Characterizing Text Difficulty with Word Frequencies

Xiaobin Chen[*] and **Detmar Meurers**
LEAD Graduate School
Department of Linguistics
Eberhard Karls Universität Tübingen
{xiaobin.chen,detmar.meurers}@uni-tuebingen.de

Abstract

Natural language processing (NLP) methodologies have been widely adopted for readability assessment and greatly enhanced predictive accuracy. In the present study, we study a well-established feature, the frequency of a word in common language use, and systematically explore how such a word-level feature is best used to characterize the reading levels of texts, a text-level classification problem. While traditionally such word-level features are simply averaged for all words of given text, we show that a richer representation leads to significantly better predictive models.

A basic approach adding a feature for the standard deviation already shows clear gains, and two more complex options systematically integrating more frequency information are explored: (i) encoding separate means for the words of a text according to which frequency band of the language they occur in, and (ii) encoding the mean of each cluster of words obtained by agglomerative hierarchical clustering of the words in the text based on their frequency. The former organizes frequency around general language characteristics, whereas the latter aims to lose as little information as possible about the distribution of word frequencies in a given text. To investigate the generalizability of the results, we compare cross-validation experiments within a corpus with cross-corpus experiments testing on the Common Core State Standards reference texts. We also contrast two different frequency norms and compare frequency with a measure of contextual diversity.

[*]Xiaobin Chen is also affiliated with the South China University of Technology, where he holds a lecturer position.

1 Introduction

Although readability research has gone through a history of more than one hundred years (DuBay, 2007), the use of Natural Language Processing (NLP) technology in readability research is a recent phenomenon. It has greatly improved the predictive accuracy by enabling a multi-dimensional characterization of a text's reading level (Benjamin, 2012; Collins-Thompson, 2014). For example, Vajjala and Meurers (2012) showed that 46 lexical and syntactic features mostly inspired by complexity measures Second Language Acquisition research support a classification accuracy of 91.3% on WeeklyReader, a collection of texts targeting children in four age groups commonly used in such readability research (Petersen and Ostendorf, 2009; Feng et al., 2010).

The readability of a text is determined by the combination of all text aspects that affects the reader's understanding, reading speed, and level of interest in the text (Dale and Chall, 1949). Recent studies explore lexical, morphological, semantic, psycholinguistic, syntactic, and cognitive features for determining the reading levels of texts (Crossley et al., 2007; Lu et al., 2014; Hancke et al., 2012; Boston et al., 2008; vor der Brück et al., 2008; Heilman et al., 2007; Feng, 2010; McNamara et al., 2014).

Among all these elements, the semantic variable of word difficulty has traditionally been found to account for the greatest percentage of readability variance (Marks et al., 1974). Word difficulty is often associated with word frequency given that the amount of exposure of a reader to the word is believed to be the major predictor of word knowledge (Ryder and Slater, 1988).

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 84–94,
San Diego, California, June 16, 2016.

In the present study, we zoom in to the question how word frequency can best be used to characterize the readability of a text. We experimented with three different methods of using frequency as a word-level feature to inform our predictions of readability at the text-level.

2 The Frequency Effect

Reading is a coordinated execution of a series of processes which involve word encoding, lexical access, assigning semantic roles, and relating the information contained in a sentence to earlier sentences in the same text and the reader's prior knowledge (Just and Carpenter, 1980). Successful comprehension of texts depends on the readers' semantic and syntactic encoding abilities (Marks et al., 1974), as well as their vocabulary knowledge in the language (Laufer and Ravenhorst-Kalovski, 2010; Nation, 2006). A general consensus of reading research is that lexical coverage/vocabulary knowledge are good predictors of reading comprehension (Bernhardt and Kamil, 1995; Laufer, 1992; Nation, 2001; Nation, 2006; Qian, 1999; Qian, 2002; Ulijn and Strother, 1990).

A reader's vocabulary knowledge is largely related to the amount of exposure they have received to words—often refered to as *frequency effect*. It is argued to be predictive of word difficulty (Ryder and Slater, 1988) and Leroy and Kauchak (2014) found that word frequency is strongly associated with both actual difficulty (how well people can choose the correct definition of the word) and perceived difficulty (how difficult a word looks). High-frequency words are usually perceived and produced more quickly and more efficiently than low-frequency ones (Balota and Chumbley, 1984; Howes and Solomon, 1951; Jescheniak and Levelt, 1994; Monsell et al., 1989; Rayner and Duffy, 1986). Consequently, a text with many high-frequency words is generally easier to understand than one with a number of rare words. Frequency of word occurrence affects not only the ease of reading, but also its acceptability (Klare, 1968).

