Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications

D i s s e r t a t i o n
zur
Erlangung des akademischen Grades
Doktor der Philosophie
in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

vorgelegt von

Sowmya Vajjala Balakrishna

aus

Vijayawada,
Indien

2015

**Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen**

Dekan: Prof. Dr. Jürgen Leonhardt

Hauptberichterstatter: Prof. Dr Detmar Meurers
Mitberichterstatter: Prof. Dr Katharina Scheiter, Prof. Dr Harald Baayen

Tag der mündlichen Prüfung: 27. Juli 2015

TOBIAS-lib, Tübingen

Abstract

Reading plays an important role in the process of learning and knowledge acquisition for both children and adults. However, not all texts are accessible to every prospective reader. Reading difficulties can arise when there is a mismatch between a reader's language proficiency and the linguistic complexity of the text they read. In such cases, simplifying the text in its linguistic form while retaining all the content could aid reader comprehension. In this thesis, we study text complexity and simplification from a computational linguistic perspective.

We propose a new approach to automatically predict the text complexity using a wide range of word level and syntactic features of the text. We show that this approach results in accurate, generalizable models of text readability that work across multiple corpora, genres and reading scales. Moving from documents to sentences, We show that our text complexity features also accurately distinguish different versions of the same sentence in terms of the degree of simplification performed. This is useful in evaluating the quality of simplification performed by a human expert or a machine-generated output and for choosing targets to simplify in a difficult text. We also experimentally show the effect of text complexity on readers' performance outcomes and cognitive processing through an eye-tracking experiment.

Turning from analyzing text complexity and identifying sentential simplifications to generating simplified text, one can view automatic text simplification as a process of translation from English to simple English. In this thesis, we propose a statistical machine translation based approach for text simplification, exploring the role of focused training data and language models in the process.

Exploring the linguistic complexity analysis further, we show that our text complexity features can be useful in assessing the language proficiency of English learners. Finally, we analyze German school textbooks in terms of their linguistic complexity, across various grade levels, school types and among different publishers by applying a pre-existing set of text complexity features developed for German.

కర్మణ్యేవాధికారస్తే మా ఫలేషు కదాచన|

మా కర్మఫలహేతుర్భూర్మా తే సఙ్గోఽస్త్వకర్మణి|| 2-47 ||

कर्मण्येवाधिकारस्ते मा फलेषु कदाचन ।
मा कर्मफलहेतुर्भूर्मा ते सङ्गोऽस्त्वकर्मणि॥ २-४७

Transliteration:
karmaṇy evādhikāras te mā phaleṣu kadācana
mā karma-phala-hetur bhūr mā te saṅgo 'stv akarmaṇi

Translation:
You have the right to work only but never to its fruits. Let not the fruits of action be your motive, nor let your attachment be to inaction.
(Chapter 2, verse 47 - Sankhya Yoga, Bhagavad Gita)

Acknowledgements

I start any piece of literature - be it fiction or scientific research, reading the acknowledgements section. Now, the time has come for me to write one. Several people supported me at various points of time from the beginning till the end of this thesis.

I would like to first thank my advisor Detmar Meurers for all the support, advice, encouragement and lively discussions during the past 4.5 years. I learnt a great deal working with him, which will definitely be useful for me in my future career as a researcher. I thank my second advisor Katharina Scheiter for the timely feedback. All the discussions with her taught me a great deal about doing interdisciplinary research and communicating it. I am thankful to Haraald Baayen for reviewing the thesis and for introducing me to Generalized Additive Mixed Models patiently, with infectious enthusiasm. I also thank Doreen Bryant and Andrea Weber for being on my thesis committee.

I thank all the faculty members, PhD students, postdocs and administrative staff at LEAD for all the fun at Gartenstraße offices, and lively discussions at the retreats and in the weekly colloquium. This surely is a unique experience and I could not have learnt what I learnt here anywhere else. I would also like to thank the CLARA[1] project administration and all its members for a friendly and stimulating research network that helped me get started with my research in Tübingen in 2011. I am very thankful to the administrative staff in CLARA/SfS and LEAD for their efficiency and promptness in responses.

I thank my current and ex-colleagues in SfS - Petra Augursky, Adriane Boyd, Serhiy Bykh, Julia Hancke, Julia Krivanek, Kaidi Lõo, Jianqiang Ma, Niels Ott, Jochen Saile and Ramon Ziai for a friendly work environment and for all the con-

[1] http://clara.b.uib.no/

versations about work and non-work. Jochen should be specially mentioned for keeping me focused by asking for daily updates about the thesis. Special thanks to Magdalena Wolska and Martí Quixal not only for the informative discussions about research but also for all those long endless conversations and emails on life, literature, nonsense fiction and everything under the sun (and also for the very special post-defense hat!)

My collaborators outside SfS - Elena Volodina and Ildikó Pilán (Gothenburg University), Karin Berendes, Doreen Bryant and Alexander Eitel should be specially thanked for making all these collaborations work successfully. I have been fortunate enough to work with and mentor several bachelors and masters students in Tübingen. I thank Doreene Amati, Ido Freeman, Sabrina Galasso, Spyridoula Georgatou, Dorothee Hoppe, Tobias Kolditz, Ulla König, Lilyana Nikolova, Eyal Schejter, and several others with whom I interacted less. Attempting to answer their questions several times gave me more clarity about several research and implementation issues apart from teaching me a lot about people management. I also thank all the students who were a part of the seminars I co-taught - discussions in the seminar sessions immensely helped me with my research.

Several of my colleagues and friends, spread across multiple locations in Europe, reviewed parts of this thesis. I found their comments coming from multiple backgrounds and perspectives immensely helpful in presenting the work clearly. I thank Hector Martínez Alonso, Petra Augursky, Serhiy Bykh, Xiaobin Chen, Simón Ruiz Hernández, Thomas Lösch, Jianqiang Ma, Carla Parra, Ildikó Pilán, Robert Reynolds, Martí Quixal, Anne-Kathrin Schumann, Michèle Suhlmann and Ramon Ziai for all the useful comments. I promise to return the favor when the time comes!

Let me now move on to people beyond my academic life now. Thanks to all my friends - Indu aka Madhuravani, Lorenzo, Madhura Panse, Purnima, Rahul, Santosh Raju, Sinem Erisken, Suneetha and Varun, for lending an ear to my rants about research and life several times. Special thanks to Lorenzo for spreading that contagious curiosity about everything and for making me read Astrophysics abstracts! Thanks to Pratibha Rani (IIIT-H) who always provided me the pdfs of locally inaccessible articles, almost instantly. Gita Ramaswamy of Hyderabad

Book Trust and Narayana Sarma of Kottapalli[2], a Telugu Childrens Magazine, need to be thanked for giving me opportunities to write, providing a welcome distraction from research. The readers and writers at pustakam.net should be remembered here for keeping me occupied in good and bad times over the past 4.5 years. I cannot thank Anu and Vara Mullapudi enough, for bringing out audio books of "Koti Kommacchi", the autobiographical trilogy written by Mullapudi Venkata Ramana[3]. More than anything, this ensured I did not feel homesick in Germany and listening to his words everyday surely made me wiser. :-).

Nothing would have been possible without the silent support from my mother Geethamani. Nothing I worry about seems to worry her and thus, she remains my pillar of strength. She has been the only one to date who never asked "How long do you need to graduate?". My brother Halley, my aunts, uncles and cousins have all been incredibly kind and supportive all through. This is the time to remember my late father, Balakrishna, who encouraged us (me and Halley) even as children to think, question and look for answers by ourselves. Lastly (but most importantly), I thank my husband Sriram, for everything. Without him in my life, I would have neither come to Tübingen to start my PhD nor would have managed to stay till the end!

Now, for everyone else, who all these days worried more than me, my advisors and my family about my thesis: Here it is, and be ready to take an oral exam from me 2 months later!!

[2] http://kottapalli.in/

[3] http://en.wikipedia.org/wiki/Mullapudi_Venkata_Ramana

Publications

Parts of this thesis appeared in the following peer-reviewed publications:

1. Sowmya Vajjala & Detmar Meurers, Readability-based Sentence Ranking for Evaluating Text Simplification, under review. 2015.

2. Sowmya Vajjala & Detmar Meurers, Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications. *International Journal of Applied Linguistics*, Special Issue on Recent Advances in Automatic Readability Assessment and Text Simplification. 65:2. John Benjamins Publishing Company. 2014. (pp. 194–222).

3. Sowmya Vajjala & Detmar Meurers, On assessing the reading level of individual sentences for text simplification. *In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association of Computational Linguistics. 2014. (pp 288–297).

4. Sowmya Vajjala & Detmar Meurers, Exploring Measures of "Readability" for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Association of Computational Linguistics. 2014. (pp. 21–29).

5. Sowmya Vajjala & Detmar Meurers, On The Applicability of Readability Models to Web Texts. *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. Association of Computational Linguistics. 2013. (pp. 59–68)

6. Serhiy Bykh, Sowmya Vajjala, Julia Krivanek & Detmar Meurers, Combining Shallow and Linguistically Motivated Features in Native Language Identification. *In Proceedings of the 8th workshop on Innovative Use of NLP for Building Educational Applications (BEA8)*. Association of Computational Linguistics. 2013. (pp. 197–206).

7. Sowmya Vajjala & Detmar Meurers, On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition, *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)*. Association of Computational Linguistics. 2012. (pp 163–173).

8. Julia Hancke, Sowmya Vajjala & Detmar Meurers, Readability classification for German using Lexical, Syntactic and Morphological features. *In Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Association of Computational Linguistics. 2012. (pp. 1063–1080)

Funding

My research has been funded by:

- 2013-2015: LEAD Graduate School (GSC 1028), a project of the Excellence Initiative of the German federal and state governments.
  (`http://purl.org/lead`)

- 2011-2013: Early Stage Researcher fellowship under European Commission's 7th Framework Program, grant agreement 238405.
  (CLARA, `http://clara.uib.no`)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Reading is one of the common modes of language learning and knowledge acquisition. Thus, anything that causes difficulty during reading is then going to affect the process of learning and comprehension. The causes of reading difficulty in a text and its effects on learners are of specific interest for educational researchers. Hence, the linguistic properties of texts that contribute to reading difficulty, like vocabulary, syntax and cohesion have been widely studied and debated for several decades now in educational psychology. The primary purpose of this research has been to decide what students should read (e.g., Lively & Pressey, 1923; Vogel & Washburne, 1928; Islam et al., 2012; Fitzgerald et al., 2015).

An approach to automatically assess the reading difficulty of a text could be useful for teachers selecting topic specific reading material for their students. In the days of availability of a wide range of information on the web, such an approach could also be beneficial for students wanting to learn about a certain topic as well, since it can suggest them relevant texts on the web that are suitable for their grade level. Apart from educational contexts, the analysis of reading difficulty can be useful in several scenarios where reading and understanding textual content plays an important role, for e.g., reading legal texts.

