Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

# Tools for Supporting Language Acquisition via Extensive Reading

**Alexandra UITDENBOGERD***

[a]*School of Computer Science and IT, RMIT University, Australia*

*alexandra.uitdenbogerd@rmit.edu.au

**Abstract:** Extensive reading, that is, the reading of large quantities of text at a comfortable level of difficulty, has been shown to be of great benefit to second and foreign language learners skills. We define seven types of reading support as management, text generation, text selection, text simplification, preparation, translation, and revision. We describe and propose a range of tools that assist learners to locate reading material of an appropriate level of difficulty and manage their progress. Our prototype implementation of two of these tools, the Bilingual eReader and the Readable Extract Search Engine, demonstrate their feasibility.

**Keywords:** CALL, extensive reading

## 1. Introduction

Imagine attempting to make sense of the following extract while reading:

> She was the youngest of the two daughters of a most XXX, XXX father; and had, in XXX of her sister's XXX, been XXX of his house from a very early XXX.

The “XXX”s above represent unknown words. For this extract, you are reading a text in which you know 81.25% of the words. If you are a native or highly fluent speaker of English, you have the advantage of excellent knowledge of the grammatical flow of English sentences, which will greatly assist you in your ability to predict the meaning of the missing words. Research has shown that for most foreign language learners to predict the meaning of words in text requires knowing at least 95% of the words (Liu and Nation, 1985). However, for most native texts, the learner would need a vocabulary of 5,000 word families in order to achieve that coverage (Hirsh and Nation, 1992). A typical 1,000 hour course of English as a second language will achieve a 2,000 word vocabulary (Laufer, 2000). This leaves a substantial shortfall in vocabulary knowledge.

It is very challenging to become proficient in a foreign language to the extent that it can be used for discourse at a high level. One useful activity for improving language skills that is readily available is extensive reading. However, the reading material needs to be at an appropriate level of difficulty as well as interesting to be the most effective. For English and French there is an extensive collection of reading material available, such as the Oxford Bookworms series. Other languages are less resource rich. However, there are large quantities of text produced in a variety of languages on the Web and elsewhere. A proportion of these are potentially suitable for reading practice. Uitdenbogerd (2006) examined a corpus of English web sites and found that the readability range of Web-sites adequately covered that of typical stories written for English language learners. However, many such web-sites provided uninteresting reading, such as navigation pages. Heilman et al. (2010) reached a similar conclusion in their work, and developed methods to eliminate web pages that don’t have sufficient prose, as well as providing categories for students to choose from to increase the chance that the material retrieved would be interesting to them.

The contributions of this paper are:

1. a taxonomy of extensive reading support that can be provided by computer (and other) systems (See Section 3)
2. tools and techniques for providing appropriate text from a corpus of literature (See Section 4)

## 2. Background

The need for language skills is becoming more relevant in globally connected societies. For example, millions of Chinese students are studying English (Zheng and Cheng, 2008), and 27 million people have taken the TOEFL English language test (TOEFL website). It is generally agreed that more exposure to language improves skills in the language. For example, various studies have shown vocabulary gain from reading (Waring and Takaki, 2003).

In a previous paper we observed that there are short extracts and sentences to be found in the classical French literature that meet strict vocabulary criteria such as consisting only of the 20 most frequently occurring words in news text, French-English cognates, and proper nouns (Uitdenbogerd, 2010). We provided an estimate of exact cognates in native text (10%) and a frequency distribution of sentence structures.

Clearly the more strict the constraints on the text, the fewer suitable extracts will be found. However, even at the strictest constraints such as 1-word sentences, or sentences consisting only of the most frequent 20 words, proper nouns and French-English cognates, extracts could be found. Relaxing constraints to allow 95% coverage provides larger quantities of extracts.

## 3. Extensive Reading Support Taxonomy

Support for extensive reading falls into seven main categories: management of an extensive reading-based programme, generation of readable text, selection of readable text, simplification of text, learner preparation, glossary or translation support, and revision.

