An “AI readability” formula for French as a foreign language

Thomas François
IRCS, University of Pennsylvania
3401 Walnut Street Suite 400A Room 423
Philadelphia, PA 19104, USA
frthomas@sas.upenn.edu

Cédrick Fairon
CENTAL, UCLouvain
Place Blaise Pascal, 1
1348 Louvain-la-Neuve, Belgium
Cedrick.Fairon@uclouvain.be

Abstract

This paper present a new readability formula for French as a foreign language (FFL), which relies on 46 textual features representative of the lexical, syntactic, and semantic levels as well as some of the specificities of the FFL context. We report comparisons between several techniques for feature selection and various learning algorithms. Our best model, based on support vector machines (SVM), significantly outperforms previous FFL formulas. We also found that semantic features behave poorly in our case, in contrast with some previous readability studies on English as a first language.

1 Introduction

Whether in a first language (L1) or a second and foreign language (L2), learning to read has been and remains one of the major concerns of education. When a teacher wants to improve his/her students’ reading skills, he/she uses reading exercises, whether there are guided or independent. For this practice to be efficient, it is necessary that the texts suit the level of students (O’Connor et al., 2002). This condition is sometimes difficult to meet for teachers wishing to get off the beaten tracks by not using texts from levelled textbooks or readers.

In this context, readability formulas have long been used to help teachers faster select texts for their students. These formulas are reproducible methods that aim at matching readers and texts relative to their reading difficulty level. The Flesch (1948) and Dale and Chall (1948) formulas are probably the best-known examples of those. They are typical of classic formulas, the first major methodological paradigm developed in the field during the 40’s and 50’s. They were kept as parsimonious as possible, using linear regression to combined two, or sometimes, three surface features, such as word mean length, sentence mean length, or proportion of out-of-simple-vocabulary words.

Later, some scholars (Kintsch and Vipond, 1979; Redish and Selzer, 1985) argued that the classic formulas suffer from several shortcomings. These formulas only take into account superficial features, ignoring other important aspects contributing to text difficulty, such as coherence, content density, inference load, etc. They also omit the interactive aspect of the reading process. In the 80’s, a second paradigm, inspired by structuro-cognitivist theories, intended to overcome these issues. It focused on higher textual dimensions, such as inference load (Kintsch and Vipond, 1979; Kemper, 1983), density of concepts (Kintsch and Vipond, 1979), or macrostructure (Meyer, 1982). However, these attempts did not achieve better results than the classic approach, even though they used more principled and more complex features.

Recently, a third paradigm, referred to as the “AI readability” by François (2011a), has emerged in the field. Studies that are part of this current share three key features: the use of a large number of texts assessed by experts (coming from textbooks, simplified newspapers or web resources) as training data ; the use of NPL-enable features able to capture a wider range of readability factors, and the combination of those features through a machine learning

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 466–477, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

algorithm. Since the work of Si and Callan (2001), this paradigm have spawn several studies for English (Collins-Thompson and Callan, 2005; Heilman et al., 2008; Schwarm and Ostendorf, 2005; Feng et al., 2010).

However, for French, the field is far from being so thriving. To our knowledge, only two "AI readability" have been designed so far for French L1 and only one for French as a foreign language (FFL) (see Section 2). This paper reports some experiments aimed at designing a more efficient readability model for FFL. In Section 2, it is further argue why a new formula was necessary for FFL. Section 3 covers the various methodological steps required to devise the model, whose results are reported in Section 4. Finally, Section 5 discusses some interesting insights gained by this work.

2 Readability models for French

Readability of French never enjoyed a large success: while readability studies on English dates back to the 20's, it is only in 1957 that the French-speaking world discovered it through the work of Conquet (1957). Since then, only a few studies focused on the topic.

The two first French L1 formulas were adaptations of the Flesch formula (Kandel and Moles, 1958; de Landsheere, 1963). It is only with Henry (1975) that French got a model fitting the particularities of the language. Henry used cloze tests to assess the level of 60 texts from primary and secondary school textbooks and trained three formulas on this corpus. It is worth mentioning that Henry's formulas have been applied to FFL by Cornaire (1988). During the same time, Richaudeau explored a different path, as a representative of the structuro-cognitivist paradigm. He used the number of words recalled by a subject after he/she has just read a sentence as a device to measure understanding and provided an "efficiency formula" of texts (Richaudeau, 1979). Although more modern in its conception, Richaudeau's hard-to-implement formula did not achieve the same recognition in the French speaking world as Henry's.

After those two major efforts, few works followed. It is worth mentioning two more authors: Mesnager (1989), who designed a classic formula for children that draw inspiration from the Dale and Chall (1948) formula and Daoust et al. (1996), who developed SATO-CALIBRAGE, a program assessing text difficulty from the first to the eleventh grade. It can be considered as the first "AI formula" for French L1, since it made use of NLP-enabled features. It is also the last formula published for French L1, if we except the adaptation of the model by Collins-Thompson and Callan (2004) to French.

