Combining a Statistical Language Model with Logistic Regression to Predict the Lexical and Syntactic Difficulty of Texts for FFL

Thomas L. François
Aspirant FNRS
CENTAL (Center for Natural Language Processing)
Université catholique de Louvain
1348 Louvain-la-Neuve, Belgium
thomas.francois@uclouvain.be

Abstract

Reading is known to be an essential task in language learning, but finding the appropriate text for every learner is far from easy. In this context, automatic procedures can support the teacher's work. Some tools exist for English, but at present there are none for French as a foreign language (FFL). In this paper, we present an original approach to assessing the readability of FFL texts using NLP techniques and extracts from FFL textbooks as our corpus. Two logistic regression models based on lexical and grammatical features are explored and give quite good predictions on new texts. The results shows a slight superiority for multinomial logistic regression over the proportional odds model.

1 Introduction

The current massive mobility of people has put increasing pressure on the language teaching sector, in terms of the availability of instructors and suitable teaching materials. The development of Intelligent Computer Aided Language Learning (ICALL) has helped both these needs, while the Internet has increasingly been used as a source of exercises. Indeed, it allows immediate access to a huge number of texts which can be used for educational purposes, either for classical reading comprehension tasks, or as a corpus for the creation of various automatically generated exercises.

However, the strength of the Internet is also its main flaw : there are so many texts available to the teacher that he or she can get lost. Having gathered some documents suitable in terms of subject matter, teachers still have to check if their readability levels are suitable for their students : a highly time-consuming task. This is where NLP applications able to classify documents according to their reading difficulty level can be invaluable.

Related research will be discussed in Section 2. In Section 3, the distinctive features of the corpus used in this study and a difficulty scale suitable for FFL text classification are described. Section 4 focuses on the independent linguistic variables considered in this research, while the statistical techniques used for predictions are covered in Section 5. Section 6 gives some details of the implementations, and Section 7 presents the first results of our models. Finally, Section 8 sums up the contribution of this article before providing a programme for future work and improvement of the results.

2 Related research

The measurement of the reading difficulty of texts has been a major concern in the English-speaking literature since the 1920s and the first formula developed by Lively and Pressey (1923). The field of readability has since produced many formulae based on simple lexical and syntactic measures such as the average number of syllables per word, the average length of sentences in a piece of text (Flesch, 1948; Kincaid et al., 1975), or the percentage of words not on a list combined with the average sentence length (Chall and Dale, 1995).

French-speaking researchers discovered the field of readability in 1956 through the work of André Conquet, *La lisibilité* (1971), and the first two formulae for French were adapted from Flesch (1948) by Kandel and Moles (1958) and de Landsheere (1963). Both of these researchers stayed quite close to the Flesch formula, and in so doing they failed to take into account some specificities of the French language.

Henry (1975) was the first to introduce specific formulae for French. He used a larger set of variables to design three formulae : a complete, an automatic and a short one, each of which

Proceedings of the EACL 2009 Student Research Workshop, pages 19–27,
Athens, Greece, 2 April 2009.

was adapted for three different educational levels. His formulae are by far the best and most frequently used in the French-speaking world. Later, Richaudeau (1979) suggested a criteria of "linguistic efficiency" based on experiments on short-term memory, while Mesnager (1989) coined what is still, to the best of our knowledge, the most recent specific formula for French, with children as its target.

Compared to the mass of studies in English, readability in French has never enthused the research community. The cultural reasons for this are analysed by Bossé-Andrieu (1993) (who basically argues that the idea of measuring text difficulty objectively seems far too pragmatic for the French spirit). It follows that there is little current research in this field: in Belgium, the Flesch formula is still used to assess the readability of articles in journalism studies. This example also shows that the French-specific formulae are not much used, probably because of their complexity (Bossé-Andrieu, 1993).

Of course, if there is little work on French readability, there is even less on French as a foreign language. We only know the study of Cornaire (1988), which tested the adaptation of Henry's short formula to French as a foreign language, and that of Uitdenbogerd (2005), which developed a new measure for English-speaking learners of French, stressing the importance of cognates when developing a new formula for a related language.