The frequency effect is based on a cognitive model assuming a higher base-level of activation for frequently-used words, so they require relatively less additional activation when they are being retrieved from the reader's mental lexicon (Just and Carpenter, 1980). This idea is supported by the findings that high-frequency words are more easily perceived (Bricker and Chapanis, 1953) and readily retrieved by the reader (Haseley, 1957). Going beyond this basic effect, in frequency-based accounts of Second Language Acquisition (Ellis, 2012), the frequency distribution of the input is a key determinant of acquisition, with regularities emerging through the learner's exposure to the distributional characteristics of the language input.

3 Word Frequency for Readability Assessment

Figure 1 illustrates how word frequency can be linked to reading comprehension. Based on a model such as this one, it is reasonable to assume that lexical frequencies can inform text-level analyses.

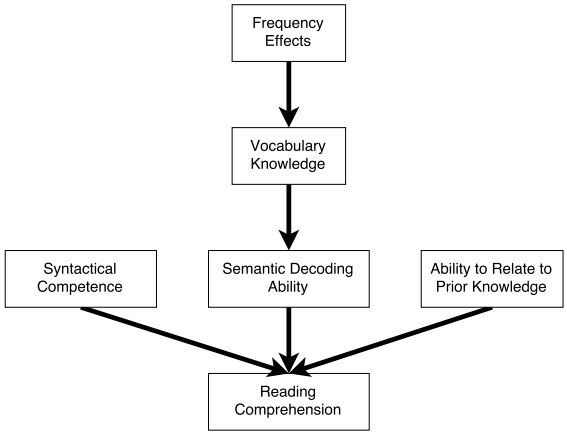

Figure 1: The frequency effect on reading comprehension

Traditional readability formulas used measures such as number of "zero-index words" (number of words that are not included in the most frequent words in English), median of index numbers (Lively and Pressey, 1923), average word weighted value (Patty and Painter, 1931), or number of words from the text that are among the first 1,000 and first 2,000 most frequent words (Ojemann, 1934) for predicting reading levels. These measures were found to be highly correlated with difficulty and effective in assessing text readability.

Modern readability assessment systems such as Lexile (Lexile, 2007), ATOS (Milone and Biemiller, 2014), and CohMetrix (McNamara et al., 2014) also made wide use of word frequencies to help determine the reading level of a text, and such systems

were found to be relatively effective (Nelson et al., 2012). However, there are several issues concerning the frequency lists used, the nature of the frequency measures, and how they are used to account for text readability that deserve more attention.

The first issue is that the frequency lists adopted by these studies were mostly drawn from written corpora. Spoken language was rarely taken into consideration when frequency lists were being composed. This runs the risk of the frequency values not being a faithful representation of the reader's actual language experience, hence being suboptimal for predicting the ease of perception and retrieval. Fortunately, the SUBTLEX frequency lists (Brysbaert and New, 2009; van Heuven et al., 2014) have been compiled on the basis of spoken language data drawn from movie and TV subtitles to obtain more faithful representations of a language typical user's experience with language. The SUBTLEX frequency lists significantly better predict word processing times than earlier norms such as Kučera and Francis (1967) and Celex (Baayen et al., 1993), or frequencies norms derived from the huge Google books corpus (cf. Brysbaert et al., 2011).

The second issue concerns how frequency is measured. Previous research generally sums up all occurrences of a word in the corpus. Yet some words may be frequent in restricted contexts but are not frequent when considering all contexts of language use. As argued by Adelman et al. (2006), a better method may be to count the Contextual Diversity (CD), the number of contexts in which a word occurs. They found the CD measure to be a better predictor of word frequency effects in lexical decision tasks, a method for probing into the word knowledge in the speaker's mental lexicon. However, to the best of our knowledge, CD measures have never been tested in text-level readability assessment. To address this gap, we experimented with both frequency and CD measures in constructing our readability models.

Finally, as for how to use word frequencies for building readability prediction models, previous research typically employed mean frequencies or the percentage of words from the top frequency bands to characterize text levels. Yet, this loses a lot of information about the distribution of word frequencies in the text. Averaging is easily affected by extreme values, and it loses information about the variability of the data. Furthermore, averaging over all occurrences of words in a text will minimize the contribution of low-frequency words—yet, it may be precisely these less-frequent words that are causing reading difficulties. In order to explore how word frequency can be better used for readability assessment, we test three different methods for characterizing texts in terms of lexical frequency: (i) complementing the mean frequency with the standard deviation, (ii) encoding separate means for the words of a text according to which frequency band of the language they occur in, and (iii) encoding the mean of each cluster of words obtained by agglomerative hierarchical clustering of the words in the text based on their frequency. The second method organizes the frequency measures around general language characteristics, whereas the third one aims to lose as little information as possible about the distribution of word frequencies in a given text. The goal of the series of experiments is to identify better methods for characterizing texts of different reading levels from the lexical perspective.