- ***Management*** via an extensive reading system frees the user from tracking their reading and their progress. It can work in conjunction with text selection and other categories of support by overseeing the different types of activity available to the user.
- ***Text Generation*** involves generating stories or other content based on strict readability constraints. For early stages that require much reading practice with a small set of vocabulary and grammar this can be a useful addition to the choices availble for reading.
- ***Text Selection*** via readability-based search enables the learner to find easier texts to read from large corpora. When combined with topic search, texts can be both relevant and readable. Text selection can also be based on recommendations in the manner of typical ratings-based recommender systems. Text selection support can help ensure that text is both readable and interesting, increasing the motivation of the learner to read and benefit from reading.
- ***Text Simplification*** has traditionally been done manually, typically for classic works of literature to make them accessible to both children and foreign language learners. Typically the works are not only simplified in language, but made significantly shorter.
- ***Preparation*** consists of materials that assist the learner with vocabulary and background knowledge before they commence reading. (See Section 4.1)
- ***Translation*** of difficult vocabulary occurring in text helps the learner to read fluently, and when a gloss is used to look up a word, the word is more remembered than if it is merely read over (Lomicka, 1998). Translations of difficult passages can also help a learner tackle more difficult texts.
- ***Revision*** typically consists of a set of questions that the learner answers once they have completed reading a text. The questions test the learner's comprehension of the text, as well as their knowledge of vocabulary and grammar seen in the text. The more involved the learner is with the language, the more they will remember. Therefore, exercises based on the reading will improve language retention (Laufer 2000).

Applications vary in their support across these categories. In Section 5 we present a range of applications and discuss them in relation to the categories of support and their applicability to language acquisition via extensive reading.

## 4. Applications

We discuss various existing and proposed applications that provide support of the types described in our taxonomy.

### *4.1 The Readable Document Search Engine*

An idea that was independently developed by several research groups, and is now a part of major search engines is to make readability a criterion for search for documents. The most developed and used system that retrieves documents for foreign language reading practice is the REAP system (Heilman et al. 2010). It allows instructors to choose target vocabulary to be studied by students, permits students to select broad interest areas, and presents the students with recommended texts of an appropriate reading level that provide practice in the target vocabulary. REAP determines the student's existing vocabulary, so that it can better estimate the reading level required. While it doesn't appear to provide preparation activities, it does provide comprehensive translation support by making any word searchable in an on-line dictionary, and target words for study are hyperlinked to a definition. Revision consists of a practice exercises on target words. The system has been shown to improve language knowledge.

