Fun with Filtering French

Alexandra L. Uitdenbogerd
RMIT University
Melbourne, Victoria, Australia

Abstract

Early use of corpora for language learning has included analysis of word usage via concordancing. In addition, some attempts have been made to use readability criteria for recommending reading to learners. In this paper we discuss various tools and approaches for enhanced language learning support, including different methods of filtering text based on vocabulary and grammatical criteria. We demonstrate the effects of various criteria on the retrieval of text, assuming the user is English-speaking and learning French. Filtering text based on a small vocabulary of frequently occurring words, a set of English-French cognates and named entities, and high coverage criteria, results in the retrieval of short readable extracts from French literature. We expect that text available from the web may yield many more documents of appropriate readability.

1 Introduction

There is a considerable need for people to learn and become proficient in foreign languages: the majority of scientific discourse is published in English; students travel to different countries to study; people migrate for career opportunities.

Language skills are often divided into four communication tasks: listening, speaking, reading, and writing. Each of these skills can be developed and practised separately to a certain extent. Improving reading skill in a language would clearly involve devoting a substantial amount of time to reading.

It has been demonstrated that *extensive* reading at a comfortable level in a foreign language is more effective for improving language acquisition than *intensive* reading at more difficult levels (Bell, 2001). Therefore there is a need for reading material at multiple language skill levels to allow learners to practise. Some publishers provide graded reading books targeted at the foreign language learner, with levels indicated either by an assumed base vocabulary size, a standard level such as that defined by the Common European Framework of Reference for Languages (COE, 2003), year of study, or an unexplained level structure. The simplest graded readers based on vocabulary size that we have seen use a base vocabulary of 150 words. Beginner readers tend to be much shorter in length than those for advanced learners. For example, the level 1 readers in the Bibliobus Collection A are approximately 150 words in length, in a comic book format, consisting of two short stories (Cowling, 1982).

Several researchers independently proposed the idea of retrieving reading material from the Web based on its readability for the purpose of reading practice or study (Collins-Thompson and Callan, 2004a; Ghadirian, 2002; Katz and Bauer, 2001; Nilsson and Borin, 2002; Uitdenbogerd, 2003). This motivated some new studies of measuring readability (Collins-Thompson and Callan, 2004b; Schwarm and Ostendorf, 2005; Si and Callan, 2001) that are more sophisticated than was possible in the initial period of readability research (Bormuth, 1966; Chall and Dale, 1995; Cornaire, 1988; Granowskey and Botel, 1974; Klare, 1974). Our earlier work demonstrated that simple techniques are still very powerful for foreign language readability measurement, but could be improved by the inclusion of automated cognate detection for a specific first and second language pair (Uitdenbogerd, 2005). Recently, the methodology of producing readability measures has been questioned, with alternative approaches defined (van Oosten et al., 2010). In a related idea, filtering according to lexical constraints was ap-

Alexandra L. Uitdenbogerd. 2010. Fun with Filtering French. In *Proceedings of Australasian Language Technology Association Workshop*, pages 58−66

plied to the results of queries to a concordancer to make the query results easier for learners to read (Wible et al., 2000).

In this paper, we explore a corpus consisting of several classic French texts, with the goal of determining the feasibility of finding reading material using strict criteria for lexical content, and with some exploration of grammatical complexity. Exploiting the considerable overlap in language pairs, such as French and English, due to their cognate content, provides a substantially larger pool of reading resources than if cognates are ignored.

2 Related Research

Early studies in readability measurement largely led to formulae that consisted of two main factors of readability: lexical and grammatical. The lexical difficulty is often approximated by word length in terms of the number of syllables or characters (Klare, 1974). Alternatively, the presence or absence of a word in a list determined its difficulty (Chall and Dale, 1995). Grammatical complexity was usually modelled by a measure of sentence length Klare (1974).

Recent years have seen an increase in output specifically on automated readability measurement for text retrieval. We discuss some contributions below.

Researchers involved with the REAP project have developed a system for delivering reading material of an appropriate level to users, where the material is retrieved from the Web (Collins-Thompson and Callan, 2004a). They have used a range of readability measurement techniques based on statistical models on lexical and grammatical features that predict the grade level of the text (Heilman et al., 2008).