As regards to French L2, the literature is even sparser. Tharp (1939) published a first formula taking into account one particularity of the L2 context: the cognates. Those are words sharing a similar form and meaning across two languages and having a facilitating effect in reading. This idea was recently replicated by Uitdenbogerd (2005), who combined a syntactic feature, the mean number of words per sentence, with the number of cognates per 100 words in her formula. Although taking into account this effect of the L1 on L2 reading is very interesting, these two studies are confined to a limited audience: English speakers learning French. As regards a more generic approach, François (2009) recently published an "AI formula" for FFL, based on logistic regression and ten features. Among those, he stressed the use of verbal tense information as a way to improve performance. However, the set of features he experimented remains limited (about 20).

From all this, it seems clear that FFL readability needs to be addressed more thoroughly, especially if we are willing to get a generic model, able to make predictions for L2 readers with any L1 background. The rest of this paper describes one such attempt.

3 Design of the formula

The design of an "AI readability" formula involves the same three steps as a classification problem. First, one need to gather a gold-standard corpus large enough to reliably train the parameters of a learning algorithm, as described in Section 3.1. The next step, covered in Section 3.2, consists in defining a set of predictors, that is to say, linguistic characteristics of the texts that will be used to predict the difficulty level of new texts. Finally, the best subset of these predictors is combined within a learning algorithm to obtain the best model possible. Experiments at this level are reported in Section 3.3.

3.1 The corpus

A gold-standard for readability consists in texts labelled according to their difficulty. For this, it is first necessary to choose a difficulty scale used for the labels (for English L1, it is usually the 12 grade levels scale), that also constrains the output of the formula. Then, each text have to be assessed with a method able to measure the reading comprehension level of the target population.

Regarding the scale, an obvious choice for the foreign language context was the beginner/intermediate/advanced continuum, recently 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 now become the reference for foreign language education, at least in Europe.

Assessing the reading difficulty of texts with respect to a target population of readers was a more challenging issue. Several techniques have been used in the literature, the most important of which are comprehension tests, cloze tests and expert judgements. They all postulate a given population of readers, although relying on expert judgements save the need for a sample of subjects to take a test. In this case, texts comes from textbooks whose content difficulty have been assessed by the publishers.

This last criterion is now mainstream in "AI readability", since it is very practical and facilitates the creation of a large corpus, but it has its own shortcomings. Studies such as van Oosten et al. (2011) found that expert agreement on a same corpus of texts might be insufficient for a classification task.

For this study, we nevertheless relied on expert judgements, since we needed a large amount of labelled texts to ensure a robust statistical learning. We selected 28 FFL textbooks, published after 2001 and designed for adults or adolescents learning FFL for general purposes. From those, we extracted 2,160 texts related to a reading comprehension task and assigned to each of them the same level as the textbook it came from.

As it was expected from van Oosten et al. (2011)'s study, differences in the publishers' conception of difficulty led to an heterogeneous labelling between textbooks. This heterogeneity was detected in three of the six levels (A1, A2, and B1) using ANOVA based on two classic readability features as independent variables: the mean number of words per sentence and the mean number of letters per word. A subsequent qualitative analysis revealed that most of the heterogeneity was coming from textbooks following the new didactic approach recommended by the CEFR: the task-oriented approach, which focuses more on the task than the text when labelling the overall reading activity. Therefore, we decided to remove those type of textbooks from our corpus, which amounted to 5 books and 249 texts. The remaining 1,852 excerpts were kept for our experiments. Their distribution is displayed in Table 1 as regard to the number of texts and tokens.

3.2 The predictors

In a second step, every text of the corpus was represented as a numeric vector of 406 features, each of them representing a linguistic dimension of the text as a single number. Their implementation drew on two different sources of inspiration: the existing predictors in the English and French literature and the psycholinguistic literature on the reading process. The complete set was classified in four families, depending on the kind of information each one is supposed to represent. These families were: "lexical", "syntactic", "semantic", and "specific to FFL context". Each of them was further divided in subfamilies, described in the rest of the section [1].

3.2.1 Lexical Features

Lexical features have been shown to be the most important level of information in many readability studies (Chall and Dale, 1995; Lorge, 1944). It is then not surprising that a wide range of lexical predictors have been developed in the literature. Our own set comprised the following subfamilies:

Statistics of lexical frequencies: frequencies of words in a text are a good indicator of the text's overall difficulty (Stenner, 1996). They are usually summarized via the mean, but we also tested the median, the interquartile range, as well as the 75^{th} and 90^{th} percentiles.

[1] Space restrictions did not enable us to formally defined each variable used in this study. The reader may consult François (2011b) for a more comprehensive description.