Because of this practical relevance, the task of automatic assessment of text complexity has attracted the attention of researchers working in several disciplines connected to language, such as education, cognitive science, psychology, linguistics and second language acquisition, to name a few. Formulae for calculating the

readability of texts, typically on the scale of school grade levels, have been proposed since the early 20th century, for various kinds of target readers. However, despite this long standing interest, most of the proposed readability formulae use surface indicators of text complexity like word length, sentence length and amount of difficult words in a text (see DuBay (2006) for a survey). To some extent, this was also because of the computational and text processing limitations of those times.

1.1 Automatic Readability Assessment

Recent developments in Computational Linguistics and Computer Science allowed the exploration of readability using more diverse linguistic indicators of text complexity and robust computational models. In the past 15 years, wide range of textual features were explored for automatic readability assessment, modeling the lexical, syntactic and semantic properties of the text. However, the traditional readability formulae remained popular and the real life applicability of the computational readability approaches was not explored much. This started to change in the recent past, with the introduction of educational standards like the Common Core[1] standards in the United States.

The Common Core State Standards (CCSSO, 2010a) Initiative is a set of guidelines for academic standards in the United States of America, which outlines what students in the primary and secondary school should know at the end of each grade in Mathematics and English/Language Arts. They were developed by the Council of Chief State School Officers (CCSSO) and the National Governors Association Center for Best Practices (NGA Center) in collaboration with teachers, school chiefs and other educational experts. In the "Reading" component of the language arts section, the standards call for a "staircase of increasing text complexity" for the students i.e., the students should be able to read more and more complex texts with increasing grade level.

A sample of 168 exemplar texts were also provided in the Appendix A of the Common Core Standards document, annotated with their appropriate grade level.

[1] `http://www.corestandards.org`

The standards were quickly adapted and endorsed by commercial systems that provide frameworks for assessing reading levels of texts (e.g., Lexile [2], DRP[3], Reading Maturity Metric[4]).

Nelson et al. (2012) compared some of the existing academic and commercial reading level prediction systems and showed that the approaches that consider diverse aspects of language complexity perform significantly better than the relatively shallow measures used by many systems, using multiple standard test sets annotated with a reading level. The Common Core Standards test set described in the previous paragraph too was one of them. Fitzgerald et al. (2015) studied texts written for primary school children in terms of their text complexity and reached the conclusion that several linguistic dimensions (word structure, word meaning, sentence and discourse level characteristics) and the interplay among them contribute to text complexity. In a different context of L2 writing assessment, Bulté & Housen (2014) also reached the conclusion that a wide range of complexity measures need to be calculated to get a comprehensive picture of writing complexity. These recent results emphasize the need to model a broader range of textual features for readability assessment.

The Elementary School Journal, a 115 year old educational research journal had a special issue on *Understanding Text Complexity* in December 2014, which prompted an in-depth discussion on approaches to automatically analyze text complexity and the methodological issues involved. The discussions in this issue stressed on the importance of modeling readability considering a broader range of linguistic properties, reader comprehension and the learning task involved (e.g., answering questions vs reading instructions).

These recent developments clearly indicate that the readability assessment systems can now play an important role in educational policy making and in choosing appropriate reading material for students.

Our approach: Against this backdrop, in this thesis, we proposed a supervised machine learning based approach to readability assessment, which models a broad

[2] https://www.lexile.com/using-lexile/lexile-measures-and-the-ccssi/

[3] http://drp.questarai.com/home/about-the-drp/drp-and-the-common-core/

[4] http://www.readingmaturity.com/rmm-web/

range of textual features that are based on research in Second Language Acquisition and Psycholinguistics along with several other linguistic characteristics of language, using state of the art methods in natural language processing research. This approach still considers only the nature of the text. However, we also conducted an eye-tracking study with a recall and comprehension task for the readers, that partially takes the reader and task into account.

1.2 Validity of Readability Assessment Approaches

While it is clear that robust automatic readability assessment approaches hold a promising future in educational applications, this discussion around the usefulness of readability models also raises questions about the empirical basis, validity and applicability of readability models for a specific application context. Since the initial days of readability research, these issues have been studied and debated extensively. As readability formulae were being used for all sorts of application scenarios irrespective of their original purpose, researchers like Bruce & Rubin (1988) cautioned about the pitfalls of this practice. They recommended that they should be used on the same population as the formula was validated against. More recently, Begeny & Greene (2013) criticized the use of readability formulae in school, arguing that they are valid only at particular grade levels, by comparing them with the oral reading fluency of children from elementary school.

Automatic readability assessment methods can be validated by either data driven or user based approaches. Another approach could be to validate the models by comparing them with theoretical results from research on human language processing.

Data driven validation: From a textual perspective, the validity of an automatic approach can be verified by testing its robustness by performing cross-corpus evaluations with corpora consisting of graded texts, intended for the same target audience.

User based validation: From the target user perspective, one has to perform an empirical evaluation and compare their performance outcomes with texts of varying difficulty.

Text Complexity and Human Processing Another approach to validate the readability assessment approaches is to explore the correlates of linguistic text complexity models with some of the research on the processing of text in the human mind. Researchers in areas like neuroscience, psycholinguistics and cognitive psychology have theoretically and experimentally studied the process of text comprehension in the human mind. Neuroscience research in the past few decades explored ways to measure brain activity, by detecting changes in the blood-flow in various regions of the brain that are affected by word-level, syntactic and semantic level characteristics of text. For example, Hruby & Goswami (2011) summarized some of the research on neural correlates of reading comprehension and difficulties.

In psycholinguistics, the notion of text complexity has been studied in the context of text comprehension, to understand what types of linguistic constructions cause comprehension problems for humans (cf. Chapter 10, Aitchison, 2011, for a summary). Text processing difficulty was also studied by computationally modeling surprisal, working memory constraints and parsing costs for a text (e.g., Boston et al., 2008; Roark et al., 2009; Boston et al., 2011). Eye-tracking is commonly used in Cognitive Psychology research to understand online processing of text by the subjects through their eye-movement patterns (Rayner, 1998). Some of the eye-tracking measures like Fixation Count and Average Fixation Duration are known to correlate with text processing difficulty for the readers (Rayner et al., 2006).

Validating our approach: In this thesis, we employed both data-driven and user-based studies to evaluate the validity of our automatic readability assessment approach. In terms of the data-driven evaluation, validated our readability model by using it to predict the reading level of other externally validated texts like Common Core exemplars. The validity of the features was studied by using the features to construct multiple readability prediction models, using datasets that cover diverse topics, and genres (written vs. spoken). To understand the effect of text complexity on readers, we conducted an eye-tracking study where the subjects read texts in two versions, prepared by external experts. Through this study, we studied both the cognitive correlates of text complexity as well as the impact of

text complexity on the performance outcomes of the subjects.

1.3 Text Simplification

While difficult to read texts pose problems to the target readers, providing them comprehensible texts can be seen as a solution to address this issue. Text simplification can be seen as one way of achieving the goal of providing comprehensible texts for users, on a given topic. The effect of text simplification on target readers was explored in psychology research before. For example, in psychology, text comprehension theories were proposed to understand the comprehension processes in humans and educational research attempted to use some of these models in instructional practice, to rewrite texts for better comprehension. Empirical investigations about the effect of the rewritten texts on the cognitive processing and performance outcomes were conducted using reading time experiments, comprehension questions and free recall.

Kintsch & van Dijk (1978) proposed a theoretical model of text comprehension, which was used by Britton & Gülgöz (1991) to rewrite texts. This model was shown to improve free recall for users. Britton et al. (1989) compared the original and revised versions of several texts prepared by five experts and concluded that the experts did not always reproduce rewritten versions reliably and had imperfect declarative knowledge about their revisions. They called for the experts to learn to be more explicit about rewrites and hoped for the development of technology that can do this automatically. Automatic text simplification and tools that provide explicit feedback to the writers about the complexity in the text produced can be seen as solutions to this issue.

Automatic Text Simplification approaches have been proposed in the recent past, both with the use of linguistic simplification rules as well as using statistical learning based approaches. There are two primary ways to evaluate automatic text simplification. From a reader perspective, one needs to conduct a user study to assess the impact of text-simplification on readers. From a data-driven perspective, readability assessment models can be used for the linguistic evaluation of simplification. They can also be used for selecting what to simplify, apart from evaluating the degree of simplification performed by the system.

Our approach: In this thesis, we explored automatic text simplification as a monolingual, statistical machine translation problem, as english to simple english translation. We primarily focused on the role of focused training data and use of various language models in improving the generation of simplified text. We also investigated the utility of readability models to compare the normal and simplified sentences in terms of their reading levels. We validated our approach by comparing the BLEU scores between machine-translated text and the original text. Though we studied the evaluation of manual text simplification in terms of readability, we did not perform any evaluation of automatically simplified text in terms of the readability yet.

1.4 Other Applications of Text Complexity Analysis

Apart from readability assessment and text simplification, the estimation of text complexity can also be useful for other educational applications. In this thesis, we used the linguistic features developed for readability assessment to two other problems specific to educational research:

- assessing the second language proficiency of adult English learners by modeling the linguistic complexity of the texts they produced
- analyzing the reading demands in German school textbooks in terms of their text complexity

1.5 Research Questions

In sum, we address the following questions in this thesis:

1. Can we build an automatic readability model that generalizes well to new texts and genres?

2. How does text complexity affect the reader's cognitive processing and performance outcomes?

3. Can we accurately quantify the readability of texts at a sentence level, for evaluating the degree of simplification performed?

4. Can we simplify texts automatically to a given reading level?

5. Can we apply the analysis of text complexity in other educational contexts?

1.6 Terminology

In the existing body of literature on this topic, the terms text complexity and text difficulty are sometimes considered different constructs. While text complexity refers to the linguistic properties of text that are believed to affect comprehension, text difficulty refers to comprehension considering the user's reading ability (Mesmer et al., 2012). While in this thesis, we focus on the text complexity aspect, we use the terms interchangeably like Sheehan et al. (2014). We use the term automatic readability assessment as the process of analyzing this construct. The difficulty that a text poses to a reader also depends on the population of learners (e.g., first versus second language learners, children versus adults), their socio cultural background and other aspects like the interaction between the reader and the text. This thesis is primarily concerned with the textual features that make it difficult to read.

1.7 Approach and Key Contributions

We applied supervised machine learning approaches with existing graded corpora as our basis for developing automatic readability and simplification models. In terms of the features, we used a combination of features from Second Language Acquisition (SLA), word-level properties from pre-existing psycholinguistic databases and other linguistic features that can be automatically extracted using publicly accessible NLP software. We evaluated the performance of the models we built by multiple cross-corpus, data driven evaluations. Apart from these, we also performed an eye-tracking study to understand the effect of text complexity on readers. The specific contributions of this thesis are:

- We compiled a corpus of web texts consisting of five grade levels, which can be used to train readability models.

