Readability of French as a Foreign Language and its Uses

Alexandra L. UITDENBOGERD
School of Computer Science and IT
RMIT
GPO Box 2476V Melbourne Australia
alu@cs.rmit.edu.au

Abstract *Reading is an important means of foreign language acquisition, particularly for vocabulary. Providing reading material that is of a suitable level of difficulty allows users to acquire vocabulary the most efficiently. Thus an on-line reading material recommender system for language learners requires a readability measure so that the difficulty of texts can be automatically assessed. However, most readability measures were developed for native child speakers of English. In this article I discuss an experiment in readability for learners of French. I conclude that using the average number of words per sentence correlates more closely with human judgements than many commonly available readability measures. I propose a new readability measure for learners of French that have English as their main language, which combines sentence length with the number of words that are similar in both languages (cognates). This measure slightly improves on sentence length for modelling French readability.*

Keywords Text readability, Information retrieval

1 Introduction

Acquiring sufficient vocabulary to read a foreign language comfortably is an ongoing problem for language learners. Once sufficient grammar is learned the student can make their way through most texts with the aid of a dictionary, but reading more naturally with native-like comprehension remains a dream. Yet, many people need to function at high proficiency in their second, third or even fourth language.

Much research effort has gone into improving language acquisition via reading. Some researchers have concluded that extensive reading at an appropriate level of difficulty is a more efficient method of language acquisition than intensive study of texts [2]. Others have discovered that in order to deduce new words in context requires knowing 95% of the words in the text [4]. This leads to the conclusion that people need to learn a vocabulary of about 5,000 words to be at that level of comfort with normal texts [5].

Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 13, 2005.

Measures of text difficulty are usually modelled on native children's knowledge of vocabulary and their text comprehension [9]. An example is the Flesch reading ease score available in Microsoft Word. Typical readability measures contain a component representing vocabulary difficulty such as word length and another representing grammatical difficulty such as sentence length. Whilst there are many readability measures developed for English native speakers, there are few for specific foreign languages, and to my knowledge only one that was designed for use across languages. Very few are specifically designed for foreign language students [9].

It is the goal of my research that the acquisition of language through reading can be made more streamlined and efficient through the building of a web text search engine — or really a recommender system — based on readability rather than relevance of topic. In earlier work I examined the issue of readability of French for English speakers [11]. My hypothesis is that current readability measures are not ideal for this purpose as most were developed by using school-age native speakers. Adult or adolescent foreign language learners have different knowledge of a foreign language and different language skills in general. In addition, their previous languages will influence their understanding of the language to be learnt.

In work to be published elsewhere I examine the question of readability of the web, that is, what is the range of readability levels of text on the web. When this is known, it will be clear at what stage the web can be used most efficiently for further language acquisition by reading.

In this paper I once again address the readability of French for English speakers. I scaled up the experiment of my preliminary work [11] by asking people with a range of skill levels in French to rank a set of 10 texts according to difficulty. I also analysed several on-line French books, as well as a corpus of spoken French in terms of vocabulary requirements. It is clear that for some texts a vocabulary of 5,000 would not be quite enough to achieve 95% word knowledge. This gives the learner few stark choices: struggle, read with insufficient understanding, or forget about it. However, there is another option: read selected texts first to build up vocabulary skills before tackling the harder

text. Various software tools can aid in the selecting and sorting of suitable text [4].

2 Literature Review

Extensively reading easier texts has been shown to be more effective for language acquisition than intensively reading more difficult texts with the aid of a dictionary [2]. Further corroboration comes from Krantz [10], in whose experiments the students with the strongest language skills gained the most vocabulary through reading a set text. Krantz found that some words can be learned purely through reading, although these words tend to be those that occur frequently in texts, implying a certain level of repetition required [10]. Pioneer of controlled vocabulary-based language teaching Michael West believed that words needed to be encountered initially at least three times before they were absorbed [12]. A later study showed that words need to occur at least five times to be retained (discussed by Ghadirian [4]). Less frequently occurring words are learnt better by looking up in a dictionary when they occur in the text, than just by reading them [10].

Given that learning a language via reading is best achieved with text that is of a suitable reading level, methods of measuring reading level are useful for selecting texts. However, as noted earlier, most of these were developed for native English-speaking schoolchildren. There have been some studies of readability for other languages. Klare mentions that some formulae were tested for English materials to be read by those of a non-English-speaking background [9]. He mentions that Tharp was the first to work on readability for languages other than English. Considerable work on French readability was completed by De Landsheere and his student Henry. More recently Cornaire tested Henry's readability formula for French as a foreign language [3]. However, I'm unaware of any formulae that consider cognates – the words that are recognisably similar to words with the same meaning in the person's native language. For example the word "methode" in French would be a cognate for an English speaker.