Therefore, we had to draw our inspiration from the English-speaking world, which has recently experienced a revival of interest in research on readability. Taking advantage of the increasing power of computers and the development of NLP techniques, researchers have been able to experiment with more complex variables. Collins-Thompson et al. (2005) presented a variation of a multinomial naive Bayesian classifier they called the "Smoothed Unigram" model. We retained from their work the use of language models instead of word lists to measure lexical complexity. Schwarm and Ostendorf (2005) developed a SVM categoriser combining a classifier based on trigram language models (one for each level of difficulty), some parsing features such as average tree height, and variables traditionally used in readability. Heilman et al. (2007) extended the "Smoothed Unigram" model by the recognition of syntactic structures, in order to assess L2 English texts. Later, they improved the combination of their various lexical and grammatical features using regression methods (Heilman et al., 2008). We also found regression methods to be the most efficient of the statistical models with which we experimented. In this article, we consider some ways to adapt these various ideas to the specific case of FFL readability.

3 Corpus description

In the development of a new readability formula, the first step is to collect a corpus labelled by reading-difficulty level, a task that implies agreement on the difficulty scale. In the US, a common choice is the 12 American grade levels corresponding to primary and secondary school. However, this scale is less relevant for FFL education in Europe. So, we looked for another scale.

Given that we are looking for an automatic way of measuring text complexity for FFL learners participating in an educational programme, an obvious choice was the difficulty scale used for assessing students' levels in Europe, that is the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001) . The CEFR has six levels: A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). However differences in learners' skills can be quite substantial at lower levels, so we divided each of the A1, A2 and B1 grades in two, thus obtaining a total of nine levels.

We still needed to find a corpus labelled according to these nine classes. Unlike traditional approaches, based on a limited set of texts usually standardised by applying a closure test to a target population, our NLP-oriented approach required a large number of texts on which the statistical models could be trained. For that reason we opted for FFL textbooks as a corpus. With the appearance of the CEFR, FFL textbooks have undergone a kind of standardisation and their levels have been clarified. It is thus feasible to gather a large number of documents already labelled in terms of the CEFR scale by experts with an educational background.

However, not every textbook can be used as a document source. Likewise, not all the material from FFL textbooks is appropriate. We established the following criteria for selecting textbooks and texts:

- The CEFR was published in 2001, so only

textbooks published since then were considered. This restriction also ensures that the language resembles present-day spoken French.

- The target population for our formula is young people and adults. Therefore, only textbooks intended for this public were used.
- We retained only those texts made up of complete sentences, linked to a reading comprehension task. So, all the transcriptions of listening comprehension tasks were ignored. Similarly, all instructions to the students were excluded, because there is no guarantee the language employed there is the same as the rest of the textbook material (metalinguistic terms and so on can be found there).

Up to now, using these criteria, we have gathered more than 1,500 documents containing about 440,000 tokens. Texts cover a wide variety of subjects ranging from French literature to newspaper articles, as well as numerous dialogues, extracts from plays, cooking recipes, etc. The goal is to have as wide a coverage as possible, to achieve maximum generalisability of the formula, and also to check what sort of texts it does not fit (e.g. statistical descriptive analyses have considered songs and poems as outliers).

4 Selection of lexical and syntactic variables

Any text classification tasks require an object (here a text) to be parameterised into variables, whether qualitative or quantitative. These independent variables must correlate as strongly as possible with the dependent variable representing difficulty in order to explain the text's complexity, and they should also account for the various dimensions of the readability phenomenon. Traditional approaches to readability have been sharply criticised with respect to this second requirement by Kintsch and Vipond (1979) and Kemper (1983), who both insist on the importance of including the conceptual properties of texts (such as the relations between propositions and the "inference load"). However, these new approaches have not resulted in any easily reproducible computational models, leading current researchers to continue to use the classic semantic and grammatical variables, enhancing them with NLP techniques.

Because this research only spans the last year, attempts to discover interesting variables are still at an early stage. We explored the efficiency of some traditional features such as the type-token ratio, the number of letters per word, and the average sentence length, and found that, on our corpus, only the word length and sentence length correlated significantly with difficulty. Then, we add two NLP-oriented features, as described below: a statistical language model and a measure of tense difficulty.