A1	A2	B1	B2	C1	C2	Total
430(58.561)	380(75.779)	552(176.973)	198(71.701)	184(92.327)	108(35.202)	1,852(510;543)

Table 1: Distribution of the number of texts and tokens per level in our corpus.

We used *Lexique3* (New et al., 2007) as our frequency database. It is a lexicon including about 50,000 lemmas and 125,000 inflected forms whose frequencies were obtained from movie subtitles. Since French has a rich morphology, we considered the probabilities of both lemma and inflected forms. Moreover, following an idea from Elley (1969), we also computed the above mentioned statistics for given POS words, such as content word, nouns, verbs, etc.

Percentage of words not in a reference list: part of the Dale and Chall (1948)'s formula, this feature is one of the most famous in readability. For our experiments, two word lists for FFL were used: the well-known – but already dated – Gougenheim et al. (1964)'s list and a second one that was found at the end of one FFL textbook: *Alter Ego* (Berthet et al., 2006). Different sizes were also experimented for both lists.

Word length: mean word length is another classic feature in readability (Flesch, 1948; Smith, 1961). We used various statistics based on the number of letters per word (mean, median, percentiles, etc.).

N-grams models: Si and Callan (2001) shown that n-grams models can successfully be applied to readability. We then used both a simple unigram approach based on the frequencies from *Lexique3*, and a more complex bigram model trained on two different corpora: the Google n-grams (Michel et al., 2011) and a corpus of newspaper articles from *Le Soir* amounting to $5,000,000$ words [2]. Both were normalized according the length n of the text as follows:

$$P(text) = \frac{1}{n}\sum_{i=1}^{n} \log P(w_i|h) \qquad (1)$$

where w_i is the i^{th} word and h a limited history of length 0 (unigram) or 1 (bigram).

Lexical diversity: the repetition effect is another factor known to affect the reading process (Bowers, 2000). It has been mainly implemented through the classic type-token ratio (TTR) that suffers from being dependent on the text length. This is why we defined a normalized TTR, which is the mean score of several TTRs, computed on text's fragments of equal length. This way, long texts were made comparable with short ones.

Orthographic neighborhood: we finally suggested a new lexical variable, based on the fact that some characteristics of the orthographic neighbors [3] of a word are known to impact the reading of this word (Andrews, 1997). Thirteen predictors were implemented to account for the number or the frequency of the orthographic neighbors of all words in a text.

3.2.2 Syntactic features

The syntactic level of information is another traditional area of investigation in readability. Although most of the scholars in the field agree that it does not lead to such efficient predictors as the lexical level, they have noticed it can be combined with the latter to improve performance of readability formulas. We therefore investigated the following subfamilies:

Sentence length: the traditional approach to syntactic difficulty relied on the number of words per sentence. We have approached it through various statistics such as the mean, the median, or several percentiles.

Part of speech ratios: Bormuth (1966) demonstrated the good predictive power of some POS ratios in a text. We computed 156 ratios based on TreeTagger's POS (Schmid, 1994). They operated as proxies for the syntactic complexity of sentences, since we did not use features based on a parser [4].

[2] Smoothing algorithms used were respectively the simple Good-Turing algorithm (Gale and Sampson, 1995) for unigrams and linear interpolation (Chen and Goodman, 1999) for the bigrams.

[3] The orthographic neighbors of a word X have been defined by Coltheart (1978) as all the words of the same length as X and varying from it only by one letter (eg. FIST and GIST).

[4] This choice was motivated as follows. Bormuth (1966), who performed a manual annotation of the syntactic structures

Verbs: although the tense and moods found in a text have been hardly considered in the field, Carreiras et al. (1997) suggested that verbal aspects are important while building a mental representation of a text and therefore impact its understanding. They help the reader to distinguish between major and minor elements associated with events described by these verbs. We therefore replicated and enhanced the feature set proposed by François (2009), considering either binary indicators or proportions of the use of tenses or moods in a text.

3.2.3 Semantic features

The importance of semantic and cognitive factors have been particularly stressed by the structuro-cognitivist paradigm, although Miller and Kintsch (1980), as well as Kemper (1983), eventually admitted not being able to demonstrate the superiority of those new predictors over traditional ones. More recent work also reported limited evidence of this alleged superiority (Pitler and Nenkova, 2008; Feng et al., 2010). In order to clarify as much as possible the situation for FFL, we implemented the following features:

Personnalization level: Dale and Tyler (1934) suggested that informal texts should be easier to read and that informality might be assessed through the type of personal pronouns found in a text. On this assumption, 13 variables were defined to take into account various personal pronouns proportions in the text.

Conceptual density: Kintsch et al. (1975) showed that the number of propositions as well as the number of different arguments in a sentence influence its reading time. Following Kintsch's propositional model, we used *Densidées* (Lee et al., 2010) to capture conceptual complexity. It is a program able to estimate the mean number of propositions per word in a text using 35 rules relying on lexical and POS clues.

