Do NLP and machine learning improve traditional readability formulas?

Thomas François
University of Pennsylvania
CENTAL, UCLouvain
3401 Walnut Street Suite 400A
Philadelphia, PA 19104, US
frthomas@sas.upenn.edu

Eleni Miltsakaki
University of Pennsylvania & Choosito!
3401 Walnut Street Suite 400A
Philadelphia, PA 19104, US
elenimi@seas.upenn.edu

Abstract

Readability formulas are methods used to match texts with the readers' reading level. Several methodological paradigms have previously been investigated in the field. The most popular paradigm dates several decades back and gave rise to well known readability formulas such as the Flesch formula (among several others). This paper compares this approach (henceforth "classic") with an emerging paradigm which uses sophisticated NLP-enabled features and machine learning techniques. Our experiments, carried on a corpus of texts for French as a foreign language, yield four main results: (1) the new readability formula performed better than the "classic" formula; (2) "non-classic" features were slightly more informative than "classic" features; (3) modern machine learning algorithms did not improve the explanatory power of our readability model, but allowed to better classify new observations; and (4) combining "classic" and "non-classic" features resulted in a significant gain in performance.

1 Introduction

Readability studies date back to the 1920's and have already spawned probably more than a hundred papers with research on the development of efficient methods to match readers and texts relative to their reading difficulty. During this period of time, several methodological trends have appeared in succession (reviewed in Klare (1963; 1984), DuBay (2004)). We can group these trends in three major approaches: the "classic studies", the "structuro-cognitivist paradigm" and the "AI readability", a term suggested by François (2011a).

The classic period started right after the seminal work of Vogel and Washburne (1928) and Gray and Leary (1935) and is characterized by an ideal of simplicity. The models (readability formulas) proposed to predict text difficulty for a given population are kept simple, using multiple linear regression with two, or sometimes, three predictors. The predictors are simple surface features, such as the average number of syllables per word and the average number of words per sentence. The Flesch (1948) and Dale and Chall (1948) formulas are probably the best-known examples of this period.

With the rise of cognitivism in psychological sciences in the 70's and 80's, new dimensions of texts are highlighted such as coherence, cohesion, and other discourse aspects. This led some scholars (Kintsch and Vipond, 1979; Redish and Selzer, 1985) to adopt a critical attitude to classic readability formulas which could only take into account superficial features, ignoring other important aspects contributing to text difficulty. Kintsch and Vipond (1979) and Kemper (1983), among others, suggested new features for readability, based on those newly discovered text dimensions. However, despite the fact that the proposed models made use of more sophisticated features, they failed to outperform the classic formulas. It is probably not coincidental that after these attempts readability research efforts declined in the 90s.

More recently, however, the development of efficient natural language processing (NLP) systems and the success of machine learning methods led to

NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations (PITR 2012)., pages 49–57,
Montréal, Canada, June 7, 2012. ©2012 Association for Computational Linguistics

a resurgence of interest in readability as it became clear that these developments could impact the design and performance of readability measures. Several studies (Si and Callan, 2001; Collins-Thompson and Callan, 2005; Schwarm and Ostendorf, 2005; Feng et al., 2010) have used NLP-enabled feature extraction and state-of-the-art machine learning algorithms and have reported significant gains in performance, suggesting that the AI approach might be superior to previous attempts.

Going beyond reports of performance which are often hard to compare due to a lack of a common gold standard, we are interested in investigating AI approaches more closely with the aim of understanding the reasons behind the reported superiority over classic formulas. AI readability systems use NLP for richer feature extraction and a machine learning algorithm. Given that the classic formulas are also statistical, is performance boosted because of the addition of NLP-enabled feature extraction or by better machine learning algorithms? In this paper, we report initial findings of three experiments designed to explore this question.