- We developed a state of the art readability model that reports second best results on the common core standards test set (Spearman's rho: 0.69). In this process, we created feature set that takes into account several linguistic properties and generalizes to spoken language too, with a classification accuracy of 96%.

- We performed an empirical analysis about the effect of text complexity and language proficiency on readers' cognitive processes and reading performance through an eye-tracking study. We show experimentally that both the variables impact the online and offline processing of the readers and in some cases, there is an interaction between them.

- We proposed a state of the art approach to rank sentences based on their difficulty level and compare the degree of simplification between them. Our approach orders the sentences correctly with $>80\%$ accuracy.

- We created a corpus of parallel simplifications of sentences across three levels of simplification, compiled from texts created by human experts.

- We explored a machine translation based text simplification approach that is trained with data containing only lexical simplifications and paraphrases that results in a better performance compared to an approach trained on a large, noisy corpus, achieving better BLEU scores in the process.

- We report on two experiments studying the application of readability oriented features to educational applications:

    1. Modeling the proficiency of L2 English writing using linguistic complexity features for real life English learner texts.
    2. Analysis of geography text books in German schools in terms of their linguistic complexity.

1.8 Outline of the Thesis

The rest of this thesis is organized as follows:

Chapter 2 first describes prior work on automated readability assessment and its applications. Later, the applications of readability assessment methods in automatic and manual text simplification are discussed. Finally, after a brief survey of research on automatic text simplification, this chapter concludes connecting these three parts and connecting the current thesis to existing research.

Chapter 3 describes our approach to readability assessment in terms of the corpora, features, modeling and evaluation. Several experiments performed to establish the validity of the proposed models across genres, topics and corpora are described in this chapter. This chapter is based on the results reported in Vajjala & Meurers (2012, 2014b,c).

Chapter 4 describes an eye-tracking experiment that studies the effect of text complexity and simplification on the cognitive processing and performance outcomes of the readers.

Chapter 5 connects readability assessment to text simplification. Here, we study readability assessment at the level of sentences to compare simplified and unsimplified versions of sentences. This chapter is partially based on the results from Vajjala & Meurers (2014a) and Vajjala & Meurers (under review).

Chapter 6 explores a Statistical Machine Translation (SMT) based approach to text simplification that handles specific operations, explores the role of various language models and studies the generalizability of the SMT approach by cross-corpus evaluation.

Chapters 7 and 8 describe two applications of readability analysis in Educational contexts. Chapter 7 uses the features described in Chapter 3 to analyze L2 learner writing samples and perform proficiency classification. Chapter 8 utilizes an existing German feature set for analyzing text complexity from Hancke (2013) to analyze the reading demands in German school textbooks, and to predict the grade level of a text book based on the linguistic features.

Finally, Chapter 9 summarizes the contributions of this thesis and discusses potential extensions to this work.

Part I

Readability Assessment and Text Simplification

Chapter 2

Literature Review

Abstract

In this chapter, we survey the contemporary and past research on readability assessment and text simplification. We also briefly discuss the key contributions of this thesis to both the areas.

2.1 Introduction

Automatic readability assessment (ARA) approaches for several application scenarios existed for almost a century now. Within educational research, the primary application of readability assessment was to assign a reading level to texts read by learners. Although early studies on readability focused on a variety of linguistic and non-linguistic textual properties, subsequent studies typically relied on easy to estimate surface measures like average length of words and sentences in a text to create formulae for readability, as they were found to correlate with other, more reliable measures. This was also motivated by the fact that there were no existing computational tools for automatic estimation of all textual properties. With the evolution of Natural Language Processing (NLP) tools for processing texts, more sophisticated approaches for readability assessment, involving statistical and Machine Learning methods started to emerge.

While traditional formulae and computational approaches continued to be pro-

posed, researchers questioned the validity of these readability approaches. The evolution of the Common Core Standards initiative (CCSSO, 2010b) in the United States created a gold standard test set for readability assessment. This was created based on the exemplar texts per grade level from the Common Core Standards document. Comparing the performance of different approaches on this test set could address the external validity of the approaches to some extent.

Once we decide that a text is difficult to read for a target reader, one way to address the issue is to simplify difficult texts to the level of the reader. Educational researchers studied the effect of simplifying texts on the comprehension of the learners and their results showed that simplified text could result in better comprehension in some scenarios. Automatic Text Simplification (ATS) could be useful in such a scenario. Interest in ATS started two decades ago to improve natural language parser efficiency and expanded to various application scenarios ranging from language learning to understanding protein-protein interactions in the Biomedical domain. In an educational context, the aim of automatic text simplification would be to provide easy to comprehend texts for language learners. Readability assessment can be very useful in such a scenario in various stages like choosing targets for simplification and to evaluate the degree of simplification. However, this relationship between readability and simplification has remained largely unexplored in the literature so far, beyond using traditional readability formulae. In this thesis, apart from developing and testing new approaches for automatic readability assessment and text simplification, we also connect the missing link between them, for a language learning scenario, by using readability assessment to evaluate text simplification.

The rest of this chapter provides an overview of the research in both ARA and ATS and puts this thesis in context. Section 2.2 first summarizes existing research on readability assessment and its applications and describes how the current thesis adds to this area. Section 2.3 summarizes the research in text simplification so far and briefly describes the contributions of this thesis to this area.

2.2 Automatic Readability Assessment (ARA)

2.2.1 Early Work and Traditional Readability Formulae

The need for assessing the difficulty of a text in an objective manner dates back to almost a century. Most of the early work on this topic focused on the lexical aspects of a text (i.e., difficult words) and was directed at school children and their reading tasks. Thorndike's approach (Thorndike, 1921; Thorndike & Lorge, 1944) can be considered the first systematic approach to assess the difficulty of texts. Thorndike (1921) compiled a list of 10,000 words in English along with their respective frequencies, providing school teachers a way to measure the difficulty of texts based on the vocabulary used. This list was followed up with another list in Thorndike & Lorge (1944), which consisted of 30,000 words. This list served as a basis for many more readability formulae during the decades that followed. Lively & Pressey (1923), while investigating the problem of selecting appropriate science textbooks for junior high school students, concluded that the median of the index numbers of words taken from Thorndike (1921) was the best indicator of vocabulary difficulty for the texts they chose and the measures they studied.

Vogel & Washburne (1928) studied multiple aspects of readability to assign appropriate grade levels for children's reading material. Apart from using Lively & Pressey (1923)'s technique for vocabulary difficulty, they also considered several aspects of the syntactic structure of the document, part-of-speech tag distribution in a sample of 1000 words, paragraph construction and physical makeup. They finally proposed the *Winnetka Readability formula* with four textual properties after removing all the correlated features. These four properties are: number of different words, number of prepositions, number of uncommon words in the 1000 word sample for a text and the number of simple sentences in a sample of 75 sentences. The scores from this formula were later mapped to the grade levels at schools. They validated their approach against 700 books (called the *Winnetka Book List*) that had been labeled as *liked* by at least 25 of around 37,000 children. This is the first formula that predicted readability by grade levels. Patty & Painter (1931) continued with the problem of measuring the vocabulary burden of textbooks and proposed an approach that combined Thorndike's list with vocabulary

diversity in the text.

Readability formulae for adult readers: In the early 30s, the focus of readability research shifted from school children to adults with reading difficulties. Dale & Tyler (1934) published a study on factors causing reading difficulties for adults of limited reading ability. They found that 10 out of the 29 significant factors for children's comprehension also had significant effects for adults. Of these, they constructed a readability formula considering three most predictive factors - number of different technical words, number of different hard non-technical words and number of indeterminate clauses. This formula was empirically validated by correlating its scores with performance in multiple-choice tests. Lorge (1944)'s formula which included three factors - average sentence length, number of prepositional phrases per 100 words and number of hard words not on the list of Dale's list of 769 easy words. It used McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs, 1926) as a basis for reading difficulty, which was also used by others in the following decades.

Gray & Leary (1935) can be considered the most comprehensive study of that time on the topic. They studied readability for adults with limited reading ability and identified 228 factors that affect the readability of a text. They divided these factors into four groups - content, style, format and organization and concentrated primarily on 64 variables of style, which can be measured reliably. The empirical basis of their reading difficulty scores was based on the results of reading comprehension tests performed with adult readers. Finally, they created a formula with five variables - average sentence length in words, number of different "hard" words, number of first, second and third person pronouns, percentage of different words and number of prepositional phrases.

The focus on adults with limited reading ability continued in the post-WWII phase, when the emphasis on clear writing increased. This resulted in the creation of several readability formulae that are still in use today (e.g., Dale-Chall Formula (Dale & Chall, 1948a; Chall & Dale, 1995), Flesch Reading Ease Formula (Flesch, 1943, 1948) and Gunning-Fog Index (Gunning, 1968)). The Dale-Chall Formula, which is a two-factor formula based on number of words not in Dale list of 3000 words and average sentence length, was shown to correlate highly

with several texts apart from the original McCall-Crabbs passages like health education materials and current affairs articles. It also included a manual (Dale & Chall, 1948b) for the application of the formula correctly.

Flesch (1948) examined readability by considering average word length (in syllables) and average sentence length (in words), average percentage of personal words and average percentage of personal sentences as features. A later version considered only the first two features and measured readability on a scale of 0-100. Variants to the Flesch formula were also proposed (Farr et al., 1951; Kincaid et al., 1975). Two other formulae of this type, which use variants of word-length, sentence-length and lists of difficult words as factors are Coleman-Liau index (Coleman & Liau, 1975) and SMOG (McLaughlin, 1969). Word lists like West's General Service List (West, 1953) and Coxhead's Academic Word List (Coxhead, 2000) were also used to create readability formulae. Bormuth (1966) also investigated readability using a large number of linguistic variables and cloze tests and concluded that more sophisticated linguistic variables are needed to improve the readability prediction. Granowsky & Botel (1974) turned the focus back to more fine-grained syntactic measures of readability by claiming that sentence length is an insufficient measure of syntactic complexity.

2.2.2 Validity of Readability Formulae:

Although more than 200 formulae have been proposed in the past century, empirical validity of the formulae and the relationship between them have been questioned by multiple researchers (e.g., Bruce & Rubin, 1988; Anderson & Davison, 1988). Klare (1974) compared several readability formulae that either were created or revised since 1960, in terms of their application scenarios and computability and concluded that the formulae do not indicate causes of difficulty in a text but only serve as an indicator of difficulty. More recently Pearson & Hiebert (2014) advocated the need for a more qualitative analysis of text complexity, to counter the unchecked quantitative approaches like readability formulae in cases where they can be invalid.

Klare (1952) compared some of the existing readability formulae at that time in terms of correlations between each other. In the discussion about coming up

with new readability formulae that included more features, he mentioned about the need for an inter-disciplinary approach combining knowledge on the topic from fields like linguistics, psychology, typography and semantics. However, most of the formulae from those days were manually estimated and thereby imposed a constraint on the creation of formulae that included more features. Quick, short hand formulae that can be easily estimated continued to be used. Klare (1974); DuBay (2004, 2006) discuss in detail on the early decades of work on readability assessment. Although computer programs to calculate readability formulae started by the late 60s (e.g., Klare (1969)), feature-rich modeling of readability started only after the evolution of natural language processing approaches.

