corpora is to collect textbooks and label each extracted document with the level of the textbook it comes from (e.g., Sato et al. (2008); Volodina et al. (2014)). We followed this procedure to create FLM-CORP, a collection of 334 texts taken from Belgian school material and representing 9 levels: *primaire* from 4 to 6 (from 9 to 11 years old) and *secondaire* from 1 to 6 (from 12 to 18 years old). The text collection was carried out by university students as part of a Corpus Linguistics course. For each level, students were asked to consult a few textbooks in three disciplines, namely French, History and Science, and to extract, as far as possible, four types of texts: *narrative*, *informative*, *argumentative* and *dialogical*. Statistics about this corpus are shown in Table 1.

Target	Texts	Words
Primaire 4	50	6.692
Primaire 5	36	7.254
Primaire 6	43	18.091
Secondaire 1	38	14.795
Secondaire 2	36	15.716
Secondaire 3	33	11.690
Secondaire 4	33	15.088
Secondaire 5	33	21.079
Secondaire 6	32	14.719
Total	334	125.124

Table 1: FLM-Corp Description

We also evaluated our toolkit for French as a Foreign Language (FFL) readability, using the CEFR scale (Council of Europe, 2001), which includes six levels: A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). Since most pedagogical materials published after 2001 indicate which CEFR level they are intended for, it is possible to build a CEFR-annotated corpus by following the same process as for the L1 corpus, as did François and Fairon (2012) to create a first FFL corpus of 1.793 texts. Yancey et al. (2021) recently expanded their collection into a large and diverse corpus extracted from 47 FFL textbooks published between 2001 and 2018. The 4.562 texts of this corpus are distributed across five CEFR levels, as the authors merged C1 and C2 levels, and contains 8 different text types: *narrative*, *informative*, *text* (other), *dialogue*, *mail/e-mail*, *sentence* (short unauthentic series of example sentences made to fit in a specific lesson), *gap filling exercise*, and a *various* category including recipe, poetry, song, advertisement, etc. In this study, we used the same corpus as Yancey et al. (2021) and were able to reproduce their sample, called FLE-CORP, totalling 2.734 texts with a balanced distribution of texts in each level, as described in Table 2.

Target	Texts	Words
A1	572	60.022
A2	574	83.294
B1	580	119.048
B2	442	130.877
C1 and C2	566	198.517
Total	2734	591.758

Table 2: FLE-Corp Description

5. Readability assessment

In order to test the effectiveness of the different types of variables, we perform Spearman correlation analyses on two different corpora (Section 5.1). To show the benefits of the proposed aggregators, we also compare classifications based on them with the readability formula by Kandel and Moles (1958) in Section 5.2.

5.1. Feature importance analysis

For each corpus, we extracted the 10 features most correlated with the texts' level. This step, portrayed in Tables 3 and 4, shows that variable importance varies according to target audience. In the FLM-Corp, composed of texts used in French-speaking schools, 7 variables out of 10 are lexical, among which 5 are representing lexical diversity (*features 1, 2, 4, 5 and 6*), and two illustrate lexical frequency in native French words lists (*8, 10*). The remaining three features in this top 10 are describing dependency relationships (*7, 9*) and language development (*3*). However, we see that correlation scores for this corpus are rather low, the top one feature only reaching an *r* of .42.

	Features	Corr
1	STTR of nouns + proper nouns	0.42
2	CTTR of content words	0.42
3	Sentence heigth	0.42
4	UberIndex of adjectives	0.42
5	CTTR of verbs	0.42
6	CTTR of all words	0.42
7	Nb of noun modifiers	0.41
8	Adjectives freq in CHILDES	0.39
9	Nb of coordinating conjunctions	0.38
10	Adjectives freq in Lexique3	0.38

Table 3: Top 10 Features on FLM-Corp

For the FLE-Corp, illustrating texts intended to FFL learners, we can see notably higher correlations between the top 10 features and the target texts' level in Table 4. On this corpus, we found that syntactic variables (*2, 4, 5, 8 and 10*), among which four reflect language development, are more important than in the FLM-Corp and related to the sentence structure in general more than a particular POS. The best lexical variables (*1, 6*) illustrate vocabulary gradation in FFL programs and rely on a dedicated FFL resource (FLELex), whereas the lexical dimension for native French readability was better portrayed by diversity and words frequency in general French or children discourse (Lexique3 and CHILDES). The classical feature of sentence length (*3*) seems more important for FFL readers than for native ones. Finally, two features appear to have a strong potential for generalisation in readability research by being in both top 10. It is no surprise that both features have been well studied. Diversity measures based on the ratio of types and tokens are widely use in readability research, and the sentence height was already identified as an important complement to readability measures since von Glasersfeld (1970).