In sum, the reviews of the frequency effect on reading comprehension and earlier research on the use of word frequency for readability assessment support the hypothesis that a lexical frequency measure reflecting the reader's language experience can play a substantial role in models of text readability. The research reported here is devoted to testing this hypothesis in a way that addresses the three problems spelled out above: the source of the frequency list, the nature of the frequency measure used, and the method for combining word-level evidence for text-level predictions.

4 Experimental Setup

Before turning to the three experiments carried out, let us introduce the resources and the general procedure used. As source of the frequency and CD[1] information, we used the SUBTLEXus (Brysbaert and New, 2009) and the SUBTLEXuk (van Heuven et al., 2014) resources. We ran all experiments with two distinct frequency resources to be able to study the impact of the choice of resource. As corpus for exploring the approach and 10-fold cross validation

[1]The CD measure we used is referred to as SUBTLCD in SUBTLEXus and as CD in SUBTLEXuk.

testing we used the leveled text corpus WeeBit (Vajjala and Meurers, 2012). For independent cross-corpus testing, we trained on WeeBit and tested on the exemplar texts from Appendix B of the Common Core State Standards (CommonCore, 2010). For machine learning, we used the basic k-nearest neighbor algorithm implemented in the R package `class` given that in our initial exploration it turned out to perform on a par or better than other commonly used algorithms such as Support Vector Machine or Decision Trees.

4.1 The SUBTLEX Lists

The SUBTLEXus (Brysbaert and New, 2009) contains 74,286 word forms with frequency values calculated from a 51-million-word corpus of subtitles from 8,388 American films and television series broadcast between 1900 and 2007. The SUBTLEXuk (van Heuven et al., 2014) is the British counterpart, consisting of 160,022 word forms with frequency values calculated from a 201.7-million-word corpus of subtitles from nine British TV channels broadcast between January 2010 and December 2012. The SUBTLEX resources provide frequency information in several forms motivated in van Heuven et al. (2014); we made use of the frequencies given on the Zipf scale (log10 of the frequency per billion words), as well as the CD values, for which each film or TV program counted as a context.

4.2 The WeeBit and Common Core Corpora

The WeeBit corpus used in a number of readability and text simplification studies (Vajjala and Meurers, 2012; Vajjala and Meurers, 2013; Vajjala and Meurers, 2014) was collected from the educational magazine Weekly Reader used in earlier readability research (Petersen and Ostendorf, 2009; Feng et al., 2010) and the BBC-Bitesize website. As summarized in Table 1, it is a 789,926-word corpus of texts labeled with five grade reading levels.

The Common Core corpus consists of exemplar texts from Appendix B of the English Language Arts Standards of the Common Core State Standards. The corpus we use for testing in our experiments is exactly the same as the one used by Nelson et al. (2012). They eliminated the lowest (K–1) level of the original six levels and removed repetition, dramas, and texts intended for teacher to read aloud, resulting in 168 remaining passages at five levels.

Grade Level	Age Group	# Articles	# Words / Article
WR Level 2	7–8	616	152.63
WR Level 3	8–9	616	190.74
WR Level 4	9–10	616	294.91
BiteSize KS 3	11–14	616	243.56
BiteSize GCSE	14–16	616	400.51

Table 1: Details of the WeeBit corpus

4.3 Experimental Procedure

The following basic procedure was followed for each of the experiments carried out:

1. Tokenize corpus texts with CoreNLP Tokenizer (Manning et al., 2014), which had also been used to compose the SUBTLEX frequency lists.
2. Characterize each text using frequency features. The nature of the features differs across the three studies, for which details are given in the following sections.
3. Train classification models on the WeeBit corpus i) in a 10-fold Cross-Validation (CV) setup or ii) using the full corpus when the Common Core data was used as test. The K-nearest neighbors algorithm of the R package `class` was used for model construction and testing.
4. Apply the trained model to the test folds or test corpus to assess model performance.
5. Report results in terms of Spearman's correlation coefficient (ρ) to allow comparison of CV and cross-corpus results. We report both 10-fold CV performance on WeeBit and the test performance on Common Core as references for model fit and generalizability. The KNN algorithm results in different models when the parameter K is set differently. The parameter K for each model was decided automatically by testing K from one up to the square root of the number of texts used for training and choosing the value that resulted in the best performing model. In this paper, we report the performance of the best models.

The complete program for feature extraction and experiment settings with R code can be obtained from `http://xiaobin.ch`.

5 Study 1: Adding Standard Deviation

In this first study, we tried the most conservative extension: in addition to the mean frequencies of the words in a given document, we computed the standard deviation (SD). So we compared +SD models trained on two frequency features (mean and SD) with the baseline −SD models trained only on the mean frequency. As explained in the previous section, we tested this using the Zipf and CD measures from two different frequency resources, SUBTLEXus and SUBTLEXuk.

We experimented with both token and type models. For token models, we considered the SUBTLEX-frequency of each word instance in a given text. For type models, each distinct word in the document was considered only once.