The paper is organized as follows. Section 2 reviews previous findings in the field and the challenge of providing a uniform explanation for these findings. Section 3 gives a brief overview of prior work on French readability, which is the context of our experiments (evaluating the readability of French texts). Because there is no prior work comparing classic formulas with AI readablity measures for French, we first report the results of this comparison in Section 3. Then, we proceed with the results of three experiments (2-4), comparing the contributions of the AI enabled features with features used in classic formulas, different machine learning algorithms and the interactions of features with algorithms. There results are reported in Sections 4, 5, and 6, respectively. We conclude in Section 7 with a summary of the main findings and future work.

2 Previous findings

Several readability studies in the past decade have reported a performance gain when using NLP-enabled features, language models, and machine learning algorithms to evaluate the reading difficulty of a variety of texts (Si and Callan, 2001; Collins-Thompson and Callan, 2005; Schwarm and Ostendorf, 2005; Heilman et al., 2008; Feng et al., 2010).

A first explanation for this superiority would be related to the new predictors used in recent models. Classic formulas relied mostly on surface lexical and syntactic variables such as the average number of words per sentence, the average number of letters per word, the proportion of given POS tags in the text or the proportion of out-of-simple-vocabulary words. In the AI paradigm, several new features have been added, including language models, parse tree-based predictors, probability of discourse relations, estimates of text coherence, etc. It is reasonable to assume that these new features capture a wider range of readability factors thus bringing into the models more and, possibly, better information.

However, the evidence from comparative studies is not consistent on this question. In several cases, AI models include features central to classic formulas which, when isolated, appear to be the stronger predictors in the models. An exception to this trend is the work of Pitler and Nenkova (2008) who reported non-significant correlation for the mean number of words per sentence ($r = 0.1637, p = 0.3874$) and the mean number of characters per word ($r = -0.0859, p = 0.6519$). In their study, though, they used text quality rather than text difficulty as the dependent variable. The data consisted solely of text from the Wall Street Journal which is “intended for an educated adult audience” text labelled for degrees of reading fluency. Feng et al. (2010) compared a set of similar variables and observed that language models performed better than classic formula features but classic formula features outperformed those based on parsing information. Collins-Thompson and Callan (2005) found that the classic type-token ratio or number of words not in the 3000-words Dale list appeared to perform better than their language model on a corpus from readers, but were poorer predictors on web-extracted texts.

In languages other than English, François (2011b) surveyed a wide range of features for French and reports that the feature that uses a limited vocabulary list (just like in some classic formulas) has a stronger correlation with reading difficulty that a unigram model and the best performing syntactic feature was the average number of words per sentences. Aluisio et al. (2010), also, found that the best correlate with difficulty was the average number of words per sentence. All in all, while there is sufficient evidence that the AI paradigm outperforms the classis formulas, classic features have often been shown to make the single strongest predictors.

An alternative explanation could be that, by comparison to the simpler statistical analyses that determined the coefficients of the classic formulas, machine learning algorithms, such as support machine vector (SVM) or logistic regression are more sophisticated and better able to learn the regularities in training data, thus building more accurate models. Work in this direction has been of smaller scale but already reporting inconsistent results. Heilman et al. (2008) considered the performance of linear regression, ordinal and multinomial logistic regression, and found the latter to be more efficient. However, Kate et al. (2010) obtained contradictory findings, showing that regression-based algorithms perform better, especially when regression trees are used for bagging. For French, François (2011b) found that SVMs were more efficient than linear regression, ordinal and multinomial logistic regression, boosting, and bagging.

Finally, it is quite possible that there are interactions between types of features and types of statistical algorithms and these interactions are primarily responsible for the better performance.

In what follows, we present the results of three studies (experiments 2-4), comparing the contributions of the AI enabled features with features used in classic formulas, different machine learning algorithms and the interactions of features with algorithms. As mentioned earlier, all the studies have been done on French data, consisting of text extracted from levelled FFL textbooks (French as Foreign Language). Because there is no prior work comparing classic formulas with AI readability measures for FFL, we first report the results of this comparison in the next section (experiment 1).