	Feature	Corr
1	Words in FLELex A1 level	-0.64
2	Sentence height	0.61
3	Nb of token/sentence	0.56
4	Nb of constituents	0.56
5	Nb of independent sentences	-0.56
6	Words in FLELex A2 level	-0.54
7	UberIndex of verbs	0.54
8	Nb of determiners	0.54
9	CTTR of all words	0.54
10	Nb of direct subordinates	0.54

Table 4: Top 10 Features on FLE-Corp

While extracting the top 10 features gives information about the most important text characteristics depending on the audience, these features can be redundant, thus do not necessarily make good candidates for a readability model or formula. We therefore also analyzed the best feature per family for each corpus, as presented in Table 5. This table shows, for each family, the mean correlation with the text level of all features from this family computed with all aggregators (*All*), in order to get an idea of the whole family importance for readability evaluation. The next column (*Avg*) presents the mean correlation of all features from the family, computed with the "average" aggregator only, which represents a common way to calculate features in the readability literature. The last column is the best family feature among the 18 types of aggregators. To verify that this approach by family would reduce the risk of information overlap, we also looked at the correlation matrix of the best feature per family in each corpus (see Appendix B).

5.2. Readability Prediction

The toolkit is designed to support readability research. We tried to assemble many variables and a more detailed description of them. In an additional effort to assess the coverage of the description provided, we used the aggregated language variables as features for modeling the target levels by using machine learning. To that end, we trained classification (linear logistic regression) and regression (support vector machine regression) models.[10] We split the data into train and test using a stratified 10-fold cross-validation approach. In the machine learning approach, we compared the classification performance in both FLE-Corp and FLM-Corp. For each variable family, we trained two models: one using the *average* aggregator and one using all aggregators. Then, cross-validation scores were averaged over all families and are reported in Table 6. In addition, we also trained a single model based on the best features (see Table 5) and, for comparison purposes, we used Kandel and Moles (1958). Aiming for a fairer comparison between the models, the same learning methods were used to map the difficulty levels of our corpora with the Kandel and Moles scores.

The results of these models, shown in Table 6, point out a better performance of the models based on all aggregators over those using only average, in both classification and regression tasks. Moreover, there is a significant performance variation when comparing at the family level as observable by the difference between the maximum and mean scores. We also highlight that the difference in the scores of the two corpora is remarkable. We believe this difference to be due to the difference in corpus size (FLE-Corp is about eight times bigger than FLM-Corp) and the number of target levels (i.e., 5 v. 9). Finally, the results of the "best" model should be regarded skeptically because the variables were selected based on the entire corpus observation. Concerning the literature for French readability, we highlight the work of Yancey et al. (2021), who compare feature-based and deep learning approaches in a similar sample of FLE-Corp. They observed accuracy of .51 when fine-tuning BERT and .56 (correlation of .77) when combining fine-tuned BERT with features. In face of these results, the richer representation provided by the aggregators, which achieved an F1 of .69 (correlation of .74), seems promising.

6. Discussion

We found that different aggregators are similarly related to the text target; e.g., for the Sentence height variable, 90P, avg, 80P, max, Q3 and median are similarly correlated in FLE-Corp. From a ML point of view this can be considered redundant, but details of data distribution can allow researchers to use the toolkit to explore new theories. Furthermore, the results presented in Table 5 indicate that several aggregators are superior to the traditional average aggregator.

We observed that models trained using only the average aggregator have lower overall performance than models trained using all aggregators for both corpora when comparing at the language variable level. This helps to support our claim that a richer representation

[10] We emphasize that the ability of different algorithms' families to take advantage of the distribution encoded by the aggregation of the variables remains to be investigated. Therefore, we opt for simple learning algorithms.

Family	FLM-CORP			FLE-CORP		
	All (std)	Avg	Best features	All (avg)	Avg	Best features corr
Length based						
Sentence Length	.24 (.06)	.28	.34 Nb of token/sentence (max)	.41 (.19)	.55	.56 Nb of token/sentence (80P)
Word Length	.20 (.05)	.19	.28 Nb of syllables/wd (var)	.26 (.14)	.36	.47 Nb of letters/wd (max)
Lexical						
Content Overlap	.15 (.04)	.12	.24 Shared (pro)noun lemmas in next sent. (90P)	.15 (.08)	.13	.30 Lemmas shared in adjacent sent. (rsd)
Lex. diversity	.30 (.10)	.21	.43 STTR of nouns + proper nouns	.29 (.17)	.14	.54 UberIndex of verb type/token
Lex. Frequency	.19 (.07)	.15	.39 Freq of adj surface (CHILDES) (skewness)	.24 (.12)	.21	.51 Verbs freq (Lexique3) (Q1)
Graded lexicons	.16 (.04)	.19	.27 Words in FLELex A1 level (Q1)	.28 (.12)	.30	.64 Words in FLELex A1 level (Daoust)
Orthog. neighbors	.20 (.06)	.24	.32 neighbor frequency (min)	.18 (.11)	.20	.44 Phonologic distance (max)
Lex. Norms	.25 (.08)	.29	.36 Imageability (avg)	.22 (.09)	.27	.44 Age of acquisition (max)
Lex. sophistication	.19 (.05)	.19	.37 Lemmas in K8 bands (FLELex) (max)	.20 (.10)	.22	.44 Lemmas in K8 bands (FLELex) (max)
Syntactic						
Dep. Relations	.20 (.06)	.21	.41 Nb of noun modifiers (max)	.24 (.14)	.31	.54 Nb of determiners (90P)
Lang. development	.19 (.07)	.20	.42 Sentence height (max)	.23 (.15)	.30	.61 Sentence height (90P)
Morph. features	.16 (.04)	.15	.27 Pronouns (90P)	.18 (.10)	.23	.5 Nb of relative pronouns (max)
POS Tag	.14 (.03)	.14	.22 Nb of subordinate conjunction (max)	.17 (.10)	.19	.41 Nb of punctuation (var)
Tense	.15 (.04)	.13	.28 Nb of Infinitives (90P)	.16 (.09)	.21	.32 Nb of Infinitives(90P)
Discourse						
Text coherence	.21 (.07)	.21	.37 LSA (lemma) in adjacent sentence (90P)	.27 (.12)	.37	.48 LSA (lemma) in all sentences (90P)
Dialogue Variables	.12 (.01)	-	.13 Exclamation/question marks (skewness)	.15 (.07)	.17	.27 Exclamation/question marks (90P)
Text likelihood	.21 (.06)	.16	.37 Mean of 1-ngram of all verbs (min)	.25 (.14)	.17	.51 Mean of 1-ngram of all verbs (min)
Ref. expressions	.04 (.03)	-	.07 Proportion of pronouns to all words	.19 (.08)	NA	.25 Proportion of def. article to all words