Table 2 sums up the results for the 10-fold CV in terms of the Spearman rank correlation ρ between model-predicted and actual reading levels of texts.[2] Table 3 shows the performance of the models trained on WeeBit and tested on Common Core.

	Token		Type	
	−SD	+SD	−SD	+SD
US-ZIPF	-.02	.40***	.26**	.42***
US-CD	-.02	.46***	.19*	.44***
UK-ZIPF	.05	.25***	.31***	.38***
UK-CD	.04	.26***	.21**	.29***

Table 2: 10-fold CV results for models without/with SD

	Token		Type	
	−SD	+SD	−SD	+SD
US-ZIPF	.03	.34***	.33***	.35***
US-CD	-.27***	.28***	.22**	.33***
UK-ZIPF	-.13	.26***	.36***	.38***
UK-CD	.00	.02	.33***	.27***

Table 3: Common Core test results without/with SD

The models trained on frequency mean and SD systematically performed better than those trained with only mean frequencies. While considering both mean and SD of word frequencies seems like an obvious choice, as far as we know no previous research made use of this option providing significantly better performance for text-level readability prediction.

The results also show that the type models uniformly outperform the token models. In order to further explore this finding, in Figure 2 we plotted

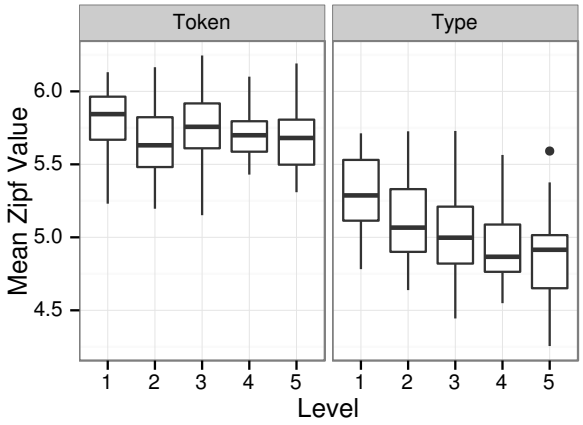

Figure 2: Mean token vs. type Zipf by Common Core text level

the mean token and type frequencies of the words in each text of the Common Core corpus by text level. As the reading level increases, the plot shows a clear pattern of decreasing mean type Zipf values. This is not observable for the token averages.

High-frequency tokens usually have multiple occurrence in a given text, inflating the sum of frequency values and obscuring the influence of low frequency words on the mean. The type-based measure eliminates multiple occurrences of tokens so that words across the frequency spectrum contribute equally. The fact that the average type frequencies in Figure 2 are so clearly associated with the readability levels transparently supports the frequency effect.

Comparing the results based on frequency (Zipf) with those using Contextual Diversity (CD), different from the lexical decision tasks (Brysbaert and New, 2009), where CD was more predictive, for text-level readability assessment, frequency performs better for readability assessment.

Finally, Table 3 also showcases that frequency lists calculated from different corpora (here: SUBTLEXus vs. SUBTLEXuk) do result in substantially

[2] Here and throughout, we mark significant differences (from the null hypothesis that there is no correlation) with *** for $p \leq .001$, ** for $p \leq .01$, and * for $p \leq .05$.

different model performance. For example, the Zipf measure from the SUBTLEXus corpus resulted in the better 10-fold CV performance than that from the SUBTLEXuk corpus, with a highly significant Spearman's correlation coefficient ($\rho = .42, p \leq .001$) for the type model.

6 Study 2: Mean Frequencies of Words from Language Frequency Bands

For the second study, frequency means[3] of words from stratified frequency lists were calculated and used as features to characterize the texts' reading levels. To stratify the frequency list, the words in the SUBTLEX lists were ordered by their frequency values. Then the list was cut into a number of frequency bands, resulting in each word being assigned a band number. Words in the same band thus occur with similar frequency in the language as represented by the corpora the SUBTLEX lists were compiled from. The words in a given text to be analyzed are matched with the words in the frequency list and grouped by the words' band numbers. The text can then be characterized by the average frequencies of the words in each band, i.e., we obtain one average per band. With both SUBTLEX lists, we experimented with up to 100 bands. As before, we used the Zipf frequency and CD measures and tested both token and type models.

Figures 3 and 4 show the performance of token and types models trained with features from both the SUBTLEX lists. The performance is given in terms of 10-fold CV ρs and cross-corpus ρs tested on Common Core. Unlike the results of Study 1, the token models did not perform significantly different from the type models for 10-fold CV. However, the type models generalized better to the Common Core test set than the token models. Word type frequency thus better captures the frequency characteristics of a text. For readability assessment purposes, calculating mean type frequency of words from each frequency band creates better prediction models.