Miyazaki and Norizuki (2008) developed a reading retrieval system more suited to Japanese learners of English, allowing the readability measurement to be learnt from a user's ratings of text.

Tanaka-Ishii et al. (2010) had a novel approach to training a classifier to measure readability. They used precisely two classes, being for easy and hard texts, trained on text for children and adults respectively. Texts are classified as pairs to determine which is more difficult.

Little work has been published specifically for French readability as a foreign language. One recent development on readability of French as a foreign language uses a machine learning approach applied to a range of features, including the verb tenses occurring in the text (François, 2009).

The only work on assessing the suitability of on-line text for learners that we know of is ours (Uitdenbogerd, 2006), in which we concluded that the percentage of web-based text (in the English language) that is in a useful range for learners is between 8 and 19%. We are unaware of any that look at extracts of larger texts.

3 Experiments

In this current piece of work we are exploring the potential of filtering text based on strict lexical and grammatical criteria within the context of two languages that have a large set of exact cognates.

Our research questions were:

- *What is the frequency distribution of distinct sentence structures in text?* If there are frequent patterns, then these could be the basis for grammatical study for beginners in the language. They could also form part of the criteria for selecting text on readability. On the usual observation that shorter sentences are more readable, we were interested in discovering whether there were useful portions of natural texts to be found that could be used for reading practice at the early stages.

- *What proportion of a French text consists of French-English cognates and vice versa?* In this work, we restrict ourselves to words that have identical spelling in both languages. Accented words were not included. As the presence of cognates allow people to understand more of a text than when there are no cognates, we wanted to estimate the cognate content of the text. When combined with a small vocabulary of frequent words, the coverage of the text should be substantially greater. This idea was again to be applied to the process of extracting potential reading material.

The tools that we used for our experiments include Tree Tagger (Schmid, 1994), and a first approximation to a cognate list using the intersection of the English and French lexicons provided as Tree Tagger parameter files for these languages.

3.1 Sentences in French Literature

Initial work was attempted with on-line French literature. One example of a long written work that is available is *Les Trois Mousquetaires* (The Three