4.1 The language model

The lexical difficulty of a text is quite an elaborate phenomenon to parameterise. The logistic regression models we used in this study require us to reduce this complex reality to just one number, the challenge being to achieve the most informative number. Some psychological work (Howes and Solomon, 1951; Gerhand and Barry, 1998; Brysbaert et al., 2000) suggests that there is a strong relationship between the frequency of words and the speed with which they are recognised. We therefore opted to model the lexical difficulty for reading as the global probability of a text T (with N tokens) occurring:

$$P(T) = P(t_1)P(t_2 \mid t_1) \cdots P(t_n \mid t_1, t_2, \ldots, t_{n-1}) \quad (1)$$

This equation raises two issues :

1. Estimating the conditional probabilities. It is well-known that it is impossible to train such a model on a corpus, even the largest one, because some sequences in this equation are unlikely to be encountered more than once. However, following Collins-Thompson and Callan (2005), we found that a simple smoothed unigram model could give good results for readability. Thus, we assumed that the global probability of a text T could be reduced to:

$$P(T) = \prod_{i=1}^{n} p(t_i) \quad (2)$$

where $p(t_i)$ is the probability of meeting the token t_i in French; and n is the number of tokens in a text.

2. Deciding what is the best linguistic unit to consider. The equations introduced above use

tokens, as is traditional in readability formulae, but the inflected nature of French suggests that lemmas may be a better alternative. Using tokens means that words taking numerous inflected forms (such as verbs), have their overall probability split between these different forms. Consequently, compared to seldom – or never – inflected words (such as adverbs, prepositions, conjunctions), they seem less frequent than they really are. Second, using tokens presupposes a theoretical position according to which learners are not able to link an inflected form with its lemma. Such a view seems highly questionable for the majority of regular forms.

In order to settle this issue, we trained three language models: one with lemmas (LM1), another with inflected forms disambiguated according to their tags (LM2), and a third one with inflected forms (LM3). The experiment was not very conclusive, since the models all correlated with the dependent variable to a similar extent, having Pearson's r coefficients of -0.58, -0.58, and -0.59 respectively. However, three factors militate in favour of the lemma model: as well as theoretical likelihood, it is the model which is most sensitive to outliers and most prone to measurement error. This suggests that, if we can reduce this error, the lemma model may prove to be the best predictor of the three.

As a consequence of these considerations, we decided to compute the difficulty of the text by using Equation 2 adapted for lemmas and, for computational reasons, the logarithm of the probabilities:

$$P(T) = exp(\sum_{i=1}^{n} \log[p(lem_i)]) \qquad (3)$$

The resulting value is still correlated with the length of the text, so it has to be normalised by dividing it by N (the number of words in the text). These operations give in a final value suitable for the logistic regression model. More information about the origin and smoothing of the probabilities is given in Section 6.

4.2 Measuring the tense difficulty

Having considered the complexity of a text's syntactic structures through the traditional factor of the "mean number of words per sentence", we decided to also take into account the difficulty of the conjugation of the verbs in the text. For this purpose, we created 11 variables, each representing one tense or class of tenses: conditional, future, imperative, imperfect, infinitive, past participle, present participle, present, simple past, subjunctive present and subjunctive imperfect.

The question then arose as to whether it would be better to treat these variables as binary or continuous. Theoretical justifications for a binary parameterisation lie in the fact that a text becomes more complex for a L2 language learner when there is a large variety of tenses, especially difficult ones. The proportion of each tense seems less significant. For this reason, we opted for binary variables. The other way of parameterising the data should nevertheless be tested in further research.

5 The regression models

By the end of the parameterisation stage, each text of the corpus has been reduced to a vector comprising the 14 following predictive variables : the result of the language model, the average number of letters per word[1], the average number of words per sentence and the 11 binary variables for tense complexity.

Each vector also has a label representing the level of the text, which is the dependent variable in our classification problem. From a statistical perspective, this variable may be considered as a nominal, ordinal, or interval variable, each level of measurement being linked to a particular regression technique: multiple linear regression for interval data; a popular cumulative logit model called proportional odds for ordinal data; and multinomial logistic regression for nominal variables. Therefore, identifying the best scale of measurement is an important issue for readability.