However, this did not diminish the popularity of old style readability formulae. Some of these formula-based approaches such as Lexile[1] (Stenner, 1996), Degrees of Reading Power[2] are still being used widely in educational applications. Flesch-Kincaid scores are being used in a range of applications including Microsoft Word. Some of the popular readability formulae are readily computable on the internet[3]. Thus, the widespread use continued despite all the criticism.

Formulae for other languages: It has to be noted that all the traditional readability research described so far are exclusively focused on English. However, such approaches with surface features were also proposed for other languages, like French (e.g., Kandel & Moles, 1958; Uitdenbogerd, 2005), Italian (Franchina & Vacca, 1986; Lucisano & Piemontese, 1988), Swedish (Jakobsen, 1971; Anderson, 1983), Chinese (Tham, 1987), Japanese (Yuka et al., 1988; Ozasa et al., 2007) etc., (Klare, 1974) also provides a brief survey of readability formulae for French, Dutch, Spanish, Hebrew, German, Hindi, Russian and Chinese.

In the past two decades, advances in the field of Natural Language Processing have lead to the development of more sophisticated readability assessment approaches using a range of language features and statistical learning models. Although most of this work still focused on English as it has more corpora to train the models and more tools to extract the features, readability research in other

[1] www.lexile.com
[2] http://drp.questarai.com/home/
[3] e.g.,http://readability.biz/Indices.html

languages is also emerging. The following section provides an overview of the methods and measures used in computational modeling of readability assessment.

2.2.3 Computational Modeling of Readability - English

Computational modeling of readability typically follows a supervised machine learning approach with the following steps:

1. **Corpus**: Select a gold standard corpus annotated with reading levels e.g., graded readers or a corpus of books across grade-levels. We assume here that the gold standard corpus takes linguistic complexity issues into account while preparing grade-appropriate content.

2. **Features**: Select the linguistic variables that are hypothesized to be indicative of text complexity.

3. **Modeling**: Select a learning algorithm and build a computational model using the corpus, features and the learning algorithm.

4. **Evaluation:** evaluate the model by a data-driven or user-based approach. In data driven approaches, the performance is usually evaluated in terms of prediction accuracy. User-based evaluations are sometimes performed by comparing recall and comprehension performance of the users on texts of varying difficulty.

Language Models The first computational models of readability were based on statistical language models. Si & Callan (2001) first proposed a model of detecting the reading difficulty of web pages using a linear combination of statistical language models and surface linguistic features. Collins-Thompson & Callan (2004, 2005) extended this approach with a larger web-corpus covering multiple US-English school grade levels and using a combination of language models followed by a feature selection step. They also performed cross-corpus evaluations and ported the approach to French and concluded that language models are better predictors of web document readability than traditional readability measures. This

work laid foundation for the REAP project [4], which provides appropriate reading materials while satisfying pre-defined lexical constraints, for language learners. REAP was also extended to work with Portuguese (dos Santos Marujo, 2009; Marujo et al., 2009).

Heilman et al. (2007) extended this approach and combined lexical, language model features with grammatical features and modeled readability using a corpus of English textbooks spanning 12 grade levels of the US education system. Their approach was focused on providing appropriate texts for language learners. Heilman et al. (2008) explored several statistical models, considering readability on nominal, ordinal and interval scales. They concluded that a Proportional Odds model, which assumes ordinal data, worked best for readability assessment. Language models were also used in (Quarteroni & Manandhar, 2006) to incorporate readability assessment into a question answering system.

Schwarm & Ostendorf (2005); Petersen (2007); Petersen & Ostendorf (2009) proposed a readability assessment approach for retrieving readable texts for English as Second Language (ESL) learners. This model relied on language models, Out of Vocabulary (OOV) word features and parse tree based syntactic features. They used a corpus of news articles written for children of various age groups by an educational news paper called Weekly Reader. They showed that a support vector machine model that combined all the features performed the best in terms of classification accuracy. The applicability of this approach for web search was also briefly explored in (Petersen & Ostendorf, 2006).

Other Word-difficulty based approaches: More recently, Kidwell et al. (2011) applied a statistical model of word acquisition for readability prediction. Flor et al. (2013) used a measure called lexical tightness, derived from word distributions in a large corpus for assessing text complexity. Shardlow (2013a,b) proposed approaches to assess the difficulty of words. Leroy & Kauchak (2013) studied the effect of word familiarity on reading difficulty. Weir & Anagnostou (2008) used collocation frequency as a measure of readability.

Since language models and word difficulty methods only look at the bag of words and n-grams, they may not always capture aspects of readability encoded in

[4] `http://reap.cs.cmu.edu`

the syntactic structure of the language. Further, issues like text coherence cannot be addressed with language models alone. Hence, several other approaches to readability explored a wide range of features.

Other Features: Kate et al. (2010) used a range of lexical and syntactic features for developing a readability approach to evaluate machine reading, which included several grammar rules that involve deeper linguistic analysis of sentences. Ma et al. (2012b) compared human and automatic feature extraction for readability assessment. Beinborn et al. (2012) describes the use of readability assessment for self-directed language learning. Medero & Ostendorf (2013) used atypical prosodic structure as a measure of reading difficulty. Green (2014) performed an eye-tracking evaluation of parser complexity metrics and showed that surprisal and entropy reduction metrics are good indicators of text readability for human comprehension. Looking at readability from a different perspective, Salama et al. (2013) showed the effect of improving the typesetting of a text based on lexical and syntactic features.

Along with various lexical, POS and parse tree based features, Feng et al. (2009); Huenerfauth et al. (2009); Feng et al. (2010); Feng (2010) used cognitively based indices for readability assessment to provide assistance for adults with intellectual disabilities. This group of features included - lexical chains, entity-density features, coreferential inference and entity grid features. Shallow discourse measures for estimating cohesion and coherence have been explored in the Coh-Metrix project (Crossley et al., 2007, 2008, 2011a; Graesser et al., 2014). Scarton & Aluísio (2010) ported the Coh-metrix approach to Brazilian Portuguese.

Genre Specificity: Most of the work on readability assessment did not consider the differences in texts of various genres. Sheehan et al. (2010, 2014) discuss the development of TextEvaluator^{TM} tool for measuring reading levels of a text and showed the effect of text genre (e.g., normal text versus spoken language) on readability prediction (Futagi et al., 2007; Sheehan et al., 2008). Sheehan et al. (2013) proposed a two-stage approach to readability consisting of a genre classification stage.

Topic Specificity: While most of the approaches proposed so far ignored topical or domain relevance, some of the approaches focused on these aspects. Qumsiyeh & Ng (2011) combined topic modeling and authorship information along with other language features for automatic readability assessment. Yan et al. (2006) proposed an approach that considers domain-specific conceptual readability as a feature. Jameel et al. (2012) proposed a model of domain specific readability based on conceptual difficulty and discourse cohesion. Zhao & Kan (2010) created a publicly accessible readability annotated corpus specifically for the mathematical domain while Kandula & Zeng-Treitler (2008) discusses the need for creating a gold standard for readability measurement of health texts.

Modeling Although almost all the proposed approaches have used either classification or regression approaches to learning, Pitler & Nenkova (2008); Tanaka-Ishii et al. (2010); Ma et al. (2012a) applied ranking and pair-wise comparison approaches to readability assessment. Brooke et al. (2012) proposed an unsupervised ranking based approach for constructing a readability lexicon.

2.2.4 Readability and Non-English languages

A majority of research into feature engineering and application of readability models focused on English. However, there is a growing body of literature in non-English languages that takes into account language specific characteristics like morphology. This section provides a short overview of some of the existing research.

Arabic: Forsyth (2014a,b) recently reported on readability classification for Modern Standard Arabic using traditional, POS based and discourse connective features. Shen et al. (2013) proposed a language independent readability assessment approach consisting of surface features and bag-of-words approaches and tested the approach for Arabic, Dari, English and Pashtun. In a follow-up study (Salesky & Shen, 2014), they showed that the addition of morphological and information theoretic features significantly improved the performance of their approach.

Bangla: Islam et al. (2012); Islam & Mehler (2013); Islam et al. (2014) explored the use of information theoretic features for English and Bangla readability assessment and showed that they perform on par with more linguistically oriented features. Sinha et al. (2012) developed new readability formulae using language specific surface features for Hindi and Bangla while Phani et al. (2014) performed an inter-rater agreement study for readability annotation which showed a moderate agreement between annotators.

Basque: Gonzalez-Dios et al. (2014b) used a range of linguistic features used in other languages along with those specific to Basque, to create two class readability models. Gonzalez-Dios et al. (2014a) proposed a tool that simplifies parenthetical structures in biographical articles in Wikipedia and also extended the approach to seven other languages. They showed that this approach improved the readability of Basque sentences.

Chinese: Lau (2006); Lau & King (2006) utilized the nature of the Chinese script to form several sub-character and character level features in addition to the common word and sentence level features for Chinese readability classification. They also explored bi-lingual website readability in this context. Chen et al. (2013) showed that lexical chains and term frequency features together were useful for the readability classification of Chinese textbooks. Hara et al. (2013) combined linguistic features with gaze information to model comma placement in Chinese text for improving the readability. Sung et al. (2014) discussed the advantage of using a multi-level modeling approach for Chinese readability assessment.

Dutch: van Oosten et al. (2010); van Oosten & Hoste (2011); Van Oosten et al. (2011); Clercq et al. (2014) studied Dutch readability assessment and investigated strategies for creating gold standard readability annotated data for both English and Dutch through crowd sourcing.

French: François (2009); Francois & Watrin (2011) used features modeling tense difficulty and multi-word expressions along with the other lexical and syntactic features for studying the readability of French as Foreign Language (FFL)

textbooks. While Francois & Fairon (2012) additionally considered semantic features, Todirascu et al. (2013) modeled coherence and cohesion for French readability assessment. François et al. (2014) developed a graded lexicon resource, which could be useful for readability assessment of French. Grabar et al. (2014) proposed an approach to diagnose the difficulty of level of medical terminology in French.

German: Vor der Brück & Hartrumpf (2007); Vor der Brück et al. (2008a,b) used comprehensive lexical, syntactic, morphological, semantic and discourse features for estimating the readability of German administrative texts. They showed that apart from the regular surface features and some of the parse tree features, aspects such as the quality of the semantic network, number of propositions per sentence and connections between network nodes representing objects in a text also played an important role in assessing the readability for German. Nietzio et al. (2012) analyzed the linguistic properties of text for preparing easy-to-read material in German. Hancke et al. (2012a) compiled a web-based two-class readability corpus for German and using a wide range of lexical, syntactic, morphological and language model features, achieved a classification accuracy of $\sim$90%. Lavalley & Berkling (2014) compared the sentence structure in various age-groups of children's writing and normal literature for German, with the aim of understanding children's reading and writing competence at word and sentence level.