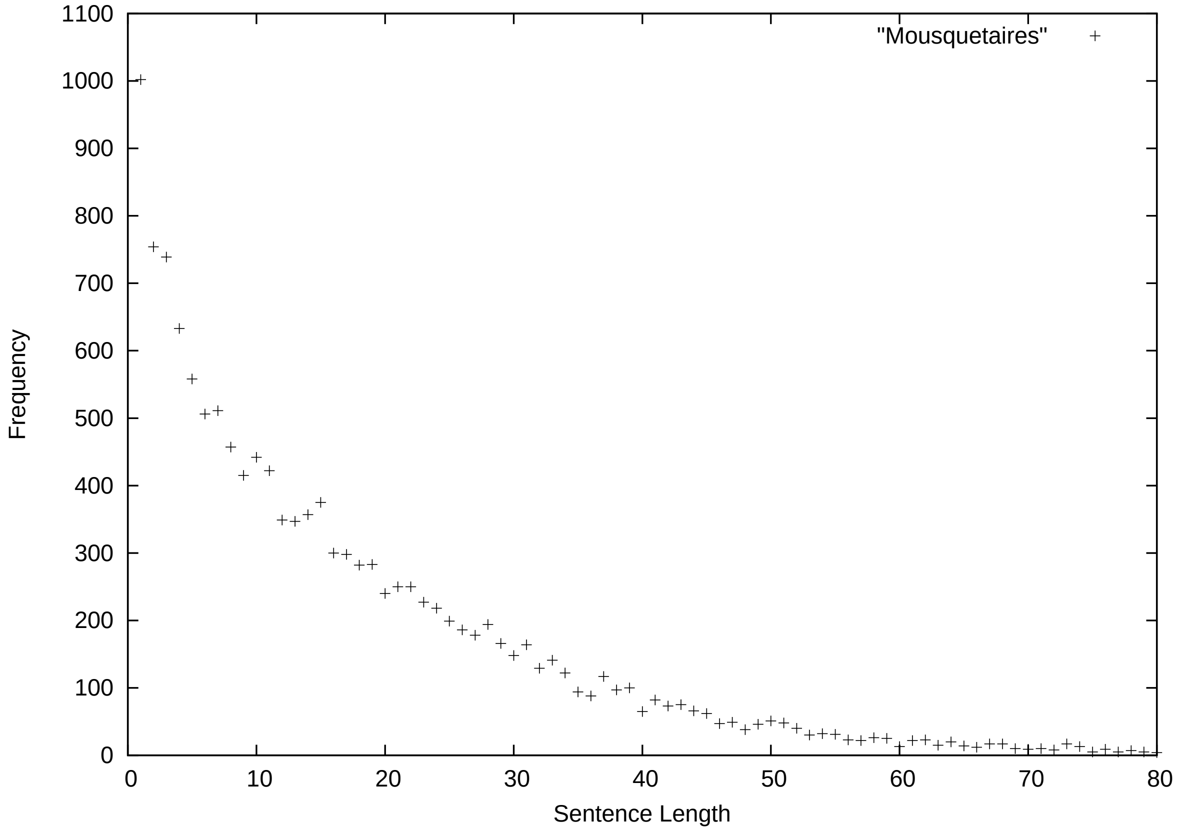

Figure 1: The frequency of each sentence length up to 80 in the text *Les Trois Mousquetaires*.

Musketeers) by Alexandre Dumas. The frequency of sentences of each length are shown in Figure 1.

We first determined the frequency of different sentence structures in the text, as described by the part-of-speech tags reported by Tree Tagger. Naturally, the majority of sentences are unique when compared in this way, however, some structures were frequent, particularly for very short utterances.

There were 11,539 different sentence structures found, of which 11,166 occurred precisely once. This figure is an overestimate, as the text file was not cleaned up, and there were a significant number of errors in the tagged data. The first four sentence types consisted of a single word, being either an interjection (263), a name (484), or a noun (65). (A sample extract from *Les Trois Mousquetaires* of single word sentences is shown in Figure 2. Figure 3 shows a sequence of single word sentences that occurs in the story.) Many frequent structures are such expressions as "said Aramis", which are usually tagged as separate sentences to the utterance of the character. The most frequently occurring sentence of length 4 is this type of phrase: "s'écria d'Artagnan", which occurs 14 times in the text. Similarly for *Le Petit Prince* (The Little Prince), the most frequent sentence structure beyond single-word interjections is the phrase "dit le petit prince" ("said the little prince"), which occurs 10 times.

Ah ! , – Non . , – Porthos ? , – Non . , – Aramis ?
, – Non . , – Oh ! , – Oh ! , – Oh ! , – Comment !
, – Oh ! , – Quoi ? , – Pardieu ! , – Oh ! , –
Silence ! , – Quoi ? , – Ah ! , – Allez . , – Ah ! , –
Silence ! , – Silence ! , – Silence ! , " Oh ! , – Ah
! , Ah ! , " Assez ? , Allez . , – Qui ? , – Peut-être
. , – Ah ! , ah ! , Pourquoi ? , – Hol ! , – Parle . ,

Figure 2: Some single-word sentences in *Les Trois Mousquetaires*.

The first interesting repeated sentence structure in Le Petit Prince is "PRO:PER ADV VER:pres ADV SENT", which occurs 5 times[1]. One example is "Elle ne change pas." (She doesn't change.) By contrast, the positive version of this sentence structure doesn't occur at all.

When grouping several French texts together

[1] PRO:PER represents personal pronouns, ADV adverbs, VER:pres present tense verbs, and SENT end of sentence

435 – Aramis !
436 – Porthos !
437 – Eh !
438 Messieurs !
439 Messieurs !

2369 – Arrêté !
2370 Athos !
2371 arrêté !
2372 pourquoi ?

Figure 3: Some single-word sentence sequences in *les Trois Mousquetaires*. Each sentence is preceded by its sentence number.

(2.8 million words, including some noise) before analysis the trend is similar, in that the most common sentence structure is the single-word utterance. Then there are two word sentences occurring frequently, such as VER:pres PRO:PER SENT, which represented "Pardonnez-moi!" (pardon me), despite being an imperative rather than simple present tense. There are several recurring sentence structures of about four words, such as "PRO:DEM VER:pres DET:ART NOM SENT", which includes "c'est l'amour" ("It's love"), so a larger collection can provide some useful simple examples for study[2]. Table 2 shows the frequency of particular sentence structures of different lengths in a corpus consisting of approximately 2.8 million words from French literature.