Table 5: Mean absolute Correlation and Best feature per family on FLM-CORP and FLE-CORP

Corpus	Feat	ACC	F1	Corr
FLE-Corp	avg	.35/.45	.49/.65	.42/.65
	agg	.40/.52	.56/.69	.56/.74
	Best	.47/.50	.60/.80	.72/.75
	KM	.36/.41	.51/.58	.43/.50
FLM-Corp	avg	.16/.23	.16/.20	.17/.32
	agg	.26/.32	.24/30.	.27/.40
	Best	.25/.35	.67/.76	.59/.69
	KM	.15/.15	.25/.26	.19/.37

Table 6: Mean/maximum classification (ACC and F1) and regression (*corr*elation) results when the models are trained by family using only the average (*avg*) aggregator and the entire set of aggregators (*agg*) and using the *best* correlated features from each family (Table 5). KM indicates Kandel and Moles (1958).

of the distribution of language variables is more suitable than using the mean as a single descriptor. In addition, our models considerably outperform the baseline model (i.e., Kandel and Moles (1958)) that was fine-tuned on the studied corpora. We emphasize that the models trained using only the best features of each family achieve an average performance higher than the other models. However, these results probably occur due to the reduction of the algorithms' search space; after all, we are already indicating the best features for each corpus. Despite this, techniques closer to the current ML state-of-the-art can probably better use the information extracted from the combination of aggregators and linguistic variables. For example, language variables may improve BERT-based deep learning results (Imperial, 2021), and they may strongly outperform deep learning models for small corpora (Deutsch et al., 2020), although not for a median-size corpus (Yancey et al., 2021).

7. Conclusion

In this paper, we presented FABRA, a new toolkit for French automatic readability assessment based on feature aggregators. The toolkit calculates more than 5k scores from the combination of linguistic features and aggregators. The toolkit is provided as a service[11], which has two major advantages: first, users do not need to install the toolkit to use it; second, the toolkit can easily be integrated into other projects. Due to its modular structure, it is easily extendable. Further, it allows for more comparable analyses in French readability research by providing a standardized set of measures. Finally, we want to enable researchers from other domains to enrich their analysis of readability, for example in text generation or in the evaluation of (neural) machine translation or automatic text simplification. We also highlight that FABRA may be used in different tasks. In this work, we target readability, although similar features have been used in different tasks and languages. Indeed, future research might explore the effectiveness of features across languages; particularly those language agnostics (e.g. length-based features) and those based on frameworks and tools available for different languages (e.g. Universal Dependencies). In the future, we plan to add more variables that would allow FABRA to be also used to assess writing production. We also plan to explore machine learning and deep learning models that better take advantage of the language variable distribution encoding.

8. Acknowledgements

This research has been funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under the grant MIS/PGY F.4518.21 and by a research convention with France Éducation International. Com-

[11] https://cental.uclouvain.be/fabra/

putational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the FNRS under Grant No. 2.5020.11 and by the Walloon Region. Finally, we would like to extend our most profound appreciation to Nami Yamaguchi, who provided a user perspective of FABRA and tested the correctness of certain features.

9. References

Al-Twairesh, N., Al-Dayel, A., Al-Khalifa, H., Al-Yahya, M., Alageel, S., Abanmy, N., and Al-Shenaifi, N. (2016). Madad: a readability annotation tool for arabic text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4093–4097.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic bulletin & review*, 4(4):439–461.

Arnaud, P. J. and Béjoint, H. (1992). *Vocabulary and applied linguistics*. Springer.

Baayen, R. H., Piepenbrock, R., and Van Rijn, H. (1993). The celex lexical database (cd-rom). linguistic data consortium. *Philadelphia, PA: University of Pennsylvania*.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The english lexicon project. *Behavior research methods*, 39(3):445–459.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J., and Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of child language*, 21(1):85–123.

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

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

Biber, D., Gray, B., and Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in l2 writing development? *Tesol Quarterly*, 45(1):5–35.

Bird, S. and Loper, E. (2004). NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, pages 69–72. Association for Computational Linguistics.