A comparison of the results with those from Study 1 shows that the models constructed using the stratification method clearly outperformed those from Study 1 using only a single mean for all words. When the Zipf measure from the US list is stratified into 60 bands, the trained model has the best performance among all the models, reaching a 10-fold CV $\rho = .83, p \leq .001$ and a cross-corpus testing $\rho = .39, p \leq .001$). Performance on the test set is rather volatile, though: when the list was cut into 20 bands, the resulting model failed to distinguish between text levels ($\rho = -.11, p \geq .05$) for the test corpus, while the with-in corpus CV correlation coefficient was $\rho = .80, p \leq .001$. The method used in this study thus needs to be fine-tuned with respect to the corpora or resources at hand to achieve the optimal results.

7 Study 3: Frequency Cluster Means

The idea behind the third study is the following: The richest frequency representation of a text would be a vector of the frequency of all words in the text. But this is too fine grained to be directly comparable across texts, and texts also differ in length.[4] We therefore incrementally group words together that differ minimally in terms of their frequency values. We can then compute the average frequencies of the words in each group. To realize this idea, we used agglomerative hierarchical clustering to construct a word frequency hierarchical cluster tree for each text in the training corpus. Concretely, we used the `hclust()` function in R with the default complete linkage method and the `dist()` function for calculating Euclidean distances as dissimilarity structure for `hclust()`.

The trees were then cut at different distances from the root to obtain an increasing number of branches, with each branch representing the set of words closest in frequency. The branch means[5] were calculated for each set and used as features to construct the prediction models.

We experimented with the Zipf measure from the SUBTLEXus frequency list with up to 100 clusters and explored type and token models. The performance of the trained models are shown in Figure 5.

[3] Without SD; adding SD did not improve performance.

[4] Orthogonal to the number of frequency values compared, note that the order of words in a given text is ignored here. The order may well provide relevant information characterizing the readability of a text. For example, a simple text may well include rare words as long as they are followed by more frequent words explaining the rare ones. This could be interesting to explore in future work.

[5] Without SD; adding SD did not improve model performance either.

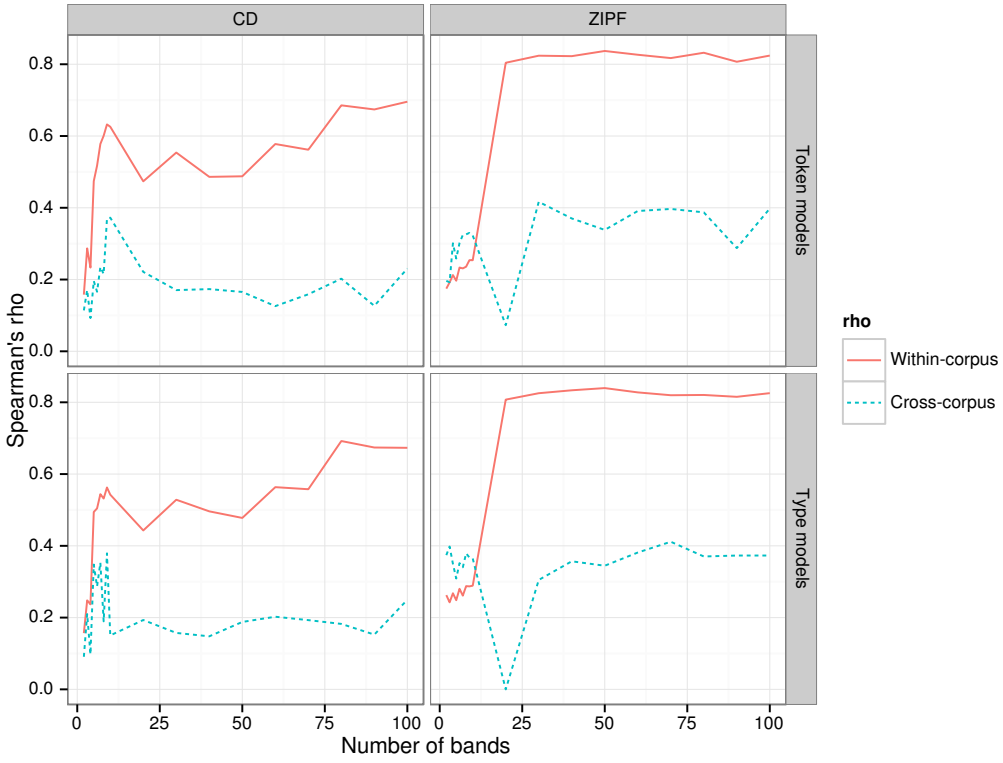

Figure 3: 10-fold CV and cross-corpus test ρs between predicted and actual text reading levels by number of SUBTLEXus bands