Our earlier work in French readability for readers with an English-speaking background revealed that average sentence length was a better indicator of readability than other standard measures. On this premise, we attempted to find extracts with a very low average sentence length. Figure 4 shows an extract from *Les Trois Mousquetaires* with a maximum sentence length of 4, as well as the first extract of at least 100 words, which is retrieved when the maximum sentence length is increased to 10, and has an average sentence length of about 5.

3734 – Connaissez -vous Athos ?

3735 – Non .

3736 – Porthos ?

3737 – Non .

3738 – Aramis ?

3739 – Non .

3740 Quels sont ces Messieurs ?

3741 – Des mousquetaires du roi .

– Ah ! fit Rochefort avec un sourire , voilà un hasard bien heureux ! et qui satisfera Son Eminence ! L' avez-vous prévenue ?

– Je lui ai écrit de Boulogne . Mais comment êtes-vous ici ?

– Son Eminence , inquiète , m'a envoyé à votre recherche .

– Je suis arrivée d'hier seulement .

– Et qu'avez-vous fait depuis hier ?

– Je n'ai pas perdu mon temps .

– Oh ! je m'en doute bien !

– Savez-vous qui j'ai rencontré ici ?

– Non .

– Devinez .

– Comment voulez-vous ? ...

– Cette jeune femme que la reine a tirée de prison .

– La maîtresse du petit d'Artagnan ?

– Oui , Mme Bonacieux , dont le cardinal ignorait la retraite .

Figure 4: Extracts from *Les Trois Mousquetaires* with a maximum length of 4 and 10 respectively.

[2] PRO:DEM for demonstrative pronouns, DET:ART for articles, and NOM for nouns

Table 1: The 20 most frequently occurring words in French newspapers and their main meanings.

1 to 10	Def.	11 to 20	Def.
de	of	que	that
le	the	dans	in
la	the	il	he/it
et	and	à	at/to
les	the	en	in
des	of the	ne	not
est	is	on	one (pronoun)
un	a/an/one	qui	who
une	a/an/one	au	at/to the
du	of the	se	him-/her-/it-self

3.2 Cognates, Named Entities, and Common Words

Our cognate (and named entity) list consisted of the intersection of the English and French lexicon provided in the parameter files with Tree Tagger. These lexicons contain all the tokens recognised by the tagger in the two languages. Clearly there are some *false friends* in this list, that is, words that look the same in both languages, but have different meanings. In addition, many cognates that are spelt differently were excluded. The initial list for English contained 358,097 words once duplicates were removed, and the French list had 475,209. Taking the intersection of the two lists gave a list of 17,908 words. Some false friends are very frequent in French. We determined the frequency of each cognate in *Les Trois Mousquetaires* and found that the majority of the highest ranked terms were either false friends, or were included due to French phrases that occur in English (eg. “fait accompli” and “laissez faire”). Highest ranked false friends in our list included: ment, pour, dans, tout, comme, plus, nous, quel, amis, fait, tend, main, voir, faire, jour, deux, ours, part, dire, sent, rend, and fort. The interjections “Ah” and “Oh” were not in the cognate list, so these were added manually.

In addition to the cognate list, a list of the 20 most frequent words in French newspapers (according to Crystal (1987), and listed in Table 1) was included in the “known” words to test the extreme case of a complete beginner. Named entities from Tree Tagger were also used in the list of permitted words. To this list we added the names of the characters from *Les Trois Mousquetaires* (Aramis, Porthos, d’Artagnan), as they were missing. Note that this procedure is to test the feasibility of the concept of retrieving useful extracts for learners, not a recommended technique for cognate generation. However, our observations discussed later provide ideas for future automation of cognate detection.

Un serpent !

Les provisions !

il est impossible !

est il possible ?

Un voyage sans fatigue et sans danger !

Impossible de continuer le commerce .

Figure 5: Some cognate-filtered sentences in the collection of several French texts, using the most frequent 125 words that occurred in the most frequent 200 words of *Les Trois Mousquetaires* and *Notre Dame de Paris*.

Using the above list provides many sentences (1409 for the larger collection, including duplicates), and some sentence sequences. On our larger collection we found 101 short sequences, including the following short fragment from *Les Misérables* that would be very easy for a beginner with English background to read:

141780 Une barricade !

141781 Ah !

141782 le bandit !

