Seeking needles in the Web haystack: Finding texts suitable for language learners

Serge Sharoff, Svitlana Kurella, Anthony Hartley
Centre for Translation Studies, University of Leeds, UK

Abstract. While modern communicative methods of language teaching rely heavily on authentic, typical and recent materials, traditional graded readers often fall short of these requirements. The project reported in this paper is aimed at (1) designing methods for retrieving web texts that are suitable for a particular group of learners, and (2) using them in actual language teaching. Given that very little is currently known about the features that constitute texts suitable for individual language learning needs in a variety of languages, the paper reviews options for selecting texts according to their lexicon, grammatical features and readability statistics on the basis of trial runs with students studying English, Chinese, German and Russian. Among other sources, we used simplified Wikipedia texts (http://simple.wikipedia.org/) and their counterparts in the main English wikipedia. In addition to selecting texts for reading exercises, we experimented with the same model applied to selecting texts suitable for grammatical gap-fill exercises. For instance, it was used for finding texts rich in modal verbs or conjunctions. The use of authentic running texts instead of artificial single-sentence examples improves students' motivation in such exercises and helps in contextualising grammatical rules.

Keywords: graded readers, LSP, multilingual resources, reading skills, text selection

Introduction

While modern communicative methods of language teaching rely heavily on reading materials that are authentic, typical and recent, traditional collections graded for reading difficulty often fall short of meeting these requirements. Graded readers for some languages, e.g., Mandarin Chinese, frequently contain adaptations of authentic texts that restrict the range of characters, simplify idiomatic expressions, rewrite syntactic constructions not covered in text books, etc. Such texts have their pedagogical value, but they do little to prepare students for reading 'real' texts. As for the typicality of genres, texts in graded readers often have a large proportion of classic literary texts – e.g., Goethe and Schiller for German, or Pushkin and Tolstoy for Russian – while texts more relevant to accomplishing everyday tasks, such as administrative or argumentative texts, are missing. Finally, for domains undergoing rapid changes, such as software or international trade, all languages, including English, are seriously lacking in up-to-date reading resources oriented to language learners.

Nowadays it is relatively easy to collect a large corpus from the Web, using either search engines (Sharoff, 2006) or web crawlers (Baroni and Kilgarriff, 2006). Modern text classification methods help in selecting subsets of web corpora belonging to particular domains or genres (Sharoff, 2007). It is also possible to use entire domain- or genre-specific collections, e.g., Wikipedia or Reuters. However, the content of what is collected is often beyond the reading skills of language learners. Authentic corpora are commonly regarded as a useful resource for tailoring reading materials to what is expected by the language learner (Leech, 1997), but as far as we know there have been no studies that have actually

implemented an automatic text selection procedure for a variety of languages and put it into teaching practice.

Existing research on automatic text selection is mostly devoted to teaching English reading skills in the context of US school education (Schwarm, Ostendorf, 2005; Collins-Thompson, Callan, 2004). Traditional readability measures, such as Flesch Reading Ease, Flesch-Kincaid and Gunning Fog have also been deployed in this context (DuBay, 2004). Some recent projects (Heilman *et al.,* 2008; Kilgarriff *et al*., 2008; Kotani *et al*., 2008) do address the problems of selecting texts (or examples) aimed at non-native speakers of English. The grading procedure is typically based on lexical coverage or frequent words, e.g. words such as *essay* are selected as predictive for higher-grade texts (Collins-Thompson, Callan, 2004). When grammatical features are used, these are based on advanced English parsers (Kotani *et al*., 2008). However, it is difficult to extend such approaches to languages other than English, since adequate parsers are rarely available and, if they are, their sets of features usually differ to such a degree that a new approach would be needed for each language. Moreover, these projects provide little information on the actual use of graded texts.

The project reported in this paper is aimed at (1) designing methods for retrieving web texts that are suitable for a particular group of learners for a variety of languages, and (2) using them in actual language teaching. The emphasis of this study was on (1) finding measures that work across a variety of languages without requiring complex resources, such as parsers, and on (2) their integration into the language learning process.