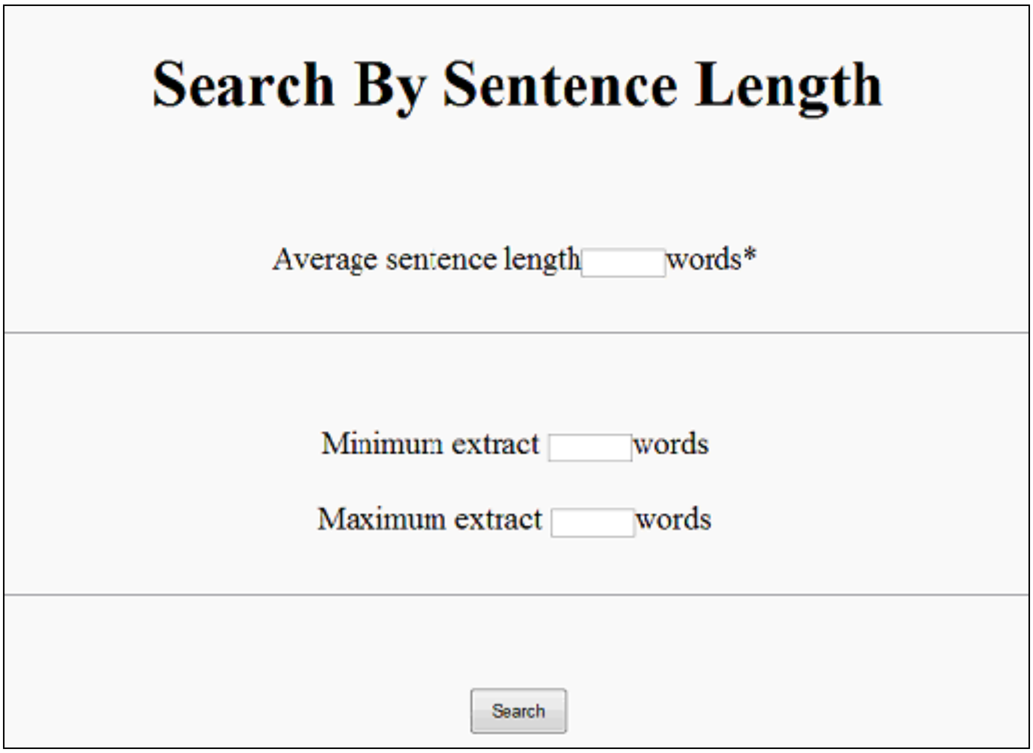
Where systems recommend native texts and also track a student's vocabulary knowledge, a personalised set of preparation material could be provided. For example, based on the word frequencies in the document and the known vocabulary, a set of 5-10 words that would improve the student's ability to understand the document the most, could be presented in a small pre-reading lesson. A simple but not necessarily optimal approach would be to select the highest frequency document words that are not in the student's current known vocabulary. These are likely to be "topic" words. For example, in a story about pirates, the words "pirate", "sail", "mast", "anchor" and "cabin" may occur frequently, but be generally unknown by a student studying the English Language. Where personalised vocabulary tracking doesn't exist, the preparation material can be selected purely by the frequency of words in the document versus their frequency in background text. Some published reading books for foreign language learners do this to some extent.

### *4.2 The Readable Extract Search Engine*

The idea behind the readable extract search engine is that a collection of native texts that are of high quality and interest, such as a collection of literature, may provide short extracts for reading. In an initial exploration we found it was possible to locate short readable extracts based on both vocabulary and grammar constraints (Uitdenbogerd, 2010). In our prototype we used sentence length as the readability measure, which has been shown to work very well for the case of French as a foreign language for English speakers (Uitdenbogerd, 2006). However, a more sophisticated readability measure, or one that is appropriate for a different language can be substituted.

Users set the sentence length criteria and the ideal extract size (Figure 1), and then retrieve a list of results (Figure 2). Users can then choose which extract is of interest from a list of titles and extract lengths.

In our prototype implementation, the extract is shown highlighted within the full text, allowing users to continue reading if they so choose (Figure 3). While the usefulness of short extracts hasn't been demonstrated as yet, extracts that provide multiple examples of a particular word or phrase has increased vocabulary knowledge (Webb, 2007).



