Automatic Characterization of Text Difficulty

Julie Medero

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Mari Ostendorf, Chair

Linda Shapiro

Lucy Vanderwende

Program Authorized to Offer Degree:
Electrical Engineering

University of Washington

Abstract

Automatic Characterization of Text Difficulty

Julie Medero

Chair of the Supervisory Committee:
Professor Mari Ostendorf
Electrical Engineering

For the millions of U.S. adults who do not read well enough to complete day-to-day tasks, challenges arise in reading news articles or employment documents, researching health conditions, and a variety of other tasks. Adults may struggle to read because they are not native speakers of English, because of learning disabilities, or simply because they did not receive sufficient reading instruction as children. In classroom settings, struggling readers can be given hand-crafted texts to read. Manual simplification is time-consuming for a teacher or other adult, though, and is not available for adults who are not in a classroom environment. In this thesis, we present a fundamentally new approach to understanding text difficulty aimed at supporting automatic text simplification. This way of thinking about what it means for a text to be "hard" is useful both in deciding what should be simplified and in deciding whether a machine-generated simplification is a good one.

We start by describing a new corpus of parallel manual simplifications, with which we are able to analyze how writers perform simplification. We look at which words are replaced during simplification, and at which sentences are split into multiple simplified sentences, shortened, or expanded. We find very low agreement with respect to how the simplification task should be completed, with writers finding a variety of ways to simplify the same content.

This finding motivates a new, empirical approach to characterizing difficulty. Instead of looking at human simplifications to learn what is difficult, we look at human reading performance. We use an existing, large-scale collection of oral readings to explore acoustic

features during oral reading. We then leverage measurements from a new eye tracking study, finding that all hypothesized acoustic measurements are correlated with one or both of two features from eye tracking that are known to be related to reading difficulty. We use these human performance measures to connect text readability assessment to individual literacy assessment methods based on oral reading.

Finally, we develop several text difficulty measures based on large text corpora. Using comparable documents from English and Simple English Wikipedia, we identify words that are likely to be simplified. We use a character-based language model and features from Wiktionary definitions to predict word difficulty, and show that those measures correlate with observed word difficulty rankings. We also examine sentence difficulty, identifying lexical, syntactic, and topic-based features that are useful in predicting when a sentence should be split, shortened, or expanded during simplification. We compare those predictions to empirical sentence difficulty based on oral reading, finding that lexical and syntactic features are useful in predicting difficulty, while topic-based features are useful in deciding how to simplify a difficult sentence.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to first thank my advisor, Mari Ostendorf, for her support, patience, and encouragement over the past six years. I'm also grateful to Lucy Vanderwende, both for her support during my internship at Microsoft Research and for her thoughtful and constructive input as a member of my committee. Thank you to Linda Shapiro, the third member of my reading committee, and to my former committee members, Jan Spyridakis, Maya Gupta, and Eve Riskin, for sharing their insights.

I've been fortunate to be part of a vibrant and supportive research lab. Thank you to Anna Margolis, Becky Bates, Jeremy Kahn, Jon Malkin, Kevin Duh, and Sheila Reynolds for their advice and teaching; to Alex Marin, Amittai Axelrod, Brian Hutchinson, Bin Zhang, Hanna Hajishirzi, Nicole Nichols, Sangyun Hahn, and Wei Wu for their support in both classwork and research; to Aaron Jaech, Hao Fang, Ji He, Trang Tran, Vicky Zayats, and Yi Luan for their input on early talks related to this work.

I owe thanks to many collaborators, both in research and in teaching. Thank you to Jared Bernstein, Jennifer Balogh-Ghosh, and Xin Chen at Pearson for their help with data, processing, discussion, and feedback. I am grateful to Emily Bender, Gina Levow, Cynthia Loe, Karen Freisem, and Theresa Ronquillo for their parts in making me a better teacher. I am deeply indebted to David Notkin for his wise and caring mentorship.

Thank you to Allen and Inger Osberg, whose fellowship funded the final year of my research.

Conducting experiments with human subjects requires a great deal of time and coordination, and I am grateful for all of the help I have had on my project. Lois Kim, Michellene Steinberg, and Patrick Dugan helped tremendously with conducting study sessions. Judy Ramey, Andy Davidson, Brian Espinosa, and Nan-Chen Chen went above and beyond in making sure that the eye tracking lab was ready for all of our sessions. I'm also grateful to

all of my research study participants, who took time out of their busy lives to contribute to the study, and to Cat Howell, Lindsey Kafer, and all of my students at Literacy Source for their help in recruiting participants.

I've gotten a great deal of support from other women in the EE and CSE departments. I am grateful to Theresa Barker, Fay Shaw, Sandra Fan, Jessica Tran, Karen Studarus, Ahlmahz Negash and Tamara Bonaci for their support and encouragement.

Being a full-time student while raising three kids has made me incredibly grateful for my family's proverbial village. Thank you to Anne Phillips, Darla Rhodes, Erik Selberg, Jake Oshins, John Hake, Julianne Hake, Madi Carlson, Mary Kaye Rodgers, Mike Corder, Samantha Moscheck, and Sue Corder for help with school-pickups, birthday parties, sports practices, and extra playdates. I am also grateful for all of the current and former teachers at the Cooperative Children's Center – amber Petersen, Bernadette Mora, Brad Belvo, Brandon Blake, Chavah Israel, Clarissa Jarem, Dan Todd, Dave Yeager, David Sienkiewicz, Denise Pilkey, Hisham Mishalani, Kevin Richardson, Linda Grigholm, Lindsey Grader, Lisa Stuhley, Madison Bruno, Meghan Finney, Mia Styant-Browne, Ranya Khalil, Sascha Mercer, Shina Kashino, Susan Grove – for giving my kids a safe, fun, nurturing place to spend their days while I was at school.

Finally, I am grateful beyond words for the support of my family. Thank you to Martin, who had no idea what he was getting into when I started this process just before his second birthday, and who will turn eight the week of my defense; to James, who has never known life without a mom who goes to school; to Maggie, who has spent the first year of her life traveling for conferences and interviews, and watching thesis-writing; and to my husband Shawn, who has picked up more slack than I would have thought possible, and has kept a smile on his face through it all.

Parts of this thesis are based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0718124 and by the National Science Foundation under Grant No. IIS-0916951. 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.

DEDICATION

To Shawn, Martin, James, and Maggie.

Chapter 1

INTRODUCTION

Despite the ever-growing availability of information in text format, 15% of U.S. adults do not read at the level necessary for completing day-to-day tasks [32], such as filling out employment forms, understanding news articles, or completing assignments for classes. In a classroom setting, texts can be manually simplified by teachers. However, the process of simplification is time-consuming. Struggling high schoolers might benefit from a simplified version of their history text book, which would allow them to keep up with their history coursework in parallel with improving their reading skills, but it is not practical for teachers to simplify entire textbooks manually. Further, adults who are not in school do not have access to a teacher who could simplify texts for them, and may need to obtain information from news articles, public health documents, and employment forms that are too difficult for them to read. These individuals would benefit from a system that could automatically simplify texts.

To automatically generate a simplified version of a text, we need to solve a number of problems. First, we need to be able to determine the difficulty level of a document, so that we can decide if it needs to be simplified for a target reader. If we determine that the document is too difficult, then we need to identify more specifically what words or sentence constructions are difficult. For the specific points of difficulty, we need to make changes to the text to improve readability while minimizing meaning changes. Difficult words might be replaced with easier synonyms, for example, or a complex sentence might be split into several simpler sentences. Finally, we need a way to evaluate the quality of the resulting simplification, so that we can verify that our resulting document is, in fact, easier for the reader. This thesis contributes to all of these areas by investigating the question of what makes a text difficult, and by exploring how to best leverage human data to guide algorithm development for automatic simplification.

Commonly used commercial software exists for determining reading level. Many of the most common formulas use easy-to-calculate metrics like word frequency, word length, and sentence length to characterize the difficulty of text. By analyzing simplified texts, recent work has shown that other factors are important in determining text difficulty [63, 24, 35, 30, 53]. Most of this work has looked at human text simplifications, which rely on the intuitions of the writer and/or on specific guidelines that specify what language should be used in simplified texts. Our work similarly questions the adequacy of classical methods of calculating text difficulty. In addition to using human text simplifications, however, we look at human oral reading performance.

Our novel approach to understanding difficulty starts from a basis of oral reading evaluation of *people*, where words correct per minute is commonly used to assess individual literacy. We turn that evaluation around, identifying features of oral readings that seem to be indicative of difficulty of *texts*. Our experimental results on a large corpus of oral readings support the hypothesis that errors, hesitations, and atypical pauses and duration patterns tend to occur at points in the text that are difficult for unskilled readers. A collection of readings of a passage by several human subjects can therefore be used to find specific locations of difficulty, as well as to empirically assess the document's overall difficulty.

Other work has looked at human silent reading performance using eye tracking, finding that features of readers' gaze patterns can be used to identify points of high cognitive load. Thus we can further validate the acoustic features by identifying relationships with those gaze features. An advantage of using oral readings over eye tracking is that audio collections are relatively low cost, since they do not require specialized equipment. They are, consequently, more practical to use in the development of new text simplification systems.

The audio features we identify can potentially be used to compare the quality of texts generated by different automatic text simplification systems, or to systematically study the relative difficulty of specific phenomena in texts. Here, we propose a method to rank words and sentences based on audio features in order to explore different sources of difficulty in texts. We learn cues to lexical and syntactic difficulty from two collections of simplified texts, then use the acoustic ranking to assess those cues.

While using oral readings from a number of people is potentially more costly than

collecting manual simplifications of a text, it is more amenable to crowd sourcing, and is arguably more reflective of actual difficulties. In addition, using oral readings to characterize text difficulty has a number of advantages. By looking at how individuals interact with a text, we can target simplifications to a specific reader or group of readers. For example, the method developed in this thesis can lead to an understanding of how difficulty is (or is not) different for native English-speaking children, second-language learners, and low-literacy (but fluent English-speaking) adults. We could even use the methods developed here to better understand what a specific reader finds difficult, allowing for highly customized simplifications, and also for detailed feedback about that individual's reading ability.

1.1 Sources of Text Difficulty

Once it is determined that a document is too difficult for a reader, it still remains to determine what the *sources* of difficulty are. A reader could struggle because a specific word is unfamiliar or difficult to decode. For example, "aching" is a difficult word for some readers to identify in its written form, even though it may be in their spoken vocabulary. "Abrogate," on the other hand, is unlikely to be a familiar word for many low-literacy adults, even if someone were to read it to them. The identity of the reader affects what is difficult as well. The word "mortgage" may well be in a fluent English-speaking adult's vocabulary, but is unlikely to be known by a third grader. Highly specialized medical concepts may be familiar to a medical doctor who is learning English as a second language (though they may not know the English word for those concepts), but are probably not familiar to fluent speakers of English who lack medical training. In automatic simplification, difficult words can be addressed by replacing them with a paraphrase (e.g. "hurting" instead of "aching") or by adding additional explanation (e.g. "The committee voted to abrogate, or get rid of, the old policy."), depending on how important the specific word is to the document's topic.

Even if all of the words in a sentence are easy enough for a specific individual, though, they can be put together in a way that is difficult to read. Long sentences, or sentences with complex syntactic structures, can make the relations between the words in a sentence more difficult to discern. Paraphrasing is not enough to deal with those difficulties; the simplification process may involve splitting a sentence into multiple pieces, getting rid of

extra details, or reorganizing the parts of the sentence to make it easier to understand. Often, those features interact, and a reader faces multiple sources of difficulty at the same time.

Discourse factors may further contribute to the overall difficulty of a document. Anaphora, rhetorical structure, and other document-level features affect readers' cognitive load beyond the difficulty of individual words or sentences. We recognize the importance of discourse factors, but do not address them in the scope of this thesis.

While human knowledge is useful for identifying features of difficult texts (e.g. length or multi-clause syntactic constructions), it is often more effective to automatically learn such features from data. In related natural language processing (NLP) problems, there is a lot of data to learn from, but that is not the case for simplification. Existing corpora of sentence-aligned explicit simplifications are too small to be used for training in a machine learning context. Comparable corpora (e.g. adult- and child-oriented encyclopedias) are aimed at many different types of readers, from young children to second language learners to low-literacy fluent adults. Further, they are not explicit simplifications, and may cover different information under the same topic heading. Our work addresses these shortcomings by learning features from these corpora, but by then using human data to understand how those features interact.

1.2 Approach and Key Contributions

This work began as an investigation of text simplification algorithms. Due to not having an adequate evaluation criterion, we began to explore how humans interact with texts, both as writers and as readers. We started with writers, developing a corpus of parallel, sentence-aligned simplifications. Analysis of this corpus showed huge variability in the simplifications generated by humans. Based on our findings from that study, and on traditions from assessment of individuals' reading abilities, we turned our focus to how readers interact with texts. In the same way that reading rate from a large number of words is used to provide information about a specific reader, we hypothesized that errors and atypical prosodic structure from a large number of readers could provide local information about the difficulty of a specific text. We analyzed a large existing corpus of oral readings, then

validated our findings by collecting eye tracking data for the same texts, since prior work has shown that eye movements are closely related to cognitive load during reading. Using that eye tracking data, we identified acoustic cues that corresponded to different types of eye movements. Finally, we explored using this acoustic measure of difficulty to learn the relative importance of different sources of text difficulty independently learned from comparable texts.

The specific contributions of this thesis are:

A new corpus of parallel human simplifications

We present a new corpus of parallel human simplifications. We report on the extent to which expert writers agree on when certain lexical and syntactic simplifications should occur. The large variation in these simplifications illustrates features of the simplification task that motivate new approaches to characterizing difficulty and to evaluating machine-generated simplifications.

A fundamentally new approach to understanding text difficulty, leveraging oral readings

We connect text assessment to individual literacy assessment methods based on oral reading. In particular, we build off of previous work in which readers' abilities are assessed by the rate at which they can read standardized texts out loud. At the passage level, this measure can be calculated as the number of words that a person can correctly read in one minute. Because we are interested in identifying specific words and constructions that are likely to be difficult (and, consequently, that should be targets of simplification), this work looks beyond passage-level signs of difficulty. We consider several ways that a reader might show that they are struggling with a text:

Errors, including substitutions, insertions, and deletions, are clear indicators of difficulty. They are relatively rare, though, and do not fully capture difficulties on their own.

Hesitations, including filled pauses and restarts (e.g. "um esc- esc- uh escapades")

Pauses that are longer than would be expected by the sentence structure can be a sign of a reader's uncertainty (e.g. "We have always been able to share our . . . escapades

and humor with our friends.")

Lengthening, or drawing out the sounds of a word, is another way that readers can slow down their reading while they struggle with a difficult word or construction (e.g. "We have always been able to share our esc a p a des and humor with our friends.")

Some of the acoustic features that we use in that method, such as lengthening and pausing, occur in fluent reading for communicative reasons. As part of developing our methods, we develop a way to account for those prosodic factors. By identifying locations that those acoustic features can be explained by prosodic boundaries, we are able to restrict our analysis to their occurrences in locations that are likely to be indicative of reading difficulty.

A comparison of acoustic and gaze-based features during oral reading

We also look at how readers' eyes move across a screen as they are reading a passage out loud. By looking at how long a reader's eyes fixate on each word, and at when the reader's eyes move back to a previous point in the passage to reread, we can get a glimpse into their cognitive process as they read. We compare the following gaze-based signs of reading difficulty to our oral reading signs:

Long fixations occur when a reader's eyes focus on a word for longer than might be expected without difficulty.

Regressions occur when a reader's gaze moves back to a previous word, which may or may not have been fixated on previously.

The use of comparable data to identify features that are related to text difficulty

Parallel corpora of texts at different reading levels are rare, so we make use of other large text corpora to identify features that are related to text difficulty. We use comparable articles from the English and Simple English Wikipedia corpora to extract information about the relative frequency of occurrence of words in the two sets. We then relate those frequencies to features extracted from Wiktionary. Finally, we use human data, in the form of the acoustic features described above, to better understand the relative importance of those features.

1.3 Thesis Overview

The rest of this thesis is organized as follows:

Reading and text difficulty have been studied in different ways by many different disciplines, and this thesis builds on findings from each of them. Chapter 2 describes prior work in characterizing and predicting text difficulty, as well as previous work related to automatically simplifying texts for human and machine use. It also shows how other natural language processing tasks have addressed evaluation challenges like the ones faced by automatic text simplification. Finally, it describes previous work that has been done in using oral readings to assess *people*, which is the basis for the work in this thesis on using similar methods to asses *texts*, and previous work connecting eye tracking to cognitive load during reading.

Chapter 3 presents a new corpus of parallel, manual simplifications. By looking at multiple sentence-aligned simplifications of the same documents, we draw conclusions about how humans simplify texts. We find, in particular, that most simplifications are more complicated than simple lexical replacement. Further, we find that participants in our study show low agreement as to which words and sentences should be targeted for simplification, motivating our use of human performance to characterize difficulty.

Chapter 4 analyzes automatic transcripts from an existing, large-scale collection of oral readings. We develop a method for controlling for prosodic structure in oral readings, then examine how different acoustic cues relate to reading difficulty. We find that pausing and lengthening in audio can be used to identify difficult points in texts, as well as to characterize the reading level of individuals.

In Chapter 5, we validate the acoustic cues from Chapter 4 through a new eye tracking collection. We describe a new data collection effort that generated eye tracking data for the same passages as the collection used in Chapter 4, along with two new passages. We find that pauses, hesitations, and lengthening in the audio are associated with long fixations and regressions in eye tracking. We also show how the audio and gaze features change over the course of a passage that is designed to incrementally increase in difficulty.

Chapter 6 develops a ranking method for characterizing the relative difficulty of words

based on acoustic cues. Using comparable corpora, we develop text-based measures of word difficulty. We then compare those difficulty measures to rankings based on the features from human data that were identified in Chapters 4 and 5. We find that the features we extract explain some of the difference in word difficulty rankings. Turning to sentence-level difficulty, we learn from manual simplifications when sentences are likely to be split into multiple pieces, shortened, or expanded during manual simplification. We get gains from lexical, syntactic, and topic-based features. We then use those features to predict sentence difficulty rankings based on the features from human data.

Finally, Chapter 7 summarizes the primary contributions of this thesis. Potential future extensions are discussed. In particular, we describe additional features that could improve the predictive power of the models in this thesis. Next steps for integrating the methods described in this thesis into a full automatic simplification system are described. We also describe ways that the methods in this thesis could be used to explore how reading difficulty varies for different types of readers, as well as how text features related to difficulty vary with genre.

Chapter 2

BACKGROUND

This chapter provides a summary of previous work related to topics covered in this thesis. Our goals are two-fold: first, to motivate the design decisions made later in this thesis, and second, to put the work described in this thesis into context. One of the primary contributions of this thesis is the presentation of a novel way of thinking about what it means for a text to be "hard." This is important for two reasons: first, to aid in deciding what should be simplified, and second, to aid in deciding whether a machine-generated simplification is a good one.

To provide the overall context for this thesis, we start by describing previous work that has been done in automatic simplification in Section 2.1. We cover two types of simplification: simplifications intended to make downstream processing of text easier for other NLP applications, and simplifications aimed at human readers.

Next, in Section 2.2 we look at previous work related to characterizing the difficulty of a passage using text-based features. We start with an overview of traditional formulas for characterizing text difficulty before moving on to describe more recent machine learning techniques for predicting text difficulty. This work informs the features that we use in Chapter 6 to predict the difficulty of words and sentences from text features.

In addition to characterizing the difficulty of an original text, we want to assess the quality of automatically-generated simplifications. To that end, Section 2.3 looks at how other NLP tasks have addressed the question of assessment. Automatic simplification is a relatively new area of study, but we can learn from what is done for assessment in tasks like machine translation and automatic summarization. The work described in this section motivates our analysis of parallel, human-generated simplifications in Chapter 3.

We finish by describing previous work related to human-based evaluation of reading. While the use of oral readings to characterize the difficulty of texts is novel in this thesis,

previous work has used oral readings to characterize the reading level of people. We describe that work, including the features that have been useful in evaluating people through oral readings in Section 2.4. We then describe previous work that makes the connection between eye tracking data and cognitive load during reading in Section 2.5. We will connect those two lines of study in Chapters 4 and 5, when we relate acoustic cues of reading difficulty to eye tracking features during oral reading.

2.1 Automatic Text Simplification

Traditionally, work on sentence simplification has focused on making sentences easier for machines to process in downstream applications. In multi-document summarization, removing certain types of syntactic constructions (e.g. appositives, gerundive clauses, non-restrictive relative clauses) can shorten sentences, which helps to keep extractive summaries within proscribed length limits [83, 72]. In semantic role labeling [84] and automatic question generation [34], learning weights on hand-generated local syntactic transformation rules can reduce the size of the parse tree feature space used in later processing. In machine translation, shorter, less complex syntactic structures can decrease the rate of long-distance reordering between the source and target languages [78].

As for downstream automatic processing, simplification to improve readability for humans can involve reducing syntactic complexity. In addition, though, the difficulty of individual words is important. The Simple English "language" resources on Wikipedia have become popular with researchers interested in simplification. Yatskar et al. [90] and Shardlow [70] both use editor comments attached to Wikipedia edit logs to extract pairs of simple and difficult words from edits aimed at simplification. Biran et al. [18] note that it is not always appropriate to replace a word with a simpler synonym, though; since harder words are sometimes needed to give the appropriate nuance of meaning, it is important to take context into account.

At the sentence level, Napoles and Dredze [54] classify sentences as "simple" or "original" depending on whether they were extracted from a Simple or Standard English Wikipedia article. They find that there are a lot of "simple" sentences in the Standard English entries, which makes the classification task difficult. These findings mirror the ones reported later

in this thesis for lexical difficulty. They are also supported by Vajjala and colleagues, who develop a model [81] to classify web documents as "easy" or "hard." They find that their classifier performance drops substantially when they apply it to Wikipedia sentences [82], and conclude that it is because there are so many easy sentences in the Standard English documents. Napoles and Dredze find, though, that training models for specific categories of articles gave better performance. Their results support the findings in this thesis on the importance of topic to predicting syntactic simplification.

Like the Wikipedia guidelines, the ASD Simplified English specification provides authors with guidelines for writing accessible content in English [58]. It was developed by Boeing as an alternative to translating manuals into multiple languages, and sepecifies a controlled subset of English vocabulary and grammar designed to make aerospace maintenance documents more accessible to readers who were not fluent English speakers [58].

Other recent work in text simplification has been focused on analyzing the kinds of changes that occur in simplification. For example, Amancio and Specia [3] report that nearly a third of all paraphrases in their corpus were part of changes that they call "abstract," which means that they would require external knowledge to generate. Their work represents a step toward understanding the challenges of text simplification. Also related to better understanding the simplification task, Pellow and Eskenazi [62] are developing a corpus of everyday English texts and crowd-sourced simplifications, but it is not yet available. We present a similar resource in this thesis, though we get our simplifications from technical writing experts rather than crowd-sourcing them. We also provide some analysis of the kinds of simplifications we see in that collection.

The approach to simplification presented in this thesis addresses many of the concerns raised above. In particular, we avoid binary distinctions between "easy" and "hard" words and sentences, since "easy" words can appear in "hard" texts and vice-versa. Instead, we characterize the difficulty of words and sentences on a continuum. Rather than using existing corpora as our gold-standard of what is and is not difficult, we use those corpora to make predictions of difficulty, and determine the relative importance of different factors by tuning to empirical evidence from readers interacting with texts.