From a theoretical perspective, viewing the levels of difficulty as an interval scale would imply that they are ordered and evenly spaced. However, most FFL teachers would disagree with this assumption: it is well known that the higher levels take longer to complete than the earlier ones. So, a more realistic position is to consider text difficulty as an ordinal variable (since the CEFR levels are

[1]Pearson's r coefficient between the language model and the average number of letters in the words was -0.68. This suggests that there is some independent information in the length of the words that can be used for prediction.

ordered). The third alternative, treating the levels as a nominal scale, is not intuitively obvious to a language teacher, because it suggests that there is no particular order to the CEFR levels.

From a practical perspective, things are not so clear. Traditional approaches have usually viewed difficulty as an interval scale and applied multiple linear regression. Recent NLP perspective have either considered difficulty as an ordinal variable (Heilman et al., 2008), making use of logistic regression, or as a nominal one, implementing classifiers such as the naive Bayes, SVM or decision tree. Such a variety of practices convinced us that we should experiment with all three scales of measurement.

In an exploratory phase, we compared regression methods and decision tree classifiers on the same corpus. We found that regression was more precise and more robust, due to the current limited size of the corpus. Linear regression was discarded because it gave poor results during the test phase. So we retained two logistic regression models, the PO model and the MLR model, which are presented in the next section.

5.1 Proportional odds (PO) model

Logistic regression is a statistical technique first developed for binary data. It generally describes the probability of a 0 or 1 outcome with an S-shaped logistic function (see Hosmer and Lemeshow (1989) for details). Adaptation of the logistic regression for J ordinal classes involves a model with $J-1$ response curves of the same shape. For a fixed class j, each of these response functions is comparable to a logistic regression curve for a binary response with outcomes $Y \leq j$ and $Y > j$ (Agresti, 2002), where Y is the dependent variable.

The PO model can be expressed as:

$$\text{logit}[P(Y \leq j \mid \mathbf{x})] = \alpha_j + \boldsymbol{\beta}'\mathbf{x} \qquad (4)$$

In Equation 4, $\mathbf{x}$ is the vector containing the independent variables, α_j is the intercept parameter for the j_{th} level and $\boldsymbol{\beta}$ is the vector of regression coefficients. From this formula, the particularity of the PO model can be observed: it has the same set, $\boldsymbol{\beta}$, of parameters for each level. So, the response functions only differ in their intercepts, α_j. This simplification is only possible under the assumption of ordinality.

Using this cumulative model, when $2 \leq j \leq J$, the estimated probability of a text Y belonging to the class j can be computed as:

$$P(Y = j \mid \mathbf{x}) = \text{logit}[P(Y \leq j \mid \mathbf{x})] - \text{logit}[P(Y \leq j-1 \mid \mathbf{x})] \quad (5)$$

When j = 1, $P(Y = 1 \mid \mathbf{x})$ is equal to $P(Y \leq j \mid \mathbf{x})$.

We said above that this model involves a simplification, based on the proportional odds assumption. This assumption needs to be tested with the chi-squared form of the score test (Agresti, 2002). The lower the chi-squared value, the better the PO model fits the data.

5.2 Multinomial logistic regression

Multinomial logistic regression is also called "baseline category", because it compares each class Y with a reference category, often the first one (Y_1), in order to regress to the binary case. Each pair of classes (Y_j, Y_1) can then be described by the ratio (Agresti, 2002, p. 268):

$$log\frac{P(Y = j \mid \mathbf{x})}{P(Y = 1 \mid \mathbf{x})} = \alpha_j + \boldsymbol{\beta}_j{}'\mathbf{x} \qquad (6)$$

where the notation is as given above. On the basis of these J-1 regression equations, it is possible to compute the probability of a text belonging to difficulty level j using the values of its features contained in the vector $\boldsymbol{x}$. This may be calculated using the equation (Agresti, 2002, p. 271):

$$P(Y = j \mid \mathbf{x}) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j{}'\mathbf{x})}{1 + \sum_{h=2}^{J} \exp(\alpha_h + \boldsymbol{\beta}_j{}'\mathbf{x})} \qquad (7)$$

Notice that for the baseline category (here, $j = 1$), α_1 and $\boldsymbol{\beta_1} = 0$. Thus, when looking for the probability of a text belonging to the baseline level, it is easy to compute the numerator, since $\exp(0) = 1$. The value of the denominator is the same for each j.