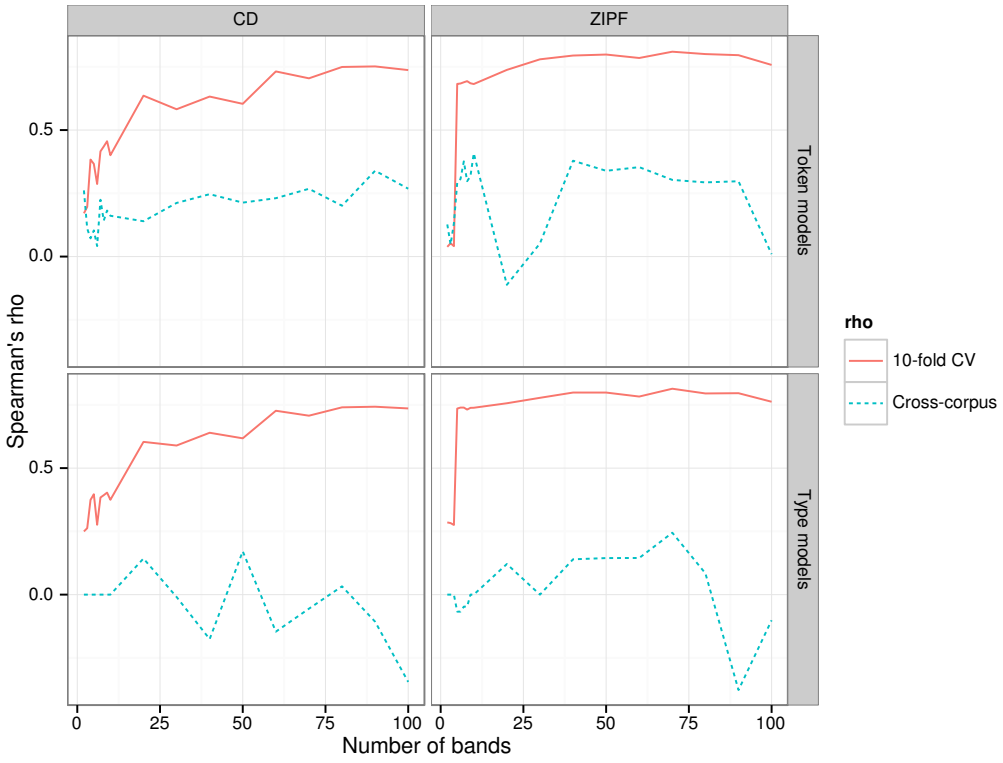

Figure 4: 10-fold CV and cross-corpus test ρs between predicted and actual text reading levels by number of SUBTLEXuk bands

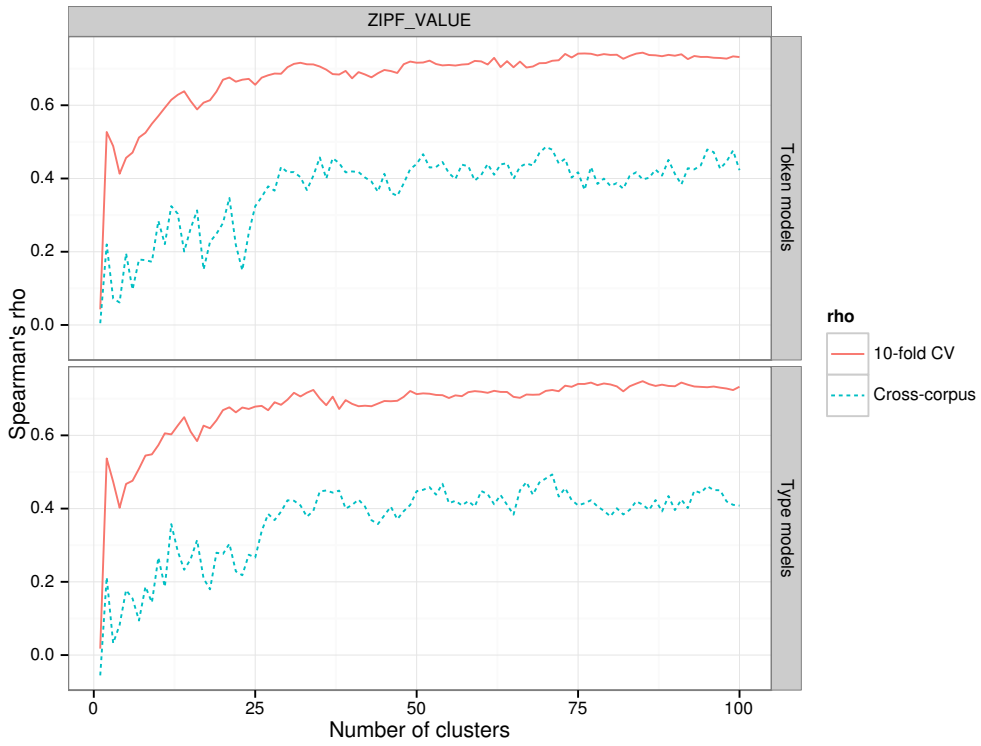

Figure 5: 10-fold CV and cross-corpus testing ρs between predicted and actual text reading levels by number of clusters

For most cutting schemes, the token and type models performed comparably. As the number of cluster increased, the trained models improved in performance, with the testing ρs peaking at 70 clusters for all models. Table 4 provides the information for the best performing models from this study.