2.2 Algorithms to Characterize Text Difficulty

Many formulas and techniques have been developed previously for characterizing the difficulty of a text passage. Early measures were designed to be simple to calculate. These measures are still used frequently. This section provides an overview of those measures, along with a description of more recent measures that have been developed using machine learning techniques. The work in this thesis aims to provide a basis for using machine learning more effectively.

2.2.1 Formulas Based on Simple Features

Previous assessment measures have varied in the complexity of their features and calculations. Traditional readability measures like the Flesch-Kincade Grade Level index [38] and the Gunning Fog index [33] rely on easily-calculated approximations to complexity based on features like sentence length and syllable counts. The Lexile [76] framework converts a function of log mean sentence length and mean log word frequency to a fixed scale. These measures are easy to calculate and easy to understand. They are based on analysis of texts that are assumed to be simple to read, however text difficulty is determined by more than the component words. Factors like syntactic structure, topicality, and discourse structure also affect reading difficulty. For example, in a medical article about arrhythmia, the word "arrhythmia" might need to be defined, but it would not make sense to completely remove it even though it is a long word. In the context of Swedish medical texts, Abrahamsson et al. [2] found that their lexical substitution system improved readability according to a reader study, but that improvement was not captured by the simple readability metrics that they considered.

2.2.2 Predicting Reading Level with Machine Learning

Recent work has explored use of data-driven learning to assign reading levels to text. Petersen and Ostendorf used features from parse trees and n-gram language models as input to a support vector machine to predict reading level, and showed higher correlation with human judgments for their machine learning-based approach than for either Flesh-Kincade

or Lexile [63]. Deane and colleagues generated automatic word clusters, and then used document counts of words from each cluster as input to factor analysis to characterize the dimensions of variance between third and sixth grade texts [24]. Heilman and colleagues used counts of lexical items and syntactic subtrees to predict reading level of labeled web texts [35].

In addition to lexical frequency and syntactic features, decodability and discourse features play a role in text difficulty. A few studies have explored how these features relate to text difficulty. Mostow and colleagues use orthographic and phonemic features of individual words to predict the likelihood of substitution errors in a child's oral reading [53]. The Coh-Metrix project includes features based on words, POS tags, argument overlap and topic cohesion, which it measures through sentence and paragraph similarity measures using LSA, in its measure of readability [30]. Nenkova et al. [56] consider lexical, syntactic and discourse (cohesion) features in predicting sentences that are disfluent or difficult for skilled readers to understand. They evaluate their features by predicting human-judged fluency ratings of machine translation output and text quality ratings for summarizations. While their goal is not to predict reading level, their work on text readability in general, and readability of machine-generated text in particular, is relevant to the task of automatic simplification.

François and Miltsakaki [29] provide an excellent comparison of "classic" and "non-classic" features for readability classification, along with a comparison of linear regression to support vector machines (SVMs). They find that the features used in classic reading metrics (e.g. word and sentence length, presence of words on hand-crafted lists of "easy" English words) are by far the most powerful in predicting reading level, but that adding the non-classic features further improves performance. They also find that an SVM outperforms linear regression in terms of pure accuracy, but not if you consider adjacent accuracy, in which reading level predictions are counted as "correct" if they are within one reading level of the target label.

Ma et al. [44] focus on predicting the fine-grained reading levels of books that are often used in elementary classrooms. They find that layout features (e.g. where pictures and text are on the page) are better predictors of readability than more complex NLP features. This

finding supports our hypothesis that different audiences will have different features that are important to difficulty. Location of images on the page is unlikely to be a relevant feature for adults wanting to read health information on the Internet, for example. Consequently, we will eventually want to develop text difficulty measures that can be tuned to a specific target population.

Other recent work has leveraged large corpora to learn what simple texts look like. This work includes the research of Sharoff and colleagues on using lexical and POS-based features [71], which used pairs of Simple and Standard English Wikipedia articles, along with the work of Brooke and colleagues [20], which extends a hand-crafted lexicon of words labeled for difficulty by rating words based on the reading level of documents they tend to appear in. Tanaka-Ishii et al. [79] use an SVM to learn a pairwise comparison that predicts which of two documents is easier. They use that comparison to sort a set of documents by difficulty, resulting in a rank ordering.

While we argue primarily for an empirical measure of text difficulty in this thesis, and focus largely on examining how readers interact with texts, automatic difficulty assessment is an important part of automatic simplification. In particular, automatic measures akin to the ones described above will be needed in automatic simplification, when we will want a predicted text difficulty score that can be optimized in a machine learning framework. We revisit the question of automatically predicting difficulty from text features in Chapter 6.

2.3 Assessment of Automatically Generated Text

A consistent problem in work on text simplification for improving readability to date is the lack of a good method for evaluating automatic simplification. A good simplification needs to make a document simpler while maintaining fluency. The difficulty of evaluating a system-generated simplification is that it cannot be scored against a single gold standard, because there are typically multiple ways to simplify a sentence. This section starts with an overview of how similar challenges have been addressed in machine translation and summarization. It then describes recent first steps that have been made in the text simplification community toward dealing with assessment questions.

2.3.1 Machine Translation

Machine translation systems must also maximize an objective (in this case, the likelihood of a translation) while generating quality, fluent text in the target language. Inter-language differences (e.g. in word order, agreement, or the presence or absence of determiners on nouns) mean that just translating individual words well is unlikely to result in a high-quality translation. As with text simplification, there is a problem of not having a single gold-standard answer for evaluation. Given the same French sentence, different native speakers may generate multiple, equally acceptable, English translations. The BLEU [61] score of a machine-generated translation is an n-gram precision score with a brevity penalty. Essentially, it measures the overlap of words and short phrases between a human reference translation and the machine output. Often, to handle variation between humans, the machine output can be scored against multiple human references once. While BLEU correlates fairly well with human judgment of translation quality and is useful during training, many studies have pointed out its limitations. It is not uncommon to use an evaluation method with a human in the loop, such as human judgments of adequacy and fluency or a human edit rate [74], to evaluate system output in formal evaluations.

2.3.2 Summarization

Automatic summarization faces similar challenges: two people asked to summarize the same document are extremely unlikely to choose exactly the same details to include in their summary. Further, systems that generate summarizations by extracting sentences from the source document(s) run the risk of generating summaries that have an unclear discourse structure, while systems that generate new summary sentences must make sure that the sentences they generate are fluent and grammatically clear. The ROUGE [43] method for evaluating automatic summarization output is similar to BLEU; it also relies on n-gram overlap between system output and optionally multiple human references. The Pyramid method [55] directly addresses the issue of language variability and human answer variability. System outputs and multiple human references are separated into Summary Content Units (SCUs). An SCU is a collection of different word spans that express the

same meaning, and evaluation is done on overlap of SCUs, not of n-grams. Further, SCUs that appear in a larger number of human summaries are weighted more heavily, so that systems are penalized most greatly for excluding a piece of information that was kept in the summary by all human summarizers. Other evaluations of automatic summaries rely on human judgments of readability [55].

2.3.3 Simplification

Like in machine translation and summarization, one common method of evaluation for much of the early work on automatic simplification has been to use human judgments. Yatskar and colleagues used human judgments to rate the relative difficulty of word and phrase pairs. Following their example, Shardlow [70] had humans rate whether sentence pairs were good examples of simplification or not. In both cases, inter-annotator agreement was a challenge for at least some of the annotators.

More recent work has attempted to find alternatives to costly and potentially unreliable human judgments. Štajner et al. [85] look at using MT evaluation metrics in place of human-judged grammaticality and meaning preservation for simplification systems. Nishikawa et al. [57] find that reading time at the document level correlates highly with subjective measures of document readability. Their work is similar to the work described in this thesis in its use of reading times, but substantially different in its focus on document-level scores.

Also similar to our goal of having an empirical characterization of difficulty based on how people interact with a text, Temnikova and Maneva [80] use comprehension questions to rate simplifications. They score both the number of participants who get comprehension questions correct and participants' response times. Some comprehension questions are more difficult than others, though, and it can be difficult to distinguish between difficult texts and difficult questions. An advantage of the approach described in this thesis is that it relies only on interactions with the text itself, so it is not sensitive to the difficulty of the comprehension questions. Further, Amancio and Specia [3] report that as many as 27% of simplifications in their Wikipedia-based corpus are the result of omissions. Using comprehension questions to measure difficulty requires knowing ahead of time that the answers to the questions will

not be removed as part of the simplification process.

2.4 Assessing Individuals: Oral Reading Assessment

While oral readings have not been used previously to assess *texts*, there have been numerous previous studies that used oral readings to assess *individuals*. Children's reading ability is commonly measured in classrooms by having a teacher count the number of words that a child can read in a minute. Researchers have shown that a customized speech recognition system can score this sort of reading assessment as reliably as a teacher [25], and that automatic systems can be used to split students into groups by reading level [26]. Other work has shown similar results for predicting fluency scores in oral reading tests for L2 learners of English [12, 15]. Project LISTEN uses automatic speech recognition of children's oral readings to assess children's reading ability, locate reading errors, and coach young readers [52].

While oral reading proficiency is not the same as comprehension, there is substantial evidence that it is a good proxy for measuring comprehension, especially for unskilled readers. A number of studies show that children's oral reading fluency is a good predictor of comprehension skills [50, 75]. Zhang and colleagues found that features of children's oral readings, along with their interactions with an automated tutor, could predict a single student's comprehension question performance over the course of a document [92].

2.5 Eye Tracking and Reading

During silent reading, our eyes do not move smoothly across a page. Instead, they make a series of jumps from one fixed location to another, typically focusing on one spot for 200-250ms [66]. The places that the eye focuses are *fixations*, and the short (10-100ms) jumps between them are *saccades*. During fixations, tiny eye movements known as *microsaccades* may occur. Microsaccades are spatially random and can be treated as additive noise in modeling eye movements [27]. With the exception of this random noise, the location of the visual focus can be treated as static during a fixation.

There may also be *regressions* in reading, when the focus point of the reader moves backward to a previous point in the text. Regressions may be a result of correcting for a

missed target word due to motor error, or due to needing to return to a previous word for additional lexical, syntactic, or semantic processing [17].

During a fixation, only a small number of characters are centrally focused, so a word may be the target of more than one consecutive fixation. A skilled reader may process 3-4 characters to the left of the fixation point and 14-15 characters to its right [67]. Additionally, though, characters to the right of the primary focus may be the subject of *parafoveal* preprocessing [68], in which a reader's brain starts processing text before it has been the target of a fixation. Researchers have taken advantage of this preprocessing to conduct controlled experiments to better understand when words are skipped [68], how acronyms and initialisms are processed differently [73], and what parts of words are processed first (e.g. vowels [9] and syllable structure [7, 5]). Above the individual word level, researchers have used eye tracking experiments to show that *and*-conjoined constituents are processed more quickly when they have parallel syntactic or semantic structure [40], and that pronouns that are inconsistent with previously read causality verbs slow down reading [41]. Bicknell and Levy analyze between word regressions (eye movements from one word to the previous word) and find evidence supporting their theory that regressions occur when the current word does not match the previously-processed words [16, 17] in terms of syntactic agreement or semantically likely phrases.

The majority of studies on reading have used skilled adult readers as subjects, but there are a few exceptions. Ashby and colleagues found that the significance of predictability on processing time was different for highly skilled readers than it was for average college readers [8]. Kemper and Liu compared the regressions of older and younger adults with a research focus on understanding declines in working memory due to aging [37]. Rayner and colleagues [67] found that poor readers had a smaller perceptual window than skilled readers; that is, they processed a smaller number of characters at a time. Weger and Inhoff [86] found that skilled readers had longer regressions during reading than unskilled readers.

Most relevant to this work, Rayner et al. [66] found that fixation duration and number of regressions both correlated well with subjective human-judged document difficulty. Inconsistencies in anaphoric resolution led to increased fixations and regressions. Based on their results, they suggest eye tracking as a way to measure an individual reader's comprehension

process without having to activate oral reading skills, which they point out are different than silent reading skills, especially for older children and adults. Green [31] examined gaze tracking in what he calls "temporarily syntactically ambiguous" sentences: sentences for which there is some ambiguity with respect to complement structure before the whole sentence has been processed. He found readers had long regression durations at these points.

While the insight into reading processing and comprehension gleaned from eye tracking experiments is substantial, such experiments remain relatively difficult to conduct. The use of specialized equipment means that subjects must be brought to an on-site testing facility. In contrast, oral readings of passages can be collected by a researcher with a laptop, or even through recordings by phone or over the Internet. Further, research shows that both word level [6, 5] and syntactic and semantic [42] processing in silent reading are influenced by prosodic features that are directly observable in oral readings, and that fluent reading is a necessary precursor to comprehension [64]. One of the primary goals of this thesis is to provide a means to capture information about reading through just oral readings, by providing an analysis of how acoustic cues of reading difficulty relate to the cues commonly used in eye tracking studies.

Chapter 3

VARIABILITY IN HUMAN SIMPLIFICATION

As discussed in the previous chapter, many NLP tasks have relied heavily on human-generated gold-standard examples to train and test automatic models. With sufficiently consistent human simplifications, a similar approach could be undertaken for evaluating text difficulty, using a process akin to the Pyramid Method for summarization. In this chapter, we present a new corpus of parallel human simplifications. We highlight specific features of how humans perform simplification that pose challenges for automatic comparison of human and automatic simplifications, which motivate the empirical studies described in detail in Chapters 4 and 5.

The rest of this chapter describes and analyzes a new corpus of parallel simplifications. Section 3.1 describes the documents used in the study, while Section 3.2 describes how the simplifications were created. Section 3.3 examines the resulting simplification corpus. Section 3.4 summarizes the findings of the chapter and how this analysis motivates the work in the rest of the thesis.

3.1 Data

To better understand the level of consistency and source(s) of variation in human simplifications, we collected multiple parallel simplifications of six articles. Two were Wikipedia articles with comparable articles in Simple Wikipedia, two were articles from Encyclopedia Britannica, and two were CNN news articles. The characteristics of the texts are summarized in Table 3.1, and the full text of each original document is included in Appendix A.

3.2 Study Design

Graduate students from the Human Centered Design & Engineering department who had expertise in technical writing were recruited to simplify the texts one sentence at a time

Title	Source	Number of Words	Number of Sentences
Managua	Britannica	314	17
Manila	Britannica	478	33
Condors	CNN	262	13
WTO	CNN	335	19
Pen	Wikipedia	953	52
PPA	Wikipedia	484	21
Total		2826	155

Table 3.1: Characteristics of the documents used in parallel manual simplification study

using a custom web-based interface. Participants were paid for each document that they simplified, and were given an extra incentive for finishing all six documents. Between six and eight parallel simplifications were obtained for each text.

Wikipedia offers a set of instructions for authors of its Simple English Wikipedia to help them write for the intended audience of that site. We gave our participants an abridged, clarified version of those guidelines. The guidelines included a set of examples of simplified sentences, along with general guidelines:

> These articles should be accessible to a variety of readers who may find aspects of English difficult, including people whose first language is not English, children who don't have much knowledge of English, or readers who have learning difficulties. You should avoid contractions and idioms, which might be more difficult for these types of readers to understand. The language should be simple, but the ideas don't have to be; don't make the articles so short that they offer little useful information.

In particular, we followed Wikipedia's guidelines in defining the audience to include second language learners, children, and low-literacy adults. Participants were instructed that they could add extra explanation to a sentence if they felt it was appropriate, and that they

could remove parts of a sentence that they felt were unnecessary and difficult to explain. Participants had access to the entire document while they were simplifying, but entered their simplifications one sentence at a time so that we would have alignments of their simplifications to the original sentences. They could enter multiple sentences as the simplification for one original sentence, so we were able to capture one-to-many simplifications.

3.3 Analysis

In this section, we analyze the resulting sets of simplified sentences. In particular, we look at features of the simplifications that present a challenge to existing NLP evaluation methods based on gold-standard annotations, and which motivate our work in the following chapters.

3.3.1 Overview of Simplifications

Dropped sentences were relatively infrequent in our dataset. Of the 155 sentences, 109 were not omitted completely by any participants. Of the remaining 46 sentences, 33 were omitted by one participant, 10 were omitted by two participants, and 3 were omitted by all participants.

Table 3.2 shows the number of sentences that were split into two or more sentences by a given number of participants. Participants agreed on 34/155 sentences having no split. There is a wide variation for the remaining sentences, with 67/155 of the sentences being split by between two and four participants, suggesting significant disagreement between the participants.

Table 3.3 shows the number of sentences that had simplifications that were the same length as the original sentence for a given number of participants. 48 sentences were not kept the same by any participants. Most sentences (130/155) were kept the same by two or fewer participants, and only two were kept the same length by at least 5 participants.[1] This suggests that the participants made changes to most sentences, either because they thought the sentences needed to be simplified or because they did not feel comfortable leaving sentences unchanged. Figure 3.1 shows how many sentences were made shorter and longer

[1]The test was only for length in terms of word count, so some cases where lengthe is the same may have involved word edits.

Number of Participants who Split	Number of Sentences
0	34
1	33
2	27
3	23
4	17
5	11
6	9
8	1

Table 3.2: Distribution of sentence splits in hand simplified data set

by different numbers of annotators. For both shortening and lengthening, the majority of sentences were changed by between 2 and 4 participants. In fact, 76 of the sentences (or just under half) were shortened by at least two participants *and* lengthened by at least two other participants. Overall, this suggests low agreement between the participants with respect to the best way to improve the sentences.

3.3.2 Lexical Variability

First, we look at words that are in the original sentences but that are removed from all (or from a majority) of the simplified sentences. These words would be ones that our study participants agreed were difficult and needed to be changed during the simplification process.

Figure 3.2 shows all of the simplifications of the sentence "When the police took back the streets, the images were just as ugly." In this sentence, the phrase "the images were just as ugly" is an abstract way of saying that violence occurred, and this corresponds to a less frequently used sense of the word "ugly." Our annotators recognized that this construction would likely be difficult, and six out of seven annotators changed that phrase in some way. There was substantial variation in *how* the annotators simplified the sentence, however.

Number of Participants who left length unchanged	Number of Sentences
0	48
1	61
2	29
3	10
4	5
5	2

Table 3.3: Distribution of unchanged length in hand simplified data set

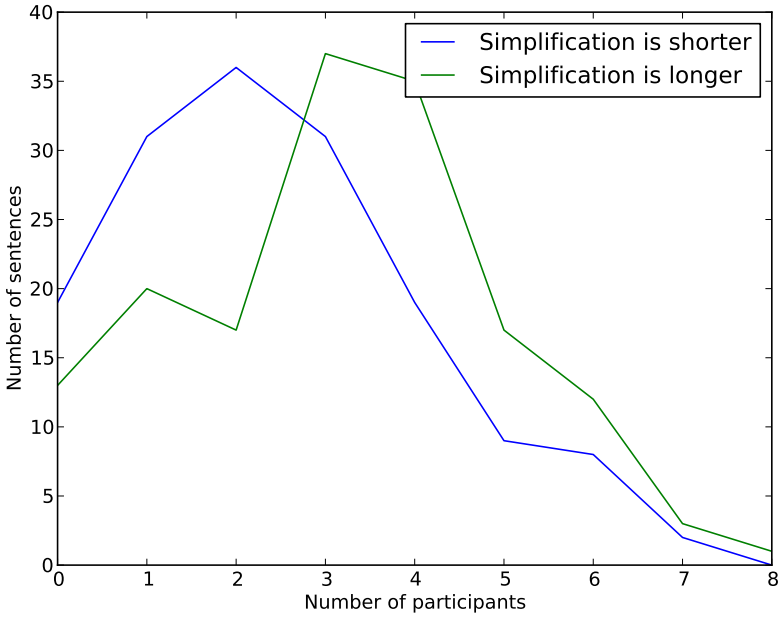

Figure 3.1: Histogram of sentence shortening and lengthening during simplification

Two used a form of "violent" in their simplification, and two used "bad," which is an easier but far less specific word. One (Simplification D) removed that clause entirely, while one (Simplification G) said that "people did not like" the police behavior, but did not describe

Original When the police took back the streets, the images were just as ugly.

A When the police took back the streets, the images were just as ugly.

B When the police gained control, a lot of violence still happened.

C The actions of the police were also violent.

D The police tried to regain control from the protesters.

E When the police took back the streets, the situation was just as bad.

F And when police confronted the protesters, the scene was just as bad.

G The violent protesters and anarchists were a problem. The police needed to gain control of the streets. The police needed to stop the violence and the anarchists. People did not like the ways that the police restored order in the streets.

Figure 3.2: Simplifications of "When the police took back the streets, the images were just as ugly." from a CNN article about the WTO protests

the behavior itself.

This sentence also shows a wide variation in handling of sentence-level simplification. One annotator (Simplification A) decided that the sentence was sufficiently easy in its original form, and did not make any changes. At the other extreme, one annotator (Simplification G) split the sentence into four different sentences, adding a lot of additional background information. The other annotators (Simplifications B-F) all made changes but kept the output at a single sentence. Across all documents, only two sentences were kept in their original form by all annotators. Nearly half (100 out of 198) were changed by at least one annotator, which could be a consequence of the annotators feeling like they needed to do *something* to improve every sentence. At the other extreme, 43 sentences were omitted by all annotators, and 110 were not omitted by any annotators. The remaining 45 sentences were omitted by one or two annotators, but kept by the rest.

Figure 3.3 shows all of the simplifications of the sentence "This aspect of the successful CSI program was just recently opened for applications." It shows a common simplification

Original This aspect of the successful CSI program was just recently opened for applications.

A This aspect of the successful CSI program was just recently opened for applications.

B Buyers can now apply for the CSI program.

C (Sentence omitted completely)

D Power providers can now submit applications to be part of the MASH program.

E The MASH program was just recently opened for applications.

F This aspect of the successful CSI program was recently opened for applications.

G People can now apply for a solar PPA under the MASH program.

Figure 3.3: Simplifications of "This aspect of the successful CSI program was just recently opened for applications." from English Wikipedia

operation, in which noun phrases are replaced with more direct, clear nominals. This change occurs in changing pronouns to nominals, but also in phrases with demonstratives. In this example, the noun phrase "this aspect of the successful CSI program" was replaced by its name ("the MASH program") by three out of seven annotators. Of the remaining four annotators, two left the sentence (effectively) unchanged and one removed it from the document completely. The final annotator kept the "CSI program" part but got rid of the "aspect" part, which changes the meaning of the sentence.

Looking at individual word removals, the words from the original sentence that were removed by the most annotators were *this*, *aspect*, *successful* and *just*. None of them were replaced with synonyms, suggesting that they were not removed because the annotators thought they were difficult lexical items. Instead, "this" and "aspect" were removed as part of the nominal change described above, while "successful" and "just" were removed completely.

Across all documents, the words that are most commonly removed by at least half of the annotators (*and, of, to, as, with, it, a, on, the, in, is, but, by, has, its*) are all function

Original A number of banks are based in Manila.

A A number of banks are based in Manila.

B Several banks are based in Manila.

C Manila is home to many banks.

D A number of banks keep their main offices in Manila.

E There are multiple banks that have their main office in Manila.

F A number of banks are based in Manila.

G Many different banks are located in Manila.

H Several banks are based in Manila.

Figure 3.4: Simplifications of "A number of banks are based in Manila."

words. This pattern shows that we cannot simply look for words that are commonly removed in simplification to identify difficult words, and informs the log difference model that we develop in Chapter 6 for describing word difficulty.

The document that the sentences in Figure 3.3 came from also showed us that topic familiarity could strongly affect the simplifications that we got from annotators. It was about power purchase agreements (PPAs), a topic related to electricity economics. One annotator admitted that she did not fully understand the article, which made it difficult to simplify. This uncertainty is evident in the wide variety of simplifications for this sentence, including one that changes the meaning of the original sentence.