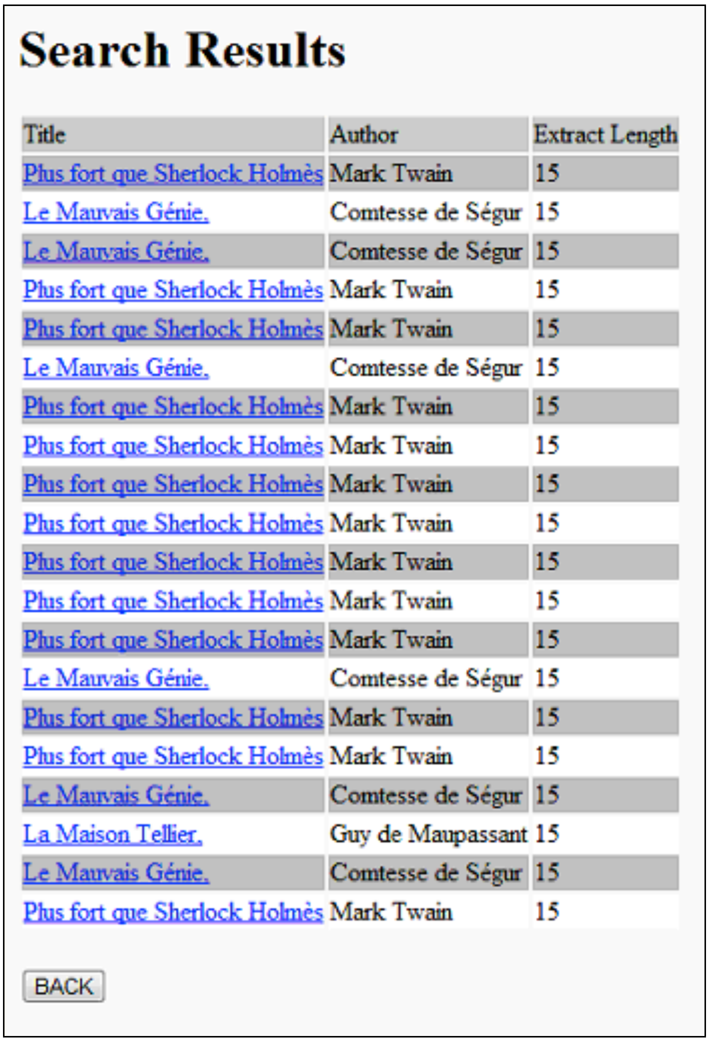
## Search By Sentence Length

Average sentence length ____ words*

Minimum extract ____ words

Maximum extract ____ words

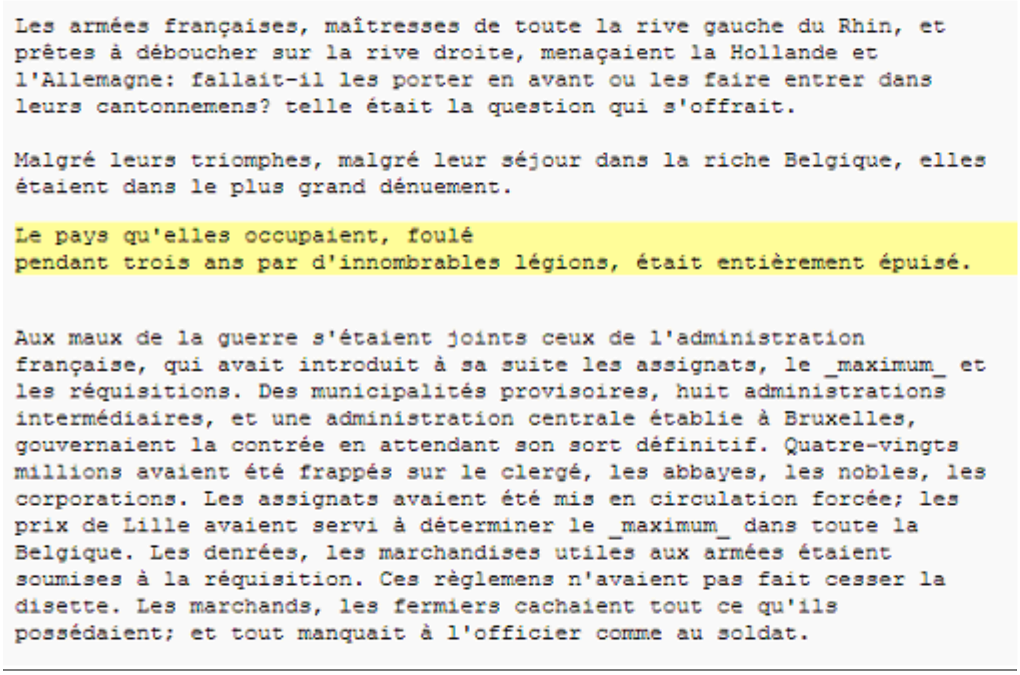
Figure 1. The Readable Extract Search Engine splash screen.