3 Experiment 1: Model comparison for FFL

To compute a classic readability formula for FFL, we used the formula proposed for French by Kandel and Moles (1958). We compared the results of this formula with the AI model trained on the FFL data used by François (2011b).

The Kandel and Moles (1958) formula is an adaptation of the Flesch formula for French, based on a study of a bilingual corpus:

$$Y = 207 - 1.015lp - 0.736lm \qquad (1)$$

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 this formula is not specifically designed for FFL, we chose to implement it over formulas proposed for FFL (Tharp, 1939; Uitdenbogerd, 2005). FFL-specific formulas are optimized for English-speaking learners of French while our dataset is agnostic to the native language of the learners.

The computation of the Kandel and Moles (1958) formula requires a syllabification system for French. Due to unavailability of such a system for French, we adopted a hybrid syllabification method. For words included in Lexique (New et al., 2004), we used the gold syllabification included in the dictionary. For all other words, we generated API phonetic representations with espeak [1], and then applied the syllabification tool used for Lexique3 (Pallier, 1999). The accuracy of this process exceeded 98%.

For the comparison with an AI model, we extracted the same 46 features (see Table 2 for the complete list) used in François' model [2] and trained a SVM model.

For all the study, the gold-standard consisted of data taken from textbooks and labeled according to the classification made by the publishers. The corpus includes a wide range of texts, including extracts from novels, newspapers articles, songs, mail, dialogue, etc. The difficulty levels are defined by the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) as follows: A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). The test corpus includes 68 texts per level, for a total of 408 documents (see Table 1).

We applied both readability models to this test corpus. Assessing and comparing the performance

[1] Available at: http://espeak.sourceforge.net/.

[2] Details on how to implement these features can be found in François (2011b).

A1	A2	B1	B2	C1	C2	Total
$68(10,827)$	$68(12,045)$	$68(17,781)$	$68(25,546)$	$68(92,327)$	$68(39,044)$	$408(127,681)$

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

of the two models with accuracy scores (acc), as is common in classification tasks, has proved challenging and, in the end, uninformative. This is because the Kandel and Moles formula's output scores are not an ordinal variable, but intervals. To compute accuracy we would have to define a set of rather arbitrary cut off points in the intervals and correspond them with level boundaries. We tried three approaches to achieve this task. First, we used correspondences between Flesch scores and seven difficulty levels proposed for French by de Landsheere (1963): “very easy” (70 to 80) to “very difficult” (-20 to 10). Collapsing the “difficult” and “very difficult” categories into one, we were able to roughly match this scale with the A1-C2 scale. The second method was similar, except that those levels were mapped on the values from the original Flesch scale instead of the one adapted for French. The third approach was to estimate normal distribution parameters μ_j and σ_j for each level j for the Kandel and Moles' formula output scores obtained on our corpus. The class membership of a given observation i was then computed as follows:

$$\arg\max_{j=1}^{6} P(i \in j \mid N(\mu_j, \sigma_j)) \qquad (2)$$

Since the parameters were trained on the same corpus used for the evaluation, this computation should yield optimal class membership thresholds for our data.

Given the limitations of all three approaches, it is not surprising that accuracy scores were very low: 9% for the first and 12% for the second, which is worse than random (16.6%). The third approach gave a much improved accuracy score, 33%, but still quite low. The problem is that, in a continuous formula, predictions that are very close to the actual will be classified as errors if they fall on the wrong side of the cut off threshold. These results are, in any case, clearly inferior to the AI formula based on SVM, which classified correctly 49% of the texts.