Heilman et al. (2008) drew attention to the fact that the MLR model multiplies the number of parameters by $J-1$ compared to the PO model. Because of this, they recommend using the PO model.

6 Implementation of the models

Having covered the theoretical aspects of our model, we will now describe some of the particularities of our implementation.

6.1 The language model: probabilities and smoothing

For our language model, we need a list of French lemmas with their frequencies of occurrence. Getting robust estimates for a large number of lemmas requires a very large corpus and is a time-consuming process. We used *Lexique3*, a lexicon provided by New et al. (2001) and developed from two corpora: the literary corpus *Frantext* containing about 15 million of words; and a corpus of film subtitles (New et al., 2007), with about 50 million words. The authors drew up a list of more than 50,000 tagged lemmas, each of which is associated with two frequency estimates, one from each corpus.

We decided to use the frequencies from the subtitle corpus, because we think it gives a more accurate image of everyday language, which is the language FFL teaching is mainly concerned with. The frequencies were changed into probabilities, and smoothed with the Simple Good-Turing algorithm described by Gale and Sampson (1995). This step is necessary to solve another well-known problem in language models: the appearance in a new text of previously unseen lemmas. In this case, since the logarithm of probabilities is used, an unseen lemma would result in a infinite value. In order to prevent this, a smoothing process is used to shift some of the model's probability mass from seen lemmas to unseen ones.

Once we had obtained a good estimate of the probabilities, we could analyse the texts in the corpus. Each of them was lemmatised and tagged using the TreeTagger (Schmid, 1994). This NLP tool allows us to distinguish between homographs that can represent different levels of difficulty. For instance, the word *actif* is quite common as an adjective, but the noun is infrequent and is only used in the business lexicon. This distinction is possible because *Lexique3* provides tagged lemmas.

6.2 Variable selection

Having gathered the values for the 14 dependent variables, it was possible to train the two statistical models.[2] However, an essential requirement prior to training is feature selection. This procedure, described by Hosmer and Lemeshow (1989), consists of examining models with one, two, three, etc., variables and comparing them to the full model according to some specified criteria so as to select one that is both efficient and parsimonious. For logistic regression, the criterion selected is the AIC (Akaike's Information Criterion) of the model. This can be obtained from:

$$\text{AIC} = -2\text{log-likelihood} + 2k \qquad (8)$$

where k is the number of parameters in the model, and the log-likelihood value is the result of a calculation detailed by Hosmer and Lemeshow (1989).

We applied the stepwise algorithm to our data, trying both a backward and a forward procedure. They converged to a simpler model containing only 10 variables: the value obtained from our language model, the number of letters per word, the number of words per sentence, the past participle, the present participle, and the imperfect, infinitive, conditional, future and present subjunctive tenses. Presumable the imperative and present tenses are so common that they do not have much discriminative power. On the other hand, the imperfect subjunctive is so unusual that it is not useful for a classification task. However, the non-appearance of the simple past is surprising, since it is a narrative tense which is not usually introduced until an advanced stage in the learning of French. This phenomenon deserves further investigation in the future.

7 First results

To the best of our knowledge, no one has previously applied NLP technologies to the specific issue of the readability of texts for FFL learners. So, any comparisons with previous studies are somewhat flawed by the fact that neither the target population nor the scale of difficulty is the same. However, our results can be roughly compared to some of the numerous studies on L1 English readability presented in Section 2. Before making this comparison, we will analyse the predictive ability of the two models.

7.1 Models evaluation

The evaluation measures most commonly employed in the literature are Pearson's product-moment correlation coefficient, prediction accuracy as defined by Tan et al. (2005), and adjacent accuracy. Adjacent accuracy is defined by Heilman et al. (2008) as "the proportion of predictions that were within one level of the human-assigned

[2] All statistical computations were performed with the MASS package (Venables and Ripley, 2002) of the R software.

