On the Contribution of MWE-based Features to a Readability Formula for French as a Foreign Language

Thomas François
Aspirant F.N.R.S.
Centre for Natural Language Processing
Institut Langage et Communication
UCLouvain
thomas.francois@uclouvain.be

Patrick Watrin
Centre for Natural Language Processing
Institut Langage et Communication
UCLouvain
patrick.watrin@uclouvain.be

Abstract

This study aims to assess the usefulness of multi-word expressions (MWEs) as features for a readability formula that predicts the difficulty of texts for French as a foreign language. Using a MWE extractor combining a statistical approach with a linguistic filter, we define 11 predictors. These take into account the density and the probability of MWEs, but also their internal structure. Our experiments show that the predictive power of these 11 variables is low and that a simple approach based on the average probability of n-grams is more effective.

1 Introduction

With the success of the communicative and action-oriented approach in the teaching of a second or foreign language (L2), teachers are encouraged to work on authentic texts in order to bring their students in contact with real linguistic data. The web is a valuable source for such documents, but the search for a document tailored to the level of the students may sometimes be tedious. In this context, readability studies may help. They aim to develop tools capable of assessing the difficulty of texts for a given population through textual features only (such as the number of letters per word, the number of words per sentence, etc.).

However, while many studies have examined the readability of English L1 (Chall and Dale, 1995), there are far fewer studies on readability in an L2, especially in French as a foreign language (FFL). In most cases, formulas for native speakers have been applied to L2 texts. However, the validity of such an approach is far from established, because it relies on three suspect assumptions : (1) the understanding of readers in the L2 is comparable to that of native speakers, (2) the textual features considered in L1 formulas are relevant to L2 reading, and (3) the weighting of these variables may be the same in a formula for L1 and L2.

If some work by Greenfield (2004) supports this vision, other authors disagree and consider that the peculiarities of the reading process in the L2, described by Koda (2005) among others, must be taken into account by designers of readability formulas. Of these dimensions, the interferences between the L1 and L2 of the learners are certainly among the most studied topics (Uitdenbogerd, 2005; Laroche, 1979). Moreover, François (2009) has shown that considering verb modes and tenses leads to the significant improvement of a L2 formula.

However, there is another textual aspect that is likely to be a good predictor of lexical difficulty for L2 readers: collocations and idioms. A good knowledge of these items is indeed associated with a fluent and appropriate use of the language (Pawley and Syder, 1983). We can therefore expect that L2 readers, especially beginners, encounter difficulties in processing these lexical chains and that texts which contain a large number of collocations and idioms are likely to be more difficult. Nevertheless, this assumption has not yet been addressed by a comprehensive study, be it for English as a second or foreign language (EFL), or for FFL. That is why we have dedicated this paper to this issue, which we explore through the specific case of FFL.

In section 2, we summarize a set of research findings about collocations and their processing, especially when reading a text in L2. Section 3 is a description of both the corpus and the lexical

Proceedings of Recent Advances in Natural Language Processing, pages 441–447,
Hissar, Bulgaria, 12-14 September 2011.

extractor we used to analyze the relationship between some characteristics of MWEs and the difficulty of the text for readers of FFL. The results of these experiments are reported and discussed in Section 4 before we conclude with some perspectives for future research.

2 MWEs and Text Difficulty

In this paper, we refer to MWEs as a set of linguistic objects the meaning and structure of which can be more or less frozen (collocations, compound words, idioms, etc.). From a statistical point of view, this class of objects commonly refers to "strings of words that are more frequently associated than it would be only by chance (Dias et al., 2000, 213).

These lexical entities have been shown to be processed by native speakers faster on average than free combinations (Underwood et al., 2004), both in reading and in oral production. This result may be interpreted as to mean that MWEs are fully or partially stored in long-term memory (Pawley and Syder, 1983) and can be recovered as such, thereby relieving short-term memory whose capacity is limited. Therefore, the processing of MWEs should be faster in reading and oral production, at least for natives, who are familiar with most of these.