Lexical cohesion : the level of cohesion in a text was measured as the average cosine of all pair of adjacent sentences in the text. Each sentence was represented by a numeric weighted vector (based on words) and projected in a vector space. As suggested by Foltz and al. (1998), two methods were used to define the vector space and weight every word: the *tf-idf* (term frequency-inverse document frequency) and the latent semantic analysis (LSA). The first approach, called "word overlap", corresponds to the "noun overlap" defined by Graesser et al. (2004, 199), except that all type of POS are taken into account. For LSA, we applied a singular value decomposition (SVD), and after comparing various sizes with a cross-validation procedure, we retained a small 15-dimensional space.

3.2.4 Features specific to FFL

Apart from the effect of cognates (Uitdenbogerd, 2005; Tharp, 1939), few features specific to the L2 context were previously investigated. It is probably because such an approach requires to train a model for each pair of language of interest and gather suitable data for evaluation. Since our study intended to design a generic model, we focused on specific predictors affecting L2 reading, whatever the learner's mother tongue is:

Multi-word expressions (MWE): MWEs are acknowledged to cause problems to L2 learners for production (Bahns and Eldaw, 1993). However, the effect of MWE on the reception side remains unclear, especially for beginners. Ozasa et al. (2007) tested the mean of the absolute frequency of all MWEs in a text as an indication of its difficulty, but it appeared non significant. In a latter experiment involving a larger set of MWE-based predictors, François and Watrin (2011) detected a significant, but limited effect. We therefore replicated this set, which includes variables based on the frequencies of MWE, their syntactic structure, their number or their length. Frequencies were estimated on the same corpora as the bigram model described above (Google and *Le Soir*).

Type of text: Finally, we defined five simple variables aiming at identifying dialogues, such as presence of commas, ratio of punctuation, etc. as suggested by Henry (1975). This focus on dialogue was

in its corpus, noticed that features based on parse trees were less efficient than classic ones, such as sentence length or part of speech ratios. Therefore, it seemed unlikely that the information collected by means of syntactic parsers, which are still committing a significant number of errors, at least for French, would belie these findings.

Level of information	Tag	Description of the variable	ρ
Lexical	PA-Alterego	Proportion of absent words from a list of easy words	0.65^3
	X90FFFC	90^{th} percentile of inflected forms for content words only	-0.64^3
	ML3	Unigram model based on lemmas	-0.55^3
	NLM	Mean number of letters per word	0.48^3
	TTR	Type-token ratio based on lemma	0.28^3
	MedNeigh+Freq	Median number of more frequent neighbor for words	-0.23^3
Syntactic	NMP	Mean number of words per sentence	0.62^3
	NWS90	Length (in words) of the 90^{th} percentile sentence	0.61^3
	LSDaoust	Percentage of sentences longer than 30 words (Daoust et al., 1996)	0.56^3
	PPres	Presence of at least one present participle in the text	0.44^3
	PRO.PRE	Ratio of pronouns on prepositions	-0.35^3
	PPres-C	Proportion of present participle among verbs	0.41^3
	PPasse	Presence of at least one past participle	0.39^3
	Impf	Presence of at least one imperfect	0.27^3
	Subp	Presence of at least one subjunctive present	0.27^3
	Cond	Presence of at least one conditional	0.23^3
	Imperatif	Presence of at least one imperative	0.02
	Subi	Presence of at least one subjunctive imperfect	0.05
Semantic	avLocalLsa-Lem	Average intersentential cohesion measured via LSA	0.63^3
	PP1P2	Percentage of P1 and P2 personal pronouns	-0.33^3
Specific	NAColl	Proportion of MWE having the structure NOUN ADJ	0.29^3
	BINGUI	Presence of commas	0.46^3

Table 2: Spearman correlation for some predictors in our set with difficulty. A positive correlation means that the difficulty of texts increases with the value of the predictor. Signification levels are the following $^1 : < 0.05$; $^2 : < 0.01$; and $^3 : < 0.001$.

explained by their extensive use in foreign language teaching, especially in the first levels. Furthermore, even for L1, various scholars stressed the fact that dialogues are often written in a simpler style and have a more mundane content (Dolch, 1948; Flesch, 1948).

3.3 The algorithms

The last step in the development of our formula was to select the most informative subset of features and combine them in a state-of-the-art machine learning algorithm. The algorithms originally considered were six: multinomial and ordinal logistic regression (respectively MLR and OLR), classification trees, bagging, boosting (both based on decision trees) and support vector machine (SVM). However, since the logistic models and the SVM clearly outperformed the others three, we will reported only about those in the next section.