Features for text selection

In consultation with language teachers we identified a variety of features that might be expected to make a text difficult to read, and tested for their effectiveness in predicting the difficulty of each text:

1. lexical coverage by word bands from respective frequency lists, i.e. top 1000, top 2000, top 3000 words; the General Service List (GSL) was used for English, frequency lists from Web corpora (Sharoff, 2006) for Chinese, German and Russian
2. average sentence length (ASL)
3. average word length in syllables (ASW) – pinyin count was used for Chinese
4. Flesch Reading Ease (FRE)
5. coverage by more frequent part of speech (POS) trigrams
6. average number of conjunctions per sentence
7. average number of lexical verbs per sentence
8. average number of passive verbs per sentence
9. average number of modal verbs per sentence
10. average number of prepositions per sentence
11. average number of punctuation marks per sentence

Since reliable part of speech taggers are available for all languages under consideration, it is possible to use them to detect known grammatical complexities without the need for full

parsers. Identifying frequent POS trigrams is an easy way to determine whether a text deviates from the standard language model, cf. the importance of language modeling in (Schwarm, Ostendorf, 2005) and (Kilgarriff *et al*., 2008). The number of lexical verbs is an indirect measure of sentence complexity, while passives and modals present well-known problems for language learners. The use of conjunctions is a way of detecting the complexity of discourse development. Kotani *et al*. (2008) assessed discourse complexity by the number of pronouns, but reported that this feature does not reduce the error rate. We used FRE (with language-specific formulas for German and Russian) to test for any correlation between the traditional readability measures mentioned earlier and our approach:

FRE_en = 206.835 - (1.015 x ASL) - (84.6 x ASW)

FRE_de = 180 - ASL - (58.5 x ASW)[1]

FRE_ru = 206.835 - (1.3 x ASL) - (60.1 x ASW)[2]

Machine Learning experiments

We calculated statistics for a range of "easy" texts from the Simple English Wikipedia website (http://simple.wikipedia.org/), and their counterparts from the main English Wikipedia website. Guidelines for contributors to the Simple English Wikipedia advise the use of Basic English vocabulary, the active voice and 'Basic English verb[s] in past, present or future only'. Further texts were added from other sources; s-sherlock is a simplified version of 'The Boscombe Valley Mystery' as published in the Penguin Readers series (sherlock is the original text). The resulting file was processed by the Weka implementation of Principal Component Analysis (PCA) to identify the most significant correlation between the features. It resulted in two main components representing a linear combination of normalised original features:

0.415prepositions+0.386lexverbs-0.352fre+0.334passiveverbs+0.32 top3000...

-0.416top2000-0.412top3000-0.41top1000+0.375punctuation+0.36 conjunctions...

These linear combinations can be roughly labelled, respectively, as the grammatical and lexical dimensions of difficulty (ranging from easy to difficult). Figure 1 displays a sample of texts in terms of the two dimensions. The majority of Simple Wikipedia texts (prefix s- in the chart) are easier on both dimensions than their main Wikipedia counterparts. For example, the simplified and standard definitions of 'induced abortion' read:

1.a An induced abortion is when a person does something to end the pregnancy.

1.b Induced abortion is the removal or expulsion of an embryo or fetus by medical, surgical, or other means at any point during human pregnancy for therapeutic or elective reasons.

In a few cases, the PCA results suggest that 'simplified' texts are not necessarily easier along both dimensions. For example, s-aesop is classified as grammatically easier yet more difficult lexically, a decision supported by the following extracts:

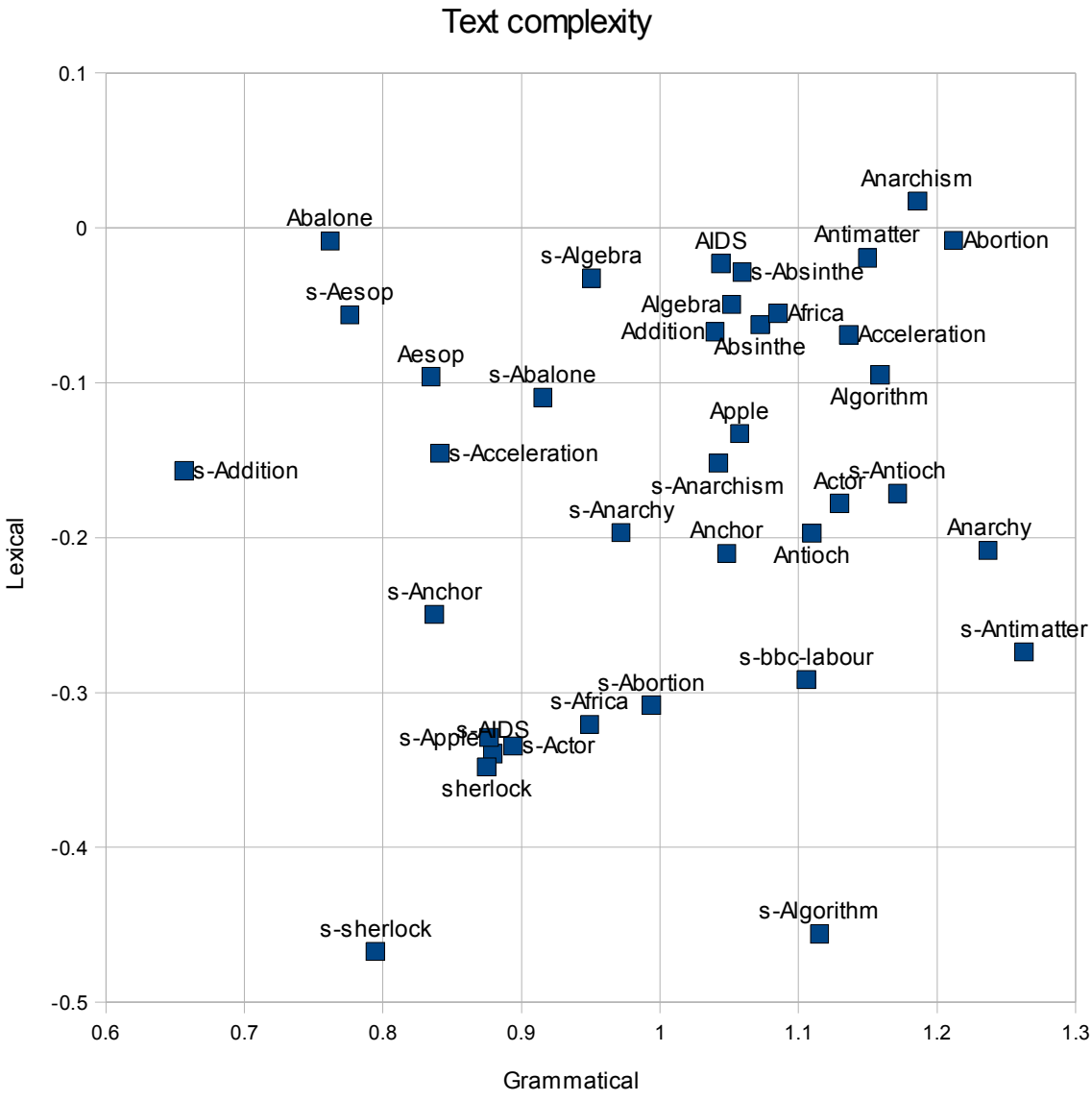
2.a Aesop's fables are still taught as moral lessons and used as subjects for various entertainments, especially children's plays and cartoons.

1http://de.wikipedia.org/wiki/Lesbarkeitsindex

2http://ru.wikipedia.org/wiki/Индекс_читабельности

2.b The various collections that go under the rubric “Aesop’s Fables” are still taught as moral lessons and used as subjects for various entertainments, especially children’s plays and cartoons.