Much recent study has been on the use of text corpora to support language learning [13, 6, 7]. Approaches include the study of parallel texts, using concordancers to understand word usage, and the development of targeted vocabularies for learning. A tool that finds web texts based on readability has also been developed [8]. Initially written to find materials of a suitable difficulty level for school-children, the concept can be applied to language learning as well.

3 Experiments

In this section I discuss two experiments. The first compares user readability assessments of 10 texts to standard readability measures and factors, as well as across different levels of language skill. The second looks at vocabulary in several on-line French texts.

3.1 Relative Readability of Different Texts

The aim of this experiment was to determine how those learning French as a foreign language perceive difficulty of texts. The research questions I raise are:

- What makes a French text easy or difficult for students of French?
- How do current readability measures compare for measuring French readability for students of the language?
- How does French readability for students of French compare to that for native speakers?

3.1.1 Method

In this experiment I wanted to ensure representative samples of various types of French text: native children's books, native adult books, reduced vocabulary books, books designed to have simple grammar, books that intentionally make use of cognates, and books that try to keep both grammar and vocabulary simple. The procedure of selection involved finding the subset from a collection of French books that met the criteria and randomly choosing one book from that subset. In addition I included the draft of a comic book that I have written in which I intentionally restricted the vocabulary to cognates and twelve of the twenty most frequently occurring words found in French newspapers.

A total of fifteen people assessed the selected texts. Table 1 shows the French language skills of the participants. Two participants were native French-speaking adults, and one had spoken French from the age of six. The remainder were students of the Alliance Française in Melbourne. The French skill-level shown is the class that the students were taking at the time of the experiment. Two of the Beginner 2 participants were of Asian descent and may have had English as a second language. The remainder of the participants seemed to have English as their main language. Due to a procedural error two of the participants in the Intermediate 6 class only assessed eight of the ten books.

Each participant was asked to rank the books from easiest to hardest using approximately the first 100 words of the text. Participants varied in the care taken over the task. Some made repeated comparisons. Some flicked through books and made judgements based on this. These rankings were compared with each other as well as with readability measurements.

Approximately the first 100 words (up to the end of the sentence after word 100) were used from each text for readability measurement. The largest number of words used (as counted by the unix utility `wc`), was

Participant Number	Skill Level	Skill Class
1–3	Beginner 2	b2/3
4	Beginner 3	b2/3
5–7	Beginner 6	b6/i1
8	Intermediate 1	b6/i1
9–12	Intermediate 6	i6
13	native	native
14	native	native
15	near-native	native

Table 1: Language skills of participants in the experiment. Beginner and Intermediate levels refer to those used at the Alliance Française. Beginner 6 is the highest beginner level.

134. The unix `style` utility was applied to each text to gather readability statistics. The statistics included: average words per sentence (WpS), average word length (W len), average number of syllables (Syll), the Kincaid formula (Kinc), the automated readability index (ARI), the Coleman-Liau formula (C.-L.), Flesch reading ease (Fles), the Fog index (Rog), Lix and the SMOG grading. The ARI formula as calculated by `style` is:

$$ARI = 4.71 * Wlen + 0.5 * WpS - 21.43 \quad (1)$$

In addition I manually counted cognates for each text (Cog). A cognate was included if it was either an exact spelling (plus or minus a trailing letter "e"), or a polysyllabic word with an obvious common root and very similar meaning to the English equivalent (eg. compliqué). Repeated cognates were counted. This mainly affected the Gnomeville and Temps des Rèves stories which had some cognates occurring at least 5 times.

I also developed a new measure that combines words per sentence with the cognate count, tuning the constant factor based on the results discussed in the next section.

$$FR = 10 * WpS - Cog \quad (2)$$

In general the cognate count for this formula would be an average per 100 words sampled, but for this experiment the samples of 100-134 were used as a basis for the count.

