An Extensible Crosslinguistic Readability Framework

Jesse Saba Kirchner
Department of Linguistics
UC Santa Cruz
1156 High Street
Santa Cruz, CA 95064
kirchner@ucsc.edu

Justin Nuger
Department of Linguistics
UC Santa Cruz
1156 High Street
Santa Cruz, CA 95064
jnuger@ucsc.edu

Yi Zhang
Baskin School of Engineering
UC Santa Cruz
1156 High Street, SOE 3
Santa Cruz, CA 95064
yiz@soe.ucsc.edu

Abstract

Automatic assessment of the readability level (i.e., the relative linguistic complexity) of documents in a large number of languages is an important problem that can be applied to many real-world applications, such as retrieving age-appropriate search engine results for kids, constructing automatic tutoring systems, and so on. Unfortunately, existing readability labeling techniques have only been applied to a very small number of languages. In this paper, we present an extensible crosslinguistic readability framework based on the use of parallel corpora to quickly create readability software for thousands of languages, including languages for which no linguists are available to define readability rules or for which documents with readability labels are lacking to train readability models. To demonstrate our idea, we developed a system based on the proposed framework. This paper discusses the theoretical and practical issues involved in designing such a system and presents the results of an experiment conducted with the system.

1 Introduction

Automatically labeling the reading difficulty of an arbitrary document is an important problem in several human language technology applications. It can, for example, be used in the next generation of personalized information retrieval systems to find documents tailored to children at different grade levels. In a tutoring system, it can be used to find online reading materials of the appropriate difficulty level for students (Heilman et al., 2006).

Of the world's more than 6,000 languages (Grimes, 2005), readability classification software exists for a striking few, and it is limited in coverage to languages spoken in countries with prominent standing in global economics and politics. A substantial number of the remaining languages nevertheless have a sufficient corpus of digital documents — a number which may already be in the hundreds and soon in the thousands (Paolillo et al., 2005). A natural idea is to create software to automatically predict readability levels (henceforth "RLs") for these documents. Such software has significant potential for applications in different areas of research, such as creating web search engines for kids speaking languages not covered by existing readability software, as described above.

There is much research on assessing the reading difficulties of texts in a particular language, and the existing work can be roughly classified as falling under two approaches. The first approach uses manually or semi-automatically crafted rules designed by computational linguists who are familiar with the language in question (Anderson, 1981). The second approach learns readability models for a particular language based on labeled data (Collins-Thompson and Callan, 2004).

Unfortunately, existing approaches cannot be easily extended to handle thousands of different languages. The first approach, using rules devised by computational linguists familiar with the languages, is impractical because for many languages, especially minority or understudied languages, there are relatively few linguists sufficiently familiar with the language to design such software. Even if these linguists exist, it is unlikely that a search engine company that wanted to serve the whole world would have the resources to

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, pages 11–18,
Suntec, Singapore, 6 August 2009.

hire all of them. The second approach, using machine learning techniques on labeled data, is very expensive because it requires the support of educated speakers of each language to provide readability labels for documents in the language. The availability of such speakers cannot always be assumed. Again, recruiting annotators for thousands of different languages is not economically feasible or practical for a company. An alternative strategy that can scale to thousands of different languages is needed.

In this paper, we propose a general framework to solve this problem based on a parallel corpus crawled from the web. To illustrate the idea, we developed an Extensible Crosslinguistic Readability system (henceforth "ECR system"), which uses a Cross-Lingual Information Retrieval (henceforth "CLIR") system that we call EXCLAIM. The ECR system functions to create RL classification software in any language with sufficient coverage in the CLIR system. We also report the promising — though very preliminary — results of an experiment that tests a real-world application of this system. Investigation of the basic assumptions and generalization of parameters and evaluation metrics are left for future work.

The rest of this paper is organized as follows. The problem setting is described in Section 2. The architecture of our ECR system is explained in Section 3. Our experimental design is laid out in Section 4, followed by experimental result analysis in Section 5. Section 6 gives an overview of related work, and section 7 concludes.