Expanding the list of frequent words to 125 (based on the most frequent 200 tokens occurring in *Les Trois Mousquetaires* and *Notre Dame de Paris*), which is approximately the size of the smallest vocabulary of published readers, a larger set of sentences is retrieved. Examples from the full collection of texts are shown in Figure 5.

Using the same level of filtering while including sentences that have at least 90% of the words in the lists, the sentence filter produces more substantial extracts. Examples are shown in Figure 6.

We calculated some general statistics to estimate the proportion of cognates in French text, as well as that of highly frequent words. Table 3 shows that based on our rough method of identifying cognates, French texts tend to consist of approximately 10% cognate content. The 20 most frequent words make up approximately 26% of the text.

270 – Ah !

271 fit d’ Artagnan .

272 – Non ; elle vous a été prise .

273 – Prise !

274 et par qui ?

887 – C’ est avec Monsieur que je me bats , dit Athos en montrant de la main d’ Artagnan , et en le saluant du même geste .

888 – C’ est avec lui que je me bats aussi , dit Porthos .

13007 Tout à coup elle jeta un grand cri de joie et s’élança vers la porte , elle avait reconnu la voix de d’ Artagnan .

13008 ” D’ Artagnan !

13009 d’ Artagnan !

Figure 6: Some cognate-filtered sentence sequences in *Les Trois Mousquetaires*.

4 Discussion

It is clear from the relative lack of repetition of sentence structures beyond those of fewer than 5 words, that either more sophisticated summaries of sentences would be required for use as sentence examples, or very large corpora. Using a chunking phase before grouping sentences may provide larger sets of related examples. The use of n-grams of different lengths would also allow learners to observe patterns of interest. For example, to better understand how adjectives are placed in French, users can look at occurrences of “ADJ NOM” as well as “NOM ADJ”.

Given that for English-speaking readers of French a good estimate of French readability is the average sentence length of the text, there is considerable scope for finding suitable extracts for reading. Filtering text based on sentence length provided extracts for reading practice that have an average sentence length of 5.

The frequency of some false friends in the cognate list suggests a simple automated technique would be to compare the relative frequency of the words in each language. Where there is a large discrepancy (for example, “aura”, which means“will have” in French), the word is highly likely to be a false friend rather than a cognate.

Applying our filter based on exact cognates, very frequent words and named entities allowed numerous sentences to be found in the corpus. Relaxing these requirements slightly by allowing some unknown words can produce extracts consisting of several sentences for reading — enough to get a sense of the moment in the story, but not as long as the shortest published stories for beginners in a language (about 75 words). Our cognate list was very restrictive in that it required words to have the exact same spelling in both languages (or as a related word, such as “arriver”, meaning “to arrive” in French). It is expected that allowing accents, typical variants such as the presence or absence of the letter “e” as a suffix, and common verb endings, will increase the size and quantity of extracts. Applying the filter to much larger bodies of text, such as found on the Web should also result in considerably more material being retrieved. Our previous work on measuring the readability of web text showed that a significant portion (8–19%) of web documents had the same readability range as stories published for those learning English Uitdenbogerd (2006). While cultural differences may mean that the range of readability of French differs from English on the web[3], we remain optimistic that many extracts can be retrieved that conform to these very strict criteria.

It should also be noted, that the texts used in the present study are relatively difficult to read. Texts written specifically for children would have smaller vocabularies, and translations into French (from English) would be likely to have larger cognate content.

Our earlier work on readability in French (Uitdenbogerd, 2005) demonstrated that sentence length was as good, if not better than the commonly used readability measures for predicting text difficulty where the person reading has an English-speaking background, and the language being read is French. Incorporating a measure of the cognate content was an even more reliable predictor of readability. The texts studied varied widely in their cognate rate, with some texts written specifically for people with an English-speaking background exploiting cognates. It might be expected that the more technical the text,

[3] An example of a relevant cultural difference is that there are many classic novels written in English for children, but none in French until relatively recently

the more common words there will be between French and English, however, our results in Table 3 showed a fairly consistent level of cognates in text. In contrast, the manual count of cognates in the samples of approximately 100 words used in Uitdenbogerd (2005) revealed a range from 5 to 42%, however, the two texts with the highest cognate count were specifically written for French learners with an English-speaking background, in the early stages of learning. When only texts written by and for native speakers of French are considered, the range was 7 to 12%, which doesn't differ too much from our estimate in the present study.