A more suitable evaluation measure for a continuous formula would be to compute the multiple correlation (R). The multiple correlation indicates the extent to which predictions are close to the actual classes, and, when R^2 is used, it describes the percentage of the dependent variable variation which is explained by the model. Kandel and Moles' formula got a slightly better performance ($R = 0.551$), which is still substantially lower that the score ($R = 0.728$) obtained for the SVM model. To check if the difference between the two correlation scores was significant, we applied the Hotelling's T-test for dependent correlation (Hotelling, 1940) (required given that the two models were evaluated on the same data). The result of the test is highly significant ($t = -19.5; p = 1.83^{e-60}$), confirming that the SVM model performed better that the classic formula.

Finally, we computed a partial Spearman correlation for both models. We considered the output of each model as a single variable and we could, therefore, evaluate the relative predictive power of each variable when the other variable is controlled. The partial correlation for the Kandel and Moles formula is very low ($\rho = -0.11$; $p = 0.04$) while the SVM model retains a good partial correlation ($\rho = -0.53; p < 0.001$).

4 Experiment 2: Comparison of features

In this section, we compared the contribution of the features used in classic formulas with the more sophisticated NLP-enabled features used in the machine learning models of readability. Given that the features used in classic formulas are very easy to compute and require minimal processing by comparison to the NLP features that require heavy preprocessing (e.g., parsing), we are, also, interested in finding out how much gain we obtain from the NLP features. A consideration that becomes important for tasks requiring real time evaluation of reading difficulty.

To evaluate the relative contribution of each set of features, we experiment with two sets of features (see Table 2. We labeled as ”classic”, not only

Family	Tag	Description of the variable	ρ	Linear
Classic	PA-Alterego	Proportion of absent words from a list of easy words from *AlterEgo1*	0.652	No
	X90FFFC	90^{th} percentile of inflected forms for content words only	-0.641	No
	X75FFFC	75^{th} percentile of inflected forms for content words only	-0.63	No
	PA-Goug2000	Proportion of absent words from 2000 first of Gougenheim et al. (1964)'s list	0.597	No
	MedianFFFC	Median of the frequencies of inflected content words	-0.56	Yes
	PM8	Pourcentage of words longer than 8 characters	0.525	No
	NL90P	Length of the word corresponding to the 90^{th} percentile of word lengths	0.521	No
	NLM	Mean number of letters per word	0.483	Yes
	IQFFFC	Interquartile range of the frequencies of inflected content words	0.405	No
	MeanFFFC	Mean of the frequencies of inflected content words	-0.319	No
	TTR	Type-token ratio based on lemma	0.284	No
	NMP	Mean number of words per sentence	0.618	No
	NWS90	Length (in words) of the 90^{th} percentile sentence	0.61	No
	PL30	Percentage of sentences longer than 30 words	0.56	Yes
	PRE/PRO	Ratio of prepositions and pronouns	0.345	Yes
	GRAM/PRO	Ratio of grammatical words and pronouns	0.34	Yes
	ART/PRO	Ratio of articles and pronouns	0.326	Yes
	PRE/ALL	Proportions of prepositions in the text	0.326	Yes
	PRE/LEX	Ratio of prepositions and lexical words	0.322	Yes
	ART/LEX	Ratio of articles and lexical words	0.31	Yes
	PRE/GRAM	Ratio of prepositions and grammatical words	0.304	Yes
	NOM-NAM/ART	Ratio of nouns (common and proper) and gramm. words	-0.29	Yes
	PP1P2	Percentage of P1 and P2 personal pronouns	-0.333	No
	PP2	Percentage of P2 personal pronouns	-0.325	Yes
	PPD	Percentage of personal pronouns of dialogue	0.318	No
	BINGUI	Presence of commas	$0,462$	No
Non-classic	Unigram	Probability of the text sequence based on unigrams	0.546	No
	MeanNGProb-G	Average probability of the text bigrams based on Google	0.407	Yes
	FCNeigh75	75^{th} percentile of the cumulated frequency of neighbors per word	-0.306	Yes
	MedNeigh+Freq	Median number of more frequent neighbor for words	-0.229	Yes
	Neigh+Freq90	90^{th} percentile of more frequent neighbor for words	-0.192	Yes
	PPres	Presence of at least one present participle in the text	0.44	No
	PPres-C	Proportion of present participle among verbs	0.41	Yes
	PPasse	Presence of at least one past participle	0.388	No
	Infi	Presence of at least one infinive	0.341	No
	Impf	Presence of at least one imperfect	0.272	No
	Subp	Presence of at least one subjunctive present	0.266	Yes
	Futur	Presence of at least one future	0.252	No
	Cond	Presence of at least one conditional	0.227	No
	PasseSim	Presence of at least one simple past	0.146	No
	Imperatif	Presence of at least one imperative	0.019	Yes
	Subi	Presence of at least one subjunctive imperfect	0.049	Yes
	avLocalLsa-Lem	Average intersentential cohesion measured via LSA	$0,63$	No
	ConcDens	Estimate of the conceptual density with *Densidées* (Lee et al., 2010)	0.253	Yes
	NAColl	Proportion of MWE having the structure NOUN ADJ	0.286	Yes
	NCPW	Average number of MWEs per word	0.135	Yes