## Search Results

| Title | Author | Extract Length |
|---|---|---|
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Le Mauvais Génie. | Comtesse de Ségur | 15 |
| Le Mauvais Génie. | Comtesse de Ségur | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Le Mauvais Génie. | Comtesse de Ségur | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Le Mauvais Génie. | Comtesse de Ségur | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |
| Le Mauvais Génie. | Comtesse de Ségur | 15 |
| La Maison Tellier. | Guy de Maupassant | 15 |
| Le Mauvais Génie. | Comtesse de Ségur | 15 |
| Plus fort que Sherlock Holmès | Mark Twain | 15 |

Figure 2. Retrieved list of extracts.



```
Les armées françaises, maîtresses de toute la rive gauche du Rhin, et
prêtes à déboucher sur la rive droite, menaçaient la Hollande et
l'Allemagne: fallait-il les porter en avant ou les faire entrer dans
leurs cantonnemens? telle était la question qui s'offrait.

Malgré leurs triomphes, malgré leur séjour dans la riche Belgique, elles
étaient dans le plus grand dénuement.

Le pays qu'elles occupaient, foulé
pendant trois ans par d'innombrables légions, était entièrement épuisé.


Aux maux de la guerre s'étaient joints ceux de l'administration
française, qui avait introduit à sa suite les assignats, le _maximum_ et
les réquisitions. Des municipalités provisoires, huit administrations
intermédiaires, et une administration centrale établie à Bruxelles,
gouvernaient la contrée en attendant son sort définitif. Quatre-vingts
millions avaient été frappés sur le clergé, les abbayes, les nobles, les
corporations. Les assignats avaient été mis en circulation forcée; les
prix de Lille avaient servi à déterminer le _maximum_ dans toute la
Belgique. Les denrées, les marchandises utiles aux armées étaient
soumises à la réquisition. Ces règlemens n'avaient pas fait cesser la
disette. Les marchands, les fermiers cachaient tout ce qu'ils
possédaient; et tout manquait à l'officier comme au soldat.
```

Figure 3. Highlighted extract in main body of text.