For L2 learners, it has been demonstrated that their collocational knowledge lags far behind their general vocabulary knowledge (Bahns and Eldaw, 1993). Surprisingly, some studies on the L2 reading of MWEs reported a facilitating effect similar to the one of native speakers for advanced L2 learners (Underwood et al., 2004). It should be noted that such studies focus on reading time, which is related to the recognition of collocations, but do not evaluate their impact on comprehension. Underwood et al. (2004) reported that some of their subjects, for which a faster processing of collocations was observed, did not know the meaning of nearly a third of them. Therefore, we assume that at beginner or intermediate level, this facilitating effect is likely to be counterbalanced by the fact that the MWEs encountered are (1) mostly unknown to readers and (2) even more difficult to elucidate using the context as their meaning can be non-compositional.

A common method to estimate to what extent a MWE is known in a given population is to use its objective frequency. However, the hypothesis that MWEs that are less frequent in the language may be more difficult to read has hardly been explored in readability. Weir and Anagnostou (2008) suggested using the mean of the absolute frequency of all MWEs in a text as an indication of its difficulty. However, they did not report any experiments related to this hypothesis. In a previous article, Ozasa et al. (2007) had presented an EFL readability formula for Japanese learners that includes, among other variables, an index of textbook-based idiom difficulty. However, this variable was not significant in its multiple linear regression model, since the p-value of the t-test for coefficient significance was $0,61$ (Ozasa et al., 2007, 4).

In view of these results, it is not clear whether MWE-based features may be effective predictors of text readability in L2. However, we believe that the studies mentioned above have approached the issue only superficially. In this paper, we investigate further how MWEs can be used within an L2 readability formula through the specific case of FFL.

3 Methodology

To conduct our experiments, it was necessary to (1) collect a corpus that was already annotated in terms of difficulty, and (2) develop an extractor of nominal MWEs.

3.1 The Corpus

The corpus used to develop a readability formula should be labelled for reading-difficulty level, a task that implies agreement on the difficulty scale. In the context of foreign language teaching in Europe, an obvious choice is the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001). The CEFR normally has six levels – A1 (Breakthrough); A2 (Waystage); B1 (Threshold); B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). However, to better reflect the evolution of learners, which is faster in the early stages of learning, we split the first three levels into two, thereby obtaining a total of nine levels.

Another positive aspect of using the CEFR is that, since its introduction, FFL textbooks have undergone a kind of standardization. It is thus feasible to gather a large number of documents that have already been labelled in terms of the CEFR scale by experts. We postulated that the level of a text is equivalent to the level of the text-

book it comes from. Following this assumption, we first gathered a corpus of 1,895 texts (about 500K words) selected from FFL textbooks, using the same criteria as François (2009). We then randomly selected 50 texts per level (thus retaining 450 texts) to establish a test corpus in which the *a priori* probability of each class is similar. To do otherwise would have resulted in a biased model.

3.2 The Extractor

Regarding the extraction process of MWEs, we use a three-step state-of-the-art procedure which draws on the work of Daile (1995) and Smadja (1993) in that it combines a linguistic filter with association measures (AM). Concretely, the texts are first POS tagged to clear most lexical ambiguities[1]. We then identify all nominal MWE candidates in the tagged text with the help of a library of transducers [2] (or syntactic patterns). Finally, the list of candidates is submitted to the statistical validation module which assigns an AM to each of them. After some experiments, we retained the *fair log-likelihood ratio* (Silva and Lopes, 1999) as our AM, since it allows to process units that are longer than bigrams.