2 Problem and Proposed Methodology

2.1 Existing Approaches to Readability Classification

In traditional approaches to computational readability classification, there is a variety of language-specific system requirements needed in order to perform the RL classification task. For some languages, this task is relatively well-studied. For example, the simple and widely-used *Laesbarhedsindex* (henceforth "LIX") calculates RLs for texts written in Western European languages[1] with the following LIX formula:

$$RL_{\mathbb{D}} = \frac{\text{words}}{\text{sentences}} + \frac{100 \times \text{words}_{char>6}}{\text{words}}$$

[1]In practice, LIX may be substituted with other metrics, such as Flesch-Kincaid.

where $\mathbb{D}$ is a document written in an unfamiliar language, and $\text{RL}_{\mathbb{D}}$ is the readability score of the document D.

The above formula relies on specific parameters which have been tuned to a certain set of languages. These include the total number of words in $\mathbb{D}$ (words), the total number of sentences in $\mathbb{D}$ (sentences), and the total number of words in $\mathbb{D}$ with more than six characters ($\text{words}_{char>6}$).

Although this formula may be successful in RL classification for languages like English and French (Björnsson and Hård af Segerstad(1979), Anderson (1981)), it remains essentially parochial in the context of other languages because the parameters overfit the data from the Western Euorpean languages for which it was designed. Since the LIX formula depends on measuring the number of characters in a word to find words greater than 6, it is ineffective in determining the readability of documents written in languages with different writing systems, such as Chinese. This is due to the fact that some languages, like Chinese, are written with characters based on semantic meaning rather than phonemes, as in English, and a large number of Chinese words consist of just one or two characters, regardless of semantic complexity (Li and Thompson, 1981). In a similar vein, many languages of the world (even some that use phonemically-based writing systems) do not adhere to the implicit assumption of the LIX formula that semantically "complex" words are longer than simpler words (Greenberg, 1954). In these languages, then, the same metric cannot be used as a valid measure of RL difficulty of documents, since word length does not correlate with semantic complexity.

One recent alternative approach has been developed for readability labeling that uses multiple statistical language models (Collins-Thompson and Callan, 2004). The idea is to train statistical language models for each grade level automatically from manually labeled training documents. However, even an approach like this is not scalable to handle thousands of languages, since it is hard to recruit annotators of all of these languages to manually label the training data.

2.2 Proposed Solution

We propose a scalable solution to the problem of labeling the readability of documents in many languages. The general idea is to combine CLIR

technology with off-the-shelf readability software for at least one well-studied language, such as English. First, off-the-shelf readability software is used to assign RLs to a set of documents in the source language, e.g. English, which serve as training data. Second, a set of key terms is selected from each group of documents corresponding to a particular RL to construct a readability model for that RL. Third, for each of these sets of terms, the cross-lingual query-expansion component of the CLIR system returns a semantically relevant set of terms in the target language. Finally, these target-language term sets are used to build the target-language RL models, which can be used to assign RLs to documents in the target language, even if language-specific readability classification software does not exist for that language. This solution plausibly extends to any of the languages covered by the CLIR system. It is possible to create a CLIR system by crawling the internet for parallel corpora, which exist for many language pairs. As a result, the proposed solution already has the potential to cover many different languages.

The success of this method relies on the assumption that readability levels remain fairly constant across syntactically and semantically parallel documents in the two languages in question, or simply across documents typified by equivalent key terms. This does not seem unreasonable: if the same information is represented in two different languages in semantically and structurally comparable ways, it is likely that the reading difficulty of the two texts should not differ much, if at all. If this assumption is true, generation of readability software really depends only on the availability of a solid CLIR system, and the problem of requiring trained computational linguists and native language speakers to design the system is mitigated.