In an encyclopedia article about Manila, the annotators found several aspects of the sentence "A number of banks are based in Manila" to simplify. Their simplified sentences are shown in Figure 3.4. The phrase "a number of" was replaced by "several", "multiple", or "many" by five out of eight annotators. The word "based," in the context of a company's location, was identified as difficult enough to change by half of the annotators. They all used different replacement wordings. Two specifically used the phrase "their main offices" to explain what it means for a company to be in a city while one just used the word

"located." The fourth switched the subject and object of the sentence, resulting in Manila being "home to" the banks. By many basic difficulty measures, "based" should not be a hard word. It is only five characters long, and is relatively common. It also has many senses, though, including four very different verb senses in Wiktionary. The sense used in this sentence is the least common and most abstract of the four, which makes this sentence a nice illustration of how word senses can interact with word decodability and frequency to affect difficulty.

The simplification from Annotator E is worth further attention. It replaces both of the potential difficulties identified by other annotators ("a number of" becomes "multiple", while "based in" becomes "have their main office in"). It additionally introduces an existential structure, though, that is not clearly motivated. Guidelines for formal English writing generally discourage this form of construction, and we don't have any evidence that it will be helpful to a reader. At least one writing expert felt it would be necessary, though. Without an empirical way to characterize difficulty, we have no good metric for confirming or denying their belief.

3.4 Conclusions

This chapter presented a new collection of parallel, human-generated, sentence-aligned simplifications. We found that given instructions like the ones used for Simple English Wikipedia, expert writers will generate very different simplifications, and show low agreement on what should be simplified. Most simplifications were more complicated than simple lexical replacement, making them difficult to extract and analyze automatically. Overall, the study motivates our work to assess difficulty empirically through struggling readers' interactions with texts, which we describe in the following chapters.

Chapter 4

LARGE-SCALE PARALLEL ORAL READINGS

By looking at word-level phenomena, this chapter develops a fine-grained analysis of oral reading. With this analysis, we can identify points of difficulty in a text. We focus in this thesis on common difficulties for a group of readers, but also look briefly at characterizing the reading of individuals. A side-benefit of the latter analysis could be a more nuanced identification of the different kinds of reading difficulties that an individual could face.

Currently, reading ability is commonly measured through words correct per minute (WCPM) on standardized texts. For the readers that we are interested in, actual reading errors (e.g. skipped or mis-read words) tend to be fairly infrequent, which means that reading rate and pausing are important contributors to WCPM. However, both can vary for communicative reasons as well. Fluent readers will commonly extend the duration of the final syllable of a major phrase or at a sentence boundary. Pausing is used for similar purposes. Since pausing and drawn-out articulation duration can both be related to difficulty *or* to prosodic boundaries in good communicative reading, we must control for boundary locations in our analysis.

This chapter looks in detail at local acoustic cues of reading difficulty. Looking at specific words and boundaries between words, we explore pauses, duration lengthening, hesitations, and reading errors using data from the Fluency Addition to the National Assessment of Adult Literacy (FAN), controlling for communicative structure. Sections 4.3 and 4.4 are based on previously published results [49], though all models have been retrained for the results in this chapter.

This chapter begins by describing the FAN data in Section 4.1. Then, the acoustic measurements we are interested in are described in Section 4.2. Section 4.3 describes two different approaches for handling the communicative factors that affect our acoustic measures. Section 4.4 analyzes our acoustic measures, controlling for the communicative factors

Title	Number of Words	Number of Sentences	Lexile Score
Amanda	155	12	700
Curly	153	14	380
Grand Canyon	166	17	570
Guide Dogs	156	13	700
Bigfoot	186	12	1020
Chicken Soup	153	10	1100
Exercise	183	11	1020
Lori Goldberg	156	8	1030

Table 4.1: Features of the eight passages used in the FAN study

addressed in Section 4.3. We use those features to predict the reading level of individuals in Section 4.5, before summarizing our conclusions in Section 4.6.

4.1 Fluency Addition to the National Assessment of Adult Literacy (FAN)

4.1.1 Oral Readings

Our study of hesitation phenomena involves empirical analysis of the oral reading data from the Fluency Addition to the National Assessment of Adult Literacy (FAN). The data collection effort was conducted by the Department of Education in 2003, with the goal of characterizing the literacy level of American adults [10]. Researchers visited participants at their homes and in prisons, and each participant completed a set of tasks to measure literacy.

The FAN's 13,000 participants were chosen to reflect the demographics of adults in the United States; thus, speakers of varying reading levels and non-native speakers were included.

FAN participants read one of four fourth-grade passages and one of four eighth-grade passage out loud [10]. The eight passages used as part of the study are included in Appendix B. The properties of the texts are summarized in Table 4.1 [32]:

4.1.2 ASR and Automatic Scoring

For each recording in the FAN data, the WCPM is computed automatically using Ordinate's VersaReader system to transcribe the speech given the target text [11]. VersaReader generates a separate language model for each passage that is tuned to give higher likelihood to the words in the text, while allowing a larger vocabulary to capture misreadings. Fragments of words in the text are also part of its vocabulary, allowing it to correctly transcribe word restarts (e.g. *esc- esc- escapades*). Word and phone-level forced alignments are provided to this project by Pearson Knowledge Technologies. (Phones correspond to the vowel and consonant sounds that would be used to specify the pronunciation of a word. We use the term "phones" rather than "phonemes" since the inventory of sounds used in speech recognition systems is generally slightly bigger than the standard set of "phonemes" used by linguists.)

The system output is automatically aligned to the target texts using the track-the-reader method of Rasmussen et al. [65], which defines weights for restarts and skipped words, and then identifies a least-cost alignment between the ASR output and a text. This alignment associates each spoken word with a word in the target text, which can be used to identify locations of reading errors word restarts, and to compute words correct per minute.

At the document level, reading scores derived from ASR output correlate very highly with manually-calculated scores; automatic calculation of WCPM has high correlation (.96-1.0) with human judgment of WCPM [12]. Since we are interested in word-level features, not just averages across documents, the accuracy of the underlying ASR system is important. The greatest problem with speech recognition errors is for very low-level readers [12]. In order to have more reliable time alignments and WCPM-based scores for individuals, participants whose BRS score was labeled "Below Basic" in the NAAL reading scale were removed, along with participants with missing or incomplete (less than a few seconds) recordings. With these exclusions, the number of speakers in our study was 11,627.

Word Error Rate (WER) for the ASR component in Ordinate's prototype reading tracker [12] may be estimated to be between 6% and 10%. In a sample of 960 passage readings, where various sets of two passages were read by each of 480 adults (160 native Spanish speakers, 160 native English-speaking African Americans, and 160 other native English speakers), the

Ordinate ASR system exhibited a 6.9% WER on the 595 passages that contained no spoken material that was unintelligible to human transcribers.

For the experiments in this chapter, we have access to word and phone times from Ordinate's speech recognition system, but not to the audio files themselves.

4.2 Acoustic Measurements

In this section, we describe the acoustic measures that we use to identify difficulties, and the normalization that we use to account for variations between individuals.

4.2.1 Duration Normalization

WCPM is generally used as a tool for assessing reading level by averaging across one or more passages. It is more noisy when comparing the readability of different texts, especially when the reading level is measured at a fine-grained (e.g. word) level. Some words are longer than others, and some phones have longer average durations than others. If longer words take longer to read orally, it may be merely a consequence of having more phones, and not of additional reading difficulty. In addition, differing speaking rates between individual readers can affect phone durations.

Normalizing for duration differences between phones

For word-level measures of lengthening, we use the standard z-score method of normalizing to account for inherent phoneme duration differences: measured duration minus phone mean divided by phone standard deviation. For the mean and standard deviation of each phone, we use speaker-independent phone statistics computed from the TIMIT Corpus,[1] which has hand-marked phonetic labels and times.

Normalizing for speaking rate differences between participants

We adopt the model used by Wightman et al. [87] to account for speaking rate in normalizing phone duration. In their model, phone durations are characterized by a Gamma distribution,

[1] Available from the Linguistic Data Consortium.

and speaker variability is characterized by a linear scaling of the phone-dependent mean parameters, where the scaling term is shared by all phones. The linear scale factor α for a speaker is estimated as:

$$\alpha = \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{\mu_{p(i)}} \tag{4.1}$$

where d_i is the duration of the i-th phone which has label $p(i)$ and where μ_p is the speaker-independent mean of phone p. We make use of the speaking rate model to adjust the speaker-independent TIMIT phone durations to the speakers in the FAN corpus by calculating the linear scale factor α for each speaker.

Thus, the phone mean and standard deviation used in the z-score normalization are $\alpha\mu_{p_i}$ and $\alpha\sigma_{p_i}$, respectively.

4.2.2 Specific Measurements

Given phone times that have been normalized for speaking rate variation and for inter-phone duration differences, we can look at a variety of measurements from transcripts of oral readings that will help us to identify points of reading difficulty:

Word Lengthening The mean normalized duration (as z-scores) of all the phones in a word.

Final Rhyme Lengthening The mean normalized duration (as z-scores) of the phones in the rhyme of the final syllable of the word. The rhyme of a syllable includes the vowel and the following consonants.

Pauses The presence and duration of pauses in oral recordings. We discard all pauses with a duration less than 200ms, since they tend to be unreliable. We identify *short pauses* as ones with a duration of at least 200ms, and *long pauses* as ones with a duration of at least 500ms.

Verbal Hesitations For the FAN study, verbal hesitations refer to non-silent pauses before a reader attempts a word. In particular, they include both filled pauses (e.g. "um") and word restarts (e.g. "esc- esc- escapades"). We note that this is not a standard use of the term "hesitations," but we use it here to be consistent with the terminology in Downey et al. [25].

4.3 Accounting for Communicative Factors

Both the lengthening and the pause features described above are sensitive to communicative factors. In fluent reading, pauses and slow average articulation rates tend to coincide with major prosodic phrase boundaries for communication reasons. In our work, we would like to account for prosodic context in using articulation rate to identify difficult words and constructions. Here, we consider two ways of characterizing prosodic context: first, with a model trained to predict prosodic boundaries from text features, and second, from the reading behavior of skilled readers on the specific texts we are interested in.

4.3.1 Using the text to predict appropriate pausing and lengthening

We trained a boosted decision tree classifier using icsiboost [28], an implementation of BoosTexter [69], on hand-annotated data from the Boston University Radio News Corpus [60] to predict the locations where we expect to see prosodic boundaries. Each word in the Radio News Corpus is labeled with a prosodic boundary score from 0 (clitic, no boundary) to 6 (sentence boundary). For each word, we use features based on parse depth and structure, POS bigrams, and word and POS break ratios [45] to predict the prosodic boundary value for the boundary following that word.

For evaluation, the break labels are grouped into: 0-2 (no intonational boundary marker), 3 (intermediate phrase), and 4-6 (intonational phrase boundary). Word boundaries with 0-2 breaks are considered non-boundary locations; 4-6 are boundary locations. We expect that, for fluent readers, lengthening and possibly pausing will be observed at boundary locations but not at non-boundary locations. Since the intermediate boundaries are the most difficult to classify, and may be candidates for both boundaries and non-boundaries for fluent readers, we omit them in our analyses. Our model has a 7.2% confusion for the boundary/no-boundary decision. The full confusion matrix is show in Table 4.2. Our model has slightly more false alarms than misses in detecting boundaries. Since our analysis will focus on reader behavior at predicted non-boundary locations, this is a satisfying result; most of the locations our model identifies as non-boundaries are correct.

Note that our 3-way prosodic boundary prediction is aimed at identifying locations where

	Ref		
hyp	No Boundary	Intermediate	Boundary
No Boundary	803	37	12
Intermediate	54	45	31
Boundary	69	53	244

Table 4.2: Confusion matrix for predicting prosodic phrase boundaries in the Radio News Corpus

fluent readers are likely to place boundaries (or not), i.e., reliable locations for feature extraction, vs. acceptable locations for text-to-speech synthesis. Because of this goal and because work on prosodic boundary prediction labels varies in its treatment of intermediate phrase boundaries, our results are not directly comparable to prior studies. However, performance is in the range reported in recent studies predicting prosodic breaks from text features only. Treating intermediate phrase boundaries as positive examples, Ananthakrishnan and Narayanan [4] achieve 88% accuracy. Treating them as negative examples, Margolis et al. [45] achieve similar results. Our model, though tuned for a different application, achieves 87% accuracy treating intermediate boundaries as positive examples and 89% accuracy treating intermediate boundaries as negative examples.

4.3.2 Using fluent readers to characterize appropriate pausing

An alternative to predicting pause and lengthening locations based on prosodic boundary information is to use empirical information about where skilled readers pause when reading a specific story. In particular, Cheng [21] stores non-parametric cumulative density functions (CDFs) for pause durations based on readings by native speakers. Given a new reader, they estimate the segmental probability of a given inter-word pause from that empirical distribution.

As suggested by Cheng's work, we characterize "appropriate" pausing at a given word boundary based on the behavior of the top 20% of our readers, where readers are sorted by

average WCPM. In particular, we compare the likelihood of a skilled reader pausing at a given boundary to our boundary predictions from the previous section.

4.3.3 Findings

Comparing an individual's reading to that of the average of many skilled readers has the advantage of accounting for the way that skilled readers actually read a text. On the other hand, predicting boundaries from text features has the advantage of being easy to apply to new texts, without needing a large-scale data collection to gather examples of new contexts. We want to ensure, however, that our boundary prediction model identifies the locations that skilled readers actually pause in our texts.

We find that fluent readers exhibit pauses of at least 200ms at 34.9% of predicted prosodic boundaries, but only at 1.9% of predicted non-boundaries. Pauses of at least 500ms occur for fluent readers at 14.8% of predicted boundaries and 0.5% of non-boundaries. Further, the majority of pausing by the top 20% of readers can be explained by prosodic structure: 91% of long pauses and 87% of short pauses occur at predicted boundaries. Thus, we conclude that our text-based model is successfully identifying locations that skilled readers are likely to pause. Given our model's advantage of being easy to apply to novel texts, we use its predicted "boundary" and "non-boundary" labels for the rest of our analysis. It is worth noting that the fact that strong readers do not always pause at predicted boundaries is not a problem for us, since many boundaries are cued by duration, which we also consider in our later analysis.

4.4 Identifying Difficulties

With our estimates of locations that pausing and lengthening are expected for communicative reasons, we can identify points in our texts where low-level readers commonly exhibit pausing and lengthening that are not prosodically motivated.

4.4.1 Pausing

As shown in the previous section, pauses align well with prosodic boundaries for fluent readers. For less-skilled readers, however, pausing may be a sign of uncertainty. Table 4.3 shows that non-fluent readers are much more likely to pause at non-boundary locations and, further, that their pauses at non-boundary locations are substantially longer than those of fluent readers.

	Boundary	No Boundary
Fluent Pause Rate $\geq$ 500 ms	15%	0.50%
Non-Fluent Pause Rate $\geq$ 500 ms	26%	7%
Fluent Pause Rate $\geq$ 200 ms	35%	2%
Non-Fluent Pause Rate $\geq$ 200 ms	41%	15%
Fluent Pause Duration (ms)	190	10
Non-Fluent Pause Duration (ms)	360	120

Table 4.3: Pausing behavior for top and bottom 20% of participants at predicted boundary and non-boundary locations

In the FAN collection, the following locations had the largest average preceding pause of all non-boundary locations for the bottom 20% of readers (by average words correct per minute):

1. One night I slept in a sleeping bag on the floor of my **...** grandma's front parlor.
2. I was secretly excited to be camping out there **...** because it would almost be like sleeping in a real forest minus the hard ground.
3. He likes to run and play with me **...** and he likes to follow my father around in the fields too.
4. At about that time, the fire went out and my **...** aching eyes dropped shut.
5. At about that time, the fire went out and **...** my aching eyes dropped shut.
6. Since dogs are **...** gentler when raised by a family, the dogs are given to children.

7. We have always been able to share our **...** escapades and humor with our friends.
8. At about that time, the fire went out **...** and my aching eyes dropped shut.
9. "I don't want the poor dears to freeze," she tells me **...** as I stare in awe at her rooms filled with greenery.
10. One night I slept in a sleeping bag on the floor of my grandma's front **...** parlor.

This list includes pauses before words like *aching*, *gentler*, *parlor*, and *escapades* that may be unfamiliar to less skilled readers. It also includes points where a low-literacy reader, who is processing smaller pieces of a sentence (and, likely, only one word at a time), might predict that a sentence boundary might occur, such as after "out" in *"At about that time the fire went out* **...** *and my aching eyes dropped shut."* That sentence is particularly interesting, since three of the most common pause locations occurred in it (*"At about that time, the fire went out* **...** *and* **...** *my* **...** *aching eyes dropped shut."*) We hypothesize that the first pause corresponds to a reader incorrectly predicting the end of a sentence. The second pause is caused by attachment ambiguity for the conjunction, and the third pause is related to lexical difficulty for the word *aching*. We see, then, that pauses are good indicators of difficulty, but that they do not distinguish between difficulty sources. It is worth noting that some of these cases (e.g. Sentences 3 and 8) would have been predicted to be boundary points if the text had included a comma. This variability in our boundary prediction method is motivation for using pause durations, rather than just pause presence, in identifying difficulty points.

4.4.2 Duration Lengthening

We also examine lengthening for all phones in a word and for just the phones in the final rhyme (from the last vowel to the end) of a word for the top and bottom 20% of readers. Table 4.4 shows that fluent and non-fluent readers have similar lengthening patterns at predicted prosodic boundaries. Non-fluent readers, however, show more lengthening at non-boundary locations.

We look at word and final rhyme lengthening for words that come before a non-break point, with the hypothesis that these words may be associated with different types of diffi-

	Boundary	No Boundary
Fluent Word Lengthening	0.18	-0.26
Non-Fluent Word Lengthening	0.16	0.06
Fluent Final Rhyme Lengthening	0.36	-0.36
Non-Fluent Final Rhyme Lengthening	0.58	-0.05

Table 4.4: Word and final rhyme lengthening for top and bottom 20% of participants, for words before predicted boundary and non-boundary locations

culties in the text. (For one syllable words, these scores will be the same.)

In the FAN collection, the underlined words in the following examples had the largest average word lengthening of all multi-syllable words preceding non-boundary locations for the bottom 20% of readers (by average words correct per minute):

1. At a b o u t that time the fire went out and my aching eyes dropped shut.
2. As a principal, Goldberg may be an a u t h o r i t y figure, but she doesn't want to instill fear in students.
3. The other day as L o r i Goldberg was walking through the halls of her school, a few whispers could be heard from the children she passed, but most greeted her with grins and hellos.
4. She brings as m a n y of her outdoor plants inside as she can for the winter.
5. Cold symptoms such as a runny nose and cough are thought to be caused by i m m u n e cells flooding into infected areas.
6. Call your doctor first if you are a man over forty or a woman over fifty and you plan to do v i g o r o u s activity instead of moderate activity.
7. The fancy exterior decorations on just a b o u t every building were carved from wood and then painted to look like stone or marble.
8. In the eighties an elderly man admitted he had made hoax B i g f o o t tracks for fifty years.
9. About one hundred thousand visitors hike along trails w i t h i n the canyon.

10. Cold symptoms such as a runny nose and cough are thought to be caused by immune cells flooding i n t o infected areas.

The list includes some words that could be expected to be difficult, like *vigorous* and *authority*. Some simple words (e.g. *within*, which is contrastive with the previous sentence) may be longer because of prosodic emphasis, which we are not controlling for. Word sense may also play a role in both instances where *about* is used with the "near to" sense. Other sources of difficulty may be related to difficult syntactic structures such as complex noun phrases, but we would need additional data to better distinguish between sources of difficulty.

The following underlined words had the largest average final rhyme lengthening of all words followed by non-boundary locations for the bottom 20% of readers:

1. She brings as ma n y of her outdoor plants inside as she can for the winter.
2. Plants and animals i n the area need this water.
3. Curly is my bi g black dog.
4. Cold symptoms such as a runny nose and cough are thought to be caused by immu n e cells flooding into infected areas.
5. Cold symptoms such as a runny nose and cough are thought to be caused by immune cells flooding into infected areas.
6. Then he remembe r e d what he had done, so he went back to the big oak tree.
7. Soon my father wanted something that was i n his coat pocket.
8. About one hundre d thousand visitors hike along trails within the canyon.
9. Interspersed in these residential areas were a variety of businesses - paint factories, lumberyards, distilleries, gasworks, mills, furniture manufacturers, warehouses, and coal distributors.
10. I was secretly excited to be camping out there because it wo u l d almost be like sleeping in a real forest, minus the hard ground.

Most of these difficulty points correspond to potential structural difficulties. A couple (*immune*, and *into* before *infected*) may correspond to either a difficult word or the word before a difficult word. We hypothesize that structural difficulties coule be a consequence of readers incorrectly predicting the location of prosodic boundaries, of complex noun phrases,

or of other difficult syntactic structures. Additional data is needed to be able to better distinguish between sources of difficulty.

4.4.3 Verbal Hesitations

We also look at where readers are likely to have verbal hesitations (filled pauses or word restarts). For the bottom 20% of FAN participants, verbal hesitations were most common before the bold words in the following sentences:

1. There is no shortage of **medicinally** active compounds such as vitamins in these ingredients.
2. Goldberg said being the principal gives her the ability to **be** part of a long term vision and larger goals for the school.
3. That sounds **illogical** because it is such a large place.
4. You can also check with local churches or synagogues, senior and civic **centers**, parks, or recreation associations for exercise wellness or walking programs.
5. You can also check with local churches or synagogues, senior and civic centers, parks, or recreation **associations** for exercise wellness or walking programs.
6. Bigfoot promoters believe it shows a genuine creature, and not a man in a **fur** suit as skeptics suspect.
7. We have always been able to share our **escapades** and humor with our friends.
8. The other day as Lori **Goldberg** was walking through the halls of her school, a few whispers could be heard from the children she passed, but most greeted her with grins and hellos.
9. A dog loves nothing better than to be with its master, and guide dogs keep their masters company all **the** time.
10. You can also check with local churches or **synagogues**, senior and civic centers, parks, or recreation associations for exercise wellness or walking programs.

The majority of these (*medicinally, illogical, associations, escapades, synagogues*) correspond to words that are either difficult (many of our participants did not seem familiar with the word *escapades*) or difficult to decode (*synagogues* has a difficult grapheme-to-phoneme

alignment for low level readers). There are still a few entries on this list that do not seem to be lexically motivated. The verbal hesitations at *to* in Sentence 2 correspond to structural confusion on the reader's part, for example. The verbal hesitation at *the* in Sentence 9 could correspond to readers expecting the word *of* to be part of that phrase.

4.4.4 Word Errors

While pausing and lengthening are hints that a reader is having difficulty, reading errors are a clearer (but rarer) signal of difficulty. Words that are commonly misread by readers are indicative of points where readers commonly struggle. In the FAN recordings, we count (but do not distinguish) the following types of errors:

- *Substitutions*: Incorrectly reading a word as another word
- *Deletions*: Skipping a word completely
- *Insertions*: Adding words that are not part of the text (excluding verbal hesitations)

These errors are relatively uncommon; only 1.5% of word utterances for fluent readers and 11.3% of word utterances for non-fluent readers in the FAN collection are errors. The most common error words are:

1. There is no shortage of **medicinally** active compounds such as vitamins in these ingredients.
2. Guide dogs, or seeing eye dogs, lead very interesting **lives**.
3. Five million people **visited** the Grand Canyon last year.
4. This is largely due **to** giant footprints discovered in mud or snow.
5. Cold symptoms such as a runny nose and cough are thought to be caused by immune cells flooding **into** infected areas.
6. Since dogs are **gentler** when raised by a family, the dogs are given to children.
7. That was three times the number of parking **places** there.
8. On the other hand, when Amanda **or** I have a rough time, we are always there for each other.
9. Soon my father **wanted** something that was in his coat pocket.
10. We have always been able to share our **escapades** and humor with our friends.