As with all measures of association, the proper functioning of this AM requires a consistent frequency mass, which was not available from the texts in our corpus. To overcome this problem, we used a frequency reference, which is a database of n-grams with their frequencies, as suggested by Watrin and François (2011). The reference allows an efficient on-the-fly computation of AMs, even in reduced contexts, provided that the frequencies stored in the database have been counted on a large corpus. For this study, we used two differents corpora as references:

- The 5-grams of Google (Michel et al., 2011), which represents the largest corpus currently available for French. Only contemporary n-grams were kept, i.e. those that relate to texts published between 2000 and 2008. We therefore obtained 1.117.140.444 5-grams. However, it must be stressed that the tokenization carried out in this resource remains very basic. It considers the following chains as 5-grams: " , l ' arbre est " or " un pique - nique . ".
- A set of newspaper articles published in 2009 in the Belgian daily *Le Soir* for a total of 5.000.000 5-grams. In this case, we were able to define our own tokenization and to consider such items as " pique-nique " or " l' " as one word.

To optimize the size of the references as well as their access time, we used a PATRICIA tree (Morrison, 1968) to store the n-grams. This data structure allows the compression of n-grams sharing a common prefix and that of nodes with only one child node, which results in queries carried out in constant time. We were then able to extract all the MWE candidates terms from the texts of our test corpus.

We then faced one last problem: what criterion should we use to decide whether a candidate term is actually a collocation ? The *log-likelihood ratio* being distributed according to a chi-square law with one degree of freedom, one possible approach is to select the MWEs for which the AM obtained is higher than 3.84 (which corresponds to $\alpha = 0.05$). However, as the size of the reference corpus increases, this solution becomes meaningless since high frequencies of occurrences generate high scores for the chi-square. Therefore, more and more phenomena appear significant (Kilgarriff, 2005).

The common solution to this issue is to empirically set a higher threshold. It has an obvious flaw: the threshold is only valid for a given corpus or one of comparable size. Once more, the use of a reference circumvents this difficulty: since the size is constant, an optimal threshold can be fixed once and for all. In our study, the selected threshold values were function of the precision of the extractor (see 4.1).

4 Results and Discussion

4.1 The Predictive Efficiency of the MWEs

From the extractor described above, it was possible to define 11 variables that aimed at taking into account various facets of MWEs. These were:

- The proportion of nominal MWEs to the number of words in the text (**NCPW**).
- The mean size (in number of words) of nominal MWEs in the text (**MSize**).
- 4 variables representing the proportion of the

[1] Tagging is done with the *TreeTagger* (Schmid, 1994).

[2] To apply our transducers to the tagged text, we use *Unitex* (Paumier, 2003). The output of the process is a file containing only the recognized sequences.

following grammatical structures : *N N* ; *N PREP (DET) N* ; *A N*, and *N A*.

- The mean probability of all nominal MWEs in the text, the probabilities used coming from our two references (**MeanP**). We also computed the 75^{th} percentile of the same probabilities distribution **P75**.

- 3 variables that are the mean probabilities of nominal MWEs of size 2 (**MP2Coll**), size 3 (**MP3Coll**), and size 4 (**MP4Coll**). Longer units were not considered, since they were too scarce.

- For the sake of comparison, we also computed two conventional variables: the number of letters per word (**NLW**) and the number of words per sentence (**NWS**.

Furthermore, we manipulated the threshold θ used for the selection of MWEs. In this way, we were able to estimate how the strength of association between the components of MWEs impacts on the predictive power of the above variables. Four thresholds were selected for each of the two references: a zero threshold where all nominal structures were considered, a second and a fourth one respectively corresponding to a 30% and 50% precision for our extractor, and an intermediate value as the third threshold. Table 1 shows the Pearson correlation coefficients (r) between the 11 aforementioned variables and the level of difficulty of the texts in our test corpus [3]

These results provide valuable lessons. First, when one roughly analyses the strength of associations, it can be noticed that several variables are significantly correlated with the difficulty of the texts, in particular **NPCW** and the NA structure. It is a interesting outcome, since neither the simple **NPCW** variable, nor structural information had been previously considered in the readability literature. Furthermore, **MeanP** mostly appeared as not being significantly correlated with difficulty, a result that is congruent with that of Ozasa *et al.* (2007).

