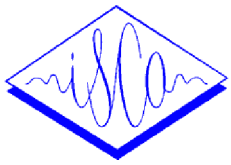
ISCA Archive
http://www.isca-speech.org/archive



Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Analysis of Vocabulary Difficulty Using Wiktionary

Julie Medero, Mari Ostendorf

Department of Electrical Engineering
University of Washington, Seattle, WA 98195, USA
{jmedero, mo}@ee.washington.edu

Abstract

Assessing vocabulary difficulty is useful for finding and creating texts at low reading levels. Prior work has focused on characteristics such as word length and word frequency. In this work, we explore whether other cues might be useful, using features extracted from Wiktionary entries. Comparing words in comparable articles in Standard and Simple English Wikipedia, we find that words that appear in Standard but not Simple English tend to have shorter definitions, fewer part-of-speech types and word senses, and fewer languages that they have been translated into.

1. Introduction

Having access to simple English texts is beneficial for young readers, second language learners, and adults with low literacy skills or learning disabilities. To serve these groups, educators may want to find or create such texts to match their science or social science curricula, particularly for people reading below their age level. Searching for text on the web by topic is now relatively successful, but the results often need to be filtered or simplified to match the desired reading level. Automatic tools using text classification algorithms are beginning to emerge for reading level detection, which can aid in text filtering. Automatic paraphrasing technology offers the possibility for aiding with text simplification. In both cases, the algorithms could benefit from new features for assessing lexical difficulty.

Much work has been done on reading level detection. Recently, a focus has been on identifying reading level-appropriate web content [11, 10, 5]. While traditional methods like the Flesch-Kincaid Grade Level index [6] and the Gunning Fog index [4] rely on easily-calculated approximations to complexity based on features like sentence length and syllable counts, more recent approaches take advantage of language modeling and statistical learning. The REAP system automatically identifies web texts for students to improve reading comprehension [5]. The Read-X project has a similar goal, aiming to categorize web documents thematically, then perform real-time difficulty analysis, to provide reading level-appropriate search results [8]. Both systems leverage a combination of lexical and grammatical features to predict reading level. Other work has focused on syntactic features and feature combination (e.g. [9], [7]). In all of these efforts, the most important feature seems to be word frequency.

In this paper, we also focus on word-level features but consider possible alternatives to frequency measures as a means of approximating lexical difficulty. The goal is to identify new features that could augment word frequency in assessing difficulty of a word, which might be useful for reading level detection but also for predicting which words should be targets for paraphrasing in automatic simplification. To do this, we make use of the data freely available online from Wiktionary. An important advantage of this approach is that it enables an adaptive indicator of word difficulty in that the Wiktionary is an evolving resource that is kept up to date by a community of writers.

While its encyclopedia counterpart, Wikipedia, has been used extensively in language processing, the Wiktionary dictionary has been used quite a bit less frequently. Zesch *et al.* use the English and German components to calculate the semantic relatedness between words, finding that their results using Wiktionary meet or exceed results using WordNet for a number of tasks [13]. Chesley *et al.* use the English dictionary to determine adjective polarity as part of a system for classifying blog post sentiment [3]. We are not aware of any previous work that has used Wiktionary as a source for predicting lexical difficulty.

In the sections to follow, we describe the Wiktionary data set used for feature extraction, and a corpus of comparable Standard and Simple English articles from Wikipedia that we leverage to group words according to whether they do vs. do not appear in the Simple English articles. We then report the results of several different analyses, including word frequency distributions, and conclude with suggestions for future work.

2. Wiktionary Data Set

Wiktionary is an online, multilingual dictionary comprised entirely of user-generated content. In addition to definitions, it includes translations of words between the 172 available languages. Because the dictionary is wiki-based and edited by users, it is constantly growing and changing. This gives it the advantage of being able to respond quickly to new lexical items or new meanings of existing lexical items. Because it is freely available and online, it is an attractive alternative to large, static word lists for tasks like reading level assessment. While the ideal end-user system would make direct, real-time use of the online dictionary, we use a static dataset for controlling experimental conditions. All of our work uses a downloaded archive of the Wiktionary content as it appeared on May 24, 2008.

Dictionary entries contain definitions for one or more senses of a word in one or more parts of speech, along with etymology, pronunciation information, related words and phrases, and other lexical information. In addition, dictionary entries may contain translations to multiple languages. For example, the definition of “paraphrase” has a total of five senses across two different parts of speech, along with three translations and a list of derived terms.