This list includes many of the difficulty points that corresponded to pauses and verbal hesitations, supporting our hypothesis that those acoustic cues are useful for identifying reading difficulty. It also includes words that are unlikely to be difficult for our readers (e.g. *to* in Sentence 4), which we hypothesize are the result of ASR system errors that are more likely for short words.

4.5 Identifying Low-Literacy Readers

Instead of trying to predict the difficulty of a word or passage, we can consider the task of trying to characterize a person's reading level. In this section, we look at how well we can predict what a person's reading rate in WCPM would be for a passage, based on how they read another passage.

4.5.1 Methodology

As part of the FAN study, each participant read one easy passage and one hard passage.

We consider three different experimental setups:

Same Only FAN participants who read the same two passages.

Any Easy Hold the hard passage constant, but consider all participants no matter which easy passage they read

Any Easy/Hard All participants, no matter which passages they read

In each case, we train three linear regression models to predict a person's WCPM on the hard passage given their reading of the easy passage. The first model has the person's WCPM on the easy passage as its only feature. The second uses the following features, extracted from each participant's recording of the simpler passage:

- Error rate
- Verbal hesitation rate
- Average final rhyme lengthening for words that are not followed by a prosodic boundary
- Standard deviation of final rhyme lengthenings across all words
- Average whole word lengthening for words that are not followed by a prosodic boundary

Configuration	WCPM	Acoustic Features	WCPM + Acoustic Features
Same	.78	.39	.80
Any Easy	.66	.38	.68
Any Easy/Hard	.52	.25	.53

Table 4.5: Variance reduction for predicting WCPM on a hard passage from a) the WCPM on a simple passage, b) acoustic features from a simple passage, or c) WCPM and acoustic features from a simple passage, for each experiment setup

- Standard deviation of whole word lengthenings across all words
- Short pause rate (pauses with duration >= 200ms)
- Long pause rate (pauses with duration >= 500ms)
- Total number of word attempts (including correct readings, errors, and verbal hesitations)

The third model has access to the reader's WCPM *and* the above features for their reading of the easy passage.

The first setup will tell us how reading correlates across two known passages. The second and third setups show how sensitive each measure is to the specific text(s) that participants read.

4.5.2 Results

The variance reduction R^2 is shown in Table 4.5 for each of the three experiment setups. In all three configurations, we get better performance using WCPM than we do from our acoustic features. We get a slight (but significant, $p < 10^{-5}$) improvement over the WCPM-only baseline by combining it with the acoustic features. One possibility is that the linear model is not sufficiently powerful. A non-linear predictor may yield better results.

4.6 Conclusions

We have shown in this chapter that pausing and lengthening in audio can be used to identify points where a reader has difficulty in oral reading when controlling for prosodic structure. These cues seem to signal both difficult words and structurally difficult points in sentences, though they seem to sometimes be the result of ASR error or prosodic emphasis, which could be addressed in future work. We have also used those features to relate individual readers' performance across passages. In the next chapter, we validate these observations with gaze data from an eye tracking study.

Chapter 5

CONNECTING ORAL READING TO GAZE TRACKING

In the previous chapter, we developed audio-based measures for characterizing the difficulty of texts at a very local scale. In this chapter, we want to verify that the audio features we found useful are related to features from gaze-tracking that are known to be related to cognitive load.

Eye tracking studies provide useful information about a reader's cognitive process. Participants' gazes are calibrated to an eye tracking device, which sits on the desk below a computer monitor. The device uses a camera to measure the angle of an ultrared light directed at a participant's eyes, and maps that measure to an (x, y) coordinate on the screen. By mapping the resulting sequence of coordinates to the location of words on the screen, we can obtain detailed information about the number of times that a participant looks at each word as well as how much time their eyes spend fixated on each word.

Knowing where readers spend the most time looking gives a valuable clue to where their cognitive load is highest. A down side to eye tracking, though, is that the studies are time-intensive and expensive to conduct. An additional challenge for our purposes is that eye tracking studies require that the participant attend a session in a laboratory. This can be both intimidating and logistically difficult for populations like the low-literacy adults that we are interested in. An audio-only collection, however, can be easily conducted in the field with a laptop and a high-quality microphone, which more easily allows future studies to be conducted at schools, libraries, non-profit education centers, or other places that are convenient for participants. Consequently, we conduct a smaller-scale eye tracking study on the same passages for which we have access to large-scale audio recordings from the FAN collection, with the aim of verifying that we can get equally valuable information from easier-to-conduct audio studies.

We expect that the difficult words and phrases identified in Chapter 4 will be more

reliably identifiable from gaze data. Since long fixations and regressions are both associated with reading difficulty [66, 31], we use them to gain insights into the relative utility of long pauses, verbal hesitations, and duration lengthening in the audio. Since skilled and unskilled readers show different gaze behaviors [8, 67, 86], we look in particular at how gaze and audio align for unskilled readers, who we expect to most benefit from difficulties being identified and, eventually, made simpler. To that end, this chapter identifies audio cues to gaze indicators of reading difficulty, and describes gaze-based signals of reading difficulty that are *not* captured by the audio measurements in this thesis.

We begin by describing the details of our eye tracking study in Section 5.1. Section 5.2 describes the processing that was done to assign fixations to their target words. In Section 5.3, we relate audio features from the FAN study to eye tracking features from our study. Section 5.4 looks in more detail at a new passage that was developed specifically for our eye tracking study, and Section 5.5 summarizes the findings.

5.1 Eye Tracking Data Collection

5.1.1 Methodology

The parallel speech and eye tracking data collection methods were developed and facilitated in collaboration with Jared Bernstein, Jennifer Balogh-Ghosh and Xin Chen at Pearson Knowledge Technologies.

Participants were recruited for our eye tracking study through university email lists, flyers posted in coffee shops and neighborhood centers, and through direct recruitment of students at a non-profit adult literacy center. All sessions took place in the LUTE lab at the University of Washington.

Participants were invited to campus to take part in a one-hour session. Each session included the following components:

Demographic Information

After consenting to participate in the study, each participant was asked the following demographic questions:

1. How old are you?
2. What language did you speak at home when growing up?
3. (If answer to (2) was not "English") At what age did you learn to speak English?
4. What language did you first learn to read and write?
5. (If answer to (4) was not "English") At what age did you learn to read English?
6. (If answer to (4) was not "English") At what age did you learn to write English?
7. What language do you usually speak now?

Our 74 subjects ranged in age from 18 to 63, with a mean age of 30.9. 59% of them spoke English as their first language, and 72% first learned to read and write in English. Of the 28 participants who did not learn to speak English first, 9 started learning English by age 5, 15 started learning English between the ages of 6 and 10, and 4 started learning English at age 11 or older.

The participants in our collection were, on average, more fluent readers than the FAN participant population. Two of our 74 participants did not successfully calibrate with the eye tracking equipment, so we were unable to collect gaze data from them. Of the remaining 72 participants, 37 were in the top 20% of the FAN population by WCPM, and 5 were in the bottom 20%.

Versant English Quick Screener

Participants next completed the Versant English Quick Screen [1], a telephone-based test of English listening and speaking ability.[1] The screening took approximately 10 minutes to complete. The screening included two parts. In the first, participants repeated 17 sentences verbatim. Participants with better English language skills are able to keep longer and more complex sentences in their working memory, and so can more accurately repeat sentences. In the second part of the test, participants answered 8 short-answer questions orally. These questions tested the participants' understanding of spoken English, and not any specific domain knowledge. For example, a participant could be asked, "Would you get water from a bottle or a newspaper," and would answer "A bottle" or "From a bottle." Participants

[1]We are grateful to our colleagues at Pearson Knowledge Technologies for providing the English screening test.

were instructed to remain silent or to say "I don't know" if they didn't know the answer to a question.

The Versant Quick Screen reports scores on a scale from 20 to 80. Of our 74 participants, all but 5 scored a 69 or above, corresponding to a "Proficient User" or upper level "Independent User" of English [1].

Eye Tracking Setup

In the second part of the study, participants read two lists of real words, two lists of nonsense words, and ten passages out loud from a computer screen. Their gaze was tracked through a Tobii X120 eye tracker, and gaze points were grouped into fixations using the Tobii Fixation Filter [59]. The eye tracking system was calibrated to each participant by having the participant follow a red dot that moved to nine locations on the screen. Participants were prompted through the study by automated delivery of recorded voice prompts via a telephone call-in system run by Pearson. The same Pearson system was used to record audio and to generate time-aligned ASR output comparable to the transcripts used in the previous chapter. All word lists and passages fit on a single screen, so participants did not need to scroll or change pages.

Participants were instructed to read the passages out loud for comprehension. In particular, they were told to pretend that they were reading the stories to someone else, and that they wanted the listener to understand the story. They were explicitly told that their goal should not be to speed read.

Survey Questions

After each passage, participants answered the following questions through a web form:

- A multiple-choice comprehension question (different for each passage)
- Was this story INTERESTING?
  Very much / Some / Not much / Not at all
- Was this story EASY?
  Very much / Some / Not much / Not at all

- Do you know much about the things in this story?
  Very much / Some / Not much / Not at all

The comprehension questions for each passage are included in Appendix C. These questions were not graded, but were included to encourage participants to read for understanding.

Collecting difficulty, interest, and familiarity information for each participant gives us the opportunity to better understand how the participants perceive the texts that they are reading, and the level of motivation that readers may have felt to read each text. Table 5.1 shows the average response for each passage, with "Not at all" corresponding to 1 and "Very much" corresponding to 4 for all questions. For comparison, the document features (word count, sentence count, and Lexile score) from Chapter 4 are repeated here. We did not observe a significant difference in participants' subjective difficulty ratings of the easier and more difficult passages. As the results show, however, the passages rated least "interesting" were also rated the least "familiar," which suggests that participants' judgments for the three questions may have been related to one another.

Title	**Interesting**	**Easy**	**Familiar**	**# Words**	**# Sentences**	**Lexile**
Amanda	2.4	3.6	2.6	155	12	700
Curly	3.0	3.8	3.4	153	14	380
GrandCanyon	3.2	3.4	2.9	166	17	570
GuideDogs	3.4	3.4	2.8	156	13	700
Bigfoot	3.3	3.3	2.8	186	12	1020
Chicken Soup	3.5	3.4	3.0	153	10	1100
Exercise	3.1	3.5	3.4	183	11	1020
Lori Goldberg	2.7	3.5	2.6	156	8	1030

Table 5.1: Average self-reported interest, ease of reading, and familiarity for participants for each passage

Eye Tracking Features

Using eye tracking equipment, we can precisely identify the length and location of fixations and regressions during the reading of a word list or passage. The following measures are commonly used [19]:

First Pass Time The length of the first fixation on a word (whether it is fixated again later or not)

Regression Path Time The total time from the first fixation on a word to the time when the eyes move rightward past the word

Total Time The sum of the lengths of all fixations on a word

Probability of a Regression The relative frequency of a first pass fixation on a word being immediately followed by a regression

The eye tracking equipment used in the studies described in this thesis has a fair amount of noise in its measurements, which make measurements like first pass time less reliable. For the experiments reported here, we make use of the total fixation time on each word as well as the probability of a regression.

5.1.2 Additional Texts

In addition to the texts included in the FAN study, participants in our data collection effort read two new passages. These passages were designed by Jennifer Balogh-Ghosh, and were included in our study to provide a point of comparison for studies that they were running at the same time. The full text of these passages is included in Appendix B. They had the following properties:

Title	**Interesting**	**Easy**	**Familiar**	**# Words**	**# Sentences**	**Lexile**
Chicago Fire	3.3	3.4	2.7	213	19	860
Grandmother's House	3.1	3.5	2.8	271	19	860

Table 5.2: "Chicago Fire" and "Grandmother's House" passages, used in the eye tracking data collection study

"Chicago Fire" Passage

The "Chicago Fire" passage is a 213-word passage designed to gradually increase in difficulty. The goal was to have a passage that most participants could read at the beginning, but that would incrementally introduce longer, less frequent words in more complex sentences. The structure of this passage allows us to analyze how individual readers' performance changes over the course of the passage. In particular, we can split the "Chicago Fire" passage into five segments:

	Sentences	Words	Syllables	Lexile Score
Segment 1	6	42	57	270
Segment 2	5	57	70	690
Segment 3	2	41	64	1210
Segment 4	2	38	54	1290
Segment 5	2	36	84	1350

Table 5.3: Segments of the "Chicago Fire" passage

A segment-level analysis of the passage is included in Section 5.4.

"Grandmother's House" Passage

Participants also read the passage "Grandmother's House." This passage was included in another data collection effort aimed at third and fifth grade children. Since the FAN data collection was limited to adults, including this passage in our collection will allow researchers to analyze differences between low-literacy adults and children.

5.2 Gaze Post-Processing

Noise in the gaze tracking data is contributed by measurement inaccuracy, calibration drift over the course of a session, and biological characteristics of the eye [36]. Observation of our data confirms the findings by Hyrskykari that vertical noise is a bigger problem

than horizontal noise in tracking reading progress. To account for vertical noise, we follow Hyrskykari's suggestion of identifying transfers from the end of a line to the beginning of the next line. By identifying locations in the gaze data where the x coordinate of one fixation is at the right side of the text and the x coordinate of the next fixation is at the left side of the text, we can align fixations to their intended row and, consequently, to their intended word. In our implementation, we followed a sliding window of 6 fixations. A row change was identified when the window contained at least two fixations within 500 pixels of the right side of the text and at least 3 fixations within 200 pixels of the left side of the text (with the pixel thresholds determined empirically), but were restricted to not occur more than once in a window. This allowed us to avoid spurious row changes from a single bad measurement.[2]

After assigning the Tobii-generated fixations to target words, as described in Section 5.2, there were frequently adjacent, short fixations that were associated with the same word. These fixations were grouped together into a single fixation with a duration equal to the sum of the durations of the component fixations.

5.3 Relating Gaze Data to Audio Data

In this section, we compare the audio features from the FAN collections to gaze-based features for the same passages in our own study. We define "fluent" and "non-fluent" readers to be ones with an average WCPM in the top and bottom 20% of FAN participants, respectively, after removing participants as described in Chapter 4. These cutoffs correspond to average WCPM reading of less than 164 word per minute ("non-fluent") and greater than 217 words per minute ("fluent"). Of the participants in our eye tracking study, 37 were fluent and 5 were non-fluent.

---

[2]Because of this vertical noise, we collected but do not use eye tracking results for the word lists that we use to predict lexical difficulty in Chapter 6. For those lists, participants were reading isolated words in columns, so we could not use the row tracking method to counteract the vertical noise in the gaze tracking measurements.

5.3.1 Regressions and Reading Rate

For our participants, we are interested in confirming Weger and Inhoff [86]'s finding of longer regressions for skilled readers. Table 5.4 shows the average number of regressions per word, the average regression length, and the average number of fixations per word for the top and bottom 20% of participants, for each story, sorted by the stories' Lexile scores. We see that for all but the easiest story, the top 20% of readers have more regressions per word than the bottom 20% of readers. We also see that for every story except one, less fluent readers have more fixations per word than fluent readers do.

Story	Lexile	Regressions/Word		Avg. Regression Length		Fixations/Word	
		Low	High	Low	High	Low	High
Curly	380	.07	.07	4.2	3.8	.90	.89
Grand Canyon	570	.07	.07	3.6	4.5	1.01	.94
Guide Dogs	700	.09	.07	3.6	4.5	.98	.94
Amanda	700	.12	.08	3.8	4.7	1.39	.94
Grandmother's House	860	.08	.06	4.0	4.6	1.05	.96
Chicago Fire	860	.07	.05	3.8	4.8	1.28	.99
Bigfoot	1020	.07	.08	3.7	4.4	1.09	.97
Exercise	1020	.08	.08	4.2	4.7	1.16	.95
Lori Goldberg	1030	.05	.09	3.6	4.6	.92	.98
Chicken Soup	1100	.08	.07	4.0	4.5	1.15	1.01

Table 5.4: Gaze features per word for top and bottom 20% of readers for each story

5.3.2 Relation of Gaze Data to Audio Cues

Here, we examine the question of what our gaze features look like for the points identified by audio cues in Chapter 4. We consider the same oral reading cues of difficulty as in that chapter:

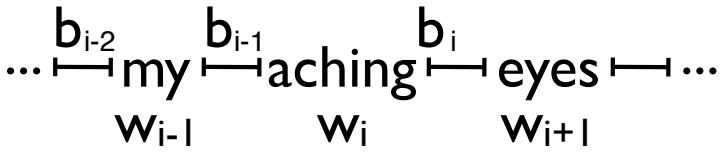

Figure 5.1: Word indices and boundaries for identifying reading difficulties

Word Lengthening Mean normalized phone duration for all phones in a word

Final Rhyme Lengthening Mean normalized phone duration for phones in the final rhyme of a word

Pauses We include all pauses of at least 200ms ("short pauses")

Verbal Hesitations Including filled pauses and restarts

We are interested in understanding how gaze and audio interact when a reader experiences difficulty. Figure 5.1 shows the context of a word w_i that might be a location of difficulty for a reader.

One possible type of difficulty would occur if the word w_i itself was difficult for the reader. In this case, we hypothesize that the reader will fixate for longer on w_i, will have longer lengthening for w_i, and will exhibit pausing and/or verbal hesitations at b_{i-1}.

Another type of difficulty would occur when a reader reached (but had not spoken) w_i and then realized that they were uncertain about the structure of the sentence, either because w_i does not match their expectations or because the structure is too complicated for them to immediately understand. In that case, we hypothesize that we will see longer lengthening for w_{i-1}, an increase in the likelihood of a regression from w_i, and/or a greater incidence of both pausing and verbal hesitations at b_{i-1}.

Thus, we are interested in making the following comparisons between audio and gaze features:

- Final rhyme lengthening for w_i and fixation duration for w_i
- Final rhyme lengthening for w_{i-1} and fixation duration for w_i
- Pause rate for b_{i-1} and fixation duration for w_i
- Pause rate for b_{i-1} and regression frequency from w_i

- Verbal hesitation rate for b_{i-1} and fixation duration for w_i
- Verbal hesitation rate for b_{i-1} and regression frequency from w_i

As discussed in Chapter 4, these acoustic cues can also be related to prosodic factors. Consequently, we eliminate from our analysis all words that might be affected by a prosodic boundary. In particular, if b_{i-1} is a predicted prosodic boundary location, then w_i is not considered for any analysis involving pausing at b_{i-1} or final rhyme lengthening on w_{i-1}. If b_i is a predicted prosodic boundary location, then w_i is not considered for any analysis involving lengthening of w_i.

Table 5.5 shows the average fixation duration on w_i and the average probability of a regression occurring right after a fixation on w_i for different acoustic conditions. We find that verbal hesitations at b_{i-1} are associated with longer fixations and an increased likelihood of a regression. Pauses at b_{i-1} correspond to longer fixations, but are not related to regressions. Lengthening of w_i and w_{i-1} are both associated with longer fixation durations, and lengthening of $w_{i=1}$ is also associated with an increased likelihood of a regression.

Acoustic Condition	**Avg. Fixation Duration**	**Regression Frequency**
w_i Rhyme Lengthening > 1.5	**423ms**	7.3%
w_i Rhyme Lengthening ≤ 1.5	**398ms**	7.8%
w_{i-1} Word Lengthening > 1.5	**448ms**	**9.4%**
w_{i-1} Word Lengthening ≤ 1.5	**399ms**	**7.4%**
b_{i-1} Pause > 200ms	**461ms**	8.3%
b_{i-1} Pause < 200ms	**397ms**	7.6%
b_{i-1} Pause > 500ms	**472ms**	8.2%
b_{i-1} Pause < 500ms	**400ms**	7.7%
b_{i-1} Verbal Hesitation	**537ms**	**13.5%**
b_{i-1} No Verbal Hesitation	**403ms**	**7.6%**

Table 5.5: Gaze features for words, grouped by acoustic features. Bolded numbers represent statistically significant differences ($p < .05$)

5.4 Analysis of Chicago Fire Passage

The "Chicago Fire" passage was designed and written by researchers at Pearson to be progressively more difficult as it goes from beginning to end. It starts with a very simple sentence ("I like people.") and gradually gets more complex; its last sentence ("This kind of industry, collocated within residential areas, was a recipe for disaster.") contains difficult words and a complex syntactic structure. Having access to gaze-tracking data for participants as they read through this passage, then, gives us a unique view into how their reading process changes. For all following results, we exclude boundary locations in the same way as described in Chapter 4.

Table 5.6 shows that as the passage gets more difficult, non-fluent readers exhibit more word-final rhyme lengthening. The exception is segment 4, which shows less lengthening than Segment 3. Similarly, fixations increase in average duration across the segments except for Segment 4, which has shorter fixations than Segment 3. The regression rate increases across segments, with Segments 3 and 4 being the same. Average regression length does not seem to correlate in any simple way with segment difficulty. The most difficult segment shows short, frequent regressions, but the other segments do not contribute to a clear pattern.

	Word Len.	Rhyme Len.	Pause Freq.	Fixation Dur.	Reg. Rate	Reg. Len.
Segment 1	-.13	-.02	.20	290	.11	6.7
Segment 2	-.07	0.18	.16	356	.12	8.1
Segment 3	0.01	0.84	.26	457	.15	7.1
Segment 4	-.07	0.45	.26	432	.15	7.5
Segment 5	0.40	1.18	.27	629	.17	4.7

Table 5.6: Reading behaviors for bottom 20% of participants by segment of Chicago Fire passage

Figure 5.2 shows final rhyme lengthening and fixation duration with error bars for each

segment for the bottom 20% of readers, showing that the variance in both values is substantial. This finding underscores the fact that not all words in even the difficult segments are difficult. Rather, the more difficult segments contain more points of difficulty interspersed with non-difficult words. Figure 5.3 reinforces this finding with a histogram of the final rhyme lengthening for the first, middle, and last segments. All of the segments have substantial mass around lengthening of 0. The most difficult segment, however, has more words with very large lengthening values, while the easiest two segments have the fewest words with very large final rhyme lengthening.

5.5 Conclusions

In this chapter, we have examined the interactions between audio and gaze-based cues of difficulty. Using the passages from the FAN study described in Chapter 4, along with two new passages, we conducted an eye tracking study with 72 participants. Each participant read ten passages out loud while their gaze was tracked. We then compared features based on the rate and duration of fixations on each word, and on the rate and length of regressions, to the features identified in the previous chapter.

We found that long fixations in the gaze data were associated with long pauses in the audio data that were not explained by prosodic phrasing. Verbal hesitations were correlated with an increased likelihood of gaze regressions. Lengthening was associated with increased regressions and longer fixations. Since long fixations and regressions are both associated with high cognitive load due to reading difficulty, these results provide evidence that some of the acoustic cues we extracted in the previous chapter are also related to difficulty. In the next chapter, we look at those acoustic cues more closely. In particular, we relate them to difficulty rankings extracted from text corpora to try to better understand how they relate to different kinds of word and sentence difficulties.