Figure 1 shows a simple process model of a system for generating RL classifiers for various languages. A set of training documents from a source language (i.e., the "L1" in Figure 1) is assigned RLs by the off-the-shelf RL classification software R(L1). Using the source langauge files and the RLs produced by R(L1), the ECR system produces a source language (L1) readability model. Through the system interface, the CLIR system (EXCLAIM) uses the L1 readability model to produce a target language (L2) readability model. The system uses the L2 readability model to produce a new RL classifier R(L2) for the target language. The newly developed classifier R(L2) can then be used to classify documents in the L2.

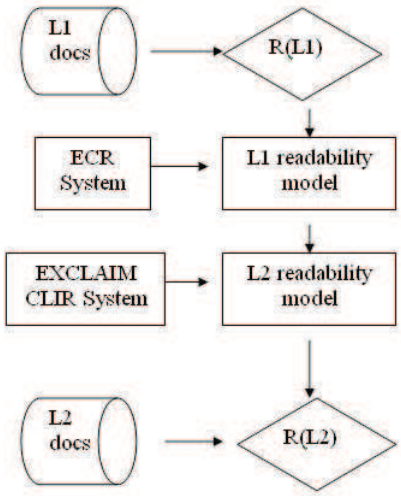

Figure 1: ECR Domain

3 System Architecture

To address any theoretical or empirical concerns and questions about the proposed solution, including those relating to the assumption that key term equivalence correlates with RL equivalence, we have developed an ECR system compatible with an existing CLIR system and have proposed evaluation metrics for this system. We developed the ECR system to meet the needs of two different kinds of users. First, higher-level *intermediate users* can build RL classification software for a given target language. Second, *end users* can use the software to classify documents in that language. In this section, we give a developer's-eye view of the system architecture (shown in Figure 2), making specific reference to the points at which intermediate and end users may interact with the system. For presentational clarity, we periodically adopt the arbitrary assumption that the source language is English, as this is the source language of our experiment described in the following section.

The ECR system has three primary tasks. The first task is to enable intermediate users to develop RL classification model for the source language. The second task is to provide the intermediate user with a toolkit to construct language-specific software that automatically tags documents in the target language with the appropriate RLs. The final task is to provide an interface module for the end

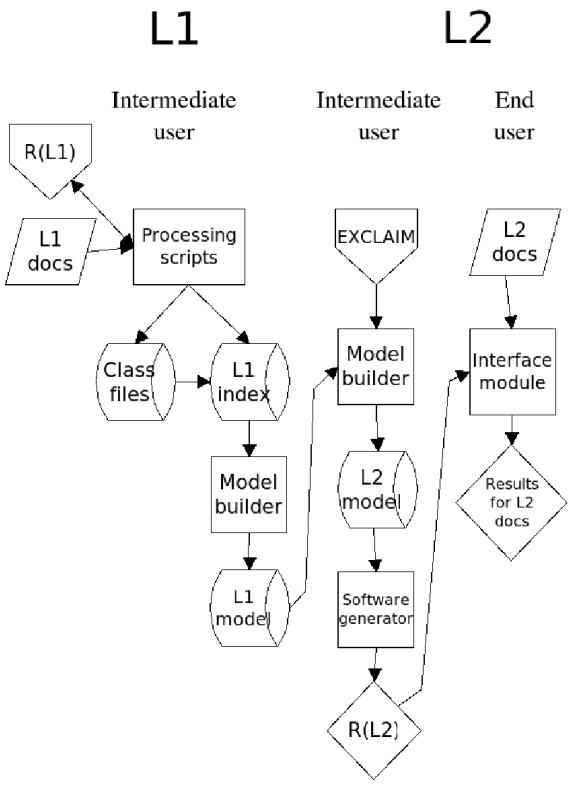

Figure 2: ECR System design

user to utilize this software.