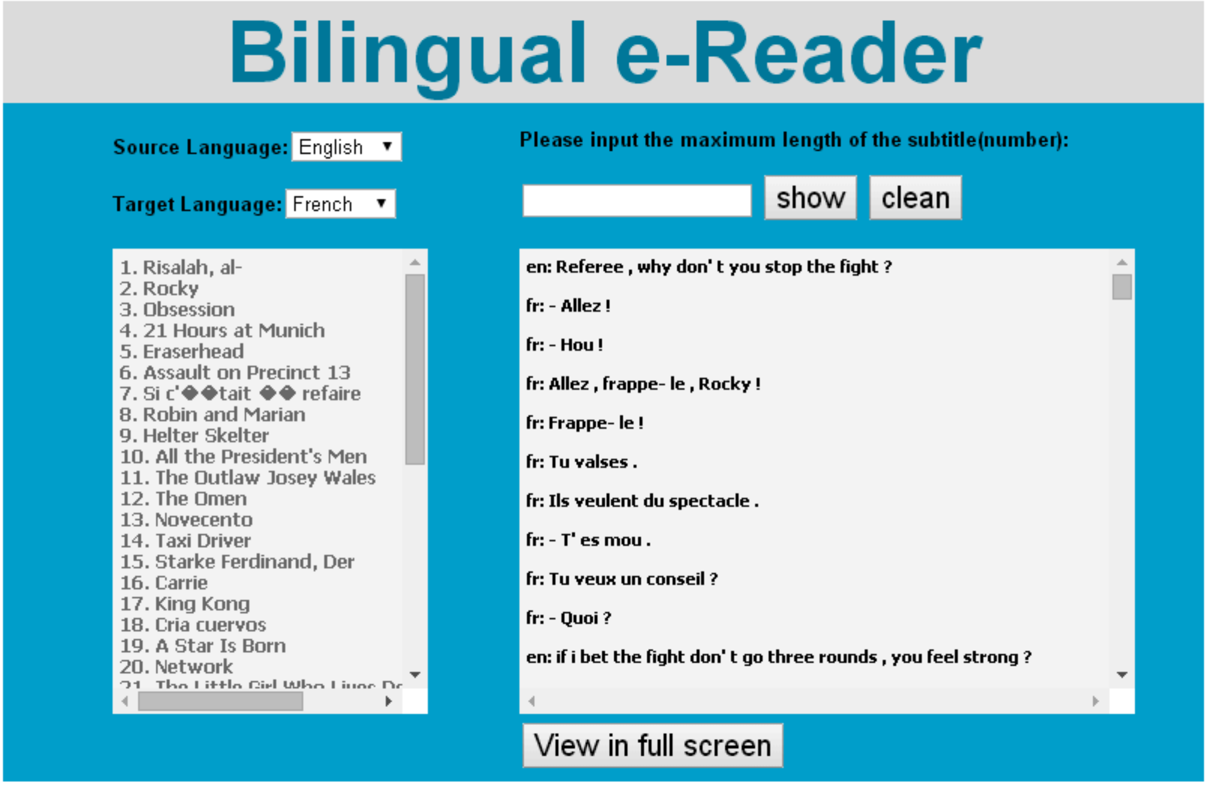
### *4.3 The Bilingual e-Reader*

The bilingual e-Reader combines simplification with translation, in that the more difficult sentences are presented in the learner's native language. The idea of mixed language reading material is not new. The approach has been used for gradually increasing the amount of target language in a story for English-German with good results compared to a normal German lesson (Weible, 1980), and has also been used to introduce Japanese kanji within an English story (Watanabe, 2002).

We developed a simple prototype using sentence length as the readability criterion, with a collection of movie subtitles (downloaded from http://opus.lingfil.uu.se/OpenSubtitles.php) as the corpus.

Figure 4 shows the main screen, including parameters for source and target language, a list of movie subtitles available, a readability parameter (maximum sentence length), and the resulting mixed language subtitles for the movie Rocky with the maximum sentence length set to 7.

The idea is a little controversial in the sense that it prevents total immersion in the target language, but it can be useful for earlier stages of language learning. Later stages would benefit from sentences that translate when clicked on. This way translations only are shown when requested.



Source Language: English

Target Language: French

Please input the maximum length of the subtitle(number):

1. Risalah, al-
2. Rocky
3. Obsession
4. 21 Hours at Munich
5. Eraserhead
6. Assault on Precinct 13
7. Si c'��tait �� refaire
8. Robin and Marian
9. Helter Skelter
10. All the President's Men
11. The Outlaw Josey Wales
12. The Omen
13. Novecento
14. Taxi Driver
15. Starke Ferdinand, Der
16. Carrie
17. King Kong
18. Cria cuervos
19. A Star Is Born
20. Network
21. The Little Girl Who Lives Do

en: Referee , why don' t you stop the fight ?

fr: - Allez !

fr: - Hou !

fr: Allez , frappe- le , Rocky !

fr: Frappe- le !

fr: Tu valses .

fr: Ils veulent du spectacle .

fr: - T' es mou .

fr: Tu veux un conseil ?

fr: - Quoi ?

en: if i bet the fight don' t go three rounds , you feel strong ?

Figure 4. The Bilingual e-Reader Search Interface.

## *4.4 The Book Bootstrapper*

This application re-orders the text of a book into readability order. This is a similar idea to the Textladder application that sorts a collection of documents into readability order (Ghadirian, 2002), but is done with a single large text. Fiction books with significant amounts of dialogue have many simple sentences and considerable variation in readability across the text. For example, the novel *Emma* by Jane Austen has 31 occurrences of the sentence "Oh!", 6 of the sentence "Ah!", and many sentences consisting of a single-word name. Its average sentence length is approximately 21.

If it is true that "meeting" a word 10 times during reading allows one to learn its meaning, then reading a novel of about 85,000 words should provide a learner with a 1,000 word reading vocabulary (Waring, 2009). However, the learner must be able to comprehend enough of the text to gain vocabulary from it. Ordering the text on readability criteria may allow vocabulary to be acquired more comfortably than reading it in the normal order. The learner can then re-read the book in the normal order with greater ease. We have developed a prototype book bootstrapper, and are currently determining its range of usefulness.