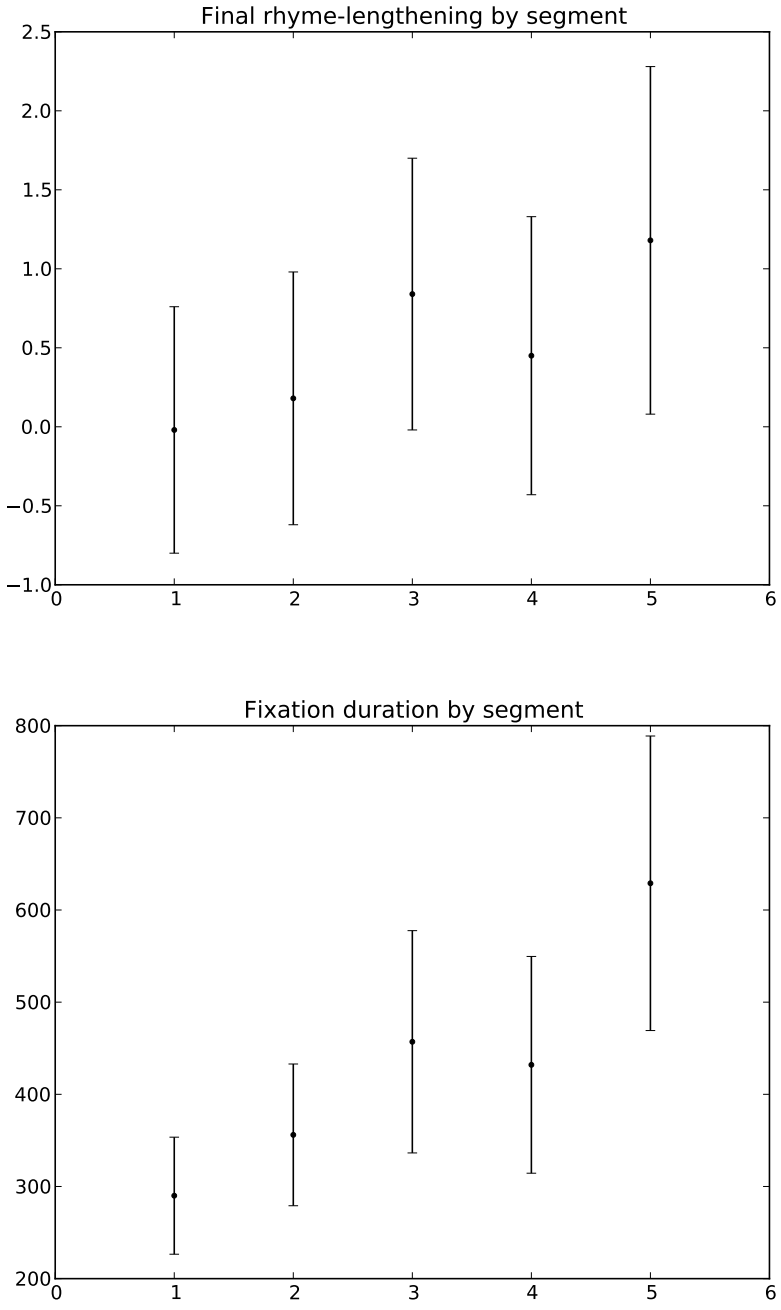

Figure 5.2: Average rhyme lengthening (top) and fixation duration (bottom) for each segment of the "Chicago Fire" passage for the bottom 20% of readers

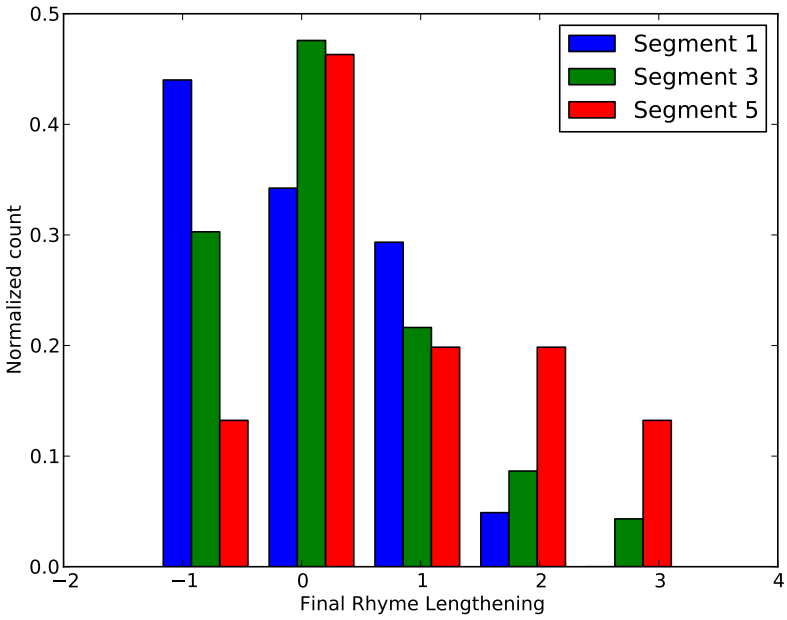

Figure 5.3: Histogram of final rhyme lengthening for words in the first, middle, and last segments of the "Chicago Fire" passage for the bottom 20% of readers

Chapter 6

PREDICTING TEXT DIFFICULTY

We would like to be able to predict the difficulty of texts automatically. Having a predictor of this type would be useful both in the context of machine learning algorithms (as a measure that could be optimized during automatic simplification) and in choosing appropriate texts for an individual. In this chapter, we look at two types of difficulty. First, we consider the difficulty of individual words, using lists of words and pseudo-words from the FAN collection. Next, we consider the difficulty of syntactic structures, and the task of predicting when the structure of a sentence should be simplified.

The rest of this chapter is organized as follows. In Section 6.1.1, we define a word difficulty ranking based on cues from oral reading, including errors, verbal hesitations, pauses, and lengthening. Section 6.1.2 uses a comparable corpus from Wikipedia to propose a text-based method for ranking words for difficulty, and report on how well that measure correlates to observations from oral reading. Section 6.1.3 computes word difficulty from a character n-gram model, and again compares to the ranking from oral reading features. In Section 6.1.5, we train a model to automatically predict word rank using features from text corpora. We turn in Section 6.2 to characterizing syntactic difficulty, before concluding with a summary of our findings in Section 6.3.

6.1 Word Difficulty

As we have seen, a binary distinction between "easy" and "hard" words can be problematic. Easy words (especially, but certainly not only, function words) are used in difficult documents. We will also see in this chapter that so-called "hard" words can be used in simple documents, when they are explained sufficiently and are relevant to the topic of the document. Further, some "hard" words will be harder than others. An early version of the study in this section defined two classes of words based on difficulty, and then analyzed

the features of the words in each class [47]. In this section, we instead characterize word difficulty with a continuous measure. As ground truth for the relative difficulty of individual words, we make use of human performance during reading.

6.1.1 Difficulty Ratings Based on Human Performance

One subtask of simplification is identifying individual words that are likely to be too difficult for a reader. Those words can be replaced with synonyms, explained in context, or changed as part of higher-order paraphrasing. Building off of our analysis from Chapter 4, we would like to characterize the difficulty of a word empirically, based on how often readers exhibit difficulty with the word. As we noted in that chapter, actual errors are uncommon, but verbal hesitations, lengthenings, and pausing can be indicators of difficulty as well. Here, we generate difficulty rankings based on several features:

$r_{e,i}$ The error rate of word w_i

$r_{p,i}$ The rate of short pauses ($\geq 200ms$) at boundary location b_{i-1} (preceding word w_i)

$r_{h,i}$ The verbal hesitation rate at b_{i-1}

$r_{l,i}$ The full word lengthening of w_i

As we noted in our analysis of the FAN passages, lexical and syntactic difficulties can interact, resulting in ambiguity as to the source of a verbal hesitation, pause, or error. In this section, we take advantage of the word lists that were collected along side the passages in the FAN collection. Participants read lists of words and pseudo- (nonsense) words out loud. Since the words were isolated, we do not have to worry about context or prosodic structure.

Excluding the first word in each of the word lists (since we cannot accurately measure preceding pauses for the first word), we examine 120 words and 83 pseudo-words from the FAN collection. For each ranking, we sort the words and label them from 1 to n, where n is the numer of words or pseudo-words. If two or more words have the same value for the feature being sorted on, they each receive a rank equal to the average of their positions on the list. For example, if the words ranked 10th through 15th all have the same value, then they each receive a rank of 12.5.

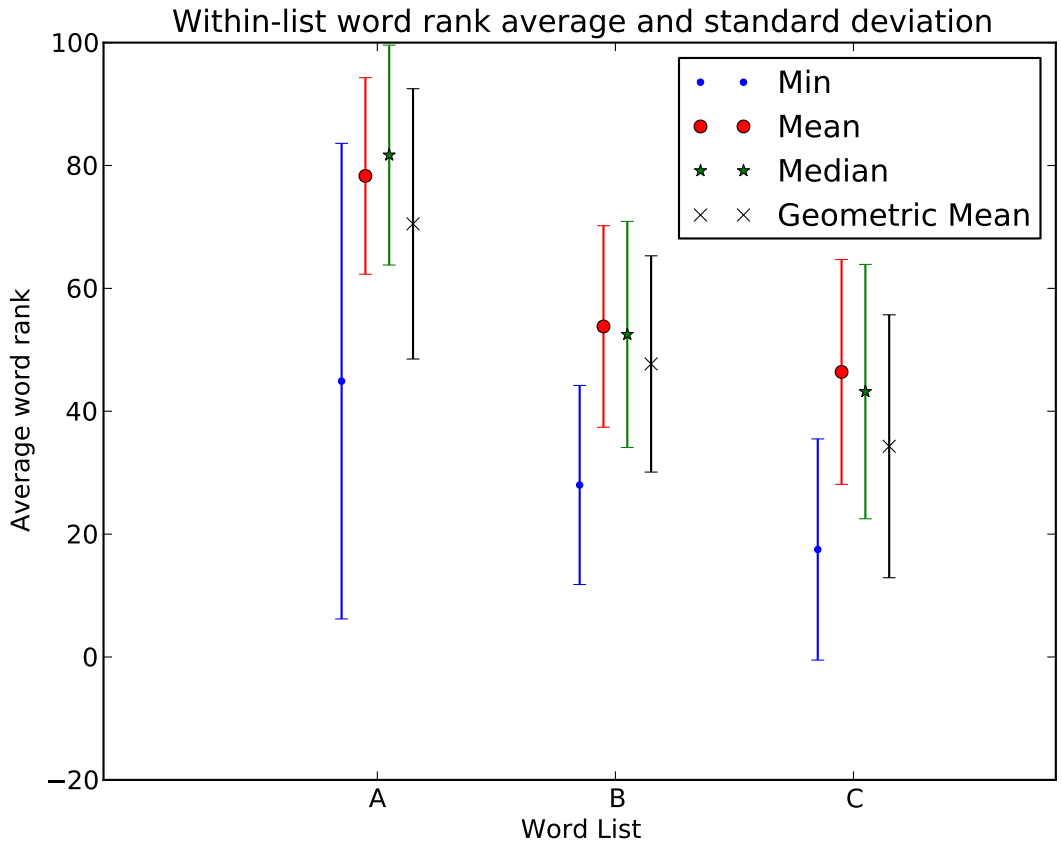

Figure 6.1: Mean and standard deviation of word ranks for each FAN word list, when final ranks are the mean, min, geometric mean, or median of the individual acoustic feature ranks

The FAN word lists are split into three separate lists, in increasing difficulty. Figure 6.1 shows the average rank (with standard deviation error bars) of the words on each list when the four ranks are combined by taking their mean, median, geometric mean, or min. All four methods sort the lists in the appropriate order, though with some overlap in the error bars. Given that, we choose to combine the ranks by taking their mean for the rest of the analysis in this chapter, since the means have the smallest within-list standard deviations. Our final difficulty ranking r_i of each word, then, is:

$$r_i = \frac{r_{e,i} + r_{p,i} + r_{h,i} + r_{l,i}}{4} \tag{6.1}$$

By this ranking, the ten hardest words in the combined FAN lists are *university*, *particular*, *carry*, *angry*, *easily*, *economic*, *primary*, *usually*, *seven*, and *likely*. The ten easiest words are *all*, *and*, *he*, *hill*, *in*, *due*, *each*, *do*, *job*, and *be*.

Next, we consider ways to predict this rank based on features extracted from text corpora.

6.1.2 Word Difficulty Cues from Comparable Corpora

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" [88]. 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. It is important to note that the reading difficulty of these simplified articles has not been verified by reading time analysis or other tests. We only know that they are believed to be easier to read by their 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.

Intuitively, we can think of "hard" words as ones that should be replaced in simplification. By looking at pairs of comparable Wikipedia articles, we can identify words that are commonly used in Standard English articles but not used in the corresponding Simple English articles.

We use a set of 22,923 pairs of comparable articles (articles on the same topic that are not necessarily equivalent in content) from the English and Simple English Wikipedia data sets. This set excludes topics that are marked as incomplete "stubs" in either language, as well as pairs where one of the documents is less than 50 words long.

As noted in Chapter 3, the most common words to be removed in simplification are function words due to the fact that many simplifications involve structural changes. Much like in topic modeling, then, we do not want to merely identify words that have high frequency. Rather, we want to identify words that are simplified more often than their overall frequency might suggest. Inspection shows that Wikipedia articles are noisy, and a large number of documents include misspellings, tags from broken markup language, and non-English words. In addition, document frequency difference is noisy for words with small overall document counts. To identify a reasonable cutoff, we label each word with whether it has an entry in Wiktionary or not. The histogram in Figure 6.2 shows how the percentage of words with a Wiktionary entry increases with document frequency. We exclude proper nouns from our analysis, so the non-words are a combination of misspellings, typos, markup tags, non-English words, and words that do not happen to have Wiktionary entries (e.g. highly technical terms or neologisms).

We choose a threshold at which 90% of words have a Wiktionary entry, which corresponds to an overall document frequency of 57. Anecdotally, it also eliminates words that have a larger Simple English document frequency than Standard English document frequency, with the exceptions of "bigger" and "lot," which seem to be likely to be used more in simpler documents. We further exclude the remaining 10% of words that do not have Wiktionary entries, resulting in a vocabulary of 13418 words.

For the remaining words, we define the log frequency difference as the log of the ratio of Standard English document count and the Simple English document count for each word.[1] Words that generally appear in a Simple English article as often as they are used in Standard English will have a log frequency difference near 0; words that frequently are used in Standard English articles but not in the corresponding Simple English articles will have a large log frequency difference.

Ranking all of the words in the FAN lists by log frequency difference correlates with our ranking from observations of readers, though not as much as we might hope; the Spearman

[1]Since the document counts are extracted from a comparable corpus, the total corpus size (in number of documents) is the same for Simple English and Standard English. This means that the log difference is the same whether we use document counts or document frequency.

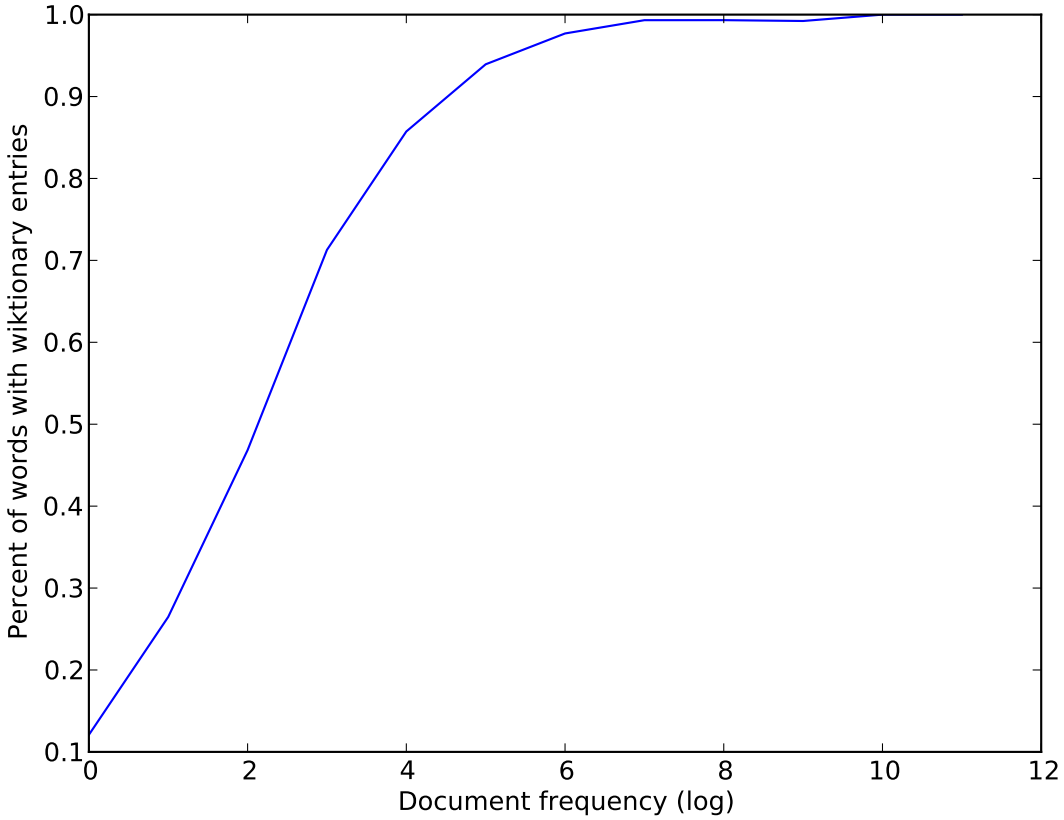

Figure 6.2: Likelihood of a word having a Wiktionary entry as a function of Wikipedia document frequency

correlation is $\rho = .38$ ($p < .0001$). The ranking based on observed difficulty cues does not correlate any better with unigram frequency, a common proxy for word difficulty. The correlation there is $\rho = .34$ ($p < .001$).

Table 6.1 shows the average unigram log frequency (from the BYU Corpus of Contemporary American English [23]) and log frequency difference of the words in each of the three word lists. The high variances in both measures make separating the lists difficult. It is not clear if that is a result of overlap in the actual difficulty of the words on the lists, or a

shortcoming of the measures. The fact that the log frequency difference is possitive for all of the words on the three lists is a consequence of the Standard English Wikipedia articles being longer and, consequently, covering more vocabulary. The higher vocabulary coverage for the Standard English articles means that using corpus frequency instead of document frequency may be an interesting alternative to pursue in future work.

Word List	**Log Unigram Frequency**	**Log Frequency Difference**
A	-7.7 (1.9)	.97 (.52)
B	-8.5 (1.1)	1.26 (.39)
C	-8.4 (0.9)	1.41 (.55)

Table 6.1: Average (and standard deviation) of log unigram frequency and log frequency difference of words in the FAN word lists

6.1.3 Word Difficulty Cues from Character n-Grams

We also want to capture how difficult a word could be to decode. Some words will be easy for readers to pronounce, even if they do know not know what the words mean. Other words will be difficult for readers to figure out how to read out loud, even if they are part of their spoken vocabulary. To examine this aspect of difficulty, we use the three lists of pseudo-words from the FAN collection. Again, the lists are intended to increase in difficulty.

We train a character n-gram model with $n = 7$ using original KN discounting, as recommended by the SRILM FAQ [77]. We compare against word length, which is used as a proxy for word difficulty in many traditional reading level formulas.

Our hypothesis is that the log likelihood from this model will be a better predictor of pseudo-word difficulty than word length. For a word of length L, character log likelihood is a sum of L log probabilities, so it scales with length L but also includes information about unlikely letter combinations. Table 6.2 shows the average word length and average word log likelihood, using the character n-gram, for each of the three lists.

Pseudoword List	Length	Character Log Likelihood
A	3.9	-6.1
B	4.6	-6.6
C	7.7	-8.8

Table 6.2: Average word length and character log likelihood of pseudo-words in the FAN pseudo-word lists

Here, we see that word length increases and likelihood decreases, as expected, with the known difficulty of the word lists. When we correlate these features with the acoustic difficulty ranking, we again do not see a high correlation across all of the words. However, looking only at the most difficult list, we obtain a correlation of $\rho = .61$ for word length and $\rho = .72$ for log likelihood computed with the character n-gram model. Character log likelihood may be most useful, then, in predicting the difficulty of longer, more difficult words.

6.1.4 Word Difficulty Cues from Wiktionary

In addition to the log likelihood ratio and character n-gram perplexity of a word, we consider features extracted from Wiktionary as a way to predict word difficulty.

In an early version of this analysis [47], we compared features of the words that occurred in at least three Simple English articles to the features of the words that occurred in only standard English articles, or in no more than two Simple English articles. In addition to standard lexical difficulty features like length and unigram frequency, we considered features extracted from each word's definition on wiktionary.org. In particular, we looked at each word's count of possible parts of speech (POS), senses (meanings), and translations into all other languages combined. Very common words (e.g. *hit*) tend to have many possible meanings and parts of speech, and to get translated into a larger number of languages. We found statistically significant differences in all three counts, but found the binary classification of words into easy vs. hard was not ideal for downstream tasks. Here, we instead

think of word difficulty as occurring on a continuum.

While its encyclopedia counterpart, Wikipedia, has been used extensively in language processing, the Wiktionary dictionary has been used quite a bit less frequently. Wu [89] develops a graph-based word similarity measure based on English Wiktionary. Similarly, 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 [91]. Chesley *et al.* use the English dictionary to determine adjective polarity as part of a system for classifying blog post sentiment [22]. We are not aware of any previous work that has used Wiktionary as a source for predicting lexical difficulty.

Wiktionary 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. We used seven of the possible parts of speech in Wiktionary (*proper noun*, *pronoun*, *noun*, *adjective*, *adverb*, *preposition*, *verb*), and collapsed all other parts of speech (e.g. *article* and *numeral*) into a single "Other" category. Word senses are enumerated for each part of speech. We used the total count, including mrare and archaic senses. In addition, dictionary entries may contain translations into any of the 172 available languages. For example, the definition of "paraphrase" has a total of four senses across two different parts of speech, along with three translations and a list of derived terms. A sample entry is shown in Figure 6.3. We do not filter out any senses from the full Wiktionary entries.

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 use of the frequent updates of the online dictionary, we use a static dataset for controlling experimental conditions. All of our work in this chapter uses a downloaded archive of the Wiktionary content as it appeared on February 2, 2014.

For the words we kept from the Wikipedia comparable corpus, we look at the POS,

Paraphrase

- Noun
  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

     Translations:

     (a) French: paraphrase (fr)

     (b) Portuguese: parfrase (pt)

     (c) Spanish: parfrasis (es)
- Verb
  1. to restate something as, or to compose a paraphrase

Figure 6.3: Sample Wiktionary entry, for the word *paraphrase*

sense, and translation counts extracted from Wiktionary entries. Figure 6.4 shows each count as a function of log document frequency difference.

As hypothesized, translation and sense count increase as document frequency difference decreases; that is, words that are nearly as common in Simple English wikipedia as they are in Standard English Wikipedia tend to have more translations and more senses than words that are much more common in Standard English Wikipedia.