In order to approach the first task, one needs a set of documents in a source language for which off-the-shelf readability software is available. This set of documents functions as a training data set; if a user is trying to assign RLs to documents in a particular domain — e.g., forestry, medical, leisure, etc. — then (s)he can already help shape the results of the system by providing domain-relevant source langauge data at this stage. To aid the intermediate user in obtaining RLs for this set of data, the ECR system has a number of parameters that may be selected, based on different models of RL-tagging — for example, we selected English as the source language and the aforementioned LIX formula due to its simplicity. The documents are then organized according to the generated RLs and separated into different RL groups.

At this point, the K most salient words are extracted from each source language RL groups ($RL_{\mathbb{S}}$) based on the following *tf*idf* term weighting:[2]

$$w_{i,j} = \left(0.5 + \frac{0.5\, freq_{i,j}}{max_l\, freq_{l,j}}\right) \times log\frac{N}{n_i}$$

[2]In principle, this choice is arbitrary and any other appropriate term-weighting formula could also be used.

The selected words $RL_{\mathbb{S}} = \{f_1, f_2, ...f_K\}$ form the basis for constructing an RL classification model for an unknown target language.

In order to construct a target language RL classification model, the cross-lingual query expansion component of a CLIR system is necessary to select semantically comparable and semantically related words in the target language. The CLIR system we developed is called EXCLAIM, or the **EX**tensible **C**ross-**L**inguistic **A**utomatic **I**nformation **M**achine. We constructed EXCLAIM from a semantically (though not structurally) parallel corpus crawled from Wikipedia (Wikimedia Foundation, 1999). All Wikipedia articles with both source and target language versions collectively function as data to construct the CLIR component. Due to Wikipedia's coverage of a large amount of languages (English being the language with the largest collection of articles at the time of writing), CLIR components for English paired with a wide number of target languages was created for EXCLAIM.

For each $RL_{\mathbb{S}}$, the query-expansion component of EXCLAIM determines a set of corresponding words for the target language $RL_{\mathbb{T}}$. Initially, each word in $RL_{\mathbb{S}}$ is matched with the source language document in EXCLAIM for which it has the highest *tf*idf* term weight. The M most salient terms in the corresponding target language document (calculated once again using the *tf*idf* formula) are then added to $RL_{\mathbb{T}}$. Therefore, $RL_{\mathbb{T}}$ contains no more than $K * M$ terms. The total set of $RL_{\mathbb{T}}$s form the base of the target language readability classification model.

Using this model, the system generates target language readability classification software on the fly, which plugs into the system's existing interface module for end users. Through the module, the end user can use the newly generated software to determine RLs for a set of target language documents without requiring any specialized knowledge of the languages or the software development process.

4 Experimental Design

We conducted an experiment to demonstrate this idea and to test our ECR system. Without loss of generality, we chose English as our source language and Chinese as our target language. While Chinese is a major language for which it would be relatively easy to find linguistic experts to write

readability rules and native speakers to label document readability for training, our goal is not to demonstrate that the proposed solution is the best solution to build readability software for Chinese. Instead, we chose these languages for the following reasons. First, we are capable of reading both languages and are thus able to judge the quality of the ECR system. Second, publicly available English readability labeling software exists, and we are not aware of such software for Chinese. Third, we had access to a parallel set of documents that could be used for the evaluation of our experiment. Fourth, the many differences between English and Chinese might demonstrate the applicability of our system for a diverse set of languages. However, the features that made Chinese a desirable target language for us are not essential for the proposed solution, and do not affect the extensibility of the approach.

We created a test set using a collection of Chinese-English parallel documents from the medical domain (Chinese Community Health Resource Center, 2004). The set comprised 65 documents in English and their human-translated Chinese translations. Although a typical user does not need to have access to sets of bilingual documents for the system to run successfully, we circumvented both the lack of off-the-shelf Chinese readability labeling software and the lack of labeled Chinese documents for the evaluation of the results of our system by using a high quality translated parallel document set. Since RLs are rough measures of semantic and structural complexity, we assume they should be approximately if not exactly the same for a given document and its translation in a different language, an extension of the ideas in Collins-Thompson and Callan (2004). Based on this assumption, we can accurately compare the RLs of the translated CCHRC Chinese medical documents to the RLs of the original English documents, which we call the "true RLs" of the testing documents.