Hindi: Sinha et al. (2012, 2014) discussed some approaches for readability assessment of Hindi using both traditional and other lexical and discourse features. They modeled readability using support vector machines, with a small set of 100 documents annotated with reading levels.

Italian: Dell'Orletta et al. (2011) worked with a corpus of Italian newspaper text at two different reading levels, with the eventual goal of performing text-simplification. They used a mixture of traditional, morpho-syntactic, lexical and syntactic features for building a two-class readability model for Italian. Among others, their feature set included verbal mood based features, which relied on the rich verbal morphology of Italian. This was later continued in Dell'Orletta

et al. (2012, 2013) where, using multiple two-class corpora spanning different genres, they showed the genre dependent nature of readability models. Tonelli et al. (2012) proposed a Coh-metrix inspired readability system for Italian that also provided a graphical representation of readability.

Japanese: Sato et al. (2008) followed a language modeling based approach for Japanese readability assessment using a corpus of textbooks sorted by their grade levels. Their method showed a strong correlation with readability in a cross corpus evaluation with web documents and also with shorter texts. Nishikawa et al. (2013) modeled readability prediction through a range of lexical, syntactic and discourse features. They used a training corpus annotated with reading time for Japanese texts. On the other hand, Sato (2014) studied the relationship between readability and word distribution in Japanese.

Polish: Broda et al. (2014) proposed a readability approach for Polish, primarily based on a corpus of web-documents and language models.

Portuguese: Readability assessment approaches for European and Brazilian Portuguese were proposed using various lexical, syntactic, discourse and language modeling features derived from English research (dos Santos Marujo, 2009; Aluisio et al., 2010; Scarton & Aluísio, 2010).

Swedish: Larsson (2006) used lexical and syntactic features to classify Swedish texts into three reading levels, with the aim of using it in a search engine to provide readable texts for students. More approaches have been proposed for Swedish in the recent past primarily relying on similar sets of features (Falkenjack et al., 2013; Heimann Mühlenbock, 2013; Falkenjack & Jonsson, 2014). Sjöholm (2012) discusses document and sentence level readability, considering readability as a ranking problem. Pilán et al. (2013, 2014) investigated both rule-based and machine-learning approaches for sentence level readability assessment of Swedish texts, for choosing appropriate sentences for language learning exercises. Pilán et al. (2015) developed an approach to classify Swedish texts based on the CEFR scale and achieved a classification accuracy of 81% for a 5 level classification.

Thai: Daowadung & Chen (2012) studied the effect of stop words in readability assessment of Thai text.

While most of the research in readability assessment for non-English languages is still in the research phase, several commercial and non-commercial applications already exist for English (e.g., SourceRater, Lexile, DRP, REAP etc.,). Some of these systems rely on a range of features to estimate the readability of a text, whereas others use a limited number of surface features. For example, SourceRater[5] (Sheehan et al., 2014) relies on a range of linguistic properties modeling compexity, concreteness, familiarity, at word level, syntactic complexity of sentences, text cohesion and stylistic aspects like narrativity, argumentation etc., Compared to this, Lexile[6] uses only two features - sentence length and word frequency.

Nelson et al. (2012) compared several existing academic and commercial English readability assessment systems using multiple standardized test datasets including those from the Common Core Standards exemplar texts, and showed that the systems based on multiple linguistic properties performed better than those that used a limited set of features. François & Miltsakaki (2012) also concluded that using machine learning models consisting of more sophisticated linguistic features improved predictions for readability assessment of French as Foreign Language. These results also motivate our approach to consider multiple aspects of readability instead of restricting to a small set of features.

2.2.5 Applications of Readability Assessment

The primary use of readability assessment so far has been to retrieve appropriate content for learners. Hence, it has been commonly used for information retrieval. More recently, readability assessment is also being used in text simplification - for picking sentences to simplify and/or to evaluate the simplified texts in terms of their readability.

[5] http://naeptba.ets.org/SourceRater3/
[6] https://lexile.com/

Readability Assessment for Information Retrieval

Readability assessment for information retrieval aims to re-rank search results and personalize them. The target users are primarily language learners and sometimes, a more general audience. Approaches to develop readability enabled search engines for language learners were explored by Bennöhr (2005); Newbold et al. (2010) using custom readability formulae. Kane et al. (2006); Ott (2009); Ott & Meurers (2010); Tan et al. (2012) used various traditional readability formulae for ranking search results . READ-X project (Miltsakaki & Troutt, 2007, 2008; Miltsakaki, 2009) too modeled readability in a similar setting using vocabulary features and considering topical relevance.

Apart from (re-)ranking of search results for target groups, approaches detecting reading level of users from query logs (Liu, Croft, Oh & Hart, 2004) and those that used linguistic features to predict query difficulty (Mothe & Tanguy, 2005) were also proposed. Kim et al. (2012) combined language modeling based reading level prediction with topic modeling for personalized search based on a user's reading level and interests.

Pera & Ng (2012) applied readability assessment for recommending books for K12 students. Bendersky et al. (2011) considered certain surface text features of readability for doing a quality based re-ranking of web documents in search results. Nakatani et al. (2010) followed a language modeling approach to rank search results considering user comprehension into account. Kanungo & Orr (2009) used search result snippet based features to predict the readability of short web-summaries.

Readability Assessment for Text Simplification

Automatic Text Simplification (which will be discussed in the next section) is one of the primary areas where Readability Assessment has been recently used. It can be useful in various stages of simplification ranging from identifying simplification targets to the evaluation of simplification outcomes. In the past decade, the use of readability assessment for simplification has mostly been restricted to using traditional readability formulae for evaluating the simplified text or as a constraint for generating simplified text (Jonnalagadda et al., 2009; Zhu et al., 2010;

Wubben et al., 2012; Klerke & Søgaard, 2013; Stymne et al., 2013). Some recent work briefly addresses issues such as classifying sentences by their reading level (Napoles & Dredze, 2010; Karpov et al., 2014; Pilán et al., 2014; Dell'Orletta et al., 2014) and identifying sentential transformations needed for text simplification using text complexity features (Medero & Ostendorf, 2011). Kauchak et al. (2014) developed a sentence level readability approach using domain specific vocabulary based features along with other features, to identify difficult to read parts in clinical texts. Some simplification approaches for non-English languages (Aluisio et al., 2010; Stajner & Saggion, 2013; Stajner et al., 2014) also touch on the use of readability assessment in identifying targets and necessary transformations for text simplification.

Other Applications

Apart from these primary application areas, readability assessment has found use in diverse domains like: understanding the readability of survey questionnaires (Lenzner, 2013), understanding the comprehensibility of patient information resources (Ellimoottil et al., 2012; Pringle et al., 2013; Hansberry et al., 2014) and segmentation of patient claims (Ferraro et al., 2014). Sitbon & Bellot (2008) proposed a readability measure for dyslexics considering grapho-phonemic consistency as a feature. Rello et al. (2012, 2014) used a notion of readability that encompasses issues like keyword highlighting and visual display for studying text comprehension in dyslexics. Readability issues have also been considered in the context of evaluating multi-document summarization (Pitler et al., 2010; Wan et al., 2010) and machine translation (Chae & Nenkova, 2009). Louis & Nenkova (2013) considered readability aspects for estimating writing quality of news articles. Kim et al. (2014) describe a device dependent readability assessment approach for a news article recommendation system. Readability assessment both from a methodical and application perspective is a very active area of research now, with several new approaches, new languages and new evaluation methods being proposed every year.

2.2.6 This thesis and readability assessment

A wide range of features has been explored in NLP approaches to readability assessment so far. However, there has been little in-depth study about the validity of the created readability models beyond the training-test sets, in terms of their cross-corpus, cross domain portability and their effect on target readers. User based evaluations, comparing the textual readability measures with reader's cognitive processing and performance outcomes were also not explored much in the past research on this topic. In this background, we view readability analysis both as a stand-alone approach as well as a useful step in the process of performing text simplification. Thus, we explore both document level (Chap 3) and sentence level readability (Chap 5) in detail. We also connect the linguistic notions of readability with cognitive psychology by reporting about a user-based eye-tracking study with texts of varying reading levels (Chapters 4).

Additionally, we used the features developed for our automatic readability assessment approach in two educational applications, which will be discussed in detail in the second part of the thesis.

1. We built classification and regression models to assess the L2 proficiency, thesis clarity and coherence in English as second language learner essays, using our readability features. Our results showed that our features are also useful in distinguishing learner texts. To our knowledge, this idea of connecting readability and proficiency has not been explored in the research before except for Hancke et al. (2012a); Hancke (2013) who used overlapping features for German readability assessment and L2 proficiency classification. This is discussed in detail in Chapter 7.

2. We used German text complexity features described in Hancke (2013) to compare Geography textbooks used in German schools. We compared the differences between the features between grades, school types and publishers. This is discussed in detail in Chapter 8.

2.3 Automatic Text Simplification (ATS)

Texts written or produced in human language can often comprise of complex syntactic constructions and difficult words, making it difficult to read. Often, other factors like a lack of coherence in the discourse and a complex nature of the topic in discussion might also contribute to this overall reading difficulty of a text. This scenario is not limited to human users and can occur during machine processing and text generation as well. Text Simplification can be defined as the process of understanding these aspects of text difficulty and devising ways to overcome them by modifying the text structure so that it will be simpler to understand for the target users.

Early research in to text simplification began in the 90s, when Chandrasekar & Srinivas (1996) described an approach for syntactic simplification to improve the performance of natural language parsers. Later research primarily focused on simplifying texts for human users with specific intellectual disabilities. Some of the approaches to text simplification were focused towards specific target user groups like aphasics, dyslexics (Devlin & Unthank, 2006), deaf people (Takahashi et al., 2001; Inui et al., 2003; Chung et al., 2013), adults with low literacy (Aluísio et al., 2008), language learners (Williams & Reiter, 2008; Petersen & Ostendorf, 2007) etc., While the application of text simplification is primarily oriented towards a target human user-group, there were approaches focusing on simplification to improve or assess the performance of machines as well. Some of the typical non-human based applications are: improving language parsing (Chandrasekar & Srinivas, 1996), text summarization (Lal & Rüger, 2002; Damay et al., 2006), question-Answering systems (Heilman & Smith, 2010), information retrieval (Klebanov et al., 2004), information extraction (Miwa et al., 2010; Evans, 2011), spoken language understanding (Tur et al., 2011), subtitle generation Daelemans et al. (2004) etc., Simplification approaches specific to domains like crisis management Temnikova (2012), health and medical text (Kandula et al., 2010; Damay et al., 2006; Abrahamsson et al., 2014), biomedical information processing (Jonnalagadda et al., 2009; Jonnalagadda & Gonzalez, 2010) were also proposed. Recently, the availability of a big parallel corpus in the form of Wikipedia-Simple Wikipedia played a major role in the development of var-

ious data-driven approaches for English text simplification. Research in other languages like Spanish, Italian, Portuguese, Basque and Danish is also starting to emerge.