4 Results

The experiments based on this methodology were twofold. First, we assessed the predictive power of each of the 406 features, considered in a bivariate relationship with difficulty. Second, we selected various subsets of features for training models and compared their performance. The two next sections summarize the main findings obtained during these two steps.

4.1 The efficiency of predictors

Spearman correlation was used to assess the efficiency of each predictor, to better account for nonlinear relationships with the criterion. Values for some variables among the four families are reported in Table 2. In accordance with the literature, it appeared that the best family of predictors were the lexical one, followed by the syntactic one. On the contrary, semantic and specific to FFL features did not perform so well, with the exception of the LSA-based feature (*avLocalLsa-Lem*).

Of all predictors, the best was surprisingly *PA-Alterego*, a list-based variable inspired by Dale and Chall (1948), but adapted to the FFL context, since the list of easy words used came from a FFL textbook (*Alter Ego 1*). This suggests that, although the predictive power of "specific to FFL" features was low, specialization to the FFL context was beneficial at other levels.

4.2 The models

Once the best single predictors were identified, it was possible to combine several of them in a readability model for comparison. This required some corpus preparation. Since preliminary experiments showed that the equal prior probabilities are required to ensure a unbiased training, the whole corpus was resampled to get the same number of texts per level (108), which amounted to a total of 648 texts. We then split this smaller corpus into two sets. 240 texts were kept for development purposes, mainly feature selection and estimation of the meta-parameters γ and C for the SVM. The remaining 408 texts were used for evaluating performance of our readability models.

4.2.1 Selection of the features

Several ways of selecting the smallest "best" subset of features were compared, given that some variables are partly redundant when combined together. The first method was based on the structuro-cognitivist assumption that readability formulas should include other features than just lexico-syntactical ones, in order to maximize variety of information. Therefore, we tried an "expert" selection, keeping either the best feature among each of the four families (set **Exp1**), or the two best features (set **Exp2**) [5].

These "expert" approaches were compared to an automatic selection, using either a stepwise procedure [6] for logistic regression (OLR and MLR) or a built-in regularization (Bishop, 2006, 10) for the SVM, based on the 46 best predictors inside each subfamily.

For the sake of comparison, we also defined two other sets: one that corresponds to a random classification (the empty subset), and a baseline, based on two classics predictors (number of letters per word and number of words per sentence), which aimed to mimic classic formulas such as those of Flesch (1948) or Dale and Chall (1948). A summary of the features included in each subset is available in Table 3.

4.2.2 Evaluation of the models

The next step then consisted in training logistic and SVM models for each of the above subsets. Their performances, reported in Table 4, were assessed using five measures: the multiple correlation ratio (R), the accuracy (acc), the adjacent accuracy [7] ($adjacc$), the root mean square error ($rmse$) and the mean absolute error (mae). It should be noted that each of these measures was estimated through a tenfold cross-validation procedure, which allowed us to compare performances of different models with a T-test.

The comparison between the models was performed in two steps. First, we computed T-tests based on $adjacc$ to compare the models based on a same set of features (either **Exp1**, **Exp2**, or **Auto**). This allowed us to pick up the best classifier for each set. In a second step, these three best models were compared the same way, which resulted in the selection of the very best classifier. The decision of adopting the adjacent accuracy as a criterion instead of the accuracy was motivated by our conviction that our system should rather avoid serious errors (i.e. larger than one level) than be more accurate, while sometimes generating terrible mistakes. However, it appeared that both metrics were mostly consistent.

The performance of the different models are displayed in Table 4. It is first interesting to note that the baseline (based on SVM) already gives interesting results. It reaches a classification accuracy of 34%, which is about twice the random. As regards the first model (**Exp1**), based on RLM and including four predictors, it outperforms the baseline by 5%, a difference close to significance ($t(9) = 1.77; p = 0.055$). Therefore, combining variables from several families seems to improve performance over the "classic" baseline, limited to lexico-syntactic features.

This finding is reinforced by the SVM model from **Exp2**, which includes eight features. It performs significantly better than the baseline ($t(9) =$

[5] For the syntactic level, since the two best variables belonged to the same subfamily (see Section 3.2) and were too highly intercorrelated, the 90^{th} percentile of the sentence length (*NWS90*) was replaced by the best feature from another subfamily: the presence of at least one present participle (*PPres*).

[6] In order to suppress as much random effects as possible, the selection process was repeated 100 times via a bootstrapping .632 procedure (Tufféry, 2007, 396-371) and only the features selected at least 50 times out of 100 were kept.

[7] Heilman et al. (2008) defined it as "the proportion of predictions that were within one level of the human assigned label for the given text".