Model Type	#Clusters	ρ
Token Model		
10-fold CV	85	.74***
Cross-corpus	70	.49***
Type Model		
10-fold CV	85	.74***
Cross-corpus	71	.49***

Table 4: Best-performing models from clustering experiment

The models constructed in this experiments performed significantly better than those from Study 1. Although the models from Study 2 had higher CV ρs, those from this study show a more stable cross-corpus testing performance, which is of major importance for using such a method in practice. It is striking that clustering the words in a text that are similar in word frequency is more reliable across corpora than grouping words by the language frequency bands as a general characteristic of language independent of the texts.

8 Comparison with Previous Work

Nelson et al. (2012) assessed the capabilities of six tools for predicting text difficulty: the commercial systems Lexile (MetaMetrics), ATOS (Renaissance Learning), DRP Analyzer (Questar Assessment, Inc.), the Pearson Reading Maturity Metric (Pearson Knowledge Technologies), SourceRater (Educational Testing Service), and the research system REAP (Carnegie Mellon University). Word frequency is a measure that is included in all these systems, though all of them incorporate additional features such as syntactic complexity. One of the evaluations reported by Nelson et al. (2012) was carried out on the freely available Common Core exemplar texts that was also used in Vajjala and Meurers (2014) and the present research, and they reported Spearman's ρ, making their results comparable to ours.

The results reported for the best systems clearly

highlight the value of rich feature sets, reaching .76 for SourceRater and .69 for Reading Maturity, which is also the level reached by the Vajjala and Meurers (2014) model.

At the same time, the approach based solely on frequency we discussed in Study 3 with a ρ of .50 is on a par with the results noted by Nelson et al. (2012) for the Lexile system, and only slightly worse than the .53 reported for DRP.

The comparison thus clearly confirms the relevance of considering how lexical frequency information is to be integrated into readability assessment.

9 Conclusions

In this paper, we explored the text readability analysis from a word-level perspective, zooming in on lexical frequency. The goal of the three experiments carried out in the research was to investigate how a text-level classification problem can be informed by a word-level feature of the text, namely the frequency of words in general language use. Word frequency is related to the difficulty level of a text given that reading comprehension is partially determined by the reader's vocabulary knowledge, which in turn is related to word frequency. The frequency effect of vocabulary on the reading levels of text is in line with a basic cognitive model positing that words of higher frequencies have a higher level of activation and require less extra effort when they are being retrieved from the reader's mental lexicon. As a result, where frequency lists faithfully represent the reader's language experience, they can predict how difficult the words used in a text are to the reader and in turn inform estimates of the readability of the text.

Three methods of using word frequency lists to predict text readability were tested and confirmed that word frequency is effective in characterizing text difficulty, especially when more than just the average frequency of the words in a text is taken into account. Characterizing text readability in terms of the overall mean and standard deviation of word frequencies performed better than models just using the mean. The model based on the frequencies of the word types occurring in the text (rather than the tokens) were better throughout and generalized much better across corpora. In terms of the nature of the measure itself, the models trained with the Zipf frequency measures were found to outperform those based on measures of Contextual Diversity. The models trained with stratified frequency measures in the second study showed the best performance for the CV evaluation using a single corpus, but generalized less well to the rather different test data set based on the Common Core texts than the clustering approach explored in the third study.

With respect to applying these methods in practical readability assessment contexts, the Zipf frequency measures from the SUBTLEX frequency lists seem to be well-suited, with the overall mean frequency and SD values computed based on the word types being easy and effective. The stratification method improves performance over the simple mean and SD, but it requires fine-tuning of the number of bands. The clustering method has the best model performance and is least sensitive to the use of different frequency lists and measures, but it is also computationally the most expensive.

While the performance of the best frequency models reaches a level that is competitive with systems such as Lexile, clearly a comprehensive approach to readability assessment will integrate a broad range of features integrating more aspects of the linguistic system, language use, and human language processing. Where texts are characterized in terms of observations of smaller units, based our results for lexical frequency it will be advisable to characterize text level readability by more than simple means when aggregating the information obtained, e.g., at the lexical or sentence level.

Acknowledgments

This research was funded by the LEAD Graduate School [GSC1028], a project of the Excellence Initiative of the German federal and state governments. Xiaobin Chen is a doctoral student at the LEAD Graduate School.

References

James S. Adelman, Gordon D. A. Brown, and José. F. Quesada. 2006. Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science*, 17(9):814–23.

R. Harald Baayen, Richard Piepenbrock, and H. van Rijn. 1993. The celex lexical database. CD-ROM.

David A. Balota and James I. Chumbley. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception & Performance*, 10(3):340–357.