2.3.1 What is simple text? – corpus and user studies

The first step towards ATS is to have an understanding about the characteristics of a simple text. This will also contribute towards the development of criterial features for performing text simplification. Corpus studies of Simplified and Unsimplified texts can provide us insights about what makes simple text simple for a given application context. In data-driven approaches, corpus analysis is primarily used to understand what linguistic features could be useful for the task. Petersen & Ostendorf (2007) studied a corpus of 104 aligned original and abridged versions of articles. Their analysis showed that simplified texts contain fewer adverbs and coordinating conjunctions. Their experiments with automatically classifying between split and un-split sentences, and decisions about which sentences to keep and drop - showed that syntactic features are useful to make these decisions. Allen (2009a,b) studied a corpus consisting of unsimplified and simplified versions of next texts in terms of relative clause distribution and incidence of various parts of speech across versions. Crossley et al. (2012) studied a corpus of 300 news texts simplified into three levels of simplification and examined the linguistic differences between them. They showed that the texts at a lower reading level are less lexically and syntactically sophisticated, and contain more cohesive features than higher level texts.

Amancio & Specia (2014) manually analyzed some of the simplifications performed by Simple Wikipedia authors and concluded that the most common transformation operations performed were paraphrasing and drop of information. They used this manual analysis to perform an automatic classification of sentences based on the transformations needed for text simplification, without much success. Medero & Ostendorf (2011) also considered the task of identifying targets for syntactic simplification as a classification task with three categories: split/no-split, omissions, expansions. They concluded that the prediction of expansions is difficult compared to prediction of splits. Such corpus studies were also re-

ported in the case of other languages, as a primary step towards the development of automatic text simplification systems. Bott & Saggion (2011a); Drndarevic & Saggion (2012a); Stajner et al. (2013) performed multiple corpus studies with Spanish texts in simplified and unsimplified versions and Aranzabe et al. (2012a) report a corpus-analysis for a text simplification system in Basque language.

The simplifications performed on texts can be broadly classified into three categories, based on the linguistic properties they address.

1. lexical simplification: identifying and replacing difficult words with simpler alternatives.

2. syntactic simplification: replacing difficult syntactic constructs with simpler versions.

3. discourse simplification: simplification in terms of the meaning.

While some Automatic Text Simplification (ATS) approaches explicitly deal with one of these categories, several approaches use an integrated approach which handles more than one of the issues from these categories. ATS approaches typically fall into two groups: rule-based and learning based. Rule-based approaches use manually or computationally derived rules for generating simplified texts and learning based approaches primarily rely on the availability of large amounts of parallel corpora instead of explicit rules. While most of the early approaches were rule based, more recent work consists of both rule-based and data-driven approaches. Several recent approaches also considered ATS as monolingual machine translation.

2.3.2 Rule based Text Simplification

Automatic text simplification was first explored in the context of improving parsing efficiency by devising rules to split sentences into simple sub-sentences. Chandrasekar et al. (1996) used finite state and tree adjoining grammars (TAG) for this purpose and showed that the TAG model did a better job at identifying the articulation points to split the sentences. They later also explored an automatic approach for simplification rule-induction (Chandrasekar & Srinivas, 1996). While this paper formalized the term automatic text simplification for a specific application

context and briefly discussed the issues that need to be addressed, it was primarily restricted to only splitting sentences and did not have a strong evaluation setup.

PSET system [7] (Practical Simplification of English Text) was developed in the late 90s for creating easy to understand newspaper texts for aphasic readers[8]. They considered lexical and syntactic simplification separately. Since aphasics have problems understanding passive sentences and long sentences with embeddings, PSET proposed a rule-based system that handled these aspects. Lexical simplification was performed on "less common words", by replacing them with alternatives obtained from Wordnet[9] and a psycholinguistic database. In the case of splitting, they resolved and replaced pronouns with the corresponding referents (Carroll et al., 1998, 1999). The syntactic simplification component of PSET is SYSTAR (SYntactic Simplification of Text for Aphasic Readers), which focused on the issue of anaphora resolution during syntactic simplification while also maintaining cohesion (Canning & Tait, 1999; Canning et al., 2000). Evaluation of this approach was performed with human users in terms of reading time. HAPPI (Helping Aphasic People Process Information) by Devlin & Unthank (2006) was an extended version of PSET.

The KURA project Inui et al. (2003); Takahashi et al. (2001)] built a text simplification system for deaf people. They followed a stepwise approach to simplification by first identifying difficult areas in a text and replacing them with simpler paraphrases. It performed a syntactic and lexical paraphrasing of the text and restricted the vocabulary to a top 2000 basic word set. They used handcrafted paraphrase rules for a broad range of transformations. Williams & Reiter (2008) described a personalized text generation and simplification system called SKILLSUM, focusing on discourse level simplification for users with low literacy.

Siddharthan (2002a,b) described a text simplification system built in two phases - syntactic simplification module and lexical simplification module. Syntactic simplification module has three stages - analysis (provided the structural representation of a sentence), transformation (applied a sequence of rules to transform this sentence and flatten the resulting structures to plain text) and regeneration

[7] http://www.informatics.sussex.ac.uk/research/groups/nlp/projects/pset.php

[8] http://en.wikipedia.org/wiki/Aphasia

[9] http://wordnet.princeton.edu/

(regenerated simplified versions). Lexical Simplification phase primarily focused on replacing difficult words with easier ones. Identification of relevant constructs for syntactic simplification was not done by full parsing but by pattern matching techniques after POS tagging and chunking. Each stage is evaluated individually, but the system was not evaluated as a whole. Siddharthan & Copestake (2002) explored the problem of generating correct referring expressions while splitting a sentence during the process of simplification. Although the algorithm was general purpose in nature, they evaluated it on text simplification. Siddharthan (2003, 2004, 2006) primarily focused on preserving the discourse structure and the document cohesion while performing sentence transformations and generating split sentences. Evaluation of the overall approach was done with human users, considering the factors of grammaticality, semantic correctness and readability (which was estimated using Flesch reading ease formula).

Although recent research in ATS has been primarily data-driven, rule based approaches are being proposed for specific application scenarios and target users. Jonnalagadda et al. (2009); Jonnalagadda & Gonzalez (2009) proposed rules for syntactic simplification based on link grammar for biomedical domain and showed that text simplification aids better extraction of protein-protein sequences. Junior et al. (2011) proposed a rule-based approach for simplifying Portuguese text by handling passive voice and subordination. Barlacchi & Tonelli (2013) described a sentence-simplification tool for improving the comprehension of factual events in Italian stories for children, by developing rules for syntactic simplification of factual events and combining them with anaphora resolution. Aranzabe et al. (2012a,b) proposed a rule-based system based on morphological properties and dependency parses to perform sentence simplification for Basque, which handled several kinds of syntactic constructs. Gasperin et al. (2009); Candido et al. (2009) described a rule-based text simplification approach for Brazilian Portuguese. Evans et al. (2014) developed a rule-based approach specifically for people with autism. Seretan (2012) proposed a French text simplification system through manual and semi-automatic creation of rules using newspaper text. Lu & Parameswaran (2009) combined ontology mapping with parse-tree based rules to perform sentence simplification.

2.3.3 Data-driven approaches

Data-driven approaches to text-simplification generally consist of some or all of the following modules:

1. complex-simple parallel corpus creation and sentence-level alignment
2. application of computational methods for learn from the data and generate simplified texts.
3. evaluation of the approach.

Corpus Creation and Alignment

Automatic approaches to text simplification require parallel, large-scale sentence aligned corpora in simplified and unsimplified versions. These corpora are created by treating the sentence alignment as a monolingual alignment problem, using cosine similarity and TF-IDF measures (e.g., Barzilay & Elhadad (2003); Nelken & Shieber (2006); Zhu et al. (2010). In English, one primary resource for this has been the Wikipedia-Simple Wikipedia articles. Zhu et al. (2010); Coster & Kauchak (2011b,a) - created publicly accessible datasets of parallel sentence level simplification corpus based on these articles, which were later used by several other researchers, to develop automatic text simplification approaches for English. Klerke & Søgaard (2012) and Klaper et al. (2013) created sentence aligned simplification corpora for Danish and German respectively. Bott & Saggion (2011b) created an unsupervised HMM based sequential model to create a Spanish text simplification corpus. Caseli et al. (2009) described an assistive system for creating a simplification system for Brazilian Portuguese text.

While the above mentioned corpus creation approaches focused on simplified texts in general, De Belder & Moens (2012) and SemEval-2012 Lexical Simplification Task [10] prepared two separate datasets specifically for lexical simplification, using the same source dataset from the English Lexical Substitution Task from SemEval-2007 (Mccarthy et al. (2007)). Both the datasets are publicly available

[10] http://www.cs.york.ac.uk/semeval-2012/task1

and contain a list of sentences, with highlighted words and their list of substitutes, ranked by ”simplicity”. The ranking was estimated with human annotator judgments.

Learning lexical simplification

Lexical simplification is primarily concerned with replacing difficult words or phrases with simpler alternatives. An automatic lexical simplification approach generally consists of the following stages:

- analyzing the sentence and picking up words or phrases that might be difficult to the target reader.
- preparing a list of possible substitutes using various methods (e.g., finding synonyms)
- ranking them in the order of their simplicity as well as suitability to the sentential context.

Yatskar et al. (2010) and Biran et al. (2011) proposed approaches based on Wikipedia corpus, which involved all the above stages. Yatskar et al. (2010) described the process of unsupervised extraction of lexical simplifications from Wikipedia, using the edit histories. Biran et al. (2011) too followed an unsupervised context-aware learning based approach, which did not require an aligned corpus. Their method also ensured meaning preservation and grammatical correctness. It involved two stages: learning simplification rules and performing sentence simplification considering both word-sentence similarity and context similarity. Evaluation was based on human judgments. In this approach, they considered only with word-word simplifications and not phrases. More recently, Horn et al. (2014) described a ranking based approach to lexical simplification based on Wikipedia corpus, which was shown to overcome the generalizability issues that existed in other approaches.

While these approaches considered all the steps of lexical simplification, some of the approaches focused only on the third step (ranking the alternatives). The English Lexical Simplification Task at SemEval-2012, introduced with the aim of

promoting research in automatic, context-aware lexical simplification and providing a common evaluation platform for such systems, is one example. The task involved producing a ranked list of alternate words for a target word in a sentence, with "simplicity" of the word as the ranking criterion. The performance of the participating systems was compared with the judgments on word "simplicity" that were given by a group of human annotators. While it is possible to have various substitutes for different target audience, this task was oriented towards fluent non-native speakers of English. The systems that participated in this task were evaluated against inter-annotator agreement as well as against three baselines. The three baselines include: contextual aptness, random word choice and most frequent word[11].