A second significant observation is that increasing θ, and thus strengthening the level of cohesion among MWEs, tends to weaken the association between most of our variables and difficulty. Faced with these results, one might conclude that MWEs are not as good predictors as the simple complex nominal structures ($\theta = 0$). However, it seems more accurate to limit this deficiency to MWEs that are detected automatically using statistical techniques. Among the best candidates of our corpus, we find MWEs such as “effet de serre (greenhouse effect) or “développement durable (sustainable development), which are relevant in the context of L2 reading, but we also found “mardi soir (late Tuesday) or “million d’euros (millions of euros), which are less relevant.

Third, as relying on correlations to conclude that a variable is a good predictor for readability does not suffice, we investigated this issue for our two best variables: **NPCW** and the NA structure. In a predictive model such as a readability formula, the informative contribution of each variable depends on the other factors in the formula. If two variables are highly correlated, they are likely to provide redundant information. In our case, although the significance level of **NPCW** and the NA structure are high, their raw correlation remains well below that of the two classic variables : **NLW** ($r = 0.58$) and **NWS** ($r = 0.578$). It is therefore not obvious that the two selected MWEs variables will be good predictors.

To clarify this issue, we compared a baseline readability formula using only **NLW** and **NWS** as predictors with the same formula which also comprised the **NPCW** and the NA structure [4]. It turns out that the contribution of the two MWE predictors is non significant ($\chi^2 = 2.98$; $p - value = 0.08$) [5], hence demonstrating that MWE-based variables do not provide really new information compared to traditional variables.

Faced with this inadequacy of variables based on automatically detected MWEs to the context of readability, we asked ourselves a second question. Would a simpler model, namely an n-gram model, be more efficient although it considers only sequences of tokens without any linguistic motivation ?

[3] In order to compute this metric, the difficulty levels A1 to C2 were converted into a discrete scale ranging from 1 to 9.

[4] The statistical model used for this comparison is based on an ordinal logistic regression, described in more detail in François (2009)

[5] The statistical technique used to compare the two models equates each of them to an explicative hypothesis of the data and calculates their log-likelihood ratio which is multiplied by the constant -2 in order to be distributed according to a chi-square law.

	Le Soir				Google			
Thresholds θ	0	15	25	43	0	139	4000	9931
NCPW	0.30^{3}	0.14^{2}	0.13^{2}	0.14^{2}	0.17^{3}	0.10^{1}	0.15^{2}	0.15^{2}
MSize	-0.02	0.03	-0.03	-0.02	-0.12^{1}	-0.19^{3}	-0.14^{2}	-0.18^{3}
NN	-0.24^{3}	-0.14^{2}	-0.01	0.03	-0.22^{3}	-0.13^{2}	0.004	0.007
NPN	0.05	0.13^{2}	0.09	0.11^{1}	0.04	0.06	0.15^{2}	0.17^{3}
AN	-0.05	-0.03	0.02	0.08	-0.07	0.03	0.08	0.09^{1}
NA	0.36^{3}	0.30^{3}	0.27^{3}	0.22^{3}	0.37^{3}	0.32^{3}	0.25^{3}	0.28^{3}
P75	-0.16^{2}	-0.10^{1}	-0.11^{1}	-0.15^{2}	-0.0001	0.02	-0.01	0.03
MeanP	-0.03	-0.03	-0.04	-0.05	0.15^{2}	0.16^{2}	0.14^{1}	0.09
MeanP2	-0.12^{1}	-0.18^{3}	-0.19^{3}	-0.20^{3}	-0.0007	-0.03	-0.06	-0.0005
MeanP3	-0.12^{1}	-0.12^{1}	-0.05	0.05	-0.02	0.03	0.02	0.02
MeanP4	-0.09	-0.07	-0.02	-0.08	-0.10^{1}	-0.05	0.01	0.02

Table 1: Pearson correlation between independent variables and text difficulty. Significance levels are noted as follows: $^{1}p < 0.05$; $^{2}p < 0.01$; $^{3}p < 0.0001$

4.2 N-gram Models

In contrast to MWEs, the use of n-gram models in readability is not new. They were first applied to the field by Si and Callan (2001) as a set of unigram models specific to every level of difficulty. Pitler and Nenkova (2008) later showed that even a single unigram model is an efficient predictor for readability. Meanwhile, higher order models have been developed by Schwarm and Ostendorf (2005) or Kate *et al.* (2010). The former authors selected the perplexity of a trigram model as one of their predictors, while the latter preferred to directly use the normalized probability outputted by the n-gram model (see Equation 1).

