Web Readability and Computer-Assisted Language Learning

Alexandra L. Uitdenbogerd
School of Computer Science and Information Technology
RMIT University
Melbourne, Australia

Abstract

Proficiency in a second language is of vital importance for many people. Today's access to corpora of text, including the Web, allows new techniques for improving language skill. Our project's aim is the development of techniques for presenting the user with suitable web text, to allow optimal language acquisition via reading. Some text found on the Web may be of a suitable level of difficulty but appropriate techniques need to be devised for locating it, as well as methods for rapid retrieval. Our experiments described here compare the range of difficulty of text found on the Web to that found in traditional hard-copy texts for English as a Second Language (ESL) learners, using standard readability measures. The results show that the ESL text readability range fall within the range for Web text. This suggests that an on-line text retrieval engine based on readability can be of use to language learners. However, web pages pose their own difficulty, since those with scores representing high readability are often of limited use. Therefore readability measurement techniques need to be modified for the Web domain.

1 Introduction

In an increasingly connected world, the need and desire for understanding other languages has also increased. Rote-learning and grammatical approaches have been shown to be less effective than communicative methods for developing skills in using language (Higgins, 1983; Howatt, 1984; Kellerman, 1981), therefore students who need to be able to read in the language can benefit greatly from extensively reading material at their level of skill (Bell, 2001). This reading material comes from a variety of sources: language learning textbooks, reading books with a specific level of vocabulary and grammar, native language texts, and on-line text.

There is considerable past work on measuring the readability of text, however, most of it was originally intended for grading of reading material for English-speaking school children. The bulk of readability formulae determined from these studies incorporate two main criteria for readability: grammatical difficulty — usually estimated by sentence length, and vocabulary difficulty, which is measured in a variety of ways (Klare, 1974). Publishers later decided to use the readability measures as a guideline for the writing of texts, with mixed success. However, new reading texts catering for foreign language learners of various languages are still being published. Most of these use specific vocabulary sizes as the main criterion for reading level. Others are based on specific language skills, such as the standard developed by the European community known as the "Common European Framework of Reference for Languages" (COE, 2003).

The goal of our research is to build an application that allows the user to improve their language skills through accessing appropriate reading material from the Web. This may incorporate personalised retrieval based on a user's level of skill in the target language, first language and specific vocabulary of interest. Greater detail about the application's requirements and potential implementation issues are discussed elsewhere (Uitdenbogerd, 2003).

In order for appropriate documents to be presented to the user for reading practice, new readability measurement techniques that are more appropriate to on-line documents will need to be developed. Measures of distance between languages that are related to reading may be useful for finer-tuned readability (as opposed to the speaking-based measure developed elsewhere (Chiswick, 2004)). For many language pairs, cognates — words that are similar in both languages, help people to understand text. There is some evidence that these affect text readability of French for English speakers (Uitdenbogerd, 2005). Automatic detection of cognates is also part of our research program. Some work exists on this topic (Kondrak, 2001), but will need to be tested as part of readability formulae for our application.

Some applications that allow the location or sorting of suitable on-line reading material already exist. One example is Textladder (Ghadirian, 2002), a program that allows the sorting of a set of texts based on their vocabulary, so that users will have learnt some of the words in earlier texts before tackling the most vocabulary-rich text in the set. However, often vocabulary is not the main criterion of difficulty (Si and Callan, 2001; Uitdenbogerd, 2003; Uitdenbogerd, 2005). SourceFinder (Katz and Bauer, 2001) locates materials of a suitable level of readability given a list of URLs. It is simply a crawler that accepts a web page of URLs such as those produced by Google, and then applies a readability measure to these to rank them. The software was developed with the aim of finding material of the right level of difficulty for school children learning in their native language.

Using English as a test case, the research questions we raise in this work are:

- What is the range of difficulty of text on the web?
- How does the range of text difficulty found on the web compare to texts especially written for language learners?

If there is overlap in the readability ranges between web documents and published ESL texts, then the combination of the two may be adequate for language learning through reading once learners are able to comfortably read published texts. In fact, we have found that ESL texts fit within the range of readability found on the Web, but that there are problems with assessing readability of the Web pages due to the types of structures found within them.