Model name	Classifieur	Set of features
Exp1	OLR, MLR and SVM	PA-Alterego + NMP + avLocalLsa-Lem + BINGUI
Exp2	OLR, MLR and SVM	PA-Alterego + X90FFFC + NMP + PPres + avLocalLsa-Lem + PP1P2 + BINGUI + NAColl
Auto-OLR	OLR	PA-Alterego + NMP + PPres + ML3
Auto-MLR	MLR	PA-Alterego + Cond + Imperatif + Impf + PPasse + PPres + Subi + Subp + BINGUI + TTR + NWS90 + LSDaoust + MedNeigh+Freq
Auto-SVM	SVM	all the 46 variables

Table 3: Results from the two selection process: expert and automatic. Description of the features can be found in Table 2.

Model	Classifier	Parameters	R	acc	adjacc	rmse	mae
Random	/	/	/	16.6	44.4	/	/
Baseline	SVM	$\gamma = 0.05; C = 25$	0.62	34.0	68.2	1.51	1.06
Exp1	RLM	/	0.70	39.4	74.2	1.34	0.97
Exp2	SVM	$\gamma = 0.002; C = 75$	0.73	40.8	77.9	1.28	0.94
Auto-OLR	OLR	/	0.71	39.6	76.1	1.33	0.96
Auto	SVM	$\gamma = 0.004; C = 5$	0.73	49.1	79.6	1.27	0.90

Table 4: Evaluation measures for the best difficulty model from each feature set (**Exp1**, **Exp2** and **Auto**), along with values for a random classification, and the "classic" baseline.

$2.36; p = 0.02$), with an accuracy gain of 7%. However, to that point, it was not clear whether this superiority was indeed a consequence of maximizing the kind of information brought to the model or merely the result of the increased number of predictor.

We thus performed another experiment to address this issue. The model **Exp1** was compared with **Auto-OLR**, the best ordinal logistic model obtained through the stepwise selection (see Tables 4 and 3), and previously discarded as a result of the T-test comparisons. Like **Exp1**, it also contains four predictors, but they are all lexical or syntactic features. Therefore, this model does not maximize the type of information. Surprisingly, we observed that **Auto-OLR** obtained similar and even slightly better performance than **Exp1** ($+2\%$ for both acc and $adjacc$). Thus, the claim that maximizing the source of information should yield better models did not stand on our data.

Finally, our best performing model was based on the **Auto** feature set and SVM. Its accuracy was increased by 8% in comparison with the **Exp2** model, which is clearly a significant improvement ($t(9) = 2.61; p = 0.01$), and outperformed the baseline by 15%. As mentioned previously, this model includes 46 features coming from our four families. It is worth mentioning that the quality of the predictions is not the same across the levels, as shown in Table 5. They are more accurate for classes situated at both ends of the difficulty scale, namely A1, C1 and C2. For A1, this is explained because texts for beginners are more typical, having very short sentences and simple words. However, the case of C1 and C2 classes is more surprising and might be due to some specificities of the learning algorithm.

	A1	A2	B1	B2	C1	C2
Adj. acc.	100%	71%	67%	71%	86%	83%

Table 5: Adjacent accuracy per level, computed on one of the 10 folds. Its adjacent accuracy was 79%, which is very similar to the average value of the model.

We also assessed the specific contribution of each family of features in two ways: on one hand, we trained a model including only the features from this family; on the other hand, we trained a model including all features except those from this family. Results for the four families are displayed at Table 6.

It appeared that the lexical family was the most accurate set of predictors (40.5%) and yielded the highest loss in performance when set aside, especially for adjacent accuracy. In fact, this was the only set whose absence significantly impacted adjacent accuracy, suggesting that the other type of predictors can only improve the accuracy of predictions, but are not able to reduce the amount of critical mistakes. The second best family was, expectedly, the syntactic one. Its accuracy closely match that of the lexical set, although more severe mistakes were made, as shown by the drop in adjacent accuracy. Finally, our two other families was clearly inferior, but they still improved slightly the accuracy of our model, although not the adjacent accuracy.

	Family only		**All except family**	
	Acc.	Adj. acc.	Acc.	Adj. acc.
Lexical	40.5	75.6	41.1	73.5
Syntactic	39.3	69.5	43.2	78.4
Semantic	28.8	61.5	47.8	79.2
FFL	24.9	58.5	47.8	79.6

Table 6: Accuracy and adjacent accuracy (in percentage) for models either using only one family of predictors, or including all 46 features except those of one family.

4.2.3 Comparaison with previous work

Comparisons with other FFL models are difficult to provide: not only there are few formulas available for FFL, but some of these focus on a different audience, making comparability low. This is why we were able to compare our results with only two previous models.

The first of them is a classic readability formula by Kandel and Moles (1958), which is an adaptation of the Flesch (1948) formula for French:

$$Y = 207 - 1.015lp - 0.736lm \qquad (2)$$

where Y is a readability score ranging from 100 (easiest) to 0 (harder); lp is the average number of words per sentence and lm is the average number of syllables per 100 words. Although it was not designed for FFL, we considered it, since it is one of the most well-known formula for French and the two features combined are very general. Their predictive power should not vary much in both contexts, as shown by Greenfield (2004) for English. We evaluated it on the same test corpus as our SVM model and obtained really lower values : a R of 0.55 and an accuracy of 33%.