In this study, we defined the 7 following variables to assess the efficiency of n-gram models in the context of readability:

- The normalized log-probability of every text (**normTLProb**), which is in keeping with Kate *et al.* (2010) and is expressed as follows:

$$\text{normTLProb} = \frac{1}{m}\sum_{i=1}^{m}\log P(w_i|h) \quad (1)$$

  where $P(w_i|h)$ is the probability of word i conditioned on the historic h limited to the $n-1$ previous words, and m stands for the number of words in the text to analyze.

- The mean (**MeanProb**) and the median (**MedianProb**) of the conditional probabilities distribution for a given text.

- Furthermore, as probabilities of MWEs were not expressed in a conditional form, but rather as a sequence's probability, we also take into consideration the probabilities of n-grams in our references. We used the arithmetic mean (**meanNGProb**), the median (**medianNGProb**), and the geometrical mean (**gmeanNGProb**) of those probabilities for a given text.

- Once more, for the sake of comparison, we developed a unigram model, based on *Lexique3* probabilites (New et al., 2007) **UnigM**.

We computed all these variables for each order of model from 2 to 5 using the frequencies stored in our two references: *Le Soir* and Google. Unfortunately, only the bigram model proved to be relevant to our approach. The discriminative capability of higher-order models suffers too much from the smoothing, since the number of unknown n-grams increases proportionally to the model order. As the probability of unknown events is always the same, the resulting variables are not discriminative enough once the order exceeds the bigram. Therefore, we only considered this level for our experimentations. The correlations of the 6 bigram-based variables with difficulty are shown in Table 2.

Again, our analyses provide some food for thought. A first observation is the complete inefficiency of variables based on a conventional bigram (r is 0.003 and -0.06 for **normTLProb**). This outcome seems highly surprising in comparison with previously reported results for English. Schwarm and Ostendorf (2005), for instance, reported successfully using n-gram models, even though they do not describe individual correlations for this variable and their good overall performance is obtained using many predictors. Such a low association is even more surprising as the

	normTLProb	MeanProb	MedianProb	meanNGProb	medianNGProb	gmeanNGProb
Google	$0,003$	$0,33^3$	$-0,04$	$0,38^3$	$-0,001$	$-0,03$
Le Soir	-0.06	$0,18^3$	$-0,01$	$0,25^3$	$-0,09$	$-0,0007$

Table 2: Correlation between the bigram-based variables and difficulty. Significance levels are noted as follows: $^1p < 0.05$; $^2p < 0.01$; $^3p < 0.0001$

unigram model **UnigM** conversely shows a strong correlation ($r = -0.57$).

However, **MeanProb**, which is also based on conditional probabilities, appears significant ($r = 0.33$ and 0.18), as does **meanNGProb** ($r = 0.38$ and 0.25). **gmeanNGProb**, where probabilities of sequences are multiplied as in the classic n-gram model, is also uncorrelated with difficulty. Therefore, this lack of association might come from the fact that we multiply probabilities instead of adding them up.

Considering our two significant variables, **meanNGProb** and **MeanProb**, one may wonder if they provide valuable information to assess the difficulty of texts. It should be noted that both features are extremely intercorrelated ($r = 0.975$) as one might expect. Therefore, it makes no sense to add them both to our baseline formula. We therefore compared this baseline with the enhanced version including only **medianNGProb** and, this time, this led to a significant improvement ($R = 0,67$; $\chi^2 = 11,66$; $p - value = 0,0006$). In relation to our research, it is particularly interesting to note that **medianNGProb** is more informative than a finer variable (**MeanP**) which requires a complex procedure to detect MWEs.