In future work we intend to develop readability formulae that take into account bulleted lists and headings. It is known from usability studies that these increase readability of text for native readers of technical documents (Redish, 2000; Schriver, 2000). We will then be in a position to better determine how the readability factors differ for people with different language backgrounds and skills within a Web context. We have already examined the case of French as a foreign language for those whose main language is English and found that standard readability formulae developed for native English speakers are less closely correlated to French reading skill than a simple sentence length measure (Uitdenbogerd, 2005). However, this work was based on prose and comic-book text samples, not HTML documents.

This article is structured as follows. We review the literature on language learning via reading, as well as describe past research on readability. We then describe our current work that examines the readability of English text on the Web. This is compared to the readability of reading books for students with English as a second language. These results are then discussed in the context of improving language skills via the Web.

2 BACKGROUND

Two main research areas are of relevance to the topic of computer-assisted language acquisition via reading: readability and language acquisition. Readability measures allow us to quickly evaluate the appropriateness of reading material, and language acquisition research informs us how best to use reading material in order to acquire language.

2.1 Readability

Readability has been studied for most of the twentieth century, and has more recently become a topic of interest to information retrieval researchers. There have been several phases in its development as a research topic. In the initial and most influential era, readability measures were developed by applying regression to data collected from children's comprehension tests. Later, Cloze tests were used as a simpler method of collecting human readability data (Bormuth, 1968; Davies, 1984). The output of this era included a vast array of formulae, mostly incorporating a component representing vocabulary difficulty, such as word length, as well as a grammatical difficulty component, which usually is represented by sentence length (Klare, 1974). The majority of published work was on English language readability for native speakers, however, some work from this era examined other languages, again in a native speaker context. More recently the language modelling approach has been applied to readability estimation of text (Si and Callan, 2001).

Despite the success of the techniques, they fell out of favour within some research and education communities due to their simplicity (Chall and Dale, 1995; Redish, 2000; Schriver, 2000) and failure to handle hand-picked counter-examples (Gordon, 1980). Other criticism was of their abuse in writing texts or in enforcing reading choices for children (Carter, 2000). Researchers tried to capture more complex aspects of readability such as the conceptual content. Dale and Chall, in response to the criticism, updated their formula to, not only use a more up-to-date vocabulary, but to allow conceptual content to be catered for. They emphasized however, that grammatical and vocabulary difficulty are still the dominant factors (Chall and Dale, 1995). In work on readability for English-speaking learners of French, we found further evidence that conceptual aspects are of minor importance compared to grammatical complexity (Uitdenbogerd, 2005). For example, the well-known fairy tale Cinderella was consistently perceived as more difficult than many unknown stories, due to the relative grammatical complexity.

The readability measures used in the experiments reported here are those implemented in the unix-based `style` utility. The measures used were the Kincaid formula, Automated Readability Index (ARI), Coleman-Liau Formula, Flesch reading ease, Fog index, Lix, and SMOG. Some readability formulae are listed below.

The ARI formula as calculated by `style` is:

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

where Wlen is the length of the word and WpS is the average number of words per sentence.

The Flesch formula for *reading ease* (RE) as described by Davies, is given as:

$$\begin{aligned} \mathrm{RE} = 206.835 \quad & -(0.846 \times \mathrm{NSYLL}) \\ & -(1.015 \times \mathrm{W/S}) \ , \end{aligned} \quad (2)$$

where NSYLL is the average number of syllables per 100 words and W/S is the average number of words per sentence (Davies, 1984).

The Dale-Chall formula (not used in our experiments) makes use of a vocabulary list in addition to sentence length:

$$S = 0.1579p + 0.0496s + 3.6365 \ , \quad (3)$$

where p is the percentage of words on the Dale list of 3,000, and s is the average number of words per sentence. The resulting score represents a reading grade.

The above formulae illustrate three ways of determining vocabulary difficulty: word length in characters, number of syllables, and membership of a list of words known by children with English as their native language. Most formulae use one of these techniques in addition to the sentence length.

More recent research into readability has involved the application of language models (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005). Using unigram models allowed very small samples of text to be used to predict a grade level for the text (Collins-Thompson and Callan, 2004). The technique was shown to be more robust than a traditional readability measure for estimating web page readability. However, the unigram approach is unlikely to be effective for the case of foreign languages, where grammatical complexity is a much more important factor than vocabulary for at least one language pair (Uitdenbogerd, 2005).