Björnsson, C. H. (1968). *Läsbarhet*. Liber.

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

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

Bormuth, J. R. (1969). Development of readability analysis. Technical report, Chicago University, IL.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3):510–526.

Carreiras, M., Carriedo, N., Alonso, M. A., and Fernández, A. (1997). The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory & Cognition*, 25(4):438–446.

Carroll, J. B. (1964). Language and thought. *Reading Improvement*, 2(1):80.

Cha, M., Gwon, Y., and Kung, H. (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*, pages 2003–2006. ACM.

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

Chaudron, C. and Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. *Studies in second language acquisition*, 12(1):43–64.

Chen, X. and Meurers, D. (2016). Ctap: A web-based tool supporting automatic complexity analysis. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pages 113–119.

Chen, X. and Meurers, D. (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, M. and Liau, T. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

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

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

Coltheart, M. (1977). Access to the internal lexicon. *The psychology of reading*.

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.

Covington, M. A. and McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Crossley, S. A. and McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

Crossley, S. A. and McNamara, D. S. (2014). Does writing development equal writing quality? a computational investigation of syntactic complexity in l2 learners. *Journal of Second Language Writing*, 26:66–79.

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior research methods*, 48(4):1227–1237.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., and Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.

Crossley, S. A., Skalicky, S., and Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3).

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

Daoust, F., Laroche, L., and Ouellet, L. (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*, pages 161–188. Springer.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Deutsch, T., Jasbi, M., and Shieber, S. M. (2020). Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17.

DuBay, W. (2004). *The principles of readability*. Impact Information.

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

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

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

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

François, T. and Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the 2012 Workshop on Predicting and improving text readability for target reader populations (PITR2012)*.

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

François, T. (2009). Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for ffl. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 19–27.

Frazier, L. (1985). Syntactic complexity. *Natural language parsing: Psychological, computational, and theoretical perspectives*, pages 129–189.

Friedman, D. B. and Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior*, 33(3):352–373.

Gillie, P. J. (1957). A simplified formula for measuring abstraction in writing. *Journal of applied psychology*, 41(4):214.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Graesser, A. C., McNamara, D. S., and Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique*, volume 2. D. Reidel.

Heaps, H. S. (1978). *Information retrieval, computational and theoretical aspects*. Academic Press.

Henry, G. (1975). Comment mesurer la lisibilité. *Éducation 2000*.

Herdan, G. (1960). *Type-token mathematics*, volume 4. Mouton.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Springer Berlin.

Hess, C. W., Sefton, K. M., and Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research*, 29(1):129–134.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*, volume 3. National Council of Teachers of English.

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of swedish. *Journal of Multilingual & Multicultural Development*, 9(1-2):67–84.

Imperial, J. M. (2021). Knowledge-rich bert embeddings for readability assessment. *arXiv preprint arXiv:2106.07935*.

Islam, Z., Mehler, A., and Rahman, R. (2012). Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553.

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

Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., Roukos, S., and Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.

Kemper, S., Jackson, J. D., Cheung, H., and Anagnopoulos, C. A. (1993). Enhancing older adults' reading comprehension. *Discourse processes*, 16(4):405–428.

Kimble, J. (1992). Plain english: A charter for clear writing. *TM Cooley L. Rev.*, 9:1.

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

Kitaev, N. and Klein, D. (2018). Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*.

Kornai, A. (2007). *Mathematical linguistics*. Springer Science & Business Media.

Kyle, K. and Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49(4):757–786.

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Ph.D. thesis, Georgia State University, Atlanta, Georgia.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, December.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Number 74 in Lund studies in English. CWK Gleerup.

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

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

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

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Mikros, G. and Voskaki, R. (2021). A modern greek readability tool. *Language and Text: Data, models, information and applications*, 356:163.

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

Nenkova, A., Chae, J., Louis, A., and Pitler, E. (2009). Structural features for predicting the linguistic quality of text. In *Empirical methods in natural language generation*, pages 222–241. Springer.

New, B. and Pallier, C., (2019). *Manuel de Lexique 3*.

Okinina, N., Frey, J.-C., and Weiss, Z. (2020). Ctap for italian: Integrating components for the analysis of italian into a multilingual linguistic complexity analysis tool. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7123–7131.

Patty, W. W. and Painter, W. I. (1931). A technique for measuring the vocabulary burden of textbooks. *The Journal of Educational Research*, 24(2):127–134.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Ryder, R. J. and Slater, W. H. (1988). The relationship between word frequency and word knowledge. *The Journal of Educational Research*, 81(5):312–317.

Sadoski, M., Goetz, E. T., and Avila, E. (1995). Concreteness effects in text recall: Dual coding or context availability? *Reading Research Quarterly*, pages 278–288.

Saggion, H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Sato, S., Matsuyoshi, S., and Kondoh, Y. (2008). Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, Morocco.

Scarton, C. and Aluısio, S. M. (2010). Coh-metrix-port: a readability assessment tool for texts in brazilian portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR*, volume 10. sn.

Schleicher, A. (2019). Pisa 2018: Insights and interpretations. *OECD Publishing*.

Schwarm, S. E. 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.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

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

Smith, E. A. (1961). Devereux readability index. *The Journal of Educational Research*, 54(8):298–303.