Several systems that participated in the competition reported the use of frequency-based approaches. Ligozat et al. (2012) used language model based probabilities from Microsoft Web corpus and Simple Wikipedia, to create a model of "simple" ranking. Johannsen et al. (2012) performed co-training with word and character n-gram features and syntactic complexity based features. While Amoia & Romanelli (2012) proposed a decompositional semantics based approach, Sinha (2012) used a combination of unigram frequencies from various sources. The best performing model is described inJauhar & Specia (2012), who used a linear weighted ranking function with three features - context sensitive n-gram frequency model, bag of words model and psycholinguistic features. Apart from the systems in this task, Shardlow (2012) too used a model consisting of Bayesian probability estimates of the frequency counts from Simple Wikipedia and Wikipedia to rank the list of possible word substitutes based on their simplicity. Thomas & Anderson (2012) explored six different approaches to lexical simplification, using Wikipedia and Wordnet and concluded that the best-performing algorithm relies on word sense disambiguation with sentence level contextual information. On a related note, Walker et al. (2011) examined the issue of automatically suggesting good lexical substitutions and compared this with human preferences.

Apart from the above mentioned approaches, lexical simplification algorithms

[11]Datasets from SemEval 2012 task: `http://www.cs.york.ac.uk/semeval-2012/task1/data/uploads/test-data.zip` and `http://www.cs.york.ac.uk/semeval-2012/task1/data/uploads/datasets/trial-dataset.zip`

were also proposed for text summarization (Blake et al., 2007) and bio-medical information extraction (Jonnalagadda et al., 2009).

While all this work in lexical simplification has been primarily English focused, some amount of recent research in other languages too exists. Keskisärkkä (2012); Keskisärkkä & Jönsson (2012) experimented with various methods of choosing alternative synonyms for performing lexical simplification in Swedish, primarily driven by word frequency and word length. Drndarevic & Saggion (2012b,a); Drndarevic et al. (2012) performed an empirical study of Spanish lexical simplification using a corpus of normal and simplified Spanish texts. Bott et al. (2012) continued in this direction and built a system that considers three aspects to lexical simplification - word frequency, word length and word vector model. Word frequency and length are used for the aspect of simplification, whereas word vectors are used for getting the context and performing sense disambiguation.

Although stand-alone lexical simplification approaches continue to be proposed, there is also an active body of research focusing more generally on syntactic simplification, which also includes lexical simplification in the process.

Learning syntactic simplification

Woodsend & Lapata (2011a,b) proposed a text simplification approach by following a data-driven approach to induce a quasi-synchronous grammar from Wikipedia and using an integer linear programming model to select appropriate simplifications. They showed that the generated simplified text had a reduced reading difficulty while preserving the grammaticality. Brouwers et al. (2014) combined a system of manually created rules with integer linear programming to generate simplified sentences in French. Feblowitz & Kauchak (2013) proposed a syntax based text simplification approach based on probabilistic synchronous tree substitution grammar and showed that the approach performed better than phrase based approaches for this task.

Bach et al. (2011) described a margin based discrimination learning approach using phrase-structure and dependency structures of sentences, to perform a sentence-level factual simplification. This approach primarily focused on splitting sentences into short sentences, primarily with a view to improve the performance of

NLP systems by providing them simpler text to process. Klerke (2012); Klerke & Søgaard (2013) described unsupervised approaches to perform text simplification in Danish that modeled simplification as a process of inserting and deleting content. Simple alternatives were sampled from possible sentence generations using heuristics and readability measures.

Siddharthan (2011) described an approach to text simplification generation by applying transformation rules on typed-dependency trees given by the Stanford parser. Siddharthan & Mandya (2014); Angrosh & Siddharthan (2014) described another text simplification approach based on synchronous dependency grammar, which allows for both manual and automatic specification of rules, enabling hybrid text simplification and showed that this approach performed better than other simplification approaches based on quasi-synchronous tree substitution grammars.

Syntactic Simplification through Machine Translation

Specia (2010) considered automatic text simplification as a machine translation problem for the first time. Using Moses toolkit[12] for training Portuguese simplification models, they showed that SMT can capture certain kind of phenomenon like lexical simplification and simple paraphrases very well. Coster & Kauchak (2011b,a) explored Phrase Based Machine Translation (PBMT) with an additional step to handle deletions using Wikipedia-Simple Wikipedia corpus as the basis. Their experiments showed a slight improvement in BLEU scores for the system with deletion handling compared to a plain Moses based SMT system. Wubben et al. (2012) too followed a PBMT approach followed by a re-ranking of generated translations based on their dis-similarity with the original version. Kauchak (2013) studied the effect of using various language models on the generation of simplified text.

Zhu et al. (2010) proposed a tree based translation model for sentence simplification, handling issues like dropping, reordering and word/phrase substitution within the model. This approach was evaluated using Machine Translation metrics like BLEU and NIST as well as traditional readability formulae, on a test-corpus

[12] `http://www.statmt.org/moses/`

consisting of 100 sentence pairs.

2.3.4 Evaluation of Text Simplification

Though different approaches have been proposed for text simplification, there exists no single evaluation framework to compare all of them on a single measure. While the evaluation of a text simplification approach is strongly depend on its target users or systems, it would be good to have a standardized evaluation method for a given purpose. Common ways of evaluating a text simplification approach used in previous research are:

- comparing the readability levels of simplified and unsimplified text, using traditional readability formulae (e.g., Jonnalagadda et al., 2009; Woodsend & Lapata, 2011a).
- using automatic measures like BLEU and ROUGE, to compare gold-standard and generated simplified texts (e.g., Coster & Kauchak (2011a); Bach et al. (2011).
- conducting user studies to read aloud and count the number of reading errors (Williams & Reiter, 2008; Devlin & Unthank, 2006)
- collecting acceptability and grammaticality judgments from users (Woodsend & Lapata, 2011a; Siddharthan, 2011)

Stajner et al. (2014) discussed the automatic evaluation of text simplification by presenting a comparison of 6 machine translation evaluation metrics and their correlations with human judgments of grammaticality and meaning preservation. Stajner & Saggion (2013) discussed the utility of more sophisticated readability assessment methods for the evaluation of Spanish text simplification systems.

Siddharthan & Katsos (2010) proposed a process of automatically assessing which reformulation of sentences are acceptable to which groups of readers using surface level features that reflect propositional complexity. Although this is not exactly an evaluation methodology for text simplification per se, it can be seen as an approach to model user-preferences about the appropriateness of a generated text. Siddharthan & Katsos (2012) extended this to include sentence quality,

magnitude estimation of acceptability judgments and sentence recall - to propose offline measures of readability for computer generated text.

More detailed survey of automatic text simplification approaches can be seen in Feng (2008); Shardlow (2014); Siddharthan (2014).

2.3.5 Text Simplification and this Thesis

In this thesis, we follow a machine translation approach to automatic text simplification. We studied the effect of training and language modeling corpora on the overall efficiency of the approach in terms of BLEU scores (Chapter 6). Additionally, as mentioned earlier, we study the usefulness of sentence level readability models to evaluate the degree of simplification performed. While studying sentence level readability, we created a new three-level sentence simplification corpus (Chapter 5) which can be used for training and testing automatic text simplification systems. We also used this corpus to perform cross-corpus evaluation of automatic text simplification. To our knowledge the effect of training corpora on the overall efficiency of a simplification system has not been explored. Further, machine translation approaches so far only relied on a single corpus and the corpus we created will enable cross-corpus evaluations in future.

Chapter 3

Automatic Readability Assessment: Approach and Evaluation

Abstract

In this chapter, we describe our Automatic Readabilty Assessment (ARA) approach and establish its validity through multiple evaluations on real world data sets. We investigate both the generalizability of the model and the features themselves. We also explore the effect of genre and topic on the performance of readability models. We show that the predictions of our model achieve a correlation of 0.9 with the actual grade level on 10-fold cross validation and the second best reported performance on the Common Core Standards dataset (rank correlation of 0.69), compared to other existing readability assessment approaches. The features we describe in this chapter also generalize well across various readability annotated datasets. The approach also works on spoken language texts (movie subtitles), achieving a classification accuracy of 96%.

This chapter is based on Vajjala & Meurers (2012, 2014b,c)

3.1 Overview

Machine learning approaches to readability assessment typically employ lexical and syntactic features that can be extracted using natural language processing tools. Some of the previous research also modeled the discourse and cognitive aspects of the task, based on the research in cognitive science on working memory (cf. discussion in Chapter 2). On the other hand, approaches to assess text complexity have also been studied in other fields such as Second Language Acquisition (SLA) and Psycholinguistics. However, there is little work on re-using insights from these areas and applying them in ARA. In this chapter, we developed features derived from SLA and psycholinguistics research, apart from other lexical and syntactic features, for readability assessment. To train and test our models, we relied on several graded corpora intended for children and language learners[2]. We modeled readability as a supervised machine learning problem, both as classification and regression. The choice between classification and regression depended on the nature of the dataset (categorical versus continuous). The primary research questions studied in this chapter are:

1. Can we build accurate readability models using linguistically motivated features?
2. Do the models thus built generalize to unseen/new texts?
3. Can the feature set result in good performance with other datasets?
4. Is there a genre bias in the model and can it be overcome?
5. Is it possible to train topic specific models?

The rest of this chapter is organized as follows. Sections 3.2, 3.3 and 3.4 detail the basic experimental setup for our modeling in terms of the corpora, features, algorithms, and the evaluation methods. Section 3.5 describes our primary readability model. Section 3.6 and Section 3.7 discuss the generalizability of the model and the feature set. Sections 3.8 and 3.9 explore genre effects in readability modeling. Section 3.10 briefly studies topic specific modeling. Section 3.11 summarizes the experiments in this chapter with pointers to future work.

[2]All the corpora are created by external sources and can be obtained for research use

3.2 Corpora

We employed four readability-annotated corpora that are annotated that have language learners as the target audience. These corpora were prepared by diverse external sources with different rubrics. While the first corpus *WeeBit* was created from web-based informative articles written for children belonging to specific grades, the *CommonCore* corpus consists of texts belonging to different genres, indicated as benchmarks for grade-wise complexity by experts. The *TASA* corpus was prepared by using a proprietary readability formula, and, finally, the *BBC-Subtitles* corpus is a collection of television subtitles classified into three age groups. Thus, the corpora we used cover different topics and genres. All of them were intended for first language English speakers.

3.2.1 WeeBit

The WeeBit corpus we compiled consists of texts at five reading levels. There are 615 documents per level intended for first language English learners of the age group 7–16 years. It is a compilation consisting of two sub-corpora: WeeklyReader and BBC BiteSize.