Schwarm and Ostendorf (2005) built a readability classifier that incorporated a wide variety of features, including traditional readability measure components, as well as n-gram models, parse-tree based features to model grammatical complexity, and features representing the percentage of unusual words. The classifier was trained and evaluated using articles written for specific grade levels. It is possible that the approach and feature set used may be applicable to foreign language learning.

2.2 Second and Foreign Language Acquisition via Reading

The idea of language acquisition via reading at an appropriate level was first formally studied by Michael West. He found that his techniques of English teaching with Bengali boys were far more successful than other approaches of the time (West, 1927). He controlled the introduction of vocabulary to no more than one new word per 60 words of text. The concept remains with us today and is known as “controlled-vocabulary”. Occasionally the reading approach falls out of favour and conversation becomes a more prominent technique. Then reading is rediscovered (Kellerman, 1981).

Studies of different ways of reading for language acquisition conclude that extensive reading at a comfortable level is superior to intensive reading at a more challenging level (Bell, 2001), and the use of glosses and multimedia improve vocabulary acquisition (Lomicka, 1998; Al-Seghayer, 2001). Looking up word meanings is more likely to lead to retention, but words can be learnt through repeated exposure and meaning inference. However, due to the need for repetition, inference is only useful for fairly common words (Krantz, 1991).

3 EXPERIMENTS

The main experiment that we discuss here is an analysis of a corpus of English web text. The corpus is a subset of the TREC web 10G collection consisting of 93,064 documents. The collection is a general snapshot of the web, including a wide variety of types of web pages.

We extracted the text and punctuation from each document in the corpus, and applied several standard readability measures to them, as implemented by the unix-based `style` utility. The measures used were the Kincaid formula, ARI, Coleman-Liau Formula, Flesch reading ease, Fog index, Lix, and SMOG.

In a second experiment we applied the same readability measures to extracts from reading books written for students of English as a second or foreign language. The statistics for the two sets of text were compared.

Results

The first part of Table 1 shows statistics describing the range of readability scores found in the collection. For the Flesch Index, the highest value represents the easiest to read, whereas for the other measures the lowest value is the easiest.

It is clear by looking at the extreme values that there are difficulties in processing web documents compared to normal text. In this section we look at the types of problem documents that are classified as very easy by a naïve application of readability formulae.

Documents with extreme scores

We examined several web documents that had extreme values for each readability measurement type.

All measures except Coleman-Liau agreed as to which was the hardest document in the collection — a large document listing access statistics of Internet domains. There were few true sentences in this document.

There were many documents (49) that had the minimum score of -3.4 using the *Kinkaid* measurement. On close inspection of a couple of these, we found they were devoid of punctuation, containing a few headings and links only. The same documents received the maximum (easiest) score of 121.2 in the *Flesch* reading ease measure.

The *Fog* measure also shared the same easiest documents, however, it also included other documents amongst those with its lowest score of 0.4. An example that was in this set of extra documents was a page of links to images, with duration times listed next to the image. The only terminating punctuation was in an email address. The *Lix* measure had a similar but not identical set of 48 documents receiving its lowest score of 1.

Two documents received the lowest value -12.1 using the *ARI* measure. In the first, the only untagged text was within `title` tags: “V.I.She: Pharmacy”. The second document contained the same title and some labelled links without punctuation.

The lowest value using the *Coleman-Liau* measure was associated with a short document in which most of the words had their letters interspersed with spaces, for example “C H A N G I N G”. The second lowest consisted of a heading, links to images, with their sizes, such as “127.1 KB” shown next to them, and a single sentence.

The SMOG score was less discriminating, giving 2,967 documents the same lowest score of 3.

Published Reading Books

The second part of Table 1 shows the readability of nine published ESL books. Interestingly, the readability results bear little resemblance to the levels advertised by the publishers. For example,

Table 1: Distribution of readability scores in the Web collection and of a range of books written for learners of ESL. For all measures except Flesch, the highest value represents the most difficult text to read. The ESL book measured as most and least difficult are indicated with a † and an asterisk ('*') symbol respectively.