Studies in people's ability to predict the meanings of words from their context indicate that a knowledge of 95% of the words in the text are needed for comprehension (Ghadirian, 2002). Using this figure as a basis, it has been concluded that a vocabulary of 5000 (relatively frequent words) is required to be able to comfortably read any text in a given language (Groot, 2000), a figure we confirmed with our study of French texts (Uitdenbogerd, 2005). The cognate content in French texts probably reduces this figure somewhat. We can expect that about 10% of the infrequent words would be known as cognates. So, using *Les Trois Mousquetaires* as an example, the 95% threshold assuming no knowledge of cognates requires a vocabulary of about 3400 frequent words. Assuming 10% of the remaining vocabulary is known, this figure drops to about 3120. However, at the early stages of learning, when a person's vocabulary is small, the gains from cognates are greater. For example, a knowledge of 20 words gives a coverage of about 31% (when combining the total of all words regardless of their part of speech), whereas the additional cognate knowledge increases that coverage to 38%.

5 Conclusion and Future Work

We demonstrated potential techniques for identifying short extracts from French literature based on lexical or grammatical criteria to allow reading practice at the very early stages without the intensive work of translation. Experiments are currently underway that attempt to apply the same technique in reverse for English web documents, that is, applying strict lexical filters based on a small frequent words list and a large list of cognates.

Future work will include incorporating inexact cognate detection (Kondrak, 2001; Inkpen et al., 2005) to allow words with slightly different spelling to be found, and more sophisticated grammatical matching. Also, the idea of determining whether a word with the same spelling in both languages is a cognate or not based on its relative frequency and other available data, such as POS tags, will be explored.

References

Bell, T. (2001). Extensive reading: speed and comprehension. *The Reading Matrix*, 1(1).

Bormuth, J. R. (1966). Readability: a new approach. *Reading Research Quarterly*, 1:79–132.

Chall, J. S. and Dale, E. (1995). *Readability revisited: the new Dale-Chall readability formula*. Brookline Books, Massachusetts, USA.

COE (2003). Common European framework of reference for languages: Learning, teaching, assessment. `http://www.coe.int/T/DG4/Linguistic/CADRE_EN.asp#TopOfPage`. Accessed 8 September, 2006.

Collins-Thompson, K. and Callan, J. (2004a). Information retrieval for language tutoring: An overview of the REAP project. In *Proc. ACM-SIGIR International Conference on Research and Development in Information Retrieval*, Sheffield, UK. Poster.

Collins-Thompson, K. and Callan, J. (2004b). A language modeling approach to predicting reading difficulty. In *Proceedings of the HLT/NAACL 2004 Conference*, pages 193–200, Boston.

Cornaire, C. M. (1988). La lisibilité: Essai d'application de la formule courte d'henry, au français langue étrangère. *Canadian Modern Language Review*, 44(2):261–273.

Cowling, S., editor (1982). *Bibliobus*. Mary Glasgow Publications Ltd, London, UK.

Crystal, D. (1987). *The Cambridge encyclopedia of language*. Cambridge University Press, New York, NY, USA.

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

Association for Computational Linguistics, Association for Computational Linguistics.

Ghadirian, S. (2002). Providing controlled exposure to target vocabulary through the screening and arranging of texts. *Language Learning and Technology*, 6(1):147–164.

Granowskey, A. and Botel, M. (1974). Background for a new syntactic complexity formula. *The Reading Teacher*, 28:21–35.

Groot, P. J. M. (2000). Computer-assisted second language vocabulary acquisition. *Language Learning and Technology*, 4(1):60–81.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications*.

Inkpen, D., Frunza, O., and Kondrak, G. (2005). Automatic identification of cognates and false friends in French and English. In Mitkov, R., editor, *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 251–257, Borovets, Bulgaria.

Katz, I. R. and Bauer, M. I. (2001). Sourcefinder: Course preparation via linguistically targeted web search. *Educational Technology and Society*, 4(3):45–49.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, X:62–102.

Kondrak, G. (2001). Identifying cognates by phonetic and semantic similarity. In Knight, K., editor, *Proceedings of NAACL 2001: 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, volume 2, pages 103–110, Pittsburgh, PA, USA. NAACL.