3.1.2 Results

Tables 2, 3, 4, 5 and 6 show the rankings of the texts by participants as well as the mean and standard deviations of these ranks. In table 3 we can see that the standard deviation of the rank is quite low for most books. However, there are a few that are greater than 2. The greatest standard deviation is found amongst the b6/i1 group for the text "La Mission de Slim Kerrigan". This text was adapted to mainly use the 1,000 most frequently occurring words, however, the number of cognates, at least in the first 100 or so words is the lowest in the set of 10 texts. Its sentences were the second longest (See the WpS statistics in Table 7). The greatest difference in average rank across the groups was for "La Grimassouille" a young children's story. The b6/i1 group ranked it three places higher than the other groups (except i6). During the experiment, two of the participants in this group apologised to me that they found this supposedly easy book rather difficult to read. The i6 group also had its largest standard deviation for this book (see Table 6).

The correlation between the different groups was quite high (Table 6), with the lowest correlation being 0.846. (Group i6 was not compared with the others due to the missing data).

Table 8 shows the correlation between standard readability measures and the average ranking for the texts given by each group of participants. In all groups except the native group the best correlated measure was a simple words per sentence count (WpS). For the native group, the ARI measure achieved a slightly higher correlation than words per sentence, and it was the best of the standard readability measures studied. The Flesch score shows a negative correlation as it is a "reading ease" score rather than a reading difficulty score, but was quite weakly correlated. Coleman-Liau gives a negative correlation despite being a reading difficulty score. This may be related to the negative correlation between word length and reading difficulty in this experiment, and indeed between word length and sentence length (-0.45).

Cognates tend to be longer words, having a correlation of 0.65 with word length and 0.68 with syllable count respectively for this collection. The negative correlation between reading difficulty and word length, particularly for non-native participants, may be related to this tendency. The word length effect may be unusually strong in this experiment due to half of the texts being written for students of French that have English as their main language, and the consequent increased use of cognates.

The new measure FR, which combines the cognate count with sentence length, achieved a slightly higher correlation than sentence length alone — except with native French speakers.

The results of the readability experiment suggest that there is a measurable difference in perceived readability between native speakers and learners of the language. The assessment of readability by non-native readers seemed to be much more based on surface features of the language and less on other factors, demonstrated by the higher correlation with sentence length and word length. Comments from two of the native speakers indicated that they took account of the conceptual difficulty of the text in their judgements in addition to other factors. This may account for much of the difference.

Book	b2	b2	b2	b3	b6	b6	b6	i1	i6	i6	nat. ad.	nat. ad.	nat. ad.	i6	i6
	1	2	3	4	5	6	7	8	9	12	13	14	15	10	11
Les Loisirs	2	3	3	2	1	2	2	2	4	4	1	2	1		
Les Miserables	10	8	7	8	3	7.5	9	6	7	7	5	10	8	5	7
La Grimassouille	3	2	2	4	7	6	4	5	2	3	3	1	3	7	1
Cendrillon	9	7	6	7	10	7.5	10	9	9	8	9	7	6	6	5
Les Tours Eiffel	4	6	4	3	4	4	5	3	3	2	4	4	4	1	6
Gnomeville: Dragon	1	1	1	1	2	1	1	1	1	1	2	5	2		
Enfants de Paris	6	4	8	6	6	4	7	7	5	5	6	6	10	2	3
Terre des hommes	7	9	9	9	9	9	8	10	10	10	10	9	9	8	8
La Mission de Slim	8	10	10	10	8	10	3	8	8	9	8	8	7	3	2
Les Temps des Reves	5	5	5	5	5	4	6	4	6	6	7	3	5	4	4

Table 2: Ranks given to each text by each participant. Where the same rank was given for two or more items the mean rank is allocated to both. For example, the rank 7.5 is given to two items that received equal rank 7 from participant 6, and their is no rank 8.

Book	all except 10 & 11		b2/3		b6/i1		native		non-native	
	ave	std dev	ave	stdev	ave	stdev	av	std	ave	std
Les Loisirs	2.2	1.01	2.5	0.58	1.75	0.50	1.33	0.58	2.50	0.97
Les Miserables	7.3	1.93	8.25	1.26	6.38	2.56	7.67	2.52	7.25	1.87
La Grimassouille	3.5	1.71	2.75	0.96	5.50	1.29	2.33	1.15	3.80	1.75
Cendrillon	8.0	1.39	7.25	1.26	9.13	1.18	7.33	1.53	8.25	1.36
Les Tours Eiffel	3.8	0.99	4.25	1.26	4.00	0.82	4.00	0.00	3.80	1.14
Gnomeville: Dragon	1.5	1.13	1	0.00	1.25	0.50	3.00	1.73	1.10	0.32
Enfants de Paris	6.2	1.63	6	1.63	6.00	1.41	7.33	2.31	5.80	1.32
Terre des hommes	9.1	0.86	8.5	1.00	9.00	0.82	9.33	0.58	9.00	0.94
La Mission de Slim	8.2	1.88	9.5	1.00	7.25	2.99	7.67	0.58	8.40	2.12
Le Temps des Reves	5.1	1.04	5	0.00	4.75	0.96	5.00	2.00	5.10	0.74

Table 3: Average and standard deviation of ranks across all participants (except 10 and 11), and across each group.