We also consider splitting words into three groups by their document frequency difference. We identify the mean and standard deviation of the frequency difference, then identify three groups of words:

Hard Words with a simple/standard ratio less than the mean

Medium Words with a simple/standard ratio between the mean and one standard deviation above the mean

Easy Words with a simple/standard ratio more than one standard deviation above the mean

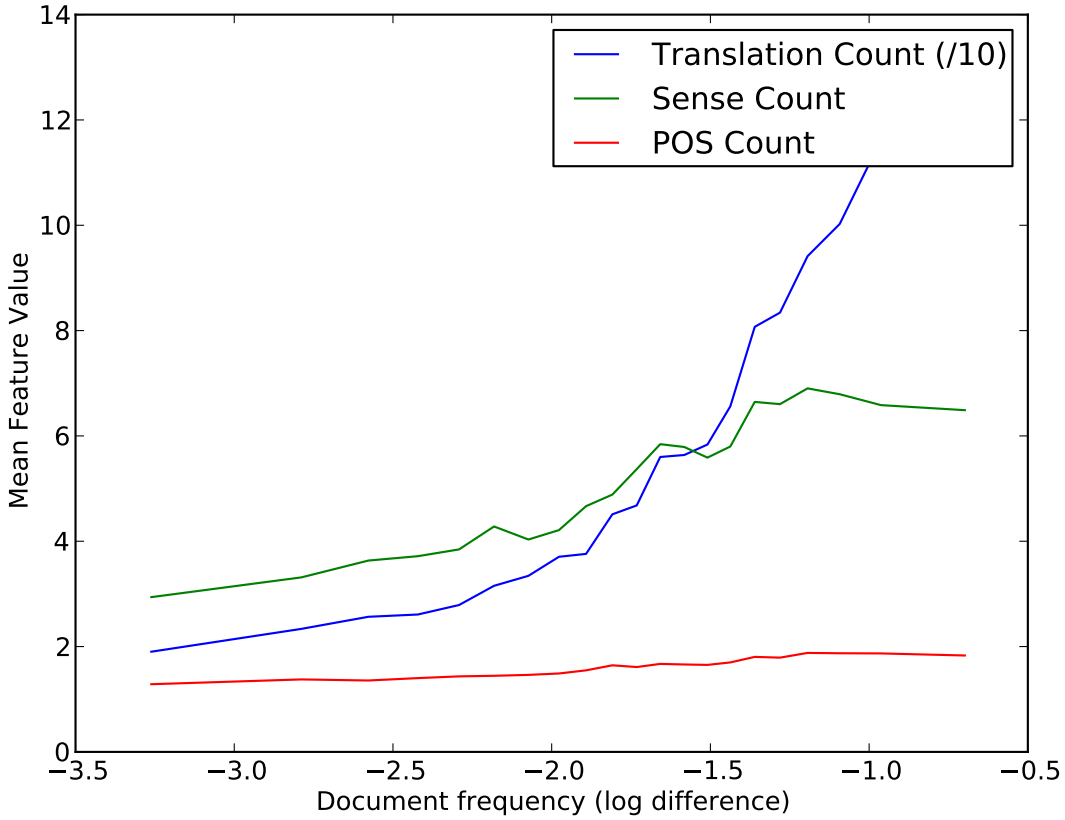

Figure 6.4: Average number of Parts of Speech, Senses, and Translations for words in Wiktionary as a function of each word's document frequency in Simple and Standard English Wikipedia

There were 7694 words in the Hard class, 4746 in the Medium class, and 1978 in the Easy class. Table 6.3 shows the POS, Translation, and Sense count averages for each class.

All three word features increase as word difficulty decreases, and all differences are statistically significant ($p < .05$). The distributions for the three classes still have substantial overlap, though, and analysis of outliers shows that "easy" words are frequent in Standard English documents and, at the same time, that "hard" words can be used in Simple English

	Translations	Senses	POS
Hard	33.4	4.2	1.5
Medium	72.9	6.2	1.7
Easy	115.9	6.8	1.9

Table 6.3: Translation, POS, and sense count for words in three-way classification by document frequency

documents, especially if they are important to the topic of the text. This finding supports the suggestion by Miltsakaki and Troutt that vocabulary difficulty is tied to the topic of the document [51].

6.1.5 Automatic Scoring of Word Difficulty

We train a linear regression model to predict, for each of the 120 words in the combined FAN real-word lists, the acoustic difficulty ranking for that word. We use the following features:

- Unigram frequency
- Word length
- Character n-gram likelihood
- Log frequency difference
- Translation count
- Sense count
- POS count
- Constant offset term

In particular, we start with a baseline model that uses just word length and unigram frequency. We then test for improvements by adding each of the other features individually. We evaluate each configuration using 10-fold cross-validation. The results, in terms of correlation (r) and rank correlation (ρ) are shown in Table 6.4.

Configuration	r	ρ
Baseline	.58	.61
+Sense Count	.60	.65
+Translation Count	.60	.62
+POS Count	.59	.62
+Log Frequency Difference	.58	.62
+Character Likelihood	.63	.65
+Sense Count, Character Likelihood	.64	**.66**

Table 6.4: Correlation (r) and rank correlation (ρ) of predicted word ranks to actual word ranks based on acoustic cues. Bold results are significant over the baseline with $p < .05$.

All of the features except log frequency difference improve on the baseline. The most effective features are sense count and character likelihood from our character n-gram model. Adding both of those features results in a small additional boost in performance. With the small number of words in the FAN lists, the improvement is not strongly statistically significant, but it is promising; we expect that we could get a more significant improvement with additional data. Manual inspection shows that our predictions do a good job of identifying the relative difficulty of the three word lists. All of the words on the easiest list are predicted to be in the easiest half, and all but 3 of the words on the hardest list are predicted to be in the hardest half. The words from the hard list that are predicted to be in the easier half ("other," "ever," "under") do, in fact, seem to be substantially easier than some of the other words on that list (e.g. "development," "substantial," "expression," which were predicted to be the three most difficult words). Much of the prediction difficulty may be related to not being able to characterize the difference between words of similar difficulty. The easy list, for example, includes words like "and," "in," and "so," which are all likely to be very easy for even the bottom 20% of readers in the FAN study. The correct ordering of those (and other similarly easy) words is much harder to predict than the fact that they are easier than the most difficult words on the list.

Configuration	r	ρ	Accuracy
Baseline	.61	.64	.72
+POS Count	.61	.65	.73
+Translation Count	.60	.62	.72
+Sense Count	.62	.64	.72
+Log Frequency Difference	.60	.62	.72
+Character Likelihood	.65	.66	.73
+Sense Count, Character Likelihood	.65	.66	.73

Table 6.5: Correlation (r) and rank correlation (ρ) to acoustic word ranks, and pair-wise accuracy of predicting the relative difficulty of pairs of words. No results were significant over the baseline at the level of $p = .05$.

We also tried a second experiment setup, in which we trained a logistic regression model to predict, for pairs of words, whether or not the first word was more difficult than the second word. We then ranked the words according to the number of words they were predicted to be more difficult than. In addition to correlation and rank correlation, this allowed us to calculate the pairwise accuracy of our model. As shown in Table 6.5, the resulting trends were similar to the results we got from predicting rank directly, but differences from the baseline were smaller.

6.2 Sentence Complexity

We saw in Chapters 3 and 4 that difficulty in sentences was based on more than just the difficulty of the component words. In this section, we predict the kinds of syntactic changes that people make in simplifying sentences.

6.2.1 Difficulty Ratings Based on Human Performance

As with words, we can characterize the difficulty of sentences by how readers interact with them. Again, we look at errors, verbal hesitations, lengthenings, and pausing at the sentence

level as indicators of difficulty. For sentences, we define the following rankings:

$r_{e,i}$ The expected number of errors in sentence s_i, calculated as the sum of the expected number of errors for each word in s_i

$r_{p,i}$ The expected number of short pauses in s_i

$r_{h,i}$ The expected number of verbal hesitations across all words in sentence s_i

$r_{l,i}$ The average full word lengthening across all words in sentence s_i

From the 8 FAN passages, we examine 97 sentences. For each ranking, we sort the sentences and label them from 1 to 97. We explore combining the four ranks by taking their mean, median, min, and geometric mean.

Four of the FAN stories ("Amanda," "Curly," "Grand Canyon" and "Guide Dogs") are intended to be appropriately difficult for late elementary-level readers. The other four ("Bigfoot", "Chicken Soup", "Exercise", and "Lori Goldberg") are intended to be appropriate for middle-level readers. Figure 6.5 shows the average rank (with standard deviation error bars) of the sentences in the easier four stories compared to the average rank of the sentences in the harder four stories for the four methods of combining. As was the case for words, all four methods agree that the sentences in the harder stories are harder (that is, lower ranked). We once again use the mean for the rest of our experiments, since it results in the lowest variance. Our final difficulty ranking r_i, then, is:

$$r_i = \frac{r_{e,i} + r_{p,i} + r_{h,i} + r_{l,i}}{4} \tag{6.2}$$

The ten most difficult sentences in the FAN corpus according to this ranking are:

- You can also check with local churches or synagogues, senior and civic centers, parks, or recreation associations for exercise, wellness, or walking programs.
- Check with your doctor first if you are a man over forty or a woman over fifty and you plan to do vigorous activity instead of moderate activity.
- There is no shortage of medicinally active compounds such as vitamins in these ingredients.
- The other day as Lori Goldberg was walking through the halls of her school, a few whispers could be heard from the children she passed, but most greeted her with grins

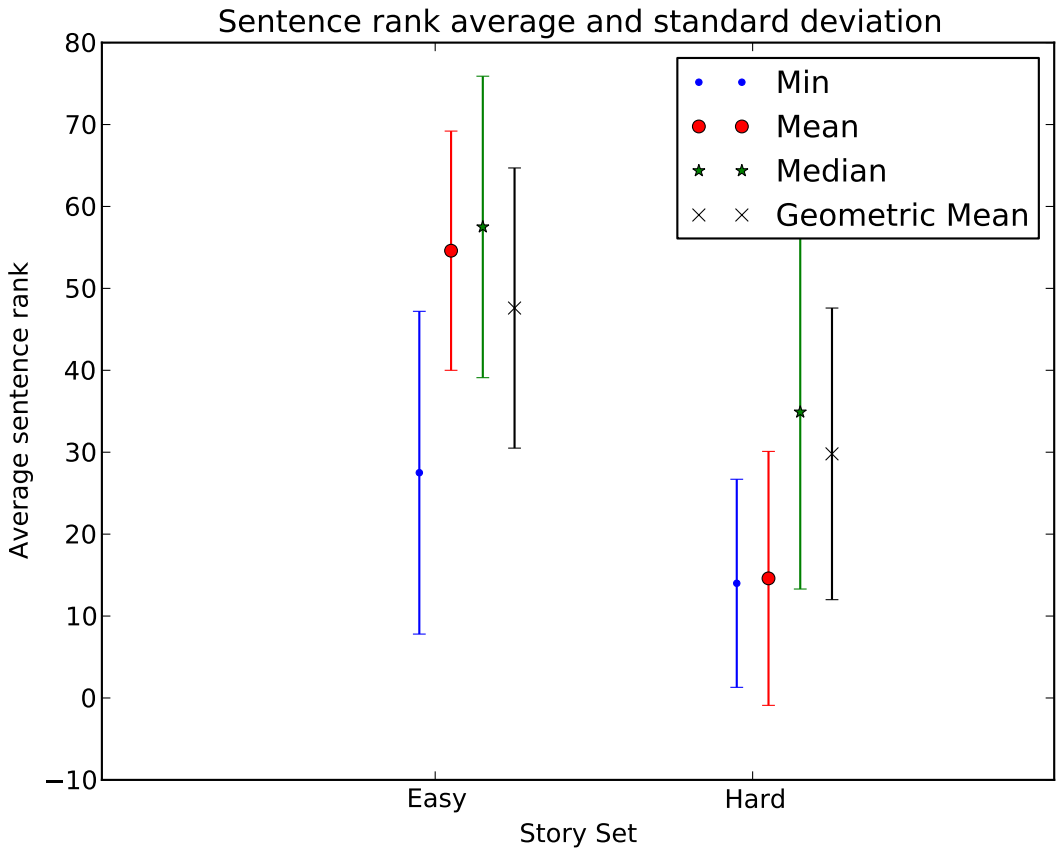

Figure 6.5: Mean and standard deviation of sentence ranks for each FAN story, when final ranks are the mean, min, geometric mean, or median of the individual acoustic feature ranks

and hellos.

- Goldberg said the door to her office is always open, and the students are encouraged to talk to her about any problems they might be having.
- Cold symptoms such as a runny nose and cough are thought to be caused by immune cells flooding into infected areas.
- For his tests, the researcher used his wife's soup recipe that includes onions, sweet potatoes, parsnips, turnips, carrots, celery stems, parsley, seasoning, and of course

chicken.

- Although not a single Bigfoot has ever been captured, killed, or found dead, many people believe in the creature's existence.
- In the eighties, an elderly man admitted he had made hoax Bigfoot tracks for fifty years.
- Finding some unattended automobiles, they left fake footprints nearby.

By contrast, the ten easiest sentences according to the ranking are:

- But the training isn't over.
- New buildings on the rim affect the springs as well.
- Curly stood watching him.
- But are the prints genuine?
- Curly is my big black dog.
- What do you think he saw?
- They come by bus, by motorcycle, and by car.
- Then the children get new puppies.
- They use a lot of water for toilets and showers.
- The Grand Canyon is running out of space.

6.2.2 Sentence Difficulty in Parallel Corpora

The number of lexical differences in simplifications that were artifacts of syntactic changes motivated us to look at common syntactic differences between sentence-aligned comparable texts. We used sentence pairs from Encyclopedia Britannica articles [14] and original/abridged pairs of news articles from the Western/Pacific Literacy Network website.[2]

This study reflects an updated version of work reported in [48]. We make the following changes from the original study:

Updated feature set The topic model has been retrained with additional data. The results in the thesis are based on a 100-topic model trained on a collection of 1.1M Standard English Wikipedia articles using the MALLET toolkit's Latent Dirichlet

[2]http://literacynet.org/cnnsf/, accessed June 15, 2004

	Omit	Expand	Neither
Split	142	257	72
No Split	454	233	864

Table 6.6: Number of sentences with Splits, Omits, and Expands in the CNN and Britannica corpus

Allocation implementation [46]. In the previous paper, a parse complexity feature was calculated based on the perplexity of parse paths. In this thesis, that has been replaced with a parse probability score from the parser.

Updated model Instead of a decision tree, we use a logistic regression model.

Based on automatic parses of the sentence pairs, we looked at three types of clause-level syntactic changes:

- **Splits**: One original sentence becomes more than one simplified sentence
- **Omissions**: The number of S nodes decreases in simplification, e.g., for less critical information
- **Expansions**: The number of S nodes increases in simplification, e.g., to define difficult but important words

The number of instances of each change is summarized in Table 6.6.

We trained classifiers to predict when each of the above changes would occur based on lexical, syntactic, and discourse features of each sentence. Those features included:

- **Lexical Features**: For each lexical feature, we create 6 bins evenly spaced across the feature value's distribution. For each sentence, we included the count of words of words in the sentence that were in the bottom n bins, for $n = 1, .., 5$. Word-level features included unigram frequency, POS count, sense count, and translation count
- **Syntactic Features**: We generate the 1-best syntactic parse for each sentence using the Stanford parser [39]. From the syntactic parse tree for each sentence, we count the number of sentence (S, SBAR, or SINV) nodes, and the number of noun phrase (NP, NNP, NPS, or NNPS) nodes, where node labels follow the Penn Treebank convention.

We also use the parse probability generated by the parser.

- **Topicality Score**: Topic similarity scores based on cosine similarity of sentences and documents represented as vectors of posterior probabilities over topics. We included the cosine similarity between each sentence and 1) the previous sentence; 2) all previous sentences in the document; 3) the entire document. We also include a sentence's topic entropy, which we define to be the entropy of the topic vector. The topic entropy gives a measure of how well a small number of topics captures the sentence. Sentences that are equally well represented by many topics will have high topic entropy, while sentences that are well represented by a single topic will have low topic entropy.

Figure 6.6 shows the classifiers' performance. We found small gains in prediction with the lexical and syntactic features, with Omits being the easiest to predict.

Table 6.7 shows the five most important features in predicting splits. Two of the top five features (parse probability and count of NP nodes) are related to the syntactic parse, suggesting that syntactic structure is important in deciding if a sentence should be split during simplification. The other three are related to document-level features. The sentence's position in a document is a function of discourse structure, with sentences earlier in a document being more likely to be split. Topic entropy and story similarity also affect the likelihood of a split. Sentences that are closer in topic to the document they come from are more likely to be split, and sentences with high topic entropy are less likely to be split. These two features are not independent, since a sentence that has high topic similarity will necessarily have low topic entropy.

Table 6.8 shows the most important features in predicting omits. As for splits, we see that syntactic features are important, with the number of S nodes being the most heavily weighted feature. Lexical features are also useful, though, with two different features based on translation counts being used. Again, topic-based features are very useful. Sentences that are similar to the sentence before them are less likely to be omitted, which matches our intuition that omissions are more likely to occur with sentences that contain asides. Sentences that are similar to all of the previous sentences in a document are more likely to be omitted, however, which may be a result of wanting to eliminate redundant or very detailed information later in a document.

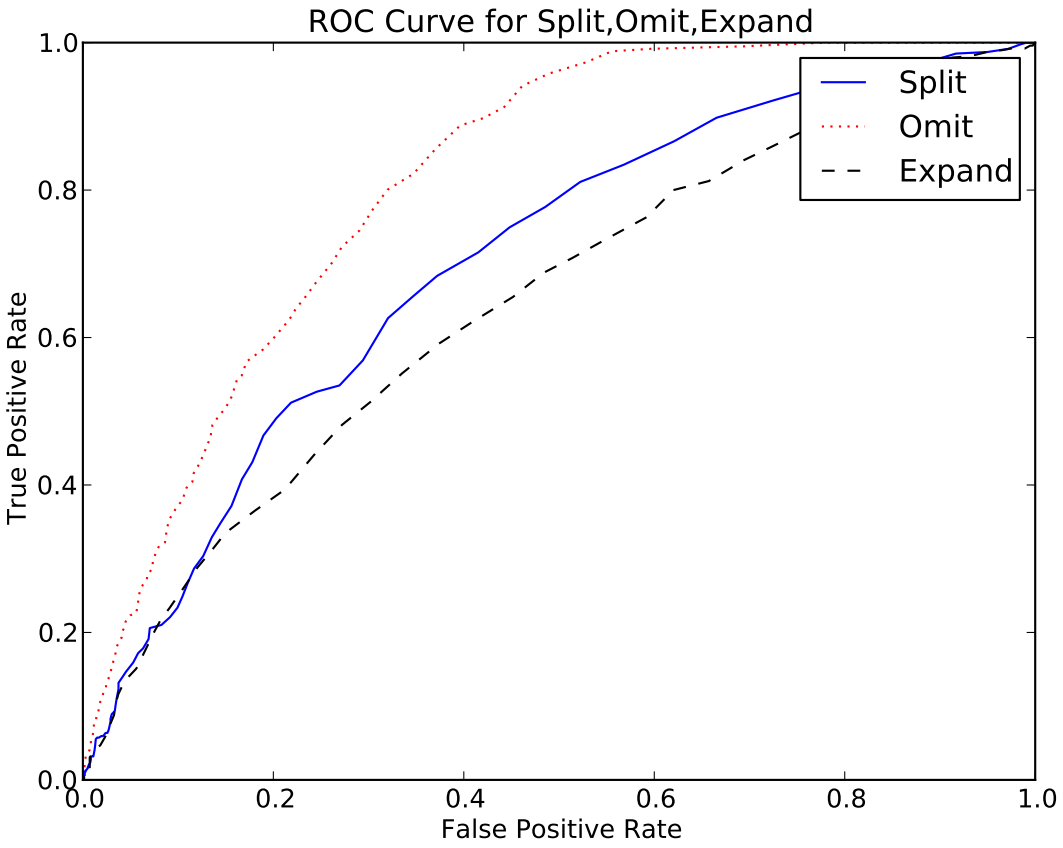

Figure 6.6: ROC Curve for predicting Splits, Omits, Expands on the CNN and Britannica dataset

Finally, Table 6.9 shows the most important features in predicting expands. Again, syntactic features are very important. Sentences with a small number of S nodes are more likely to be expanded. Sentences with a large number of NP nodes are also more likely to be expanded. These findings makes intuitive sense; sentences with many NP nodes under a small number of S nodes may correspond to ones with a very dense representation of content. Lexical features are also useful. Sentences with many words with a small number of translations (an indication of a potentially difficult or domain-specific word) are more

Feature	**Weight**
NP Node Count	.36
Topic Entropy	-.30
Parse Score	-.22
Position in Document	-.22
Topic Similarity to Story	.21

Table 6.7: Most heavily weighted features for predicting Splits in the CNN and Britannica corpus

Feature	**Weight**
S* Node Count	1.07
Count of words with ≤ 11 Translations	-0.9
Topic Similarity to Previous Sentence	-.39
Count of words with ≤ 5 Translations	.33
Topic Similarity to All Previous Sentences	.32

Table 6.8: Most heavily weighted features for predicting Omits in the CNN and Britannica corpus

likely to be expanded, which may be the result of needing to explain those words. On the other hand, sentences with a large number of words with a single possible POS are *less* likely to be expanded, which may be a consequence of some difficult or specialized content being omitted. Topic-based features do not show up here, which is surprising since they were useful for predicting omissions. We hypothesize that this may be because the syntactic features overwhelm them. A very long or dense sentence is going to be split if additional information needs to be added, and our method of identifying expansions from node counts is somewhat limited when splits are also involved. To have a better idea of what is going on in these cases, it would be useful to have a set of hand-labeled sentence pairs indicating

Feature	**Weight**
S* Node Counts	-.87
Count of words with $\leq$ 11 Translations	.28
NP* Node Counts	.27
Count of words with $\leq$ 1 Senses	-.25
Parse Depth	.18

Table 6.9: Most heavily weighted features for predicting Expands in the CNN and Britannica corpus

what, if any, information is new in the simplified sentences.

Overall, features of all three types (syntactic, lexical, and topical) were useful in predicting sentence-level changes.

6.2.3 Automatic Prediction of Sentence Difficulty

We train a linear regression model to predict, for each sentence in the FAN data, the acoustic difficulty ranking for that sentence. We use the same lexical, syntactic, and topic-based features as in Section 6.2.2.

We start with a baseline model that uses just word features. We then test for improvements by adding syntactic and topical features. We evaluate each configuration using 10-fold cross-validation. The results, in terms of correlation (r) and rank correlation (ρ) are shown in Table 6.10.

The syntactic features improve slightly on the baseline configuration, even for the small number of sentences in the FAN collection. Adding the topic features, on the other hand, actually hurts the model's performance. This finding does not necessarily contradict the results we saw when using topic features to predict types of syntactic changes, though. In that experiment, we were simultaneously predicting *if* a sentence should be simplified and *how* it should be simplified. A sentence that was a negative instance of an Omit could have been kept because it did not need to be changed at all, or it could have been expanded.

Configuration	r	ρ
Baseline	.59	.58
+Syntactic	.60	**.64**
+Topic	.55	.52

Table 6.10: Correlation (r) and rank correlation (ρ) of predicted word ranks to actual word ranks based on acoustic cues. Bold results are significant over the baseline with $p < .05$.

In this experiment, we are only predicting whether a sentence is difficult or not, with no attempt to decide how to address individual difficulty points. It may be the case, then, that syntactic and lexical features are useful in identifying where difficulties are likely to occur, and topic features are then useful in deciding whether to omit or expand the content around the identified difficulties.

As an alternative, we consider the possibility of using the split prediction model from Section 6.2.2 to rank sentences. Since that model was trained on separate data, we do not need to use cross-validation. Ranking the FAN sentences in order of the likelihood of a split gives us a correlation of $r = .53$ and a rank correlation of $rho = .55$. While this is less than we got from training a model to predict difficult directly, it is from a model that was trained for a separate task. This experiment indicates, then, that the parallel data is useful for learning cues to sentence difficulty.