Quartile	Kinkaid	ARI	Coleman-Liau	Flesch	Fog	Lix	SMOG
Min	-3.4	-12.1	-8.3	-62809.2	0.4	1	3
LQ	6.4	7.4	11.1	46.2	9.4	37	8.9
Med	9.1	10.7	13.2	60.8	12.3	46	10.9
UQ	12.2	14.3	15.6	73.4	15.8	55.3	13.2
Max	24174.6	30988	130.4	121.2	24798	61997.5	219.6
Average	11.809	14.20	13.4176	54.09	15.13571	52.1665	11.28505
Book	Kinkaid	ARI	Coleman-Liau	Flesch	Fog	Lix	SMOG
card	5.3	†6.1	†9.0	†84.7	†8.2	28.6	†7.8
christmas	3.5	2.9	8.2	88.5	5.8	22.1	6.6
dead	1.1	-0.0	6.4	100.6	3.8	16.6	5.2
ghost	3.4	3.1	7.6	91.3	6.1	24.0	6.5
lovely	2.5	2.7	7.5	96.9	5.0	21.5	5.2
murders	†5.4	5.9	7.7	86.7	7.8	28.7	6.5
presidents	*0.2	-0.1	6.2	*107.0	3.2	14.0	*4.2
simon	1.3	*-0.9	*5.9	97.0	*3.1	*12.6	4.7
thirty	5.2	5.3	7.2	87.7	7.9	†28.9	7.0
Min	0.2	-0.9	5.9	84.7	3.1	12.6	4.2
Max	5.4	6.1	9.0	107.0	8.2	28.9	7.8

Table 2: Web pages out of $93,064$ with readability scores within the range of sample ESL texts.

	Kinkaid	ARI	Coleman-Liau	Flesch	Fog	Lix	SMOG
Count	16123	17321	7863	8176	14748	8166	11344
Percent	17	19	8	9	16	9	12

The Card is described as a level 3 story with 1,000 headwords, making it in the middle of the range of 5 levels of difficulty. However, five of the readability measures identified it as the most difficult book of the set. In contrast, Simon the Spy and The President's Murder are both identified as easy texts, which is in agreement with the advertised beginner level of these stories.

When the levels are compared to those of the analysed web pages, it is clear that the ranges fall well within the extremes found on the web. However, as we have already seen, these extremes are often pathological cases, and not usually of interest for reading practice. As a percentage, the set of suitable texts for those that require the reading level found in ESL books, is probably quite small, given that the lower quartiles of web readability exceed the maximum scores for the range of books tested. In fact, depending on the reference readability measure, the percentage of web texts falling within the same range of readability is in the range 8 to 19% (See Table 2 for details).

In Figure 1 we show a few examples of web text that fall in the range of readability found in the ESL texts. These examples illustrate a few types of content found in web pages: links and message headers.

Discussion

While there is a wide range of values for the readability measures in the web collection studied, a very large proportion of documents with low scores are arguably not very useful for reading practice. The `style` utility assumes that the input consists of normal text in the form of sentences. If these are not found, or if there are too many non-sentences in the document, then the utility fails. In addition, documents that do contain sufficient text may still consist largely of headings, links, and lists. It is unclear how useful these documents would be for reading practice.

For a reading recommender to be successful, further criteria than just a readability score will be needed. Some preprocessing of the documents for better readability assessment may be necessary. It was observed in the documents receiving low scores that there were sentences without punctuation. Web authors often include instructions without punctuation, both within links and in normal displayed text. Some examples found in the low-scoring documents are “Click on books”, “To view Shockedmovies, you need to have Netscape 2.0 and the Shockwave plug-in”, “Last modified on December 10, 1995”, and “Update Your Profile”. Inserting punctuation may make readability scores more reliable, however, automatic techniques for deciding when to do so are not completely obvious. As mentioned in the introduction, readability measures that take into consideration the use of bulleted lists and headings would be of utility for web page assessment, since these structures are frequently used in web pages, and often are the only textual content within a page. Collins-Thompson and Callan’s approach avoids this issue by using unigram word models exclusively to measure readability (Collins-Thompson and Callan, 2004). However, for the bilingual case, particularly in language pairs such as French and English, this is likely to be ineffective (Uitdenbogerd, 2005).

An alternative approach is to filter the pool of web pages to be analysed, either by crawling suitable subdomains, or by applying a set of rules to ensure sufficient suitable text on a page before inclusion.

Another important consideration is how interesting the document will be to the user, as a person’s comprehension skills vary with their interest in the text. Indeed, the documents should be sufficiently interesting for users to want to use the proposed system. An existing technique for increasing the chance of interesting documents being presented to the user is collaborative filtering, which relies on user feedback, whether explicit or implicit. Another possibility involves the examination of content words and phrases within documents.