LIX-based RLs can be roughly mapped to grade levels, e.g., a text that is classified with an RL of 8 is appropriate for the average 8th grade reader. Since we can assign RLs to the English versions of the 65 CCHRC documents, these RLs can serve as targets to match when generating RLs for the corresponding Chinese versions of the same documents.

An advantage of our system arises from a complete vertical integration which allows a user with knowledge of the eventual goal to help shape the development of the target language RL classification model and software. In our case, the target language (Chinese) test set was from the medical domain, so we selected the OHSU87 medical abstract corpus as an English data set. We automatically classified the OHSU87 documents using the LIX mapping schema assigned by the UNIX *Diction and Style* tools,[3] given in the following Table.

LIX Index	RL	LIX Index	RL
Under 34.0	4	48.0-50.9	9
34.0-37.9	5	51.0-53.9	10
38.0-40.9	6	54.0-56.9	11
41.0-43.9	7	57.0 and over	12
44.0-47.9	8		

Table 1: Mapping of LIX Index scores to RLs as assigned by *Diction*

Then, we concatenated the English OHSU87 documents in each RL group. The *tf*idf* formula was used to select the K English words most representative of each RL group.

Next, we automatically selected a set of Chinese words for each RL class to create a corresponding Chinese readability model by passing each English word through the CLIR system, EXCLAIM, to retrieve the most relevant English document in the Wikipedia corpus, where relevance is measured using the *tf*idf* vector space model. The top M Chinese words from the corresponding Chinese document in the parallel Wikipedia corpus were added to $RL_{\mathbb{T}}$. By repeating this process for each word of each RL class, the Chinese readability model was constructed. In our experiment, we set $K = 50$ and $M = 10$ arbitrarily. The ECR system then automatically generated the subsequent RL classification software for Chinese.

Finally, we assigned a RL to each document in the test set. At this point the procedure is essentially similar to document retrieval task. Each RL group's set of words $RL_{\mathbb{T}}$ was treated as a document (d_j), and each test document to be labeled was treated as a query (q). RLs were ranked based on the cosine similarity between $RL_{\mathbb{T}}$ and q. Finally, the top-ranked RL was assigned to each test document.

[3] Available online at `http://www.gnu.org/software/diction/diction.html`.

5 Empirical Results

The results are presented below in Table 2. The RL assigned to each Chinese document is compared to the "true RL" of the English document, on the assumption that translation does not affect the readability level. Although only 7.8% of the RLs were predicted accurately (i.e., the highest ranked RL for the Chinese document corresponded identically to the RL of the translated English document), over 50% were either perfectly accurate or off by only one RL.

Correctly predicted RL	7.8%
RL off by 1 grade level	43.1%
RL off by 2 grade levels	18.4%
RL off by 3 grade levels	18.4%
RL off by 4 grade levels	6.1%
RL off by 5 grade levels	3.1%
RL off by 6 grade levels	0%
RL off by 7 grade levels	3.1%
RL off by 8 grade levels	0%

Table 2: Distribution of RLs as predicted by our ECR system

This table motivates us to represent the results in a more comprehensive fashion. Intuitively, the system tends to succeed at assigning RLs *near* the correct level, though not necessarily at the exact level. To quantify this intuition, we used Root Mean Squared Error (RMSE) to evaluate the experimental results. We compared our results to two kinds of baseline RL assignments. The first method was to randomly assign RLs 1000 times and take the average of the RMSE obtained in each assignment; this yielded an average RMSE of 3.05. The second method used a fixed equal distribution of the nine RLs, applying each RL to each document an equal number of times, and taking the average of these results. This baseline returned an average RMSE of 3.65. The average RMSE of our ECR system's performance on the CCHRC Chinese documents is 2.48. This number compares favorably against both of the baseline algorithms.