Tardif, T., Shatz, M., and Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of english, italian, and mandarin. *Journal of Child Language*, 24(3):535–565.

Tonelli, S., Tran, K. M., and Pianta, E. (2012). Making readability indices readable. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 40–48.

Truitt, T. and Zwaan, R. (1997). Verb aspect affects the generation of instrument inferences. In *38th annual meeting of the Psychonomic Society, Philadelphia*.

Vajjala, S. and Meurers, D. (2013). On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68.

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

Volodina, E., Pilán, I., Eide, S. R., and Heidarsson, H. (2014). You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 128–144.

von Glasersfeld, E. (1970). The problem of syntactic complexity in reading and readability. *Journal of Reading Behavior*, 3(2):1–14.

Washburne, C. and Morphett, M. V. (1938). Grade placement of children's books. *The Elementary School Journal*, 38(5):355–364.

Yancey, K., Pintard, A., and Francois, 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.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466.

10. Language Resource References

Alario, F.-X. and Ferrand, L. (1999). A set of 400 pictures standardized for french: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, 31(3):531–552.

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

Bonin, P., Méot, A., Aubert, L.-F., Malardier, N., Niedenthal, P., and Capelle-Toczek, M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots. *L'année Psychologique*, 103(4):655–694.

Bonin, P., Méot, A., Ferrand, L., and Roux, S. (2011). L'imageabilité: normes et relations avec d'autres variables psycholinguistiques. *LAnnee psychologique*, 111(2):327–357.

Bonin, P., Méot, A., and Bugaiska, A. (2018). Concreteness norms for 1,659 french words: Relationships with other psycholinguistic variables and word recognition times. *Behavior research methods*, 50(6):2366–2387.

Desrochers, A. and Bergeron, M. (2000). Valeurs de fréquence subjective et d'imagerie pour un échantillon de 1,916 substantifs de la langue francaise. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 54(4):274.

Desrochers, A. and Thompson, G. L. (2009). Subjective frequency and imageability ratings for 3,600 french nouns. *Behavior research methods*, 41(2):546–557.

Ferrand, L., Bonin, P., Méot, A., Augustinova, M., New, B., Pallier, C., and Brysbaert, M. (2008). Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic french words and their relation with other psycholinguistic variables. *Behavior research methods*, 40(4):1049–1054.

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

Gougenheim, G., Michea, R., Rivenc, P., and Sauvageot, A. (1964). L'elaboration du français fondamental (1er degré). *Étude sur l'établissement d'un vocabulaire et d'une grammaire de bas, París: Didier*.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Lawrence Erlbaum Associates, Mahwah, NJ.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, 28(4):661–677.

New, B. (2006). Lexique 3: Une nouvelle base de données lexicales. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006)*.

Sagot, B. and Fišer, D. (2008). Building a free french wordnet from multilingual resources. In *OntoLex*.

A. Overview of all the variables

Table 7 describes all variables available in the toolkit.

	Variable	Description
Length based		
Word length	LENwrdSTEM LENwrdLETTERS LENwrdSYL	Number of letters per word stem. Number of letters per word. Number of syllables per word.
Sentence length	LENsntWRD	Number of token per sentence, including punctuation.
Lexical Variables		
Graded lexicons	LEXgrdBA1 to LEXgrdBB2 LEXgrdBOOV LEXgrdFA1 to LEXgrdFC2	Proportion of words in Beacco's French Reference Level Descriptors for each CEFR level (A1 to B2). Words out of vocabulary in Beacco's French Reference Level Descriptors. Frequency of words in FLELex resource for each CEFR level (A1 to C2).
Orthographic neighbors	LEXnghORT LEXnghPHO LEXnghNUM LEXnghNUMH LEXnghFRQH LEXnghFRQ LEXnghNUMH LEXnghAVGF	Mean Orthographic Levenstein distance, computed on Lexique3. Mean Phonologic Levenstein distance, computed on Lexique3. Number of lexical neighbors. Number of lexical neighbors more frequent that the word in the text. Neighbors cumulative frequency in a reference corpus considering neighbors with higher frequency than the word in the text. Neighbors cumulative frequency in a reference corpus. Number of lexical neighbors with frequency in the list. Neighbors average frequency in a reference corpus.
Lexical Norms	LEXnrmCNCR LEXnrmIMG LEXnrmFAM LEXnrmAOA LEXnrmCNCROOV LEXnrmIMGOOV LEXnrmFAMOOV LEXnrmAOAOOV	Words level of concreteness. Imageability. Words familiarity, also called subjective frequency. Age of acquisition of each word. Words out of vacabulary on lexical norms dictionary (concreteness). Words out of vacabulary on lexical norms dictionary (Imageability). Words out of vacabulary on lexical norms dictionary (familiarity). Words out of vacabulary on lexical norms dictionary age of acquisition).
Lexical diversity	LEXdvrHLL LEXdvrHLT LEXdvr[*A/AD/F/M/N/V/W*] [*L/S*][*T/U/C/L/M/R/S*]*(W)*	Number of distinct lemmatized hapax in each sentence. Number of distinct hapax (wordform) in each sentence. Adjectives (A), adverbs (AD), function words (F), modifiers (adjectives and adverbs; M), nouns and proper nouns (N), verbs (V), or all words (W) of the text either in lemma (L) or wordform (S), with different versions of their type-token-ratio : [*TTR (T), CTTR (C), RTTR (R), LogTTR (L), UberIndex (U), SquaredTTR (S), MATTR (M)*]. For verbs, the final W indicates that the ratio is normalized over all word tokens, while without this specification, the ratio is normalized over verb tokens.
Lexical sophistication	LEXsopGK1 to LEXsopGK9 LEXsopCK1 to LEXsopCK9 LEXsopLWK1 to LEXsopLWK9 LEXsopLLK1 to LEXsopLWK9 LEXsopFK1 to LEXsopFK9 LEXsopTKOG LEXsopTYOG	Number of words in the first 9 frequency bands of 1000 words of Gougenheim vocabulary list. Number of surface form words in the first 9 frequency bands of 1000 words of CHILDES. Number of surface form words in the first 9 frequency bands of 1000 words of Lexique3. Number of lemmas in the first 9 frequency bands of 1000 words of Lexique3. Number of lemmas in the first 9 frequency bands of 1000 words of FLELex. Number of "sophisticated" tokens by total number of tokens ("sophisticated" = not in Gougenheim) Number of "sophisticated" types by total number of types