Table 2: List of the 46 features used by François (2011b) in his model. The Spearman correlation reported here also comes from this study.

the features that are commonly used in traditional formulas like Flesch (length of words and number of words per sentence) but also other easy to compute features that were identified in readability work. Specifically, in the ”classic” set we include number of personal pronouns (given as a list) (Gray and Leary, 1935), the Type Token Ratio (TTR) (Lively and Pressey, 1923), or even simple ratios of POS (Bormuth, 1966).

The ”non-classic” set includes more complex NLP-enabled features (coherence measured through LSA, MWE, n-grams, etc.) and features suggested by the structuro-cognitivist research (e.g., information about tense and variables based on orthographical neighbors).

For evaluation, we first computed and compared the average bivariate correlations of both sets. This test yielded a better correlation for the classic features ($\bar{r} = 0.48$ over the non-classic features $\bar{r} = 0.29$)

As a second test, we trained a SVM model on each set and evaluated performances in a ten-fold cross-validation. For this test, we reduced the number of classic features by six to equal the number of predictors of the non-classic set. Our hypothesis was the SVM model using non-classic features would outperform the classic set because the non-classic features bring richer information. This assumption was not strictly confirmed as the non-classic set performed only slightly better than the classic set. The difference in the correlation scores was small (0.01) and non-significant ($t(9) = 0.49; p = 0.32$), but the difference in accuracy was larger (3.8%) and close to significance ($t(9) = 1.50; p = 0.08$). Then, in an effort to pin down the source of the SVM gain that did not come out in the comparison above, we defined a SVM baseline model (b) that included only two typical features of the classic set: the average number of letter per word (NLM) and the average number of word per sentence (NMP). Then, for each of the i remaining variables (44), we trained a model m_i including three predictors: NLM, NMP, and i. The difference between the correlation of the baseline model and that of the model m_i was interpreted as the information gain carried by the feature i. Therefore, for both sets, of cardinality N_s, we computed:

$$\frac{\sum_{i=1}^{N_s} R(m_i) - R(b)}{N_s} \tag{3}$$

where $R(m_i)$ is the multiple correlation of model m_i.

Our assumption was that, if the non-classic set brings in more varied information, every predictor should, on average, improve more the R of the baseline model, while the classic variables, more redundant with NLM and NP, would be less efficient. In this test, the mean gain for R was 0.017 for the classic set and 0.022 for the non-classic set. Although the difference was once more small, this test yielded a similar trend than the previous test.

As a final test, we compared the performance of the SVM model trained only on the “classic” set with the SVM trained on both sets. In this case, the improvement was significant ($t(9) = 3.82; p = 0.002$) with accuracy rising from 37.5% to 49%. Although this test does not help us decide on the nature of the gain as it could be coming just from the increased number of features, it shows that combining ”classic” and ”non-classic” variables is valuable.