Recall that the actual RL-tagging procedure has been treated as a document retrieval task, using Vector Space Cosine similarity. As such, RLs are not simply "picked out" for each document: each document receives a cosine similarity score for each RL, calculated on the basis of its similarity to the language model word set constructed for each RL. For the results above, only the top ranked RL was considered, as this would be the RL yielded if the user wanted a discrete numeric value to assign to the text. If we allow for enough flexibility to select the better of the two top-ranked RLs assigned to each document by our ECR system, the results are as given in Table 3.

Correctly predicted RL	10.8%
RL off by 1 grade level	49.2%
RL off by 2 grade levels	27.7%
RL off by 3 grade levels	7.7%
RL off by 4 grade levels	1.5%
RL off by 5 grade levels	0%
RL off by 6 grade levels	3.1%
RL off by 7 grade levels	0%
RL off by 8 grade levels	0%

Table 3: RL Distribution (Best of Two Top-Ranked RLs)

While this extra selection is certain to improve the RMSE, what is surprising is the extent to which the RMSE improves. Once again, RMSE can be calculated in the following way. The two top-ranked RLs for each document are taken into consideration, and of these two RLs, the RL nearest to the true RL is selected. Selecting the best of the two top-ranked RLs causes the RMSE to drop to 1.91.

6 Related Work

The method described above builds on recent work that has exploited the web and parallel corpora to develop language technologies for minority languages (Trosterud (2002), *inter alia*).

Yarowsky et al. (2001) describe a system and a set of algorithms for automatically deriving autonomous monolingual POS-taggers, base noun-phrase bracketers, named-entity taggers, and morphological analyzers for an arbitrary target language. Bilingual text corpora are treated with existing text analysis tools for English, and their output is projected onto the target language via statistically derived word alignments. Their approach is especially interesting insofar as the system does not require hand-annotation of target-language training data or virtually any target-language-specific knowledge or resources.

Martin et al. (2003) present an English-Inuktitut aligned parallel corpus, demonstrating superior

sentence alignment via Pointwise Mutual Information (PMI). Their approach provides broad coverage of cross-linguistic morphology, which has implications for dictionary expansion tasks; problems encountered in dealing with the agglutinative morphology of Inuktitut are suggestive of the myriad issues arising from cross-language comparisons.

Rogati et al. (2003) present an unsupervised learning approach to building an Arabic stemmer, modeled on statistical machine translation. The authors use an English stemmer and a small parallel corpus as training resources, with no parallel text necessary after the training phase. Additional monolingual texts can be incorporated to improve the stemmer by allowing it to adapt to a specific domain.

While Yarowsky et al. (2001), Martin et al. (2003) and Rogati et al. (2003) all focus on aligned *parallel* corpora, our approach differs in that we use *comparable* documents from Wikipedia are linked thematically on the basis of semantic content alone: there is no presumed structural or lexical alignment between parallel documents. We have adapted the methods used in conjunction with aligned parallel corpora for use with non-aligned parallel corpora to handle the task pursued by Collins-Thompson and Callan (2004), which presents a new approach to predicting the RLs of a document by evaluating readability in terms of statistical language modeling. Their approach employs multiple language models to estimate the most likely RL for each document.

This approach contrasts with other previous monolingual methods of calculating readability, such as Chall and Dale (1995), which assesses the readability of texts by calculating the percentage of terms that do not appear on a 3,000 word list that 80% of tested fourth-grade students were able to read. Similarly, Stenner et al. (1988) use the word frequency information from a 5-million-word corpus.