4 CONCLUSIONS

The purpose of this preliminary work towards a utility for assisting users in improving foreign language skills via reading, was to find evidence that sufficient documents of suitable readability are likely to exist on the web. We determined the readability of over 90,000 web pages written in English, using the unix `style` utility and found a considerable range of readability scores. The range of readability scores found in ESL books fell within the lower quartile of web page readability scores, representing 8 to 19% of documents in the collection. This could mean that there are many suitable pages for reading practice which a readability-based reading recommender system could retrieve for users. However, due to the artifacts of web pages and the readability measures, not all pages with low scores in readability are suitable for reading practice. The automated location of those that *are* suitable is part of the future research plans of this project. An additional factor that must be incorporated is prediction of how interesting the documents are likely to be for users.

Our analysis used web pages written in English and compared these to ESL texts under the broad assumption that similar distributions of readability would occur in other languages. However, cultural and political differences of the countries speaking different languages may influence the types of text available, and hence the readability range.

Learners of English are relatively fortunate in that there are many reading books specifically written for them. This is not the case for many other languages. It is possible that the Internet may be an even more important reading resource for languages other than English.

Acknowledgements

We thank Vasundhara Srikantha for her preliminary work on web text analysis.

References

K. Al-Seghayer. 2001. The effect of multimedia annotation modes on L2 vocabulary acquisition: a comparative study. *Language Learning and Technology*, 5(1):202–232, January.

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

J. R. Bormuth. 1968. Cloze test readability: criterion reference scores. *Journal of Educational Measurement*, 5:189–196.

B. Carter. 2000. Formula for failure. *School Library Journal*, 46(7):34–37, July.

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

B. R. Chiswick. 2004. Linguistic distance: A quantitative measure of the distance between English and other languages. Technical Report 1246, Institute for the Study of Labor (IZA).

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.

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

A. Davies. 1984. Simple, simplified and simplification: what is authentic? Applied linguistics and language study, chapter 9, pages 181–198. Longman.

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

R. M. Gordon. 1980. The readability of unreadable text. *English Journal*, pages 60–61, March.

J. Higgins. 1983. Can computers teach? *Calico Journal*, pages 4–6, September.

A. P. R. Howatt. 1984. *A history of English language teaching*. Oxford University Press, Oxford, UK.

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

M. Kellerman. 1981. *The forgotten third skill*. Pergamon Press, Oxford.

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

G. Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In K. Knight, 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, June. NAACL.

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

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

J. Redish. 2000. Readability formulas have even more limitations than Klare discusses. *Journal of Computer Documentation*, 24(3):132–137, August.

K. A. Schriver. 2000. Readability formulas in the new millennium: What's the use? *Journal of Computer Documentation*, 24(3):138–140.