Paraphrase

- Noun
  - paraphrase (plural paraphrases)
    1. a restatement of a text in different words, often to clarify meaning
    2. a similar restatement as an educational exercise
    3. restatement of a text
  - Derived terms
    1. paraphrastic
    2. paraphrastical
    3. paraphrastically
  - Translations
    - restatement of a text
      1. French: paraphrase (fr) f
      2. Portuguese: paráfrase (pt) f
      3. Spanish: paráfrasis (es)
- Verb
  - to paraphrase (third-person singular simple present paraphrases, present participle paraphrasing, simple past and past participle paraphrased)
    1. to restate something as, or to compose a paraphrase

3. Wikipedia Comparable Articles

Wikipedia is an online source of encyclopedia data in a variety of languages. One of the "languages" offered by Wikipedia is "Simple English," which seeks to provide articles on the same topics as the Standard English Wikipedia but using "fewer words and easier grammar," making it more accessible to "students, children, adults with learning difficulties and people who are trying to learn English" [12]. Because the Simple English Wikipedia covers the same topics as the Standard English Wikipedia, it is a source of comparable texts manually simplified by a wide variety of authors. [1]

We collected archives of the Simple English and Standard English Wikipedia content on June 13, 2008 for topics that are covered by both encyclopedias. In our analysis here, we look at a 2475 document subset of those topics. All documents were processed to remove tables, bulleted lists, and other non-textual information. The remaining words were stemmed and part-of-speech (POS) tagged using the RASP tagger [2]. Because the set of POS tags used by the parser is substantially more fine-grained than the POS tags used in Wiktionary, we developed a table for mapping tags from the parser to the most appropriate tag from Wiktionary.

We are interested in exploring whether we can characterize the difference between words that are used in both the Simple English and the Standard English articles and words that are only used in the Standard English articles, as we believe that this will help us to identify "difficult" lexical items. Consequently, in the following experiments, we look at two sets of words: the unique stemmed word types that appear in the Simple English articles ("Simple" – 19927 words), and the unique stemmed word types that appear in the Standard English articles but not in any Simple English articles ("Standard" – 30376 words). Words that appeared only once were almost exclusively misspellings, formatting errors, etc., so they were excluded from further analysis. We further exclude any words that do not have Wiktionary entries.

[1] It is important to note that the reading difficulty of these simplified articles has not been verified by reading time analysis or other tests, but rather is believed to be easier to ready by authors. However, Wikipedia does have administrators who work to ensure that all articles meet their quality standards, and our work shows that the distinction between Simple and Standard English articles is, in fact, consistent with standard difficulty measures like unigram frequency.

4. Analysis

In the following sections, we investigate properties of words in the Simple vs. Standard English subsets using standard approaches (word length and unigram frequency) as well as properties of words extracted from Wiktionary entries (definition length and POS, sense, and translation counts). For all significance tests reported, we use a Wilcoxon sign rank test.

In extracting definition characteristics related to word sense and translations, some of our analyses consider only definition senses that match the POS tagger's labeling. In other words, in determining how difficult the lexical item "affect" is as a noun, we might not want to consider the dictionary entries treating its more frequent use as a verb. Experiments requiring exact POS matches are labeled "exact." Because the POS tagger makes mistakes, however, requiring an exact POS match may be too restrictive. Manual examination found instances of the "Number" POS tag being assigned to words like "article(s)," billionth" and the letter "l." To minimize the number of dictionary look-up misses that are a result of these tagging errors, we consider the case of allowing for only the tag mismatches described in Table 1; these experiments will be labeled "close." For the scenario where no POS restriction is made, the experiments are labeled "all."

Original POS Tag	Allowed POS Tags
Proper Noun	Noun
Noun	Proper Noun
Adverb	Adjective
Number	Adjective, Symbol, Noun

Table 1: Allowed POS mismatches in "close" POS matches

4.1. Standard Measures

A standard indicator of word difficulty is **word length** – short words tend to be easier to read. Indeed, the average length of the words in the Standard English vocabulary set is greater than that of the words in the Simple English set: 7.6 vs. 6.8 characters, and 2.9 vs. 2.6 vowels for Standard vs. Simple words respectively. Both results are statistically significant with $p < 0.05$, though the difference is smaller than we expected.

It is generally thought that difficult words tend to be lower frequency, while words that appear in texts aimed at lower reading levels will tend to be shorter in length and higher in frequency. Certainly there are exceptions, since children's books may have some long dinosaur names that are infrequent in general English. However, we expect the percentage of low frequency words to be smaller for Simple English. In order to better understand how **word frequency** might be used in identifying words that are candidates for paraphrasing in simplification, we looked at the distribution of unigram probabilities for words in the two sets. We chose to use unigrams from the Google

n-grams, because it has been argued that text on the web is representative of general English and because of the large vocabulary represented. The Google n-grams corpus gives counts of unigrams, bigrams, trigrams, fourgrams and fivegrams of over thirteen million unique words that appeared at least 200 times in a large collection of web data [1].