## *4.5 The Text Simplifier*

An alternative to the careful selection of text, or the efforts of writing simple text, is the automatic text simplifier. Using similar techniques to document summarisation in addition to word substitution, the text simplifier both shortens a given text and reduces the vocabulary and grammar difficulty. The usefulness of simplification over translation was demonstrated by Eskenazi, Lin and Saz (2013).

## *4.6 The Text Generator*

For extensive reading practice, the text generator should generate interesting stories on a small vocabulary and grammar repertoire. Shim and Kim (2002) developed a story generator that uses autonomous agents. To our knowledge, this idea hasn't been applied to foreign language learning yet. While not exclusively reading-based, chat-bots can also provide engaging reading practice.

## 5. Conclusion and Future Work

Reading extensively in the target language improves language skills, and this happens most efficiently when the reading material is both interesting and at a suitable difficulty level. Therefore systems that can provide appropriate reading material in sufficient quantities will be of great benefit to the language learner. We identified seven areas of support that applications can provide for extensive reading: management, generation of readable text, selection of readable text, simplification of text, learner preparation, glossary or translation support, and revision.

We described simple prototype systems that allow the user to locate suitable reading material, either as extracts, or as mixed language texts, exploiting the variability in language difficulty across a typical native text. Other systems were described that are either already in existence or may be worthwhile additions to the range of tools for the language learner. We are currently determining the applicability of one of these: the book bootstrapper. Future work will include determining the effectiveness of the proposed applications for language learning.

## Acknowledgements

Software development for the: Readable Extract prototype was by Christopher Dore, Len Wang and XinYang Yao; and the Bilingual eReader was by Wei Yan, Yueyao Zhuang and Ziteng Zhang. The Book Bootstrapper is currently being developed by Daniel Collins.

The Bilingual eReader concept was originally discussed with Laurianne Sitbon, who suggested the use of movie subtitles.

Support for this project has been provided by the Australian Government Office for Learning and Teaching. The views in this project do not necessarily reflect the views of the Australian Government Office for Learning and Teaching.

## References

Eskenazi, M., Lin, Y., Saz, O., 2013, Tools for non-native readers: the case for translation and simplification, *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, NAACL2013, Atlanta, p.20-28

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

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenzi, M. (2010) Personalization of Reading passages Improves Vocabulary Acquisition. *International Journal of Artificial Intelligence in Education*, 20, 73-98.

Hirsh, D. & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, 8(2):689–696.

B. Laufer. (2000). Task effect on instructed vocabulary learning: the hypothesis of involvement. In *Selected Papers from AILA ’99* Tokyo, pages 47–62. Waseda University Press, Tokyo.

Liu, N. & Nation, I.S.P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, 16(1), 33-42.

Lomicka, L.L. (1998). To gloss or not to gloss: an investigation of reading comprehension online, *Language Learning and Technology*, 1(2), 41-50.

Shim, Y. & Kim, M. (2002). Automatic Short Story Generator Based on Autonomous Agents. *Intelligent Agents and Multi-Agent Systems*. Lecture Notes in Computer Science, 2413, pp 151-162

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.

Uitdenbogerd, A.L. (2010). Fun with Filtering French. *Australasian Language Technology Association Workshop*.

Waring, R. (2009). The inescapable case for extensive reading. *Extensive Reading in English Language Teaching*, Andrzej Cirocki (ed.).

Waring, R. & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, 15(2).

Watanabe, H. (2002). Coco & the Gold Flute. *Kanji Dojo*, North Balwyn, Vic. Australia.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, 28(1), 46-65.

Weible, D.M. (1980). Teaching reading skills through linguistic redundancy. *Foreign Language Annals*, 6, 487-493.

Zheng, Y., Cheng, L. (2008), Test review: College English Test (CET) in China. *Language Testing*, 25(3), 408-417.