	Variable	Description
	LEXsopTKOGc	Number of “sophisticated” lexical tokens by total lexical tokens
	LEXsopTYOGc	Number of “sophisticated” lexical types by total number of lexical types
	LEXsopTKOGv	Number of “sophisticated” verb tokens by total verb tokens
	LEXsopTYOGv	Number of “sophisticated” verb types by total number of verb types
Lexical Frequency	LEXfrq[*C/F/L*] [*A/C/F/NOTN/N/V/W*][*L/S*]	Frequencies in CHILDES (C), FLELex (F) or Lexique3 (L) by word type [*adjective (A), common noun (N), grammatical (F), lexical (C), non noun (NOTN), verb (V), all (W)*] for lemma (L) or wordform (S).
Content Overlap	LEXcovLGAL	Any lemma is shared in any sentences.
	LEXcovLGAR	Pronoun lemmas are shared in any sentences.
	LEXcovLGCO	Word lemmas are shared in any sentences.
	LEXcovLGNO	Noun lemmas are shared in any sentences.
	LEXcovLGST	Noun and pronoun lemmad are shared in any sentences.
	LEXcovLLAL	Any lemma is shared in adjacent sentences.
	LEXcovLLAR	Pronoun lemmas are shared in adjacent sentences.
	LEXcovLLCO	Word lemmas are shared in adjacent sentences.
	LEXcovLLNO	Noun lemmas are shared in adjacent sentences.
	LEXcovLLST	Noun and pronoun lemmas are shared in adjacent sentences.
Syntactic Variables		
POS Tag	SYNpos[+tag]	Number of different POS types in the text, following universal guidelines.
Dependencies	SYNdep[+dep]	Number of different dependency types, following universal guidelines.
Morphology	SYNmor[+type]	Number of different morphological types, following universal guidelines.
Language development	SYNdevAFT	Words quantity after the main verb in each sentence.
	SYNdevBFR	Words quantity before the main verb in each sentence.
	SYNdevAVGPRSHGT	Deepness of constituents in the text.
	SYNdevSUB	Number of words directly subordinate in the dependency tree.
	SYNdevSIM	Global (average of the relation between one sentence v. all other sentences) syntactic similarity.
	SYNdevSIMA	Local (relation between one sentences and its next sentence) syntactic similarity.
	SYNdevNPHRS	Number of constituents.
	SYNdevNPRS[+type]	Number of different types of constituents in the text, following this guideline.
	SYNdevHGT	Deepness of sentences in the text.
	SYNdevTU	T-units.
	SYNdevVG[*1/2/3*]	Number of verbs from French 1st, 2nd and 3rd groups in the text.
	SYNdevYNGVE	Yngve index.
Tense	SYNtnsPRT	Number of verbs at present tense in the text.
	SYNtnsCND	Number of verbs at conditional tense in the text.
	SYNtnsFUT	Number of verbs at future tense in the text.
	SYNtnsIMPF	Number of verbs at imperfect tense in the text.
	SYNtnsPST	Number of verbs at past tense in the text.
	SYNtnsPRSP	Number of present participles in the text.
	SYNtnsPSTP	Number of past participles in the text.
	SYNtnsIMPR	Number of verbs at imperative mode in the text.
	SYNtnsINF	Number of verbs at infinitive mode in the text.
	SYNtnsSUBJ	Number of verbs at subjunctive mode in the text.
	SYNtnsSUBJPRT	Number of verbs at subjunctive present in the text.
	SYNtnsSUBJIMPF	Number of verbs at subjunctive imperfect in the text.
	SYNtns[+tense]U	Binary measure of the presence/absence of tenses in the text.
Discourse Variables		