While our work has drawn from several techniques employed in prior research, we have mainly hybridized the technique of using parallel corpus employed by Yarowsky (2001) and the language modeling approach employed by Collins-Thompson and Callan (2004). Our approach relies on parallel corpora to build a readability classifier for one language based on readability software for another language. Rather than focusing on language-specific readability classification based on training data drawn from the same language as the testing data (Collins-Thompson and Callan, 2004), we have constructed a radically extensible tool that can easily create readability classifiers for an arbitrary target language using training data from a source language such as English. The result is a system capable of allowing a user to construct readability software for languages like Indonesian, for example, even if that user does not speak Indonesian — this is possible due to the large parallel English-Indonesian corpus on Wikipedia.

7 Conclusion

We have proposed a general framework to quickly construct a standalone readability classifier for an arbitrary (and possibly unfamiliar) language using statistical language models based both on monolingual and non-aligned parallel corpora. To demonstrate the proposed idea, we developed an Extensible Crosslingual Readability system. We evaluated the system on the task of predicting readability level of a set of Chinese medical documents. The experimental results show that the predicted RLs were correct or nearly correct for over 50% of the documents. This research is important because it is the only technique we are aware of that is capable of straightforwardly creating readability labels for hundreds, or theoretically even thousands, of different languages.

Although the general framework and architecture of the proposed system are straightforward, the details of implementation of the system modules could be further improved to achieve better performance. For example, all target language words are selected from a single "best-matching document" using EXCLAIM in this paper. Further experimentation might discover a better word selection module. Future work may also reveal delineation points for over- and under-specialized sets of training data. The OHSU87 data set was selected on the basis of its medical domain coverage, however it may not have provided broad enough coverage of the appropriate domain-independent vocabulary in the CCHRC documents. And finally, we conducted the experiment using our own CLIR system, EXCLAIM, while other CLIR systems might yield better results.

Acknowledgements

The research reported here was partly supported by NSF Grant #BCS-0846979 and the Institute of Education Sciences, US Department of Education, through Grant R305A00596 to the University of California, Santa Cruz. Any opinions, findings, conclusions or recommendations expressed in this paper are the authors', and do not necessarily reflect those of the sponsors.

References

Jonathan Anderson. 1981. Analysing the readability of English and non-English texts in the classroom with Lix. Paper presented at the Annual Meeting of the Australian Reading Association.

C. H. Björnsson and Birgit Hård af Segerstad. 1979. *Lix på Franska och tio andra språk*. Pedagogiskt centrum, Stockholms skolförvaltning.

Jeanne S. Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline, Cambridge, Mass.

Chinese Community Health Resource Center. 2004. CCHRC Medical Documents. Retrieved December 9, 2006, from `http://www.cchphmo.com/cchrchealth/index_E.html`.

Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*. ACL.

Joseph H. Greenberg. 1954. A quantitative approach to the morphological typology of language. In *Method and Perspective in Anthropology: Papers in Honor of Wilson D. Wallis*, pages 192–220, Minneapolis. University of Minnesota Press.

Barbara Grimes. 2005. *Ethnologue: Languages of the World, 15th ed.* Summer Institute of Linguistics.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2006. Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *Proceedings of the Ninth International Conference on Spoken Language Processing*.

Charles N. Li and Sandra Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 workshop on building and using parallel texts: Data driven machine translation and beyond*. ACL.

John Paolillo, Daniel Pimienta, and Daniel Prado. 2005. *Measuring Linguistic Diversity on the Internet*. UNESCO, France.

Monica Rogati, Scott McCarley, and Yiming Yang. 2003. Unsupervised learning of arabic stemming using a parallel corpus. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*. ACL.

A.J. Stenner, I. Horabin, D.R. Smith, and M. Smith. 1988. *The Lexile Framework*. Metametrics, Durham, NC.

Trond Trosterud. 2002. Parallel corpora as tools for investigating and developing minority languages. In *Parallel corpora, parallel worlds*, pages 111–122. Rodopi.

Wikimedia Foundation. 1999. Wikipedia, the free encyclopedia. Retrieved May 8, 2006, from `http://en.wikipedia.org/`.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 161–168.