In a small number of cases, the PCA results suggest that the ‘simplified’ text is in fact more difficult along both the lexical and the grammatical dimension. However, closer inspection of the texts in question indicates that the classifier has in fact detected true complexity that is belied by the ‘simplified’ label. For example, the difficulty of s-absinthe is shown to be slightly greater than that of absinthe in the main English Wikipedia. This has been indeed acknowledged by its readers, who have added the following tag to s-absinthe: ‘The English used in this article may not be easy for everybody to understand.’ Similarly, inspection of s-antioch reveals that it has been edited from antioch by simple deletion. None of the complexity of the main entry has been modified and, indeed, some of the deleted text is grammatically simpler than the retained text, even if its content has been judged to be not worth preserving.



We conducted a similar exercise for Chinese and Russian, this time selecting original texts published in quality newspapers for comparison with texts published on the Chinese and Russian BBC websites and considered by language teachers to be significantly easier for native English students. Their PCA transforms yielded the following combinations of parameters:

Chinese:

0.522asl+0.475lexverbs+0.422prepositions-0.388fre+0.387conjunctions...

-0.495top500-0.494top1000-0.482top2000+0.379asw-0.324fre...

Russian:

0.461asl-0.444fre+0.433lexverbs+0.332conjunctions+0.316asw...

-0.556top1000-0.55top500-0.533top2000-0.238conjunctions-0.135asw...

This supports the intuition of language teachers that texts with a higher number of less frequent words, conjunctions, prepositions and longer sentences tend to be more difficult for language learners.

Language teaching experiment

To ground our research in the practice of language teaching we selected some texts classified as moderately difficult according to the dimensions reported above. These texts were presented as reading exercises to language students – British students for all languages except English, and foreign students attending pre-sessional English language courses. Language teachers confirmed that the texts selected were appropriate for their students. They also designed multiple-choice questions to test understanding of key aspects of the content and the argumentation in each text. For instance, a text from http://news.bbc.co.uk/1/hi/magazine/4149835.stm (shown as s-bbc-labour in Figure 1) was assessed by ten questions, of which the following is one example:

1 What is “unpaid overtime” (lines 3-4)?

- **a** extra work which no-one pays them for
- **b** getting no pay when they are away from work
- **c** rest and holiday periods which are not paid
- **d** longer holidays without pay
- **e** work which they should not be doing

The main task of the test was to check the assumptions of the language teachers about the appropriateness of certain texts according to students’ level of language competence. The test results confirm that the teachers’ predictions of the difficulty level correspond to the automatic score: the average number of correct answers for the English was 6.2 (σ=2.03). Questions to test global comprehension (Alderson, 2000) in our test proved more difficult than those for testing local comprehension: the former gave 40% vs. 90% of correct answers for the latter.

Applications of automatically graded texts

The proposed method of grading texts by their difficulty finds various potential applications in the process of learning and teaching foreign languages. The first application is creating

graded readers for extra-curricular reading. Teachers' experience shows that successful students regularly read authentic texts in foreign languages. Weaker learners, on the other hand, struggle to find texts suitable for their level of linguistic competence and therefore are often put off by the excessive difficulty of the majority of authentic texts available on-line. However, reading outside the classroom can be crucial for making progress in language learning at the intermediate level, especially outside the country of the target language. Thus, this use meets the needs of many students.

A second application will be beneficial for the language teachers. Selecting texts for the classroom is a well-known problem which, until now, has largely relied on intuitive decision-making and teachers' experience. This requires a teacher to solve many tasks at once to find: a text on a desired topic; a text that is suitable for certain grammar or/and lexical tasks; a text suitable for a certain group of students; a text that can be discussed; a text of a certain genre; etc. This list of requirements can vary and is open-ended. Most teachers are happy if the text can fit at least two of these requirements, and if they do not have to amend it. Finding suitable texts within the short time that is usually available for lesson preparation is a very demanding process. Automatic text selection and grading should relieve and support teachers to a great extent. Our future work will include developing a multipurpose tool not only to give a teacher the opportunity of selecting suitable texts according to subject and difficulty (that alone would be a great advantage), but also to give them the possibility of picking up texts with specific grammatical and lexical phenomena on which they are working in class.