The second model was that of François (2009), which is based on a multinomial logistic regression including ten features: a unigram model similar to *ML3*, the number of letters per word, the number of words per sentence, and binary variables indicating the presence of a past participle, present participle, imperfect, infinitive, conditional, future and present subjunctive tenses in the text. To our knowledge, this model is the best current generic model available for FFL. On our data, it yielded an accuracy of 41% and an adjacent accuracy of 72.7%, both estimated through a 10-fold cross-validation procedure. Therefore, our new approach achieved an accuracy gain of 8% over this state-of-the-art model, which was considered as a significant difference ($t(9) = 3.72; p = 0.002$).

Apart of those two studies, Uitdenbogerd (2005) also developed recently a FFL formula. However, as explained previously, this work focused on a specific category of L2 readers, the English-speakers learning FFL, which resulted in a different problem. She reported a higher R than us (0.87 against 0.73). However, this value might be the training one and was estimated on a small amount of novel beginnings. It is therefore likely that our model generalize better, especially across genres and L2 readers with different L1 backgrounds.

5 Discussion and conclusion

In this paper, we introduced a new "AI readability" formula for FFL, able to predict the level of texts according to the largely-spread CEFR scale. Our model is based on a SVM classifier and combines 46 features corresponding to several levels of linguistic information. Among those, we suggested some new features: the normalized TTR and the set of variables based on several characteristics of words' neighbors. Comparing our approach with two previously published formulas, our model significantly outperformed both these works. Therefore, it represent a robust generic solution for FFL readers willing to find various kind of texts that suit their linguistic abilities.

Besides the creation of a new FFL readability formula, this study produced two valuable insights. First, we showed that maximizing the type of linguistic information might not be the best path to go, since a model based on four lexico-syntactic features yielded predictions as accurate as those of a model relying on our *Exp1* set of variables. However, this finding might be partly accounted by the lower predictive power of the features from the semantic and specific-to-FFL family, with the notable exception of the LSA-based predictor (*avLocalLsa-Lem*), which is the third best predictor when considered alone.

This leads us to our second finding, relative to the set of semantic features. Yet their importance was largely praised in the structuro-cognitivist paradigm and in most of the recent works, our experiments cast serious doubts about their efficiency, at least in a L2 context. Not only the expert models, to which we imposed the presence of one or two semantic predictors, did not perform the best, but none of the features from our semantic set was retained during the automatic selection of the variables for the logistic models. On the contrary, in some subsets, the LSA-based feature was sometimes considered as collinear with the other variables. Finally and foremost, we showed that dropping the semantic features did not impact significantly the performance of our best model.

With reservations one may have because of the limited number of semantic predictors in our set, these results however raise some concerns about whether the information coming from semantic variables is really different from that carried by lexico-syntactic features. Our results clearly show that this may not be the case. This conclusion contradicts the assumptions of the structuro-cognitivist paradigm, but corroborates Chall and Dale (1995)'s view that the information carried by semantic predictors is largely correlated with that of lexico-syntactical ones.

Further investigation on this issue would definitely be worthwhile, since several facts could explain these contradictory findings. First, it might be that semantic and lexical predictors are correlated because the methods used for the parameterization of the semantic factors heavily relie on lexical information. This is the case for the LSA, as well as for the propositional approach of the content density.

Alternatively, this difference with other work in L1 could be due to the L2 context. Chall and Dale (1995) explained that the lexicon and the syntax are more important for children learning to read than for more advanced readers, who then become more sensitive to organisationnal aspects. From the threshold hypothesis (Alderson, 1984), we know that before reaching a sufficient level of proficiency, L2 learners struggle mostly with the lexicon and the syntactic structures. This might explain why lexico-syntactic predictors were so predominant in our experiments. Some further experiments are thus needed to investigate which of these facts better account for our findings on the semantic features.

A last avenue of research worth mentioning would be to develop the family of specific-to-FFL predictors, to determine whether taking into account the impact of a given L1 language on the readability of L2 texts would increase performance over a generic model enough so that tuning efforts are worthwhile.

Acknowledgments

Thomas François was an Aspirant F.N.R.S. when this work was performed. The writing of this paper was done while being a recipient of a Fellowship of the Belgian American Educational Foundation. We thank both for their support. We would also like to acknowledge the invaluable help of Bernadette Dehottay for the collection of the corpus used in that study.

References

J.C. Alderson. 1984. Reading in a foreign language : a reading problem or a language problem ? In J.C. Alderson and A.H Urquhart, editors, *Reading in a Foreign Language*, pages 1–24. Longman, New York.