	Variable	Description
Text coherence	DIScohLSAL	The LSA, based on frWaC occurrence, considering the word lemma at global (average of the relation between one sentence v. all other sentences) level.
	DIScohLSALADJ	The LSA, based on frWaC occurrence, considering the word lemma at local (relation between one sentences and its next sentence) level.
	DIScohLSAS	The LSA, based on frWaC occurrence, considering the word surface at global (average of the relation between one sentence v. all other sentences) level.
	DIScohLSASADJ	The LSA, based on frWaC occurrence, considering the word surface at local (relation between one sentences and its next sentence) level.
	DIScohPPMIL	The PPMI, based on frWaC occurrence, considering the word lemma at global (average of the relation between one sentence v. all other sentences) level.
	DIScohPPMILADJ	The PPMI, based on frWaC occurrence, considering the word lemma at local (relation between one sentences and its next sentence) level.
	DIScohPPMIW	The PPMI, based on frWaC occurrence, considering the word surface at global (average of the relation between one sentence v. all other sentences) level.
	DIScohPPMIWADJ	The PPMI, based on frWaC occurrence, considering the word surface at local (relation between one sentences and its next sentence) level.
Dialogue variables	DISdiaPPEI1	Percentage of exclamation and question marks considering all sentence stops.
	DISdiaPPEI2	Percentage of exclamation and question marks considering all sentence stops and colons.
	DISdiaBINGUI	Presence of dialogue quotes.
Referential expressions	DISrefPN	Proportion of pronouns to all nouns.
	DISrefPW	Proportion of pronouns to all words.
	DISrefPRSW	Ratio of personal pronouns in the text.
	DISrefPOSSW	Ratio of possessives in the text.
	DISrefDW	Proportion of definite article to all words.
	DISrefDN	Proportion of definite article to all nouns.
Text likelihood	DISlkh[*A/C/F/N/V/W*] [*L/S*][*ML/GM/M/*][+ngram]	Probability of adjectives (A), content words (C), function words (F), nouns (N), verbs (V), or all words (W), in lemma (L) or wordform (S), using different types of N-gram models : [*log of mean, geometric mean, mean*].

Table 7: Variables description

B. Correlation matrices

	DISlkhVLM	DIScohLSALADJ	DISdiaPPEI2	LEXdvrNLS	LEXsopLLK8	LEXfrqCAS	LEXgrdFA1	LEXcovLLST	LEXnghFRQ	LEXnrmIMG	SYNdepNMOD	SYNmorPRON_REL	SYNtnsINF	SYNposSCONJ	SYNdevHGT	LENwrdSYL	LENsntWRD
DISlkhVLM	-	-0.24	-0.18	-0.56	-0.40	-0.48	0.26	-0.19	0.39	0.05	-0.32	-0.24	-0.14	-0.15	-0.37	-0.24	-0.32
DIScohLSALADJ	-0.24	-	0.04	0.27	0.27	0.31	-0.34	0.24	-0.17	-0.17	0.48	0.30	0.15	0.17	0.39	0.24	0.36
DISdiaPPEI2	-0.18	0.04	-	0.32	0.12	0.21	0.13	0.18	-0.11	-0.02	-0.01	0.06	0.14	0.18	0.08	-0.05	0.09
LEXdvrNLS	-0.56	0.27	0.32	-	0.43	0.62	-0.06	0.31	-0.45	-0.02	0.30	0.29	0.30	0.31	0.47	0.08	0.41
LEXsopLLK8	-0.40	0.27	0.12	0.43	-	0.42	-0.28	0.22	-0.32	0.08	0.31	0.23	0.22	0.20	0.30	0.25	0.33
LEXfrqCAS	-0.48	0.31	0.21	0.62	0.42	-	-0.38	0.22	-0.36	-0.07	0.45	0.35	0.15	0.18	0.40	0.39	0.40
LEXgrdFA1	0.26	-0.34	0.13	-0.06	-0.28	-0.38	-	-0.01	0.14	0.05	-0.52	-0.28	-0.01	0.07	-0.25	**-0.75**	-0.28
LEXcovLLST	-0.19	0.24	0.18	0.31	0.22	0.22	-0.01	-	-0.22	-0.05	0.22	0.17	0.15	0.33	0.27	0.02	0.15
LEXnghFRQ	0.39	-0.17	-0.11	-0.45	-0.32	-0.36	0.14	-0.22	-	0.02	-0.26	-0.20	-0.24	-0.31	-0.29	-0.13	-0.32
LEXnrmIMG	0.05	-0.17	-0.02	-0.02	0.08	-0.07	0.05	-0.05	0.02	-	-0.15	-0.08	-0.12	-0.10	-0.14	-0.24	-0.10
SYNdepNMOD	-0.32	0.48	-0.01	0.30	0.31	0.45	-0.52	0.22	-0.26	-0.15	-	0.34	0.14	0.13	0.54	0.54	0.63
SYNmorPRON_REL	-0.24	0.30	0.06	0.29	0.23	0.35	-0.28	0.17	-0.20	-0.08	0.34	-	0.04	0.19	0.38	0.22	0.33
SYNtnsINF	-0.14	0.15	0.14	0.30	0.22	0.15	-0.01	0.15	-0.24	-0.12	0.14	0.04	-	0.35	0.27	0.09	0.27
SYNposSCONJ	-0.15	0.17	0.18	0.31	0.20	0.18	0.07	0.33	-0.31	-0.10	0.13	0.19	0.35	-	0.22	-0.05	0.20
SYNdevHGT	-0.37	0.39	0.08	0.47	0.30	0.40	-0.25	0.27	-0.29	-0.14	0.54	0.38	0.27	0.22	-	0.27	0.67
LENwrdSYL	-0.24	0.24	-0.05	0.08	0.25	0.39	**-0.75**	0.02	-0.13	-0.24	0.54	0.22	0.09	-0.05	0.27	-	0.29
LENsntWRD	-0.32	0.36	0.09	0.41	0.33	0.40	-0.28	0.15	-0.32	-0.10	0.63	0.33	0.27	0.20	0.67	0.29	-