This point relates to a third practical application of the proposed method: automatic creation of grammatical or lexical exercises which can help the teacher to develop meaningful tasks or to support the student in exploratory activities on the basis of the authentic content. We took our model already used for selecting texts for reading exercises and experimented with applying it to the selection of texts suitable for grammatical gap-fill exercises. For instance, it was used to find texts rich in modal verbs or conjunctions. The use of authentic running texts in such exercises instead of artificial single-sentence examples improves students' motivation and helps in contextualising grammatical rules. For instance, a text from http://teacher.scholastic.com/activities/wwatch/hurricanes/witnesses.htm was found to have a significant coverage by a large variety of modal verbs (some legal texts had greater coverage, but little variety of lexical items), so that it allowed a gap fill exercise like:

I screamed, "Mom, Dad, we _____ get out! The water's rising!" I packed one outfit in my book bag, and my parents grabbed a few things. We _____n't find our dog, Bear. But we just _____ leave.

Conclusions and further research

The project will be developed further along the two main lines of enquiry identified above. In terms of feature selection, we would like to experiment with other features indicative of text difficulty, such as nominalisations or the number of different participants in a text, as well as language-specific features, e.g. the use of oblique cases in Russian or genitive in German. Such features are not always marked explicitly, e.g. a list of nouns is needed to find nominalisations, we are not aware of a reliable German tagger that marks genitive constructions. Also, more research is needed on finding the most discriminative features. In our experiments the PCA transform did not select the coverage by POS trigrams, a feature considered to be capturing the language model. In terms of using selected texts in reading exercises, we would like to make more experiments with testing text comprehension, for instance, using MCQs to check understanding of texts that are considerably more or less

difficult according to our model. Also it is important to investigate the balance between local and global comprehension required of an individual text. For instance, reading an instruction for operating a safety-critical device implies paying attention to exact understanding of every step, while reading a magazine might be less demanding. Automatic text selection has to take such purposes into account as well.

References

Alderson, J. Charles. 2000. Assesing Reading. Cambridge.

Baroni, M., Kilgarriff, A. 2006. Large linguistically-processed Web corpora for multiple languages. In: *Companion Volume to Proc. of the European Association of Computational Linguistics*, Trento, 87-90.

Collins-Thompson, K., Callan, J. 2004. A language modeling approach to predicting reading difficulty. In *Proc. of HLT/NAACL,* 193-200

DuBay, W. 2004. The principles of readability. Impact Information, California. http://www.impact-information.com/impactinfo/readability02.pdf

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*. Association for Computational Linguistics.

Kilgarriff. A., Husak, M, McAdam, K, Rundell, M, Rychly, P. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Euralex08.*

Leech, G. 1997. Teaching and language corpora : A convergence. In A. Wichmann, S. Fligelstone, A. M. McEnery, & G. Knowles (eds), *Teaching and Language Corpora*, London, 1-23.

Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus*, M. Baroni and S. Bernardini (eds.). Bologna: Gedit, 63-98. http://wackybook.sslmit.unibo.it/

Sharoff, S. 2007. Classifying Web corpora into domain and genre using automatic feature identification. In *Proc. of the Third Web as Corpus Workshop*, Louvain-la-Neuve, September, 2007.

Schwarm S., Ostendorf. S. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. *Association for Computational Linguistics*, 2005.

The authors

Serge Sharoff is a lecturer in the Centre for Translation Studies (CTS), University of Leeds. He is involved in several projects related to corpus collection and corpus-based technologies for language learning and translation.

Svitlana Kurella is a PhD student in CTS. Her project aims at developing an effective corpus-based methodology for acquiring reading abilities in Polish and Ukrainian based on the knowledge of a second language (L2, here Russian). She is also involved in teaching Russian at Leeds.

Anthony Hartley is the Director of CTS. His research interests are in Machine Translation, controlled languages and quality of translation and interpreting.