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

J. Bahns and M. Eldaw. 1993. Should We Teach EFL Students Collocations? *System*, 21(1):101–14.

A. Berthet, C. Hugot, V. Kizirian, B. Sampsonis, and M. Waendendries. 2006. *Alter Ego 1*. Hachette, Paris.

C.M. Bishop. 2006. *Pattern recognition and machine learning*. Springer, New York.

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

J.S. Bowers. 2000. In defense of abstractionist theories of repetition priming and word identification. *Psychonomic bulletin and review*, 7(1):83–99.

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

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

S. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4):359–393.

K. Collins-Thompson and J. Callan. 2004. A language modeling approach to predicting reading difficulty. In

Proceedings of HLT/NAACL 2004, pages 193–200, Boston, USA.

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

M. Coltheart. 1978. Lexical access in simple reading tasks. In G. Underwood, editor, *Strategies of information processing*, pages 151–216. Academic Press, London.

A. Conquet. 1957. *La lisibilité*. Assemblée Permanente des CCI de Paris, Paris.

C.M. Cornaire. 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.

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

E. Dale and R.W. Tyler. 1934. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly*, 4:384–412.

F. Daoust, L. Laroche, and 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.

G. de Landsheere. 1963. Pour une application des tests de lisibilité de Flesch à la langue française. *Le Travail Humain*, 26:141–154.

E.W. Dolch. 1948. *Problems in reading*. The Garrard Press, Champaign : Illinois.

W.B. Elley. 1969. The assessment of readability by noun frequency counts. *Reading Research Quarterly*, 4(3):411–427.

L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *COLING 2010: Poster Volume*, pages 276–284.

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

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

T. François and P. Watrin. 2011. On the contribution of MWE-based features to a readability formula for French as a foreign language. In *Proceedings of the International Conference RANLP 2011*.

T. François. 2009. Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. In *Proceedings of the 12th Conference of the EACL : Student Research Workshop*, pages 19–27.

T. François. 2011a. La lisibilité computationnelle : un renouveau pour la lisibilité du français langue première et seconde ? *International Journal of Applied Linguistics (ITL)*, 160:75–99.

T. François. 2011b. *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.

W.A. Gale and G. Sampson. 1995. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237.

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

A.C. Graesser, D.S. McNamara, M.M. Louwerse, and Z. Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

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

M. Heilman, K. Collins-Thompson, and M. Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–8.

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

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

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

W. Kintsch and D. Vipond. 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.

W. Kintsch, E. Kozminsky, W.J. Streby, G. McKoon, and J.M. Keenan. 1975. Comprehension and recall of text as a function of content variables1. *Journal of Verbal Learning and Verbal Behavior*, 14(2):196–214.

H. Lee, P. Gambette, E. Maillé, and C. Thuillier. 2010. Densidées: calcul automatique de la densité des idées dans un corpus oral. In *Actes de la douxime Rencontre des tudiants Chercheurs en Informatique pour le Traitement Automatique des langues (RECITAL)*.

I. Lorge. 1944. Predicting readability. *the Teachers College Record*, 45(6):404–419.

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

B.J.F. Meyer. 1982. Reading research and the composition teacher: The importance of plans. *College composition and communication*, 33(1):37–49.

J.B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, The Google Books Team, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

J.R. Miller and W. Kintsch. 1980. Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory*, 6(4):335–354.

B. New, M. Brysbaert, J. Veronis, and C. Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(04):661–677.

R.E. O'Connor, K.M. Bell, K.R. Harty, L.K. Larkin, S.M. Sackor, and N. Zigmond. 2002. Teaching reading to poor readers in the intermediate grades: A comparison of text difficulty. *Journal of Educational Psychology*, 94(3):474–485.

T. Ozasa, G. Weir, and M. Fukui. 2007. Measuring readability for Japanese learners of English. In *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*.

E. Pitler and A. Nenkova. 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.

J.C. Redish and J. Selzer. 1985. The place of readability formulas in technical communication. *Technical communication*, 32(4):46–52.

F. Richaudeau. 1979. Une nouvelle formule de lisibilité. *Communication et Langages*, 44:5–26.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

S.E. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530.

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

E.A. Smith. 1961. Devereaux readability index. *The Journal of Educational Research*, 54(8):289–303.

A.J. Stenner. 1996. Measuring reading comprehension with the lexile framework. In *Fourth North American Conference on Adolescent/Adult Literacy*.

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

S. Tufféry. 2007. *Data mining et statistique décisionnelle l'intelligence des données*. Éd. Technip, Paris.

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

P. van Oosten, V. Hoste, and D. Tanghe. 2011. A posteriori agreement as a quality measure for readability prediction systems. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6609 of *Lecture Notes in Computer Science*, pages 424–435. Springer, Berlin / Heidelberg.