Figure 1 shows the relative frequency and cumulative distribution of words in the simple and Standard English vocabulary sets, as a function of their unigram probabilities (at the low end of the scale). The figure shows that a large proportion of words are low frequency – consistent with Zipf's law – though the relative percentage and distribution peak is smaller for Simple English. In the Simple set, roughly 16% of the words have a unigram frequency less than 10^{-8}, but this holds for 38% of the Standard words. Thus, there is a significant difference in distributions, but low unigram frequency alone is not a reliable indicator of a difficult word. The low frequency words in both sets include misspellings ("elecction"), proper nouns, and rare words. There may be some bias in the particular comparable corpus used here, since even Simple Wikipedia has encyclopedic entries for rare words, e.g. "acanthopterygii" (with expanded explanations), but this is consistent with the introduction of new vocabulary in reading texts. The data provides guidance on the proportion of low frequency words that are appropriate in simplified texts.

The histogram continues to decay for both data sets. We were surprised to find so many high frequency words in the Standard English subset, since presumably most high frequency words are represented in the Simple set. From anecdotal inspection, we found that many were associated with stemming errors, but there were also some words (e.g. "products," "services," "sites," "porn") that might reflect a bias in the distribution of documents on the web. Thus, high frequency words according to these unigrams cannot be assumed to be uniformly good for low reading level text.

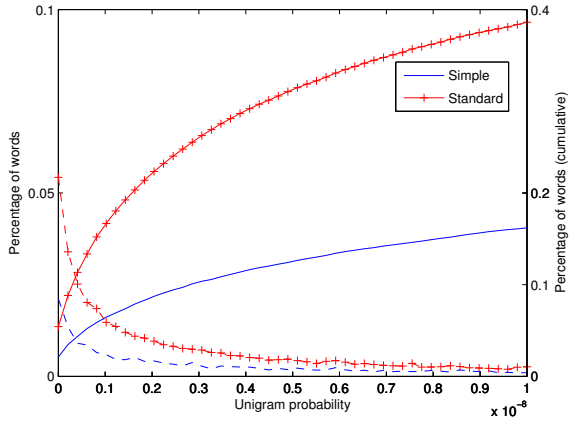


Figure 1: Percentage (and cumulative percentage) of words in the Simple vs. Standard English vocabulary sets with different unigram frequencies.

4.2. Definition length

A simple Wiktionary feature to compute is the length of the entry in characters. Manually inspecting some dictionary entries leads us to believe that very common, simple words tend to be used in multiple contexts and, as a result, to have multiple senses. For example, while the word "cat" has twenty senses in three different parts of speech, the word "feline" has only three senses in two parts of speech. Because translations are user-contributed, we also expect that simple, common words will have been translated into more languages. Because of these factors, we expect to find that the words in the Simple set will, in general, have longer dictionary entries than words in the Standard set. Indeed, the difference in mean definition length between the Simple set (810 characters) and the Standard set (619 characters) is statistically significant with $p<0.05$.

4.3. POS Counts

Next, we look at the total number of parts of speech associated with a word. Again, our expectation is that very common words will be more likely to have multiple parts of speech. Figure 2 shows the percentage of words in the Simple and Standard English vocabularies as a function of the word's POS label counts. Comparing the Simple and Standard sets, we find that the difference between the two sets is significant with $p<0.05$, with an average of 2.7 different POS tags for words in the Simple English set and 2.2 tags for words in the Standard set.

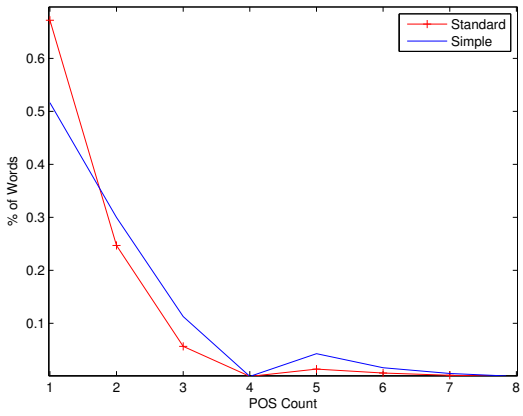


Figure 2: Percentage of words in the Simple vs. Standard English vocabulary sets with different numbers of POS labels.