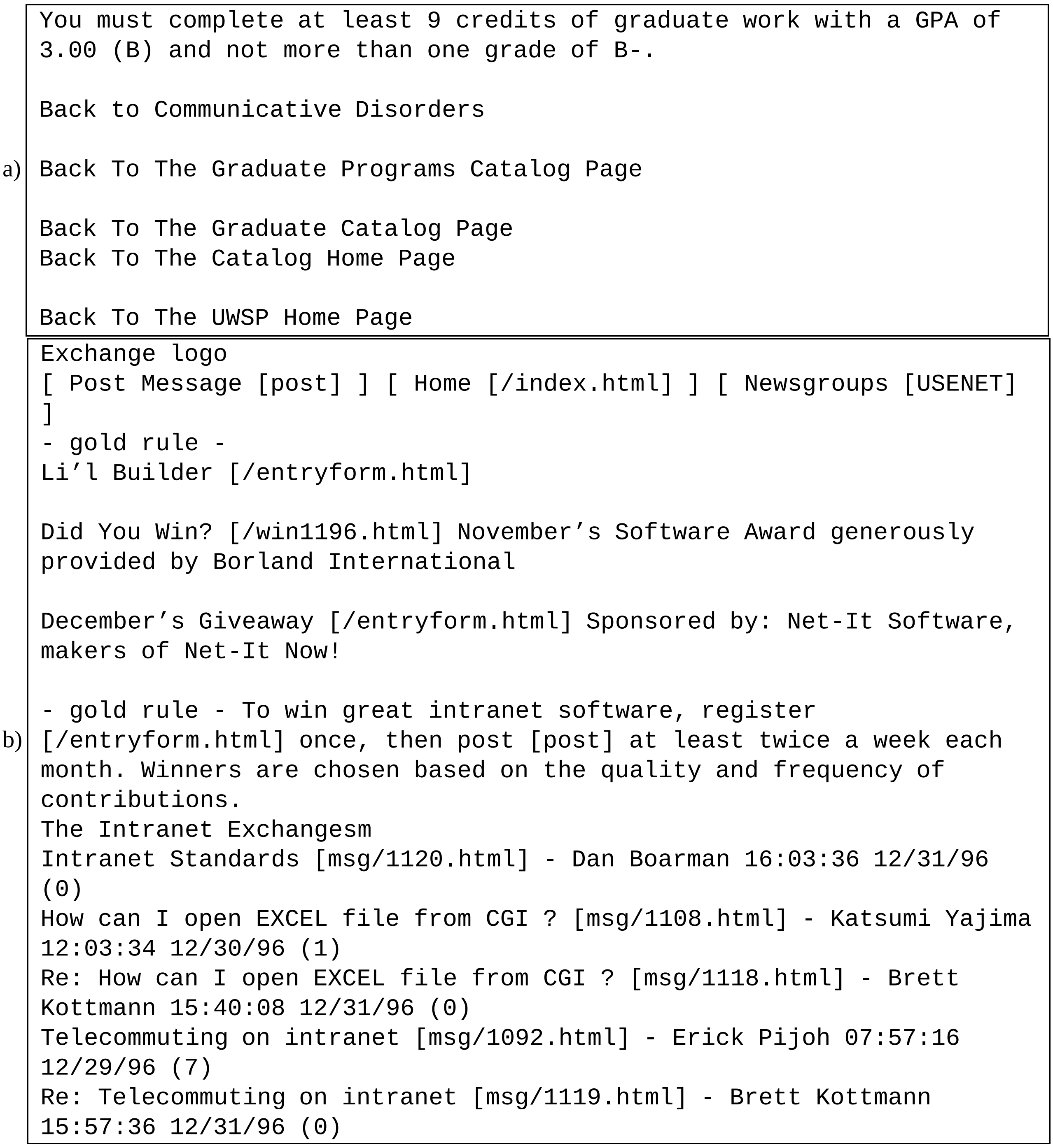
S. E. Schwarm and M. Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the Association for Computational Linguistics*.

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

A. L. Uitdenbogerd. 2003. 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. Springer.

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

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



a)

```
You must complete at least 9 credits of graduate work with a GPA of
3.00 (B) and not more than one grade of B-.

Back to Communicative Disorders

Back To The Graduate Programs Catalog Page

Back To The Graduate Catalog Page
Back To The Catalog Home Page

Back To The UWSP Home Page
```

b)

```
Exchange logo
[ Post Message [post] ] [ Home [/index.html] ] [ Newsgroups [USENET]
]
- gold rule -
Li’l Builder [/entryform.html]

Did You Win? [/win1196.html] November’s Software Award generously
provided by Borland International

December’s Giveaway [/entryform.html] Sponsored by: Net-It Software,
makers of Net-It Now!

- gold rule - To win great intranet software, register
[/entryform.html] once, then post [post] at least twice a week each
month. Winners are chosen based on the quality and frequency of
contributions.
The Intranet Exchangesm
Intranet Standards [msg/1120.html] - Dan Boarman 16:03:36 12/31/96
(0)
How can I open EXCEL file from CGI ? [msg/1108.html] - Katsumi Yajima
12:03:34 12/30/96 (1)
Re: How can I open EXCEL file from CGI ? [msg/1118.html] - Brett
Kottmann 15:40:08 12/31/96 (0)
Telecommuting on intranet [msg/1092.html] - Erick Pijoh 07:57:16
12/29/96 (7)
Re: Telecommuting on intranet [msg/1119.html] - Brett Kottmann
15:57:36 12/31/96 (0)
```

Figure 1: Sample Web Documents with Readability Matching Typical ESL Texts. Both the above documents received a score of 5.9 on the Coleman-Liau readability measure, thus equivalent to the easiest ESL texts in our study. Item a) shows the complete text extracted from the document. Item b) is an extract of the document with the largest number of words and a score of 5.9.