5 Experiment 3: Comparison of statistical models

In this section, we explore the hypothesis that AI models outperform classic formulas because they use better statistical algorithms. We compare the performance of a“classic” algorithm, multiple linear regression, with the performance of a machine learning algorithm, in this case SVM. Note that an SVMs have an advantage over linear regression for features non-linearly related with difficulty. Bormuth (1966, 98-102) showed that several classic features, especially those focusing on the word level, were indeed non-linear. To control for linearity, we split the 46 features into a linear and a non-linear subset, using the Guilford’s F test for linearity (Guilford, 1965) and an $\alpha = 0.05$. This classification yielded two equal sets of 23 variables (see Table 2). In Table 3, we report the performance of the four models in terms of R, accuracy, and adjacent accuracy. Following, Heilman et al. (2008), we define ”adjacent accuracy” as the proportion of predictions that were within one level of the assigned label in the corpus.

	Model	**R**	**Acc.**	**Adj. acc.**
Linear	LR	0.58	27%	72%
	SVM	0.64	38%	73%
Non-Linear	LR	0.75	36%	81%
	SVM	0.70	44%	76%

Table 3: Multiple correlation coefficient (R), accuracy and adjacent accuracy for linear regression and SVM models, using the set of features either linearly or non linearly related to difficulty.

Adjacent accuracy is closer to R as it is less sensitive to minor classification errors.

Our results showed a contradictory pattern, yielding a different result depending on type of evalution: accuracy or R and adjacent accuracy. With respect to accuracy scores, the SVM performed better in the classification task, with a significant performance gain for both linear (gain = 9%; $t(9) = 2.42; p = 0.02$) and non-linear features (gain = 8%; $t(9) = 3.01; p = 0.007$). On the other hand, the difference in R was non-significant for linear (gain = 0.06; $t(9) = 0.80; p = 0.22$) and even negative and close to significance for non-linear (gain = -0.05; $t(9) = 1.61; p = 0.07$). In the light of these results, linear regression (LR) appears to be as efficient as SVM accounting for variation in the dependant variable (their R^2 are pretty similar), but produces poorer predictions.

This is an interesting finding, which suggests that the contradictory results in prior literature with regard to performance of different readability models (see Section 2) might be related to the evaluation measure used. Heilman et al. (2008, 7), who compared linear and logistic regressions, found that the R of the linear model was significantly higher than the R of the logistic model ($p < 0.01$). In contrast, the logistic model behaved significantly better ($p < 0.01$) in terms of adjacent accuracy. Similarly, Kate and al. (2010, 548), which used R as evaluation measure, reported that their preliminary results "verified that regression performed better than classification". Once they compared linear regression and SVM regression, they noticed similar correlations for both techniques (respectively 0.7984 and 0.7915).

To conclude this section, our findings suggest that (1) linear regression and SVM are comparable in accounting for the variance of text difficulty and (2) SVM has significantly better accuracy scores than linear regression.

6 Experiment 4: Combined evaluation

In Experiment 2, we saw that ”non-classic” features are slightly, but non-significantly, better than the ”classic” features. In Experiment 3, we saw that SVM performs better than linear regression when the evaluation is done by accuracy but both demonstrate similar explanatory power in accounting for the variation. In this section, we report evaluation results for four models, derived by combining two sets of features, classic and non-classic, with two algorithms, linear regression and SVM. The results are shown in Table (4).

The results are consistent with the findings in the previous sections. When evaluated with accuracy scores SVM performs better with both classic ($t(9) = 3.15; p = 0.006$) and non-classic features ($t(9) = 3.32; p = 0.004$). The larger effect obtained for the non-classic features might be due to an interaction, i.e., an SVM trained with non-classic features might be better at discriminating reading levels. However, with respect to R, both algorithms are similar, with linear regression outperforming SVM in adjacent accuracy (non-significant). Linear regression and SVM, then, appear to have equal explanatory power.