6.3 Conclusion

In this chapter, we have explored automatic predictions of word and sentence difficulty. Using rankings of words and pseudo-words based on acoustic cues of difficulty, we ordered words by observed difficulty. We compared those observed difficulty rankings to difficulty measures extracted from text corpora, including log frequency difference in comparable Wikipedia articles, character n-gram likelihood, word length, unigram frequency, and features extracted from Wiktionary entries. We found that word length, unigram frequency, sense count, and character n-gram likelihood were all useful in predicting the difficulty

ranking of a word.

We also looked at predicting when simplifying a sentence should include certain syntactic changes. In particular, we looked at when sentences are split into two or more simplified sentences, when parts of a sentence are omitted, and when a sentence should be expanded as part of the simplification process. We found that measures of topicality were useful in distinguishing between omission and expansion. Lexical and syntactic features were also useful in predicting syntactic changes, though the task remains a difficult one. Lexical and syntactic features were also useful in predicting the difficulty of a sentence, while topic features were less useful for that task.

Chapter 7

SUMMARY AND FUTURE DIRECTIONS

7.1 Summary

This thesis has presented a fundamentally new approach to understanding text difficulty. Leveraging human performance, we have connected text assessment to individual literacy assessment methods based on oral reading. In order for our human-based methods to be effective, we have developed a method for accounting for prosodic factors that affect acoustic cues to reading difficulty. We have validated those cues against results from a new eye tracking study. Finally, we have used comparable data to extract features that could be used to predict text difficulty, and used the human data to understand the relative importance of those features.

The specific contributions of this thesis include:

A new corpus of parallel human simplifications

To address the question of how well and how reliably people simplify texts, we created a new corpus of parallel human simplifications, with between 6 and 8 sentence-aligned simplifications per document. We found that expert technical writers had low agreement with respect to when an original sentence should be split into multiple pieces, as well as to when a sentence should be shortened as part of the simplification process. We also found that the most common lexical removals in these simplifications were function words, which was an indication that most of the simplifications involved structural changes, not just direct lexical substitution.

A fundamentally new approach to understanding text difficulty, leveraging oral readings

Instead of using labeled data, we present an approach to understanding text difficulty that relies on observations of how people interact with a text. We first developed an approach for

accounting for prosodic boundaries in oral readings, and then identified hesitations, errors, and pauses and duration lengthenings that were not explained by prosodic boundary effects.

We used those acoustic cues to predict the reading level of individuals. We also used their frequency of occurrence to rank words and sentences by difficulty.

A comparison of acoustic and gaze-based features during oral reading

We conducted an eye tracking study with 72 participants, using texts for which large-scale parallel recordings were available. Each participant read 10 passages out loud, and we collected both audio and gaze information during their reading. We compared fixation and regression information in their gaze path to hesitations, pauses, duration lengthening, and errors in the audio. We found that long pauses, lengthening, and hesitations in the audio were associated with long fixations. Lengthening and long pauses were also associated with regressions.

The use of comparable data to identify features that are related to text difficulty

Since large, parallel corpora of texts at different reading levels are rare, we made use of other large text corpora to identify features that are related to text difficulty.

We created a corpus of 22,923 comparable articles from English and Simple English Wikipedia. All placeholder ("stub") articles were removed, as were any topics for which one of the articles was less than 50 words long. The resulting pairs of plain-text articles were made available online[1] for use by other researchers. We characterized words by the number of documents that they occurred in in each document set. We similarly processed a new corpus of 1.1M Standard English articles, and used those to build a topic model for measuring the similarity between sentences and documents.

We presented the first work to use word features extracted from Wiktionary to predict lexical difficulty. We made use of sense, POS, and translation counts in Wiktionary definitions to predict how difficult individual words would be for readers. frequency and word length.

[1] http://tial.ee.washington.edu/ jmedero/wikiComparable.tgz

We then used features extracted from text corpora to predict the difficulty of words and sentences. Both the character likelihood model and the Wiktionary features improved the performance of our baseline model, which used unigram frequency and word length only, for predicting word difficulty. Syntactic and lexical features were useful in ranking sentences by difficulty, while topic-based features were helpful in predicting when sentences should be split or omitted during simplification.

7.2 Future Directions

Automatic simplification is a relatively new area of NLP. There are, as a result, a number of interesting directions of future work related to the methods presented in this thesis. We highlight some of them here. First, we look at potential enhancements to the models used in this thesis. Next, we discuss how the work described in this thesis might be integrated into a full automatic simplification system. We then discuss how the human assessment-based techniques developed in this thesis could be used to explore how the text features that influence difficulty vary for different reader types and for different genres of texts.

7.2.1 Model Enhancements

The models developed in this thesis explain some of the variance in word and sentence difficulty. We expect that we could get even better performance through some additions to our models. In particular, in Chapter 4, we develop a method for predicting and accounting for prosodic breaks. We do not, however, account for prosodic emphasis. Developing a method to account for that might particularly improve our results related to duration lengthening.

The layout of text on the screen can affect the reading process. Words at the beginning of a line have longer average fixations (485ms, compared to 382ms for other words). Words at the end of a line have a higher probability of regressions (19% of the time, compared to 6% for other words). In addition to accounting for prosodic structure, future work should take into account the location of line breaks.

For the FAN texts used in Chapter 4, we had access to phone-aligned automatic transcripts, but not to the audio recordings themselves. Consequently, we focused on features that could be extracted from the transcripts. Audio recordings might provide additional

cues of reader uncertainty or reading difficulty, though, and we could explore how readers' pitch and energy might help in characterizing word and sentence difficulty.

Our model for predicting syntactic changes (splits, omits, and expands) relied on S node counts as a proxy for identifying when content was added or removed. A more careful look at those changes would be possible with pairs of sentences that were hand-labeled for the content that was unique to each sentence.

We could also include additional text-based features in our models for predicting difficulty. Difficulties that are a consequence of semantic surprise or of idiomatic language might be captured by an n-gram language model, so adding language model surpisal features to words and sentences could improve our prediction of difficult words and sentences. Similarly, anaphora can be a source of increased cognitive load, so including features indicating anaphoric relations could improve our predictions.

Finally, we saw promise in our results on predicting word and sentence difficulty in Chapter 6, but the word lists we were testing on were short. It would be interesting to conduct an expanded word difficulty study by collecting oral readings of additional texts, which we could then compare with the same text-based features used in our Chapter 6 analyses.

7.2.2 Automatically Simplifying Texts

To generate a simplified text, we need to balance two competing demands. The first is to make the words and sentences in a passage as easy as possible. The second, though, is to minimize disruptions to the meaning of the text. Since even synonym replacements of words are likely to result in at least nuanced changes in meaning, we accept that any simplification will change the meaning of the original document somewhat. Depending on the target reader, though, we may want to emphasize simplicity more or less over meaning preservation.

In natural languages, an idea can be expressed in multiple ways. A paraphrase can be a single word synonym (e.g. "courageous" → "brave") or a multi-word phrase (e.g. "was later shortened" → "was made shorter later on"). Statistical systems for learning paraphrases

can also learn paraphrase patterns by replacing words or phrases with slots. For example, a paraphrase rule for changing a passive voice sentence to active voice could look like "NP_1 was VBN_2 by $NP_3 \rightarrow NP_3$ VBD_2 NP_1", indicating that the noun phrases NP_1 and NP_3 change places, while the verb VBN_2 changes from past participle form to past tense.

We can use existing technologies for generating paraphrases to generate potential simplifications. Each paraphrase of a word, phrase, or construction in an original text is a potential simplification. Existing paraphrase systems (e.g. [13]) could be used to generate candidate phrases, and those potential simplifications (along with the non-simplification option of keeping all word as-is) could be represented as a graph. The difficulty prediction scores and topicality features from this work could be used as weights on the edges of that graph. By choosing a trade-off factor between simplicity and meaning preservation (as calculated from a topic model), the resulting edge weights could be used to generate simplifications.

7.2.3 Genre Differences in Text Difficulty

Our analysis in this thesis has focused on short stories and informational texts. Future work could explore the extent to which the features identified here predict difficulty for other types of texts. This line of work would explore changes that would need to be made to reliably simplify word problems for a math class, articles on consumer-directed health website, or documents on a government website. To understand these differences, we would need to conduct additional studies with human subjects, resulting in new collections of oral readings. We could then analyze the extent to which the features that correlated with difficulty in this study were useful to other genres, and the extent to which additional features were necessary to capture difficulty in other genres.

7.2.4 Simplifying for Different Types of Readers

A number of different types of readers can benefit from simplification technology. This work has focused on fluent English-speaking adults with no learning disabilities. However, second language learners, individuals with a variety of learning disabilities, and children

of different ages could also benefit from automatically simplified texts. Further work is needed to test the extent to which the features identified in this work are also reliable for predicting difficulty for these different audiences. As for genre differences, this would require additional human subjects experiments, and would result in new collections of oral readings by different types of readers. With readers who struggle for different reasons (e.g. age, first-language literacy, English speaking proficiency, learning disabilities), we could compare the effectiveness of our existing features, and explore the need for additional features.

BIBLIOGRAPHY

[1] Versant English Test: Test Description & Validation Summary. Technical report, Pearson Education, 2008.

[2] Emil Abrahamsson, Timothy Forni, Maria Skeppstedt, and Maria Kvist. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 2014.

[3] Marcelo Amancio and Lucia Specia. An Analysis of Crowdsourced Text Simplifications. In *Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 123–130, 2014.

[4] S Ananthakrishnan and S S Narayanan. Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. *IEEE Trans. Audio, Speech, and Language Processing*, 16(1):216–228, 2008.

[5] Jane Ashby. Prosody in skilled silent reading: evidence from eye movements. *Journal of Research in Reading*, 29(3):318–333, 2006. ISSN 01410423.

[6] Jane Ashby and Charles Clifton. The prosodic property of lexical stress affects eye movements during silent reading. *Cognition*, 96(3):B89–B100, 2005.

[7] Jane Ashby and Keith Rayner. Representing syllable information during silent reading: Evidence from eye movements. *Language & Cognitive Processes*, 19(3):391–426, 2004.

[8] Jane Ashby, Keith Rayner, and Charles Clifton. Eye movements of highly skilled and average readers: differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology*, 58(6):1065–1086, 2005.

[9] Jane Ashby, Rebecca Treiman, Brett Kessler, and Keith Rayner. Vowel processing during silent reading: evidence from eye movements. *Journal of experimental psychology Learning memory and cognition*, 32(2):416–424, 2006.

[10] Justin Baer, Mark Kutner, John Sabatini, and Sheida White. Basic Reading Skills and the Literacy of Americas Least Literate Adults: Results from the 2003 National Assessment of Adult Literacy (NAAL) Supplemental Studies. Technical report, NCES, 2009.

[11] Jennifer Balogh, Jared Bernstein, Jian Cheng, and Brent Townshend. Final Report Ordinates Scoring of FAN NAAL Phase III: Accuracy Analysis. Technical report, Ordinate, 2005.

[12] Jennifer Balogh, Jared Bernstein, Jian Cheng, Alistair Van Moere, Brent Townshend, and Masanori Suzuki. Validation of Automated Scoring of Oral Reading. *Educational and Psychological Measurement*, 2011.

[13] Colin Bannard and Chris Callison-Burch. Paraphrasing with bilingual parallel corpora. In *Proc. ACL*, pages 597–604. Association for Computational Linguistics, ACL, 2005.

[14] R Barzilay and N Elhadad. Sentence alignment for monolingual comparable corpora. In *Proc. EMNLP*, pages 25–32, 2003.

[15] Jared Bernstein, Jian Cheng, and Masanori Suzuki. Fluency Changes with General Progress in L2 Proficiency. In *Proc. Interspeech*, pages 877–880, 2011.

[16] Klinton Bicknell and Roger Levy. A Rational Model of Eye Movement Control in Reading. In *Proc. ACL*, pages 1168–1178. Association for Computational Linguistics, 2010.

[17] Klinton Bicknell and Roger Levy. Why readers regress to previous words: A statistical analysis. In *Proc. Cognitive Science Conference*, 2011.

[18] Or Biran, Samuel Brody, and Noémie Elhadad. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *ACL*, 2011.

[19] Julie E Boland. Linking Eye Movements to Sentence Comprehension in Reading and Listening. In M Carreiras and Charles E Clifton, editors, *The On-line Study of Sentence Comprehension: Eyetracking, ERPs and Beyond*, chapter 4, pages 51–76. Psychology Press, 2004.

[20] Julian Brooke, Vivian Tsang, David Jacob, Fraser Shein, and Graeme Hirst. Building readability lexicons with unannotated corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 33–39, 2012.

[21] Jian Cheng. Automatic assessment of prosody in high-stakes English tests. In *Proc. Interspeech*, pages 1589–1592, 2011.

[22] Paula Chesley, Bruce Vincent, Li Xu, and RK Srihari. Using verbs and adjectives to automatically classify blog sentiment. *Training*, page 8, 2006.

[23] M Davies. The Corpus of Contemporary American English (COCA): 425 million words, 1990-present. http://www.americancorpus.org, 2008. URL `http://corpus.byu.edu/coca/`.

[24] Paul Deane, Kathleen M Sheehan, John Sabatini, Yoko Futagi, and Irene Kostin. Differences in Text Structure and Its Implications for Assessment of Struggling Readers. *Scientific Studies of Reading*, 10(3):257–275, July 2006.

[25] Ryan Downey, David Rubin, Jian Cheng, and Jared Bernstein. Performance of Automated Scoring for Children's Oral Reading. In *Proc. Workshop on Innovative Use of NLP for Building Educational Applications*, pages 46–55, Portland, Oregon, June 2011.

[26] Jacques Duchateau, Leen Cleuren, Hugo Van, and Pol Ghesqui. Automatic Assessment of Childrens Reading Level. *Proc. Interspeech*, pages 1210–1213, 2007.

[27] Andrew T Duchowski. *Eye tracking methodology: theory and practice*. Springer, 2007. ISBN 1846286085.

[28] Benoit Favre, Dilek Hakkani-Tür, and Sebastien Cuendet. Icsiboost, 2007. URL `http://code.google.come/p/icsiboost`.

[29] Thomas François and Eleni Miltsakaki. Do NLP and Machine Learning Improve Traditional Readability Formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 49–57, 2012.

[30] A C Graesser, Danielle S McNamara, M M Louwerse, and Z Cai. Coh-Metrix:Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36:193–202, 2004.

[31] Matthew Green. An eye-tracking evaluation of some parser complexity metrics. In *Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 38–46, 2014.

[32] Elizabeth Greenberg, Justin Baer, Eric Dunleavy, Barbara Forsyth, and Jared Bernstein. Technical Report and Data File Users Manual For the 2003 National Assessment of Adult Literacy. Technical report, Institute of Education Sciences, 2009.

[33] R Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[34] Michael Heilman and Noah A Smith. Tree Edit Models for Recognizing Textual Entailments, Paraphrases, and Answers to Questions. In *Proc. NAACL-HLT*, pages 1011–1019, 2010.

[35] Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. An analysis of statistical models and features for reading difficulty prediction. *Proc. Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79, 2008.

[36] Aulikki Hyrskykari. Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading. *Computers in Human Behavior*, 22(4):657–671, 2006.

[37] Susan Kemper and C J Liu. Eye movements of young and older adults during reading. *Psychology and Aging*, 22:84–93, 2007.

[38] J. P. Jr. 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.

[39] D Klein and C Manning. Accurate Unlexicalized Parsing. In *Proc. ACL*, pages 423–430, 2003.

[40] Pia Knoeferle and Matthew W Crocker. Constituent order and semantic parallelism in online comprehension: eye-tracking evidence from German. *The Quarterly Journal of Experimental Psychology (2006)*, 62(12):2338–2371, 2009.

[41] A Koornneef and J Vanberkum. On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, 54(4):445–465, 2006.

[42] Hamutal Kreiner. The Role of Reading Prosody in Syntactic and Semantic Integration: Evidence from Eye Movements. In *International Symposium on Discourse and Prosody as a complex interface*, pages 1–17, 2002.

[43] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summarization Branches Out*, pages 25–26. Association for Computational Linguistics, 2004.

[44] Yi Ma, Ritu Singh, Eric Fosler-Lussier, and Robert Lofthus. Comparing Human Versus Automatic Feature Extraction for Fine-grained Elementary Readability Assessment. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 58–64, 2012.

[45] Anna Margolis, Mari Ostendorf, and Karen Livescu. Cross-genre training for automatic prosody classification. In *Proc. Speech Prosody Conference*, 2010.

[46] Andrew Kachites McCallum. MALLET: A Machine Learning for Language Toolkit, 2002. URL `http://mallet.cs.umass.edu`.

[47] Julie Medero and Mari Ostendorf. Analysis of Vocabulary Difficulty Using Wiktionary. In *Proc. SLaTE Workshop*, 2009.

[48] Julie Medero and Mari Ostendorf. Identifying Targets for Syntactic Simplification. In *Proc. Speech and Language Technology in Education Workshop*, 2011.

[49] Julie Medero and Mari Ostendorf. Atypical Prosodic Structure as an Indicator of Reading Level and Text Difficulty. In *Proc. NAACL-HLT*, 2013.

[50] J Miller and Paula J Schwanenflugel. Prosody of Syntactically Complex Sentences in the Oral Reading of Young Children. *Journal of Educational Psychology*, 98(4): 839–843, 2006.

[51] Eleni Miltsakaki and Audrey Troutt. Real-time web text classification and analysis of reading difficulty. *Proc. Workshop on Innovative Use of NLP for Building Educational Applications*, pages 89–97, 2008.

[52] J Mostow, A G Hauptmann, L L Chase, and S Roth. Towards a reading coach that listens: automated detection of oral reading errors. In *Proc. AAAI*, pages 392–397. Publ by AAAI, 1993.

[53] J Mostow, J Beck, S Winter, S Wang, and B Tobin. Predicting Oral Reading Miscues. In *Proc. ICSLP*, 2002.

[54] Courtney Napoles and Mark Dredze. Learning simple Wikipedia: a cogitation in ascertaining abecedarian language. In *Proc. Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 42–50, 2010.

[55] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. The Pyramid Method. *ACM Transactions on Speech and Language Processing*, 4(2):4–es, 2007.

[56] Ani Nenkova, Jieun Chae, Annie Louis, and Emily Pitler. Structural Features for Predicting the Linguistic Quality of Text Applications to Machine Translation , Automatic Summarization and Human-Authored Text. *Empirical Methods in NLG*, pages 222–241, 2010.

[57] Hitoshi Nishikawa, Toshiro Makino, and Yoshihiro Matsuo. A Pilot Study on Readability Prediction with Reading Time. In *Proceedings of The 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 78–84, 2013.

[58] AeroSpace of Europe and Defence Industries Association. ASD Simplified Technical English: Specification ASD-STE100, 2005.

[59] Anneli Olsen. Tobii I-VT Fixation Filter - Algorithm Description. Technical report, Tobii Technology, 2012.

[60] M Ostendorf, P J Price, and S Shattuck-Hufnagel. The Boston University Radio News Corpus. Technical report, Boston University, March 1995.

[61] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, number July in ACL '02, pages 311–318. Association for Computational Linguistics, 2002.

[62] David Pellow and Maxine Eskenazi. An Open Corpus of Everyday Documents for Simplification Tasks. In *Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 84–93, 2014.

[63] Sarah E Petersen and Mari Ostendorf. A Machine Learning Approach to Reading Level Assessment. *Computer, Speech and Language*, 23:89–106, 2009.

[64] Timothy Rasinski. Reading fluency instruction: Moving beyond accuracy, automaticity, and prosody. *The Reading Teacher*, 59(7):704–706, April 2006.

[65] MH Rasmussen, Jack Mostow, Zheng-hua Tan, Bø rge Lindberg, and Yuanpeng Li. Evaluating Tracking Accuracy of an Automatic Reading Tutor. In *Proc. Speech and Language Technology in Education Workshop*, 2011.

[66] Keith Rayner, K H Chace, Timothy J Slattery, and Jane Ashby. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10(3): 241–255, 2006.

[67] Keith Rayner, Timothy J Slattery, and Nathalie N Bélanger. Eye movements, the perceptual span, and reading speed. *Psychonomic bulletin & review*, 17(6):834–839, 2010.

[68] Keith Rayner, Timothy J Slattery, Denis Drieghe, and S.P. Liversedge. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):514, 2011.

[69] R.E. Schapire and Y. Singer. BoosTexter: a boosting-based system for text categorization. *Machine Learning*, 39(2):135–168, 2000.

[70] Matthew Shardlow. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of The 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, 2013.

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

[72] Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. Syntactic simplification for improving content selection in multi-document summarization. In *Proc. ACL*, pages 407–414, 2004.

[73] Timothy J Slattery, Elizabeth R Schotter, Raymond W Berry, and Keith Rayner. Parafoveal and foveal processing of abbreviations during eye fixations in reading: making a case for case. *Journal of experimental psychology. Learning, memory, and cognition*, 37(4):1022–31, July 2011.

[74] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proc. ACM*, pages 223–231, 2006.

[75] Louise Spear-Swerling. Childrens Reading Comprehension and Oral Reading Fluency in Easy Text. *Reading and Writing*, 19(2):199–220, 2006.

[76] AJ Stenner. Measuring reading comprehension with the Lexile framework. In *Proc. North American Conference on Adolescent/Adult Literacy*, 1996.

[77] Andreas Stolcke. SRILM – An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, pages 901–904, 2002.

[78] Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Tsutomu Hirao, and Masaaki Nagata. Divide and Translate: Improving Long Distance Reordering in Statistical Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 418–427, 2010.

[79] Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada. Sorting Texts by Readability. *Computational Linguistics*, 36(2):203–227, 2010.

[80] Irina Temnikova and Galina Maneva. The C-Score – Proposing a Reading Comprehension Metrics as a Common Evaluation Measure for Text Simplification. In *Proceedings of The 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 20–29, 2013.

[81] Sowmya Vajjala and Detmar Meurers. 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, 2013.

[82] Sowmya Vajjala and Detmar Meurers. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 2014.

[83] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management*, 43(6):1606–1618, November 2007.

[84] David Vickrey and Daphne Koller. Sentence Simplification for Semantic Role Labeling. In *Proc. ACL-HLT*, pages 344–352, 2008.

[85] Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. One Step Closer to Automatic Evaluation of Text Simplification Systems. In *Proceedings of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 1–10, 2014.

[86] U.W. Weger and A.W. Inhoff. Attention and eye movements in reading: Inhibition of return predicts the size of regressive saccades. *Psychological Science*, 17:187–191, 2006.

[87] C. W. Wightman, Shattuck-Hufnagel, S., Ostendorf, M., and P. J. Price. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3):1707—-1717, 1992.

[88] Wikipedia. Simple English Wikipedia, 2009. URL `http://simple.wikipedia.org/`.

[89] Wei Wu. *No Title*. Ph.d., University of Washington, 2012.

[90] Mark Yatskar, Bo Pang, Cristian Danescu Niculescu Mizil, and Lillian Lee. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proc. NAACL-HLT*, pages 365–368, Los Angeles, California, 2010.

[91] Torsten Zesch, Christof Müller, and Iryna Gurevych. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of AAAI*, volume 8, pages 861–866, 2008. ISBN 9781577353683.

[92] X N Zhang, J Mostow, and J E Beck. Can a Computer Listen for Fluctuations in Reading Comprehension? *Artificial Intelligence in Education*, 158:495–502, 2007.