4.4. Sense Counts

We expect highly technical words to have very specific meanings and, thus, few senses. In contrast, we expect common words to be more likely to have multiple senses. If so, the average number of senses for the Simple English dataset should be higher than the average number of senses for the Standard English data set. Table 2 gives the results for each of the POS restriction options described earlier. In all three cases, the Simple English mean number of senses is higher than the Standard English mean; all results are statistically significant with $p<0.05$.

Because each part of speech in a dictionary entry has at least one sense, it is possible that the higher sense count for words in the Simple data set is a direct result of the higher POS count. To account for that, we also looked at the number of senses divided by the number of parts of speech. While the difference in means decreased, the statistical significance still held.

POS Restriction	Standard English	Simple English
exact	1.63	1.72
close	1.66	1.81
all	2.17	2.69

Table 2: Mean number of senses per word for Standard English and Simple English

4.5. Translation Counts

Finally, we consider the number of translations available for a given word. Users can contribute translations of an English word into any of the 171 other languages covered by Wiktionary. Our intuition is that common words will be more familiar to a large number of users and, consequently, will be more likely to have been translated into multiple languages. At the same time, common words may be more likely to have more than one sense, with different senses being translated into a given target language differently. Overall, we expect words only used in the Standard English dataset to have fewer available translations than words used in the Simple English dataset. Table 3 shows the results for each POS restriction option. All differences are statistically significant with $p<0.05$.

POS Restriction	Standard English	Simple English
exact	7.23	11.89
close	7.53	12.00
all	8.22	13.03

Table 3: Mean number of translations per word for Standard English and Simple English

5. Discussion

In summary, we have confirmed that simple English texts have fewer low frequency words, but still contain a large percentage of low frequency words. Histograms of words in terms of their frequencies in general English provide an indicator of the relative number of low frequency words that are appropriate in simplified text. We have also shown that some easily-extracted features of user-generated Wiktionary entries can be useful in distinguishing word types that appear in Simple English Wikipedia articles from words that only appear in the Standard English comparable articles. POS counts, translation counts and definition length were all shown to be useful in characterizing the two classes. However, like word frequency, these cues may best be used in describing the distribution of words in a text. There will still be a need for some "non-simple" words in simple texts.

Future work may look at the relative frequency of different POS tags/senses for a given word and whether rare POS tags or senses are more likely to occur in the Standard English data. In this work, we looked at the features of unique word types without considering how many times they appeared in each data set. Another possible analysis would look at dictionary entry statistics as they relate to the number of word tokens in the Simple English and Standard English articles, distinguishing between frequent and infrequent words within those genres. Other word features that would be worth investigating in the future include concreteness, imageability, distribution across genres, and neighborhood density. It would be interesting to explore how our difficulty measures correlate with the reading time data available from the elexicon project. Finally, we plan to integrate these cues in automatic algorithms for reading level detection or text simplification.

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0326276. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank the anonymous reviewers for several useful comments.

7. References

[1] T. Brants and A. Franz. *Web 1T 5-gram Version 1*, 2006.

[2] E. J. Briscoe, J. Carroll, and R. Watson. The second release of the RASP system. In *Proc. of the ACL-Coling'06 Interactive Presentation Session*, pages 77–80, 2006.

[3] P. Chesley, B. Vincent, L. Xu, and R. K. Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *Proc. AAAI Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[4] R. Grunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[5] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *Proc. ICSLP*, pages 829–832, 2006.

[6] Jr. J. P. Kincaid, R.P. Fishburne, R.L. Rodgers, and B.S. Chisson. Derivation of new readability formulas for Navy enlisted personnel. *Research Branch Report 8-75*, 1975.

[7] K. Collins-Thompson M. Heilman and M. Eskenazi. An analysis of statistical models and features for reading difficulty prediction. In *The Workshop on Innovative Use of NLP for Building Educational Applications*. ACL, 2008.

[8] E. Miltsakaki and A. Troutt. Real-time web text classification and analysis of reading difficulty. In *Workshop on Innovative Use of NLP for Building Educational Applications*, 2008.

[9] S. Petersen and M. Ostendorf. A machine learning approach to reading level assessment. *Computer, Speech and Language*, 23(1):89–106, 2009.

[10] S. E. Petersen and M. Ostendorf. Assessing the reading level of web pages. In *Proc. of ICSLP*, pages 833–836, 2006.

[11] S. Kurella S. Sharoff and A. Hartley. Seeking needles in the web haystack: Finding texts suitable for language learners. In *TaLC-8*, 2008.

[12] Wikipedia. Simple English wikipedia - simple English wikipedia, the free encyclopedia, 2009. [Online; accessed 11-May-2009].

[13] T. Zesch, C. Müller, and I. Gurevych. Using wiktionary for computing semantic relatedness. In *Proc. AAAI Conference on Artificial Intelligence*, 2008.