As regards the type of features, the explanatory power of both models seems to increase with non-classic features as shown in the increased R, although significance is not reached ($t(9) = 0.49; p = 0.32$ for the regression and $t(9) = 1.5; p = 0.08$ for the SVM).

7 General discussion and conclusions

Recent readability studies have provided preliminary evidence that the evaluation of readability using NLP-enabled features and sophisticated machine learning algorithms outperform the classic readability formulas, such as Flesch, which rely on surface textual features. In this paper, we reported a number of experiments the purpose of which was to identify the source of this performance gain.

Specifically, we compared the performance of classic and non-classic features and the performance

	Model	R	Acc.	Adj. acc.
Classic	LR	0.66	30.6%	78%
	SVM	0.67	37.5%	76%
Non-classic	LR	0.68	32%	76%
	SVM	0.68	41.8%	73%

Table 4: Multiple correlation coefficient (R), accuracy and adjacent accuracy for linear regression and SVM models with either the classic or the non-classic set of predictors.

of two statistical algorithms: linear regression (used in classic formulas) and SVM (in the context of FFL readability). Our results indicate that classic features are strong single predictors of readability. While we were not able to show that the non-classic features are better predictors by themselves, our findings show that leaving out non-classic features has a significant negative impact on the performance. The best performance was obtained when both classic and non-classic features were used.

Our experiments on the comparison of the two statistical algorithms showed that the SVM outperforms linear regression by a measure of accuracy, but the two algorithms are comparable in explanatory power accounting for the same amount of variability. This observation accounts for contradictory conclusions reported in previous work. Our study shows that different evaluation measures can lead to quite different conclusions.

Finally, our comparison of four models derived by combining linear regression and SVM with “classic” and “non-classic” features confirms the significant contribution of “non-classic” features and the SVM algorithm to classification accuracy. However, by a measure of *adjacent accuracy* and explanatory power, the two algorithms are comparable.

From a practical application point of view, it would be interesting to try these algorithms in web applications that process large amounts of text in real time (e.g., READ-X (Miltsakaki, 2009)) to evaluate the trade-offs between accuracy and efficiency.

Acknowledgments

We would like to acknowledge the invaluable help of Bernadette Dehottay for the collection of the corpus, as well as the Belgian American Educational Foundation that supported Dr. Thomas François with a Fellowship during this work.

References

S. Aluisio, L. Specia, C. Gasperin, and C. Scarton. 2010. Readability assessment for text simplification. In *Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles.

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

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.

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.

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

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

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.

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 franais langue étrangère*. Ph.D. thesis, Université Catholique de Louvain. Thesis Supervisors : Cédrick Fairon and Anne Catherine Simon.

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

W.S. Gray and B.E. Leary. 1935. *What makes a book readable*. University of Chicago Press, Chicago: Illinois.

J.P. Guilford. 1965. *Fundamental statistics in psychology and education*. McGraw-Hill, New-York.

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.

H. Hotelling. 1940. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *The Annals of Mathematical Statistics*, 11(3):271–283.

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

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

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.

G.R.. Klare. 1963. *The Measurement of Readability*. Iowa State University Press, Ames, IA.

G.R. Klare. 1984. Readability. In P.D. Pearson, R. Barr, M. L. Kamil, P. Mosenthal, and R. Dykstra, editors, *Handbook of Reading Research*, pages 681–744. Longman, New York.

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)*.

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

E. Miltsakaki. 2009. Matching readers' preferences and reading skills with appropriate web texts. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 49–52.

B. New, C. Pallier, M. Brysbaert, and L. Ferrand. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516.

C. Pallier. 1999. Syllabation des représentations phonétiques de brulex et de lexique. Technical report, Technical Report, update 2004. Lien: http://www. pallier. org/ressources/syllabif/syllabation. pdf.

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.

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.

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

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

M. Vogel and C. Washburne. 1928. An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, 28(5):373–381.