With respect to the models based on bigrams, one last surprising observation is the direction of the correlation. In our data, more complex texts are, on average, composed of more frequent units. This result is completely opposed to that of the classic unigram model: **UnigM** shows a strong negative correlation ($r = -0.57$) which is consistent with the assumption that more frequent words are easier. For this assumption to be applicable to higher order models, it would require that a similar pattern be found: less frequent word sequences should be more complex to read. Unexpectedly, this is not what we obtained. Although this result questions the validity of such an assumption, there may be other explanations. One is that the language used in beginner texts might be less likely, since it often use an "unnatural" style.

5 Conclusion

In this study, we investigated what would be the contribution of variables based on automatically extracted MWEs for a FFL readability formula. These were found to be negligible, both in absolute terms and compared with a simpler approach based on n-grams models. This replicates and extends the results of Ozasa et al. (2007) on English. Our experiment emphasizes how taking into account linguistic notions through an automatic approach may not always lead to satisfactory results in the context of L2 readability. Indeed, the NLP processing we used seems to generate too many approximations (coverage issue of the references, extraction errors, etc.) that reduce the effectiveness of our variables.

Regarding the n-grams, we found two interesting predictors for a readability formula: **meanNGProb**, and **MeanProb**. Besides, some of our results appeared surprising: (1) the conventional n-gram models proved ineffective on our data ($r = -0,06$), yet they are widely used in the field; (2) the negative association between objective frequency and difficulty, observed for unigram models, was not replicated for longer sequences. These two issues need to be further investigated to determine whether they are due to peculiarities of our data or not.

Finally, we wonder whether these results would be replicated (1) if verbal MWEs were taken into account instead of nominal ones ; (2) if the detection of MWEs were done manually (although it would be a huge work), and (3) if only idioms, semantically more opaque, were considered. This last perspective, intellectually attractive, must be tempered since it is likely that this kind of MWEs is too rare in texts to be analyzed with a statistical approach.

References

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

J.S. Chall and E. Dale. 1995. *Readability Revisited:*

The New Dale-Chall Readability Formula. Brookline Books, Cambridge.

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.

Béatrice Daille. 1995. Combined approach for terminology extraction: lexical statistics and linguistic filtering. Technical report, Lancaster University.

G. Dias, S. Guilloré, and J.G.P. Lopes. 2000. Extraction automatique dassociations textuelles à partir de corpora non traités. In *Proceedings of 5th International Conference on the Statistical Analysis of Textual Data*, pages 213–221.

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.

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

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.

A. Kilgarriff. 2005. Language is never ever ever random. *Corpus linguistics and linguistic theory*, 1(2):263–276.

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

J.M. Laroche. 1979. Readability measurement for foreign-language materials. *System*, 7(2):131–135.

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.

D.R. Morrison. 1968. PATRICIA - practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM*, 15(4):514–534.

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.

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*.

Sébastien Paumier. 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Ph.D. thesis, Université de Marne-la-Vallée.

A. Pawley and F.H. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. Richards and R. Schmitt, editors, *Language and Communication*, pages 191–225. Longman, London.

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.

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.

J.F. Silva and G.P. Lopes. 1999. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Sixth Meeting on Mathematics of Language*.

F. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177.

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

G. Underwood, N. Schmitt, and A. Galpin. 2004. The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt, editor, *Formulaic sequences: acquisition processing and use*, pages 155–172. John Benjamins, Amsterdam.

P. Watrin and T. François. 2011. N-gram frequency database reference to handle MWE extraction in NLP applications. In *Proceedings of the 2011 Workshop on MultiWord Expressions: from Parsing and Generation to the Real World (ACL Workshop)*, pages 83–91.

G.R.S. Weir and N.K. Anagnostou. 2008. Collocation frequency as a readability factor. In *Proceedings of the 13th Conference of the Pan Pacific Association of Applied Linguistics*.