All except 10 & 11		Native		Non-native	
	mean		mean		mean
Gnomeville: Dragon	1.5	Les Loisirs	1.33	Gnomeville: Dragon	1.10
Les Loisirs	2.2	La Grimassouille	2.33	Les Loisirs	2.50
La Grimassouille	3.5	Gnomeville: Dragon	3.00	La Grimassouille	3.80
Les Tours Eiffel	3.8	Les Tours Eiffel	4.00	Les Tours Eiffel	3.80
Le Temps des Reves	5.1	Le Temps des Reves	5.00	Le Temps des Reves	5.10
Enfants de Paris	6.2	Cendrillon	7.33	Enfants de Paris	5.80
Les Miserables	7.3	Enfants de Paris	7.33	Les Miserables	7.25
Cendrillon	8.0	Les Miserables	7.67	Cendrillon	8.25
La Mission de Slim	8.2	La Mission de Slim	7.67	La Mission de Slim	8.40
Terre des hommes	9.1	Terre des hommes	9.33	Terre des hommes	9.00
	std dev		std dev		std dev
Terre des hommes	0.86	Les Tours Eiffel	0.00	Gnomeville: Dragon	0.32
Les Tours Eiffel	0.99	La Mission de Slim	0.58	Le Temps des Reves	0.74
Les Loisirs	1.01	Les Loisirs	0.58	Terre des hommes	0.94
Le Temps des Reves	1.04	Terre des hommes	0.58	Les Loisirs	0.97
Gnomeville: Dragon	1.13	La Grimassouille	1.15	Les Tours Eiffel	1.14
Cendrillon	1.39	Cendrillon	1.53	Enfants de Paris	1.32
Enfants de Paris	1.63	Gnomeville: Dragon	1.73	Cendrillon	1.36
La Grimassouille	1.71	Le Temps des Reves	2.00	La Grimassouille	1.75
La Mission de Slim	1.88	Enfants de Paris	2.31	Les Miserables	1.87
Les Miserables	1.93	Les Miserables	2.52	La Mission de Slim	2.12

Table 4: Sorted mean and standard deviation of ranks for each group

	b2/3	b6/i1	i6	nat
b2/3	1	0.85		0.92
b6/i1		1		0.85
i6				
nat				1

Table 5: Correlation between different groups

Book	i6				average	stddev
	9	12	10	11		
Les Tours Eiffel	2	1	1	6	2.5	2.38
La Grimassouille	1	2	7	1	2.75	2.87
Enfants de Paris	3	3	2	3	2.75	0.50
Le Temps des Reves	4	4	4	4	4	0.00
La Mission de Slim	6	7	3	2	4.5	2.38
Les Miserables	5	5	5	7	5.5	1.00
Cendrillon	7	6	6	5	6	0.82
Terre des hommes	8	8	8	8	8	0.00

Table 6: Rankings of the eight texts by the i6 group of participants.

Book	WpS	W len.	Syll.	Kinc.	ARI	C.-L.	Fles.	Fog	Lix	SMOG	Cog
Cendrillon	20.0	3.74	1.29	7.4	6.2	6.2	77.6	9.5	28.8	7.7	7
Enfants de Paris	8.8	4.28	1.46	5.1	3.2	9.4	74.2	7.7	24.9	8.2	11
Gnomeville: Dragon	4	4.44	1.48	3.4	1.4	10.3	77.4	5.3	28.1	6.3	42
Grimassouile	9.2	4.23	1.36	4.1	3.1	9.1	82.4	5.1	25.5	6.2	11
Les Loisirs	4.8	3.47	1.25	1.0	-2.7	4.6	96.6	3.1	12.4	5.0	6
Les Mis. (adapted)	11.6	4.1	1.28	4.0	3.6	8.3	86.9	6.5	28.9	7.1	7
Les Tours Eiffel	13.0	3.92	1.31	4.9	3.5	7.3	83.0	7.9	27.5	8.2	15
Slim Kerrigan	17.7	3.67	1.15	4.9	4.7	5.8	9.5	7.8	21.4	6.2	5
Les Temps des Reves	13.8	3.83	1.25	4.5	3.5	6.7	87.5	8.4	24.7	8.5	12
Terre des Hommes	15.5	3.78	1.24	5.1	4.1	6.5	86.0	6.8	25.2	5.7	12

Table 7: Readability statistics for the text samples used in the experiment.