Miyazaki, Y. and Norizuki, K. (2008). Developing a computerized readability estimation program with a web-searching function to match text difficulty with individual learner's reading ability. In *WorldCALL*.

Nilsson, K. and Borin, L. (2002). Living off the land: The web as a source of practice texts for learners of less prevalent languages. In *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation*, pages 411–418, Las Palmas. ELRA.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Somers, H. and Jones, D., editors, *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the Association for Computational Linguistics*.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In Liu, L. and Grossman, D., editors, *Proc. International Conference on Information and Knowledge Management*, volume 10, pages 574–576, Atlanta, Georgia, USA. ACM, ACM.

Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010). Sorting texts by readability. *Computational Linguistics*, 36(2):203–227.

Uitdenbogerd, A. L. (2003). Using the web as a source of graded reading material for language acquisition. In Zhou, W., Nicholson, P., Corbitt, B., and Fong, J., editors, *International Conference on Web-based Learning*, volume 2783 of *Lecture Notes in Computer Science*, pages 423–432, Melbourne, Australia. Springer.

Uitdenbogerd, A. L. (2005). Readability of French as a foreign language and its uses. In Turpin, A. and Wilkinson, R., editors, *Australasian Document Computing Symposium*, volume 10.

Uitdenbogerd, A. L. (2006). Web readability and computer-assisted language learning. In Cavedon, L. and Zukerman, I., editors, *Australasian Language Technology Workshop*, pages 99–106.

van Oosten, P., Tanghe, D., and Hoste, V. (2010). Towards an improved methodology for automated readability prediction. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Wible, D., Kuo, C.-H., Chien, F.-Y., and Wang, C. (2000). Toward automating a personalized concordancer for datadriven learning: a lexical difficulty filter for language learners. In Ketteman, B. and Marko, G., editors, *Conference on Teaching and Language Corpora*, volume 4, pages 147–154, Graz. Rodopi.

Table 2: Example sentence structures of each length in the corpus of French literature. Note that frequencies for items are approximate due to the inaccuracy of the tagger and the noise in the data. Note also that some tags are incorrect, such as the sentence of length 2, which should really have been labelled as “imperative” instead of “present” tense.

Len	Structure	Freq
	Example	
	Translation	
1	INT SENT	
	Ah!	
	Ah!	943
2	VER:pres PRO:PER SENT	
	Sauvez-moi!	
	Save me!	204
3	VER:pres DET:ART NOM SENT	
	Videz le vase!	
	Empty the vase	123
4	PRO:PER VER:simp DET:ART NOM SENT	
	Il leva les yeux.	
	He raised his eyes. 73	
5	PRO:PER ADV PRO:PER VER:pres ADV SENT	
	Elle ne la sait pas.	
	She doesn’t know her/it.	51
6	PRO:PER VER:simp DET:ART NOM PRP NOM SENT	
	Elle resta un moment sans parler.	
	She remained speechless for a moment.	15
7	PRO:PER ADV PRO:PER VER:pres ADV PRP NOM SENT	
	Il n’ y a pas de jardin.	
	There is no garden	11
8	DET:ART NOM VER:pres DET:ART NOM PRP DET:ART NOM SENT	
	La philosophie est le microscope de la pensée.	
	Philosophy is the microscope of thought	9
9	PRO:PER VER:impf DET:ART NOM KON PRO:PER VER:impf DET:ART NOM SENT	
	Elle était la lumière et il était l’ ombre	
	She was the light and he was the shade	4
10	DET:ART NOM VER:impf DET:ART NOM PUN DET:ART NOM VER:impf DET:ART NOM SENT	
	Les assaillants avaient le nombre ; les insurgés avaient la position.	
	The assailants had the numbers; the insurgents had the position.	3

Table 3: Statistics of occurrence of cognates (including named entities and false friends), and highly frequent words for French texts

Text	Types	Tokens	Cognates	Top 20 News Words
Le Petit Prince	2,614	16,484	1,773 (11%)	4,214 (26%)
Les Méditations	3,040	29,976	3,030 (10%)	9,111 (30%)
Les Trois Mousquetaires	16,029	235,056	23,137 (9.8%)	61,439 (26%)
Notre Dame de Paris	18,100	176,245	18,451 (10%)	51,880 (29%)