Rebekah George Benjamin. 2012. Reconstructing readability: recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Elizabeth B. Bernhardt and Michael L. Kamil. 1995. Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, 16(1):15–34.

Marisa Ferrara Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Peter D. Bricker and Alphonse Chapanis. 1953. Do incorrectly perceived stimuli convey some information? *Psychological Review*, 60(3):181–188.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Marc Brysbaert, Emmanuel Keuleers, and Boris New. 2011. Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2(March):1–8.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of past, present, and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

CommonCore. 2010. *Common Core State Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects*. Common Core State Standards Initiative.

Scott Crossley, David Dufty, Philip McCarthy, and Danielle McNamara. 2007. Toward a new readability: A mixed model approach. In *Proceedings of the 29th annual conference of the Cognitive Science Society*, pages 197–202, Nashville, Tennessee, USA.

Edgar Dale and Jeanne Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

William H. DuBay. 2007. *Unlocking Language: The Classic Readability Studies*. BookSurge Publishing.

Nick C. Ellis. 2012. Frequency-based accounts of SLA. In Susan M. Gass and Alison Mackey, editors, *Handbook of Second Language Acquisition*, pages 193–210. Routledge.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Nomie Elhadad. 2010. A comparison of features for automatic readability assessment. In *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.*

Lijun Feng. 2010. *Automatic Readability Assessment*. Doctoral dissertation, The City University of New York.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1063–1080, Mumbai, India.

Leonard Haseley. 1957. *The relationship between cue-value of words and their frequency of prior occurrence*. Unpublished master's thesis, Ohio university.

Michael J. Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT 2007*, pages 460–467, Rochester, NY. Association for Computational Linguistics.

Dowes H. Howes and Richard L. Solomon. 1951. Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology*, 41(6):401–410.

Jörg D. Jescheniak and Willem J. M. Levelt. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 20(4):824–843.

Marcel Adam Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329–354.

George R. Klare. 1968. The role of word frequency in readability. *Elementary English*, 45(1):12–22.

Henry Kučera and W. Nelson Francis Francis. 1967. *Computational Analysis of Present-day English*. Brown University Press, Providence, RI.

Batia Laufer and Geke Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.

Batia Laufer. 1992. How much lexis is necessary for reading comprehension? In H. Bejoint and P. Arnaud, editors, *Vocabulary and applied linguistics*, pages 126–132. Macmillan, Basingstoke & London.

Gondy Leroy and David Kauchak. 2014. The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association*, 21(e1):1–4.

Lexile. 2007. The Lexile Framework for reading: Theoretical Framework and Development. Technical report, MetaMetrics, Inc., Durham, NC.

Bertha A. Lively and Sidney L. Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational Administration and Supervision*, 9:389–398.

Xiaofei Lu, David a. Gamson, and Sarah Anne Eckert. 2014. Lexical difficulty and diversity of American elementary school reading textbooks: Changes over the past century. *International Journal of Corpus Linguistics*, 19(1):94–117.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Carolyn B. Marks, Marleen J. Doctorow, and M. C. Wittrock. 1974. Word frequency and reading comprehension. *The Journal of Educational Research*, 67(6):259–262.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press, New York, NY.

Michael Milone and Andrew Biemiller. 2014. The development of ATOS: The Renaissance readability formula. Technical report, Renaissance Learning, Wisconsin Rapids.

S. Monsell, M. C. Doyle, and P. N. Haggard. 1989. Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118(1):43–71.

I. S. Paul Nation. 2001. *Learning vocabulary in another language*. Cambridge University Press, Cambridge.

I. S. Paul Nation. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1):59–82.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, The Gates Foundation.

Ralph J. Ojemann. 1934. The reading ability of parents and factors associated with reading difficulty of parent education materials. *University of Iowa Studies in Child Welfare*, 8:11–32.

Willard W. Patty and W. I. Painter. 1931. A technique for measuring the vocabulary burden of textbooks. *Journal of Educational Research*, 24(2):127–134.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:86–106.

David Qian. 1999. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review*, 56(2):282–308.

David Qian. 2002. Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3):513–536.

Keith Rayner and Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Randall James Ryder and Wayne H. Slater. 1988. The relationship between word frequency and word knowledge. *The Journal of Educational Research*, 81(5):312–317.

Jan M. Ulijn and Judith B. Strother. 1990. The effect of syntactic simplification on reading EST texts as L1 and L2. *Journal of Research in Reading*, 13(1):38–54.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to Web texts. In *Proceedings of The 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68, Sofia, Bulgaria. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics*, 165(2):194–222.

Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–90.

Tim vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep syntactic and semantic indicators. In Tomaž Erjavec and Jerneja Žganec Gros, editors, *Proceedings of the 11th International Multiconference: Information Society—IS 2008—Language Technologies*, pages 92–97, Ljubljana, Slovenia.