Measure	PO model	MLR model
Results on training folds		
Correl.	0.786	0.777
Exact Acc.	32.5%	38%
Adj. Acc.	70%	71.3%
Results on test folds		
Correl.	0.783	0.772
Exact Acc.	32.4%	38%
Adj. Acc.	70%	71.2%

Table 1: Mean Pearson's r coefficient, exact and adjacent accuracies for both models with the ten-fold cross-validation evaluation.

label for the given text". They defended this measure by arguing that even human-assigned reading levels are not always consistent. Nevertheless, it should not be forgotten that it can give optimistic values when the number of classes is small.

Exploratory analysis of the corpus highlighted the importance of having a similar number of texts per class. This requirement made it impossible to use all the texts from the corpus. Some 465 texts were selected, distributed across the 9 levels in such a way that each level contained about 50 texts. Within each class, an automatic procedure discarded outliers located more than 3σ from the mean, leaving 440 texts. Both models were trained on these texts.

The results on the training corpus were promising, but might be biased. So, we turned to a ten-fold cross-validation process which guarantees more reliable values for the three evaluation measures we had chosen, as well as a better insight into the generalisability of the two models. The resulting evaluation measures for training and test folds are shown in Table 1. The similarity between them clearly shows that, with 440 observations, both the models were quite robust. On this corpus, multinomial logistic regression was significantly more accurate (with 38% of texts correctly classified against 32.4% for the PO model), while Pearson's R was slightly higher for the PO model.

These results suggest that the exact accuracy may be a better indicator of performance than the correlation coefficient. However they conflict with Heilman et al.'s (2008) conclusion that the PO model performed better than the MLR one. This discrepancy might arise because the PO model was less accurate for exact predictions, but better when the adjacent accuracy by level was taken into account. However, the data in Table 2 do not support this hypothesis; rather they confirm the superiority of the MLR model when adjacent accuracy is considered. In fact, PO model's lower performance seems to be due to a lack of fit to the data, as revealed by the result of the score test for the proportional-odds assumption. This yielded a p-value below 0.0001, clearly showing that the PO model was not a good fit to the corpus.

There remains one last issue to be discussed before comparing our results to those of other studies: the empirical evidence for tense being a good predictor of reading difficulty. We selected tenses because of our experience as FLE teacher rather than on theoretical or empirical grounds. However we found that exact accuracy decreased by 10% when the tense variables were omitted from the models. Further analysis showed that the tense contributed significantly to the adjacent accuracy of classifying the C1 and C2 texts.

7.2 Comparison with other studies

As stated above, it is not easy to compare our results with those of previous studies, since the scale, population of interest and often the language are different. Furthermore, up till now, we have not been able to run the classical formulae for French (such as de Landsheere (1963) or Henry (1975)) on our corpus. So we are limited to comparing our evaluation measures with those in the published literature.

With multinomial logistic regression, we obtained a mean adjacent accuracy of 71% for 9 classes. This result seems quite good compared to similar research on L1 English by Heilman et al. (2008). Using more complex syntactic features, they obtained an adjacent accuracy of 52% with a PO model, and 45% with a MLR model. However, they worked with 12 levels, which may explain their lower percentage.

For French, Collins-Thompson and Callan (2005) reported a Pearson's R coefficient of 0.64 for a 5-classes naive Bayes classifier while we obtained 0.77 for 9 levels with MLR. This difference might be explained by the tagging or the use of better-estimated probabilities for the language model. Further research on this point to determine the specificities of an efficient approach to French readability appears very promising.

Level	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2	Mean
PO model	91%	91%	67%	68%	53%	55%	56%	86%	68%	70%
MLR model	93%	90%	69%	51%	59%	56%	64%	88%	73%	71%

Table 2: Mean adjacent accuracy per level for PO model and MLR model (on the test folds).

8 Discussion and future research

This paper has proposed the first readability "formula" for French as a foreign language using NLP and statistical models. It takes into account some particularities of French such as its inflected nature. A new scale to assess FFL texts within the CECR framework, and a new criteria for the corpus involving the use of textbooks, have also been proposed. The two logistic models applied to a 440-text corpus gave results consistent with the literature. They also showed the superiority of the MLR model over the PO model. Since Heilman et al. (2008) found the opposite, and the intuitive view is that levels should be described by an ordinal scale of measurement, this issue clearly needs further investigation.