Table 8: Correlation matrix on FLM-Corp best feature per family

	LEXfrqLVS	LEXnghPHO	LEXsopFK8	LEXnrmAOA	LEXdvrVSU	LEXcovLLCO	LEXgrdFA1	SYNdevHGT	SYNposPUNCT	SYNmorPRON_REL	SYNdepDET	SYNtnsINF	DIScohLSAL	DISlkhVLM	DISdiaPPEI2	DISrefDW	LENwrdLETTERS	LENsntWRD
LEXfrqLVS	-	-0.36	-0.35	-0.34	-0.52	0.24	0.64	-0.49	0.34	-0.36	-0.53	-0.11	-0.40	0.56	0.31	-0.39	-0.40	-0.50
LEXnghPHO	-0.36	-	0.42	0.46	0.47	-0.31	-0.49	0.50	-0.34	0.40	0.49	0.28	0.42	-0.47	-0.22	0.26	0.62	0.47
LEXsopFK8	-0.35	0.42	-	0.41	0.45	-0.28	-0.44	0.43	-0.24	0.38	0.42	0.24	0.39	-0.46	-0.20	0.17	0.42	0.43
LEXnrmAOA	-0.34	0.46	0.41	-	0.51	-0.31	-0.42	0.51	-0.34	0.42	0.49	0.32	0.41	-0.48	-0.21	0.20	0.45	0.49
LEXdvrVSU	-0.52	0.47	0.45	0.51	-	-0.39	-0.47	0.58	-0.32	0.50	0.52	0.45	0.45	-0.56	-0.21	0.20	0.46	0.53
LEXcovLLCO	0.24	-0.31	-0.28	-0.31	-0.39	-	0.25	-0.38	0.31	-0.24	-0.36	-0.23	-0.51	0.35	0.35	-0.21	-0.30	-0.38
LEXgrdFA1	0.64	-0.49	-0.44	-0.42	-0.47	0.25	-	-0.56	0.39	-0.43	-0.56	-0.22	-0.44	0.52	0.34	-0.38	-0.49	-0.53
SYNdevHGT	-0.49	0.50	0.43	0.51	0.58	-0.38	-0.56	-	-0.57	0.53	**0.78**	0.40	0.62	-0.53	-0.36	0.31	0.51	**0.84**
SYNposPUNCT	0.34	-0.34	-0.24	-0.34	-0.32	0.31	0.39	-0.57	-	-0.36	-0.54	-0.28	-0.48	0.35	0.24	-0.34	-0.36	-0.55
SYNmorPRON_REL	-0.36	0.40	0.38	0.42	0.50	-0.2-4	-0.43	0.53	-0.36	-	0.46	0.31	0.39	-0.41	-0.20	0.18	0.37	0.49
SYNdepDET	-0.53	0.49	0.42	0.49	0.52	-0.36	-0.56	**0.78**	-0.54	0.46	-	0.30	0.62	-0.49	-0.40	0.53	0.50	**0.88**
SYNtnsINF	-0.11	0.28	0.24	0.32	0.45	-0.23	-0.22	0.40	-0.28	0.31	0.30	-	0.30	-0.36	-0.06	0.03	0.27	0.33
DIScohLSAL	-0.40	0.42	0.39	0.41	0.45	-0.51	-0.44	0.62	-0.48	0.39	0.62	0.30	-	-0.45	-0.31	0.37	0.41	0.64
DISlkhVLM	0.56	-0.47	-0.46	-0.48	-0.56	0.35	0.52	-0.53	0.35	-0.41	-0.49	-0.36	-0.45	-	0.23	-0.21	-0.47	-0.50
DISdiaPPEI2	0.31	-0.22	-0.20	-0.21	-0.21	0.35	0.34	-0.36	0.24	-0.20	-0.40	-0.06	-0.31	0.23	-	-0.29	-0.23	-0.41
DISrefDW	-0.39	0.26	0.17	0.20	0.20	-0.21	-0.38	0.31	-0.34	0.18	0.53	0.03	0.37	-0.21	-0.29	-	0.25	0.35
LENwrdLETTERS	-0.40	0.62	0.42	0.45	0.46	-0.30	-0.49	0.51	-0.36	0.37	0.50	0.27	0.41	-0.47	-0.23	0.25	-	0.49
LENsntWRD	-0.50	0.47	0.43	0.49	0.53	-0.38	-0.53	**0.84**	-0.55	0.49	**0.88**	0.33	0.64	-0.50	-0.41	0.35	0.49	-

Table 9: Correlation matrix on FLE-Corp best feature per family