Appendix A

PARALLEL SIMPLIFICATION STUDY TEXTS

A.1 Condors

It had all the makings of a soaring success story. The endangered California condor, the largest bird in North America, (had) more than tripled in population, after dwindling to just 30 birds two decades ago. But the program that's attempting to save this endangered species may be putting the bird, and even humans, "in" danger.

Steve Beissinger is a conservation biologist at UC Berkeley who says, "Some birds have come into a local California town and perched on roofs of residents' houses and pulled off the shingles and damaged television antennas. And a group of eight even went into one resident's bedroom and started pulling chunks out of his mattress."

In the mid-1980's, all wild condors were captured, and bred. Condor-like puppets raised many of their young so the parents could go on breeding. But some scientists say those young, now released into the wild, crave human interaction and are approaching people, cars and buildings. Birds have been shot, poisoned by drinking anti-freeze and electrocuted by overhead wires.

Meanwhile, the condor faces another threat: lead poisoning. Steve Beissinger says, "Birds feed on dead animals, carcasses of hunter killed deer, Jack rabbits, (and) livestock and often times there are fragments or pieces of whole bullets left behind." To keep the condor population from heading into a tailspin, Beissinger suggests alternative ammunition for hunters and an end to puppet rearing practices.

The authors of the study say their next step is to call for a committee of independent scientists to review the condor reintroduction program and make sure it stays on the right track.

A.2 WTO

On the very first day there was a tense moment and a hint of things to come. Blocked at every turn, a World Trade Organization (WTO) delegate couldn't get through the human chain of protesters. Surrounding the convention center where the WTO meeting was to be held, peaceful protesters stood their ground trying to delay the meetings inside.

They promised to stall the World Trade Organization. They succeeded at least temporarily. The first day's events were delayed for hours, due to the blockade and a security scare in the convention center.

Few could have anticipated so much anger towards such an obscure organization. A steelworker in the labor march had this to say, "A lot of people didn't know what that was about; no concept, no idea of what the WTO is." Labor groups, environmentalists and farmers formed unlikely coalitions, fearing the WTO could upset their world. Rep. George Miller says, "It affects our environment, it affects our workers, it affects our trade policy, it affects our habitat."

While tens of thousands rallied peacefully, those with more violent agendas stole the headlines. A band of about 150 self-avowed anarchists smashed windows, sprayed graffiti and looted storefronts. When the police took back the streets, the images were just as ugly. Protesters were arrested, tear gas and pepper spray was used, and finally police cordoned off the downtown area to keep protesters away from the meeting.

By the week's end there was an uneasy truce. Seattle Mayor Paul Schell says, "Ninety-nine percent of the protests went peacefully when people had their say, and it will have an impact on the WTO."

On the inside, some delegates say the cries from the outside were tough to ignore. "Maybe it has sensitized a bit, the structure of the WTO." Seemingly out of nowhere came an issue that inspired massive protests right through the final day. With the presidential race in preliminary heats, you can bet that from now on when voters want to talk trade, the politicians will listen.

A.3 Managua

city, capital of Nicaragua, lying amid small crater lakes on the southern shore of Lake Managua. One of Central America's warmest capitals, the city is only 163 feet (50 m) above sea level. Throughout the Spanish colonial period, Managua was recognized only as an Indian town, outranked by the relatively nearby Spanish cities of Len and Granada. Its choice as a permanent capital in 1857 came after partisans of those two rival cities had exhausted themselves in internecine conflict. Much of Managua was rebuilt after 1931, when it was ravaged by earthquake and fire. After the disastrous earthquake of 1972, the business section was rebuilt 6 miles (10 km) away (to the south and west) from the former city centre. It was the scene in 1978-79 of general strikes against the Somoza government and of heavy fighting, particularly in the Sandinista-held slum areas. Notable landmarks include Daro Park, with its monument to Nicaragua's famed poet Rubn Daro (see photograph); the National Palace; and the 20th-century cathedral. In 1952 the University of Managua became part of the National University of Nicaragua. Other universities are the Central American (1961) and the Polytechnic (1968; university status 1978).

Managua, the largest city in the country, is also its centre of commerce and culture. It produces a variety of small manufactures, including processed meat, furniture, metal, and textiles; it also has an oil refinery. Coffee and cotton are the principal crops grown in the agricultural hinterland. The city has railroad and highway connections with the Pacific port of Corinto and with the cities of Len and Granada. The Pan-American Highway and an international airport tie it to other Central and North American cities.

The city is surrounded by rich agricultural lands devoted primarily to the cultivation of coffee, cotton, and corn (maize). The importance of sugarcane, rice, sorghum, cattle, and horses is decreasing. Pop. (1985 est.) city, 682,111.

A.4 Manila

The capital of the Philippines, Manila gets its name from the nilad plant, a flowering shrub. The city was first known as Maynilad, but the name was later shortened. Manila is located on the eastern shore of Manila Bay, lying on the low, narrow plain of the Pasig River. The river runs through the middle of the city, dividing it into two sections. The two parts are linked by a number of bridges.

Places of interest

The Malacaang Palace used to be the home of the president. The ruins of the Spanish fortress city of Intramuros (the name means "within walls") is known for the San Agustin Church and other historic sites.

Rizal Park, the main area for outdoor recreation, features several gardens and a playground. Other public parks include the Manila Zoological and Botanical Gardens, the Mehan Garden, and Paco Park.

Major theaters in Manila include the Folk Arts Theater, the Metropolitan Theater, and an open-air theater in Rizal Park. The National Library, the National Museum, and the National Institute of Science and Technology are other cultural centers. The University of Santo Tomas was established in 1611 and is the oldest university in the Far East. There are many other universities in the city as well.

Economy

Manila is a center for trade and commerce. The city is home to textile, publishing and printing, and food industries. Its factories produce paints, medicines, aluminum items, rope, shoes, coconut oil, soap, and lumber. A number of banks are based in Manila.

History

In the late 16th century, Manila was a walled Muslim settlement. In 1571, the Spanish destroyed the settlement and founded Intramuros. Manila became the capital of a new

Spanish colony. Some scattered villages stood outside the city walls, each ruled by a local chief. As the Spaniards established colonial rule, churches and convents were built.

Manila was invaded by other countries several times. The Chinese invaded the city in 1574, and the Dutch raided it in the mid-17th century. In 1762, the city was captured and held by the British, but it was returned to Spain a year later.

Manila became the center of anti-Spanish feelings in the 1890s. The execution of a Filipino patriot, Jos Rizal, in 1896 started a rebellion against the Spanish that lasted for a year. In August 1898, United States forces captured the city during the Spanish American War. Shortly after that the United States took control of the whole country and made Manila their headquarters in the Philippines. The city developed into a major trading and tourist center.

During World War II, the Japanese captured Manila. The city was destroyed when United States forces successfully fought to get it back in 1945. In 1946, Manila became the capital of the independent Republic of the Philippines, and the city was rapidly rebuilt with aid from the United States. Population (2000 census), 1,581,082.

A.5 Pen

A pen (Latin pinna, feather) is a long, thin, rounded device used to apply ink to a surface for the purpose of writing or drawing, usually on paper. There are several different types, including ballpoint, rollerball, fountain, and felt-tip. Historically, reed pens, quill pens, and dip pens were used. Modern-day pens come in a variety of colors, shapes and assortments. The most common contain black or blue ink.

Types

The main modern types of pens can be categorized by the kind of writing tip or point:

- A ballpoint pen dispenses viscous oil-based ink by rolling a small hard sphere, usually 0.7-1.2 mm and made of brass, steel or tungsten carbide. The ink dries almost immediately on contact with paper. This type of pen is generally inexpensive and reliable. It has replaced the fountain pen as the most popular tool for everyday writing.
- A rollerball pen dispenses a water-based liquid or gel ink through a ball tip similar to that of a ballpoint pen. The less-viscous ink is more easily absorbed by paper than oil-based ink, and the pen moves more easily across a writing surface. The rollerball pen was initially designed to combine the convenience of a ballpoint pen with the smooth "wet ink" effect of a fountain pen. Gel inks are available in a range of colors, including metallic paint colors and glitter effects.
- A fountain pen uses water-based liquid ink delivered through a nib. The ink flows from a reservoir through a "feed" to the nib, then through the nib, due to capillary action and gravity. The nib has no moving parts and delivers ink through a thin slit to the writing surface. A fountain pen reservoir can be refillable or disposable, this disposable type being an ink cartridge. A pen with a refillable reservoir may have a mechanism, such as a piston, to draw ink from a bottle through the nib, or it may require refilling with an eyedropper. Refillable reservoirs are available for some pens designed to use disposable cartridges.
- A marker, or felt-tip pen, has a porous tip of fibrous material. The smallest, finest-tipped markers are used for writing on paper. Medium-tip markers are often used

by children for coloring. Larger markers are used for writing on other surfaces such as corrugated boxes, whiteboards and for chalkboards, often called "liquid chalk" or "chalkboard markers." Markers with wide tips and bright but transparent ink, called highlighters, are used to mark existing text. Markers designed for children or for temporary writing (as with a whiteboard or overhead projector) typically use non-permanent inks. Large markers used to label shipping cases or other packages are usually permanent markers.

Historic types

These historic types of pens are no longer in common use:

- A dip pen (or nib pen) consists of a metal nib with capillary channels, like that of a fountain pen, mounted on a handle or holder, often made of wood. A dip pen usually has no ink reservoir and must be repeatedly recharged with ink while drawing or writing. The dip pen has certain advantages over a fountain pen. It can use waterproof pigmented (particle-and-binder-based) inks, such as so-called India ink, drawing ink, or acrylic inks, which would destroy a fountain pen by clogging, as well as the traditional iron gall ink, which can cause corrosion in a fountain pen. Dip pens are now mainly used in illustration, calligraphy, and comics (notably manga).
- A quill is a pen made from a flight feather of a large bird, most often a goose. Quills were used as instruments for writing with ink before the metal dip pen, the fountain pen, and eventually the ballpoint pen came into use. The shaft of the feather acts as an ink reservoir, and ink flows to the tip by capillary action. Quill pens were used in medieval times to write on parchment or paper. The quill eventually replaced the reed pen.
- A reed pen is cut from a reed or bamboo, with a slit in a narrow tip. Its mechanism is essentially similar to that of a quill. The reed pen has almost disappeared but it is still used by young school going students in some parts of Pakistan, who learn to write with them on small timber boards known as "Takhti". Popular belief has it that writing with a reed pen improves handwriting.

- The ink brush is the traditional writing implement in East Asian calligraphy. The body of the brush can be made from either bamboo, or rarer materials such as red sandalwood, glass, ivory, silver, and gold. The head of the brush can be made from the hair (or feathers) of a wide variety of animals, including the weasel, rabbit, deer, chicken, duck, goat, pig, tiger, etc. There is also a tradition in both China and Japan of making a brush using the hair of a newborn, as a once-in-a-lifetime souvenir for the child. This practice is associated with the legend of an ancient Chinese scholar who scored first in the Imperial examinations by using such a personalized brush. Calligraphy brushes are widely considered an extension of the calligrapher's arm. Today, calligraphy may also be done using a pen, but pen calligraphy does not enjoy the same prestige as traditional brush calligraphy.

United States

Statistics on writing instruments (including pencils) from WIMA (the U.S. Writing Instrument Manufacturers Association) show that in 2005, retractable ball point pens were by far the most popular in the United States (26

A.6 PPA

A Power Purchase Agreement (PPA) is a legal contract between an electricity generator (provider) and a power purchaser (host). The power purchaser purchases energy, and sometimes also capacity and/or ancillary services, from the electricity generator. Such agreements play a key role in the financing of independently owned (i.e. not owned by a utility) electricity generating assets.

The seller under the PPA is typically an independent power producer, or "IPP." Energy sales by regulated utilities are typically highly regulated, so that no PPA is required or appropriate.

The PPA is often regarded as the central document in the development of independent electricity generating assets (power plants), and is a key to obtaining project financing for the project.

Under the PPA model, the PPA provider would secure funding for the project, maintain and monitor the energy production, and sell the electricity to the host at a contractual price for the term of the contract. The term of a PPA generally lasts between 5 and 25 years. In some renewable energy contracts, the host has the option to purchase the generating equipment from the PPA provider at the end of the term, may renew the contract with different terms, or can request that the equipment be removed.

One of the key benefits of the PPA is that by clearly defining the output of the generating assets (such as a solar electric system) and the credit of its associated revenue streams, a PPA can be used by the PPA provider to raise non-recourse financingfrom a bank or other financing counterparty.

Commercial PPA providers can enable businesses, schools, governments, and utilities to benefit from predictable, renewable energy.

In the United States, the solar power purchase agreement (SPPA) depends heavily on the existence of the solar investment tax credit, which was extended for eight years under the Emergency Economic Stabilization Act of 2008. The SPPA relies on financing partners with a tax appetite who can benefit from the federal tax credit. Typically, the investor and the solar services provider create a special purpose entity that owns the solar equipment.

The solar services provider finances, designs, installs, monitors, and maintains the project. As a result, solar installations are easier for customers to afford because they do not have to pay upfront costs for equipment and installation. Instead, customers pay only for the electricity the system generates. With the passage of the American Recovery and Reinvestment Act of 2009 the solar investment tax credit can be combined with tax exempt financing, significantly reducing the capital required to develop a solar project. Moreover, in certain circumstances the federal government will provide a cash grant in lieu of an investment tax credit where a financing partner with a tax appetite is not available.

Solar PPAs are now being successfully utilized in the California Solar Initiative's Multifamily Affordable Solar Housing (MASH) program. This aspect of the successful CSI program was just recently opened for applications.

Appendix B

TEXTS

B.1 Amanda

Amanda and I have been friends since we were nine. We have always shared our most enjoyable and most trying times. For instance, every time Amanda and I are together, we always have fun. We have always been able to laugh at ourselves. We always think of some new joke to tell the next day at school. We have always been able to share our escapades and humor with our friends. Just sharing our stories brings back memories of more of our escapades. On the other hand, when Amanda or I have a rough time, we are always there for each other. When Amandas grandfather died, I was there for her. She was very upset and I was there for her to talk to. Then, when my grandfather was very ill and in the hospital, I was worried about my family. She was there for me the whole time my grandfather was in the hospital.

B.2 Bigfoot

Bigfoot is supposedly a big, hairy creature that lives in remote regions of North America. Although not a single Bigfoot has ever been captured, killed, or found dead, many people believe in the creature's existence. This is largely due to giant footprints discovered in mud or snow. But are the prints genuine? In the eighties, an elderly man admitted he had made hoax Bigfoot tracks for fifty years. He carved huge feet from a piece of wood and had a friend fasten them to his own feet. Finding some unattended automobiles, they left fake footprints nearby. When the cars' owners returned, they discovered the tracks and reported their experience. A famous Bigfoot encounter or hoax occurred in the sixties when the creature was captured on film. Two Bigfoot hunters were riding horseback in California when, suddenly, they encountered a hairy monster that frightened their horses, causing them to rear. One of the hunters jumped off, grabbed his movie camera, and filmed the creature as it strode away. Bigfoot promoters believe it shows a genuine creature and not a man in a fur suit, as skeptics suspect.

B.3 Chicago Fire

I like people. Most people feel the same way. So most people live in towns and cities. One city in the U.S. is Chicago. Many years ago, there was a big fire in Chicago. The city had a large number of buildings. Many of these buildings were very beautiful. The trouble was that most of these buildings were made of wood. Wood can easily catch on fire and burn. Even buildings that appeared to be made of stone and brick burned down. They were thought to be fireproof, but in fact their frames and floors were made of wood. In the old days, builders commonly disguised wood to make it look like other kinds of building materials. The fancy exterior decorations on just about every building were carved from wood and then painted to look like stone or marble. The poorest districts in Chicago were the ones that suffered the most damage from the great Chicago fire. Lot sizes were small, and owners usually filled them up with cottages, barns, sheds, and outhousesall made of fast- burning wood, naturally. Interspersed in these residential areas were a variety of businessespaint factories, lumberyards, distilleries, gasworks, mills, furniture manufacturers, warehouses, and coal distributors. This kind of industry, collocated within residential areas, was a recipe for disaster.

B.4 Chicken Soup

What's the age-old remedy for a cold? You've guessed it, chicken soup. Until recently, no one really studied why that homemade broth is good for the cold and soul. But a researcher at the University of Nebraska Medical Center has found why chicken soup really does help. Cold symptoms such as a runny nose and cough are thought to be caused by immune cells flooding into infected areas. The cells kill germs but cause inflammation in the process. But chicken soup helps to reduce inflammation in the nose, throat, and lungs by countering the immune cells that cause the inflammation. For his tests, the researcher used his wife's soup recipe that includes onions, sweet potatoes, parsnips, turnips, carrots, celery stems, parsley, seasoning, and, of course, chicken! There is no shortage of medicinally active compounds, such as vitamins, in these ingredients. It isn't known exactly which compounds in the soup affected the immune cells.

B.5 Curly

Curly is my big black dog. He is so strong that he can carry me on his back. He likes to run and play with me, and he likes to follow my father around in the fields too. One day my father took off his coat and laid it on the ground under a big oak tree. Curly stood watching him. My father told him to watch his coat. Curly sat down on the coat. My father forgot all about his coat and went home without it. Late in the evening I missed my dog. I looked everywhere for him, calling him, but Curly did not come. Soon my father wanted something that was in his coat pocket. Then he remembered what he had done, so he went back to the big oak tree. What do you think he saw? Curly was sitting on the coat so that nobody could carry it away.

B.6 Exercise

Just about anyone, at any age, can do some type of activity to improve his or her health. Even if you have a chronic disease, you can still exercise. In fact, physical activity may help your condition, but only if its done during times when your condition is under control. During flare-ups, exercise could be harmful. You should talk to your doctor for guidance. Check with your doctor first if you are a man over forty or a woman over fifty and you plan to do vigorous activity instead of moderate activity. Vigorous activity is the kind that makes you breathe and sweat hard. Your doctor might be able to give you a go ahead over the phone, or he or she might ask you to come in for a visit. It's important to check with your doctor before increasing your physical activity. Local gyms, universities, or hospitals can help you find a teacher or program that works for you. You can also check with local churches or synagogues, senior and civic centers, parks, or recreation associations for exercise, wellness, or walking programs.

B.7 Grand Canyon

The Grand Canyon is running out of space. That sounds illogical because it is such a large place. However, many people visit the Grand Canyon. They come by bus, by motorcycle, and by car. Some people fly over the canyon in airplanes. Five million people visited the Grand Canyon last year. Most people visited the South Rim of the canyon. Six thousand cars visited that area every day. That was three times the number of parking places there. About twenty-one thousand visitors to the Grand Canyon take raft trips down the Colorado River. This river is at the bottom of the Grand Canyon. About one hundred thousand visitors hike along trails within the canyon. However, the visitors to the South Rim area are the main problem. They use a lot of water for toilets and showers. Plants and animals in the area need this water. The water comes from springs on the side of the rim. New buildings on the rim affect the springs as well.

B.8 Grandmother's House

It's like a jungle in my grandmother's house because she has so many plants. Even though she has enough plants out in her front yard, she still insists on having more. There are rows of tulips near her house, big clumps of ferns, and hedges of roses in the back. There are also pots of houseplants inside. She brings as many of her outdoor plants inside as she can for the winter. "I don't want the poor dears to freeze," she tells me, as I stare in awe at her rooms filled with greenery. "Besides," she likes to say, "a house full of plants is much cozier than a house without. And mark my words, there's more magic in a house filled with plants." "Okay, Grandma," I say because I don't want to argue with her. One night I slept in a sleeping bag on the floor of my grandma's front parlor. The front parlor by far has the most plants in the house. My two older brothers call it the jungle room because we can no longer see the wallpaper. All we see when we walk in the door are leaves and colorful flowers. It actually smells quite nice. I was secretly excited to be camping out there because it would almost be like sleeping in a real forest minus the hard ground. Grandma made a fire in the fireplace that night so I could roast marshmallows and read books. I read until around midnight. At about that time, the fire went out and my aching eyes dropped shut. I closed my book and laid my head on the pillow.

B.9 Guide Dogs

Guide dogs, or seeing eye dogs, lead very interesting lives. For at least ten or twelve years they are responsible for leading a blind person. To do this job, they must be intelligent, gentle, and very well trained. Most guide dogs are born at a kennel. Since dogs are gentler when raised by a family, the dogs are given to children. When the dogs are about fourteen months old, they come back to the kennel to be trained. Then the children get new puppies. The dogs train in groups for three months. They know more at the end of that time than most dogs ever learn. But the training isn't over. Their new masters arrive and they train together for one more month. At the end of that time, they are ready for the world. A dog loves nothing better than to be with its master, and guide dogs keep their masters company all the time.

B.10 Lori Goldberg

The other day as Lori Goldberg was walking through the halls of her school, a few whispers could be heard from the children she passed, but most greeted her with grins and hellos. As a principal, Goldberg may be an authority figure, but she doesn't want to instill fear in students. She wants their trust. Goldberg said the door to her office is always open and the students are encouraged to talk to her about any problems they might be having. Some even get the chance to read to the principal from the comfort of a small rocking chair. Goldberg said once a child can read a book, they read to her and get a certificate in return. Her interest in children led her to pick education over other career options years ago. Goldberg said being the principal gives her the ability to be part of a long term vision and larger goals for the school.

B.11 Word List 1

up	day	angry
it	true	major
be	fall	door
off	match	speak
in	none	while
was	just	
fee	any	
all	state	
air	level	
lay	tore	
he	news	
my	index	
had	drive	
so	city	
red	work	
cut	very	
as	needs	
do	except	
ten		

B.12 Word List 2

respect
since
training
force
noble
took
truth
design
little
army
likely
even
into
other
under
ever
before
people
around
always
number
family
easily
without
himself
despite
another
example
history
several
federal
primary
usually
evidence
business
personal
together
economic
something
including
essential
committee

B.13 Pseudoword List 1

mip
leb
het
jad
wom
dag
vod
zun
bim
fub
kig
tup
dac
nyd
blaff
praw
glay
cuge
teeld
slipe
ceft
hape
stoy
floud
frew
trode
griep
roce
phraw
clebe

B.14 Pseudoword List 2

cheme
cilf
drobe
choop
dyte
plix
shune
doin
necy
legute
fabit
rupper
prattle
mullow
plinnor
sible
corelic
platic
setric
drousing
constrieve
quorisian
tronic
stapely
vortastic
manious
recilf
decrift
custalian
rofedian

Appendix C

COMPREHENSION QUESTIONS

Passage	**Comprehension Question**
Amanda	Who was ill and in the hospital? • Amanda's grandfather • My grandfather • Amanda • Me
Bigfoot	Where did two horseback riders see Bigfoot? • Canada • Washington • California • Texas
ChicagoFire	What were the buildings in Chicago made of? • Wood • Brick • Stone • Glass
ChickenSoup	What is the age-old remedy for a cold? • Water • Chicken Soup • Aspirin • Sleep

Curly	What color is Curly? • Black • Brown • White • Gray
Exercise	Who can do exercise? • Just about anyone • Healthy people • Young people • Old people
GrandCanyon	What part of the Grand Canyon is visited by the most people? • The South Rim • The Colorado River • The Trails • The Buildings
GrandmothersHouse	What does my grandmother have a lot of? • Plants • Pets • Cookies • Books
GuideDogs	How long do the dogs train in groups? • 10 years • 10 months • 3 years • 3 months

LoriGoldberg	Where does Lori Goldberg work? • A Zoo • A Library • A Store • A School