Group	WpS	W len.	Syll.	Kinc.	ARI	C.-L.	Fles.	Fog	Lix	SMOG	Cog	FR
all	0.83	-0.31	-0.51	0.67	0.73	-0.30	-0.39	0.63	0.27	0.15	-0.56	0.84
b2/3	0.79	-0.36	-0.60	0.56	0.66	-0.36	-0.48	0.59	0.18	0.13	-0.61	0.82
b6/i1	0.85	-0.23	-0.41	0.78	0.81	-0.22	-0.29	0.64	0.36	0.17	-0.56	0.86
native	0.70	-0.10	-0.30	0.67	0.72	-0.09	-0.35	0.66	0.40	0.24	-0.33	0.69
non-native	0.85	-0.36	-0.57	0.66	0.72	-0.36	-0.39	0.62	0.22	0.12	-0.61	0.87

Table 8: Correlation between readability measures and mean user rankings.

3.2 Vocabulary of French Texts

In this experiment I examined the vocabulary size required to be able to understand 95% of the text of several books and corpora. The texts examined were the French bible, a corpus of spoken French, Consuelo by nineteenth century author George Sand, and Le Petit Prince by Saint-Exupéry. Figure 1 shows the vocabulary (types) required for different portions of the given text sources.

The vocabulary required for the children's book Le Petit Prince is comfortably less than 1500 words, however Consuelo requires somewhat more than 5,000. The spoken French corpus requires a vocabulary that is less than 2,000, and the French bible needs about 4,500 words. This suggests that children's books and conversational vocabulary may be achievable, but that long adult texts will be a challenge. The figure 5,000 for required vocabulary size seems to be supported in these examples, but obviously this experiment is somewhat small in scale for any extrapolation to other texts. It also emphasises that vocabulary requirements grow with the text size, making shorter texts a good choice for learners. This is reflected in current practice, as most reading books for learners of a language are quite short (for example, those published by Hachette and Oxford for French and English learners respectively).

4 Conclusions

With the aim of providing an on-line reading recommender for language learners based on readability, I've explored various factors that affect readability. In a study of vocabulary size of French text, my results confirm the previously cited figure of 5,000 as a required vocabulary size for ease of reading. Smaller vocabularies are likely to be required for reading children's books and for general conversation, but even these vocabularies grow with the amount of text.

In my study into French readability for English native speakers, participants ranked a set of texts. These were compared to some readily available readability measures as well as the number of cognates. Amongst the set of measures, the best was a simple average of the number of words per sentence — a similar finding to my earlier work [11]. Combining this with a cognate count gave slightly better results.

There were slight differences between native and non-native speaker assessments of texts but there was stronger correlation between these groups than there was between most group's rankings and the readability measures.

It may be thought that other factors, such as familiarity with the topic discussed, or the story outline would be important for readability. While there is evidence that a rich reading environment that incorporates images aids comprehension [1], storyline familiarity was clearly not a strong factor in this experiment, since both Cinderella (Cendrillon) and Les Miserables were rated as quite difficult by participants in this study. However, two native-speaking participants commented that their readability assessments incorporated the conceptual difficulty of the texts — a factor that would be difficult to measure in text.

The number of cognates in the text was reasonably highly correlated with readability, and when combined with sentence length predicts readability well, as assessed by English-speaking learners of French. This work did not clearly distinguish the relative importance of cognates and sentence length, however, as there was only one text with a markedly different number of cognates and it also had the shortest average number of words per sentence. Future work should probably include the exploration of this aspect of readability.

If cognates are important for measuring French readability for English speakers, then for it to be useful for a text recommender, automated means of identifying cognates in text will need to be developed. This is expected to be the next step in this research project.

5 Acknowledgements

I thank the Alliance Française of Melbourne for their assistance with experiments. I also thank Betsy Kerr for providing the French spoken language corpus.