This research is still in progress, and further analyses are planned. The predictive capacity of some other lexical and grammatical features will be explored. At the lexical level, statistical language models seems to be best, and tagging the texts to work with lemmas turned out to be efficient for French, although it has not been shown to be superior to disambiguated inflected forms. Moreover, due to their higher sensibility to context, smoothed n-grams might represent an alternative to lemmas.

Once the best unit has been selected, some other issues remain: it is not clear whether a model using the probabilities of this unit in the whole language or probabilities per level (Collins-Thompson and Callan, 2005) would be more efficient. We also wonder whether the L1 frequencies of words are similar to those in L2 ? FFL textbooks use a controlled vocabulary, linked to specific situational tasks, which suggests that it is highly possible that the frequencies of words in FFL differ from those in mother-tongue French.

Grammatical features have been taken into account through simple parameterisation. More complex measures (such as the presence of some syntactic structures (Heilman et al., 2007) or the characteristics of a syntactic-parsing tree) have been explored in the literature. We hope that including such factors may result in improved accuracy for our model. However, these techniques are probably dependent on the quality of the parser's results. Parsers for French are less accurate than those for English, which may generate some noise in the analysis.

Finally, we intend to explore the performance of other classification techniques. Logistic regression was the most efficient of the statistical models we tested, but as our corpus grows, more and more data is becoming available, and data mining approaches may become applicable to the text-categorization problem for FFL readability. Support vector machines have already been shown to be useful for readability purposes (Schwarm and Ostendorf, 2005). We also want to try aggregating approaches such as boosting, bagging, and random forests (Breiman, 2001), since they claim to be effective when the sample is not perfectly representative of the population (which could be true for our data). These analyses would aim to illuminate some of the assets and flaws of each of the statistical models considered.

Acknowledgments

Thomas L. François is supported by the Belgian Fund for Scientific Research (FNRS), as is the research programme from which this material comes.

I would like to thank my directors, Prof. Cédrick Fairon and Prof. Anne-Catherine Simon, my colleagues, Laure Cuignet and the anonymous reviewers for their valuable comments.

References

Alan Agresti. 2002. *Categorical Data Analysis. 2nd edition*. Wiley-Interscience, New York.

J. Bossé-Andrieu. 1993. La question de la lisibilité dans les pays anglophones et les pays francophones. *Technostyle, Association canadienne des professeurs de rédaction technique et scientifique*, 11(2):73–85.

L. Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

M. Brysbaert, M. Lange, and I. Van Wijnendaele. 2000. The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, 12(1):65–85.

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

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.

A. Conquet. 1971. *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 and Education Committee and Council for Cultural Co-operation. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

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

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

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

S. Gerhand and C. Barry. 1998. Word frequency effects in oral reading are not merely age-of-acquisition effects in disguise. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 24(2):267–283.

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

M. Heilman, K. Collins-Thompson, and M. Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. *Association for Computational Linguistics*, The 3rd Workshop on Innovative Use of NLP for Building Educational Applications:1–8.

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

D.W. Hosmer and S. Lemeshow. 1989. *Applied Logistic Regression*. Wiley, New York.

D.H. Howes and R.L. Solomon. 1951. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40):1–4.

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.

J. Kincaid, R.P. Fishburne, R. Rodgers, and B. Chissom. 1975. Derivation of new readability formulas for navy enlisted personnel. *Research Branch Report*, 85.

W. Kintsch and D. Vipond. 1979. Reading comprehension and readability in educational practice and psychological theory. *Perspectives on Memory Research*, pages 329–366.

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

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

B. New, C. Pallier, L. Ferrand, and R. Matos. 2001. Une base de données lexicales du français contemporain sur internet: LEXIQUE. *LAnnée Psychologique*, 101:447–462.

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.

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.

P.-N. Tan, M. Steinbach, and V. Kumar. 2005. *Introduction to Data Mining*. Addison-Wesley, Boston.

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

W.N. Venables and B.D. Ripley. 2002. *Modern Applied Statistics with S*. Springer, New York.