References

[1] K. Al-Seghayer. The effect of multimedia annotation modes on l2 vocabulary acquisition: a comparative study. *Language Learning and Technology*, Volume 5, Number 1, pages 202–232, January 2001.

[2] T. Bell. Extensive reading: speed and comprehension. *The Reading Matrix*, Volume 1, Number 1, April 2001.

[3] C. M. Cornaire. La lisibilité: Essai d'application de la formule courte d'henry, au français langue étrangère. *Canadian Modern Language Review*, Volume 44, Number 2, pages 261–273, January 1988.

[4] S. Ghadirian. Providing controlled exposure to target vocabulary through the screening and arranging of texts. *Language Learning and Technology*, Volume 6, Number 1, pages 147–164, January 2002.

[5] P. J. M. Groot. Computer-assisted second language vocabulary acquisition. *Language Learning and Technology*, Volume 4, Number 1, pages 60–81, May 2000.

[6] S. Hunston. *Corpora in applied linguistics*. The Cambridge Applied Linguistics Series. Cambridge University Press, Cambridge, 2002.

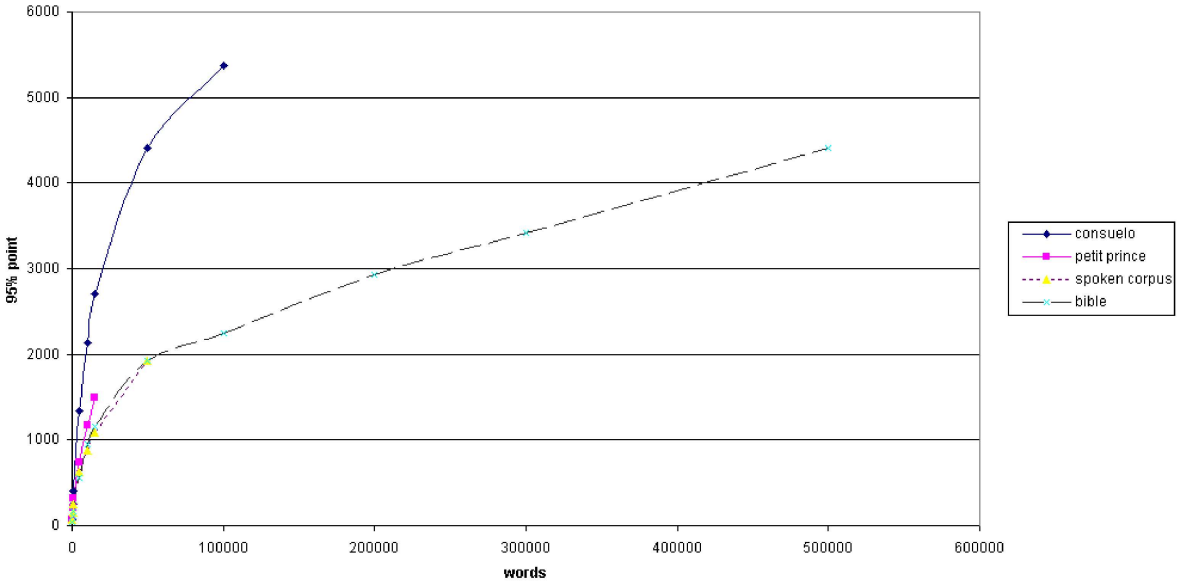

Figure 1: Graph of the 95% vocabulary point for four different French texts. The horizontal axis refers to the number of words from the start of the text that were included in the calculation.

[7] E. St. John. A case for using a parallel corpus and concordancer for beginners of a foreign language. *Language Learning and Technology*, Volume 5, Number 3, pages 185–203, September 2001.

[8] I. R. Katz and M. I. Bauer. Sourcefinder: Course preparation via linguistically targeted web search. *Educational Technology and Society*, Volume 4, Number 3, pages 45–49, 2001.

[9] G. R. Klare. Assessing readability. *Reading Research Quarterly*, Volume X, pages 62–102, 1974.

[10] G. Krantz. *Learning vocabulary in a foreign language: a study of reading strategies*. Ph.D. thesis, University of Göteborg, Sweden, 1991.

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

[12] M. West. The construction of reading material for teaching a foreign language. *Dacca University Bulletin*, Number 13, 1927. Republished as part of "Teaching English as a foreign language, 1912 – 1936: Pioneers of ELT, Volume V: Towards Carnegie", R. Smith editor.

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