WeeklyReader (WR) is an educational newspaper[3], with articles targeted at four grade levels (Level 2, Level 3, Level 4 and Senior), corresponding to children between ages: 7-8, 8-9, 9-10 and 9-12 years respectively. The articles cover a wide range of non-fiction topics from science to current affairs, written according to the grade level of the readers. The exact criterion of graded writing, is not published by the magazine, though. We obtained permission to use the graded articles in the magazines and downloaded WR2, WR3, WR4 and WRSenior texts available in the archives in November 2011. In addition to the main articles, the online WeeklyReader magazine issues included teacher guides, student quizzes, images and brainteaser games. While preparing the corpus for classification, we removed articles that had only pictures, games and quizzes and no text content. Apart from that, we also removed teaching instruction articles from each issue, since they are

[3] `http://www.weeklyreader.com`

not relevant to build a classifier.

Though we used the same WeeklyReader text base as the previous works (e.g., Petersen, 2007; Feng, 2010), the corpus is not identical to what they used, since we downloaded our version more recently and thus the archive contained more articles per level. It is also not entirely clear what pre-processing was performed in the previous work to remove pages from the corpus that were not the actual reading material.

BBC-Bitesize consists of a collection of articles classified into four grade levels[4] (KS1, KS2, KS3 and GCSE), corresponding to children between ages 5–7, 7–11, 11–14 and 14–16 years. The Bitesize corpus is freely available on the web, and we used a crawled version from 2009 which was used in Ott & Meurers (2010). Most of the articles at KS1 grade consisted of images and flash files and other audio-visual material, with little text. Hence, we did not consider KS1, in combining the two corpora. On the other grades, we removed pages that contained only images, audio or video files.

To cover a broad range of non-overlapping age groups, we used Level2, Level3 and Level4 from WeeklyReader and KS3 and GCSE from Bitesize data respectively and built a combined corpus, which covers learners aging 7 to 16 years. It must be noted that while KS2 in Bitesize covers the age group of 7-11 years, levels 2, 3, 4 in WeeklyReader together cover ages 7-10 years. Similarly, the WRSenior Level has an overlap with WRLevel4 and Bitesize-KS3. Hence, we excluded KS2 and WR-Senior from the combined corpus. We will refer to the combined 5-level corpus we created as *WeeBit*.

Since we later model readability assessment as regression, which assumes that the data falls on an interval scale with evenly spaced reading levels, we used numeric values from 1–5 as reading levels instead of the original class names in the *WeeBit* corpus. In this thesis, we used 616 articles from each category as our corpus, instead of the entire corpus described in Vajjala & Meurers (2012), to ensure equal representation of all reading levels[5]. The texts on an average had 23.4 sentences at the lowest level and 27.8 sentences at the highest level. The grade levels

[4] `http://www.bbc.co.uk/bitesize`

[5] The exact file ids can be shared for replication purposes

in the corpus, the age groups they represent, and the numeric reading level are shown in Table 3.1.

WeeBit class	Age (years)	Reading level
Level 2	7–8	1
Level 3	8–9	2
Level 4	9–10	3
KS3	11–14	4
GCSE	14–16	5

Table 3.1: *WeeBit* Corpus

In Vajjala & Meurers (2013), we found that some of the initial readability models we built using the WeeBit corpus generalized well to various web corpora and were useful in identifying diverse reading levels for search engine result pages. Hence, this will be the primary training corpus for most of our document level readability models explained below, unless mentioned otherwise.

3.2.2 Common Core Standards Corpus

The Common Core Standards corpus consists of 168 English texts belonging to four genres. This corpus was created from the exemplar texts given in the Appendix-B of the Common Core Standards description document, excluding the items categorized as poetry[6]. These texts were classified into grade bands of the US education system, by experts in education. This corpus was introduced as an evaluation corpus for readability models in the recent past (e.g., Sheehan et al., 2010; Nelson et al., 2012; Landauer & Way, 2012; Flor et al., 2013; Flor & Klebanov, 2014). We used this corpus to test our readability model and to evaluate its performance across genres, to compare our model with other existing readability systems and to train a model to verify the generalizability of the feature set used.

[6] `http://www.corestandards.org/assets/Appendix_B.pdf`

3.2.3 TASA corpus

The TASA corpus consists of about 37,000 texts annotated with their reading level in terms of Degrees of Reading Power (DRP)[7] scale assigned by Touchstone Applied Science Associates Inc. (TASA). The score typically ranges from 30-80. The corpus was created in 1995 from 6,333 textbooks, fiction and non-fiction works used in schools and colleges throughout the United States, with the aim of estimating the frequency of words at different grade levels. It consists of texts with a mean length of 250-300 words covering nine content areas: business, health, home economics, industrial arts, language arts, miscellaneous, science, social studies, and uncategorized. The corpus is widely used in Latent Semantic Analysis[8] and was used as an evaluation corpus in some of the Coh-Metrix[9] readability studies (cf. Graesser et al., 2012). We use this corpus for evaluation of our model and to test the adaptability of our features to different topic categories.

3.2.4 BBC Subtitles corpus

The BBC started subtitling all the scheduled programs on all of its main channels in 2008, implementing UK regulations designed to help the hearing impaired. Van Heuven et al. (2014) constructed a corpus of subtitles from the programs run by nine TV channels of the BBC, collected over a period of three years, from January 2010 to December 2012. They used this corpus to compile an English word frequency database, SUBTLEX-UK[10], as a part of the British Lexicon Project (Keuleers et al., 2012). The subtitles of four channels (CBeebies, CBBC, BBC News and BBC Parliament) were annotated with the channel names.

While CBeebies targets children aged less than 6 years, CBBC telecasts programs for children 6–12 years old. The other two channels (News, Parliament) are not assigned to a specific age group, but it seems safe to assume that they target a broader, adult audience. In sum, we used the BBC subtitle corpus with a three-way categorization: CBeebies, CBBC, Adults. The full corpus consists of

[7] http://drp.questarai.com/home/
[8] http://lsa.colorado.edu/spaces.html
[9] http://cohmetrix.memphis.edu/
[10] http://crr.ugent.be/archives/1423

4846, 4840 and 3776 documents respectively for each category. For our experiments, we use a balanced subset of this corpus with 3776 instances for each class, to avoid a bias towards majority classes while classification. Table 3.2 shows the basic statistics for the corpus we used[11].

Program category	Age group	avg. tokens	avg. sentence length
CBEEBIES	< 6 years	1144	4.9
CBBC	6–12 years	2710	6.7
Adults (News + Parliament)	> 12 years	4182	12.9

Table 3.2: BBC Subtitles Corpus

3.3 Features

We explored a wide range of features for developing our readability model. They can be broadly classified into three categories: lexical richness and POS features, syntactic complexity features and word characteristics features. The word characteristics features consist of morphological features, psycholinguistic features and semantic features.

3.3.1 Lexical Richness and POS Features

We adapted some of the measures of lexical richness from Second Language Acquisition research for readability assessment. These measures consisted of two variations of type-token ratio and the measures of lexical variation (noun, verb, adjective, adverb and modifier variation). In addition, this feature set also includes the density of different Parts Of Speech (POS) in the texts that study the relation between POS tag density and reading level. The POS information was extracted using the Stanford Tagger (Toutanova et al., 2003). The formulae for type-token ratio as well as lexical variation features and the definition of lexical words were obtained from Lu (2012). Table 3.3 lists all the features belonging to this group along with the notation used to indicate the feature in the rest of this thesis.

[11]The exact file ids we used can be shared for replication experiments

Feature Code	Feature Description
POS tag density based features	
nouns	(nouns + proper nouns)/all words
propernouns	proper nouns/all words
pronouns	pronouns/all words
conj	conjunctions/all words
adj	adjectives/all words
ver	non-modal verbs/all words
interj	interjections/total sentences
adverbs	adverbs/total sentences
modals	modal verbs/total sentences
perpro	personal pronouns/total sentences
whpro	wh- pronouns/total sentences
numfuncwords	function words/total sentences
numdet	determiners/total sentences
numvb	VB tags/total sentences
numvbd	VBD tags/total sentences
numvbg	VBG tags/total sentences
numvbn	VBN tags/total sentences
numvbp	VBP tags/total sentences
Lexical Richness Features from SLA research	
lexicals	lexical words/ words
advvar (Adverb Variation)	adverbs/lexical words
adjvar (Adjective Variation)	adjectives/lexical words
modvar (Modifier Variation)	(adverbs + adjectives)/lexical words
nounvar (Noun Variation)	nouns/lexical words
verbvar (Verb Variation-II)	(verbs + aux. verbs)/lexical words
ttr (Type Token Ratio (TTR))	types/tokens
cttr (Corrected TTR)	types/$\sqrt{2 * tokens}$

Table 3.3: Lexical Richness and POS features

All features are ratios of counts unless indicated otherwise. This group of features together will be referred to as LEX in this thesis.

3.3.2 Syntactic Complexity Features

We extracted a range of syntactic features based on phrase structure trees from the Berkeley Parser (Petrov & Klein, 2007), which encoded the frequency of occurrence and length of some of the phrase groups. In addition, we adapted measures used for calculating the syntactic complexity of L2 writing from Lu (2010) for this task. These measures from SLA research proved to be useful for readability classification in our previous work (Vajjala & Meurers, 2012, 2013). We adapted the following measures for this study: mean lengths of various production units, measures of co-ordination and sub-ordination, the presence of particular syntactic structures, number of phrases of various categories, average lengths of phrases, parse tree height and number of constituents per subtree. We used the Tregex (Levy & Andrew, 2006) pattern matcher to count the occurrence of various syntactic patterns. Table 3.4 lists all the syntactic complexity features used in this thesis along with the notations. All the syntactic features together will be referred to as SYN in this thesis.

Feature Code	**Feature Description**
Non-parse tree based features	
senlen	average sentence length
commas	commas/sentences
General parse tree features	
numnp	NPs/sentences
numvp	VPs/sentences
numpp	PPs/sentences
numsbar	SBARs/sentences
avgnpsize	average length of an NP
avgvpsize	average length of an VP

avgppsize	average length of an PP
avgparsetreeheight	average height of a parse Tree
depwords	average dependents per word
numconstituents	constituents/sentences
numsubtrees	subtrees/sentences
numwh	wh-phrases/sentences
rrc	reduced relative clauses/sentences
conjp	conjunction phrases/sentences
unparsable	sentences where the parser failed/sentences
features from SLA research (Lu, 2010)	
numclauses	clauses/sentences
numtunits	t-units/sentences
mlc	average length of a clause
mlt	average length of a t-unit
cnperc	complex nominals/clauses
cnpert	complex nominals/t-units
depcperc	dependent clauses/clauses
depcpert	dependent clauses/t-units
coordperc	co-ordinate clauses/clauses
coordpert	co-ordinate clauses/t-units
vppert	verb phrases/t-units

Table 3.4: Syntactic Complexity Features

3.3.3 Word Characteristic Features

While the previous two feature sets are primarily based on POS tags and phrase structure trees, we additionally explored word level features. We hypothesized that the information about the morpho-syntactic, psycholinguistic, semantic, and age-of-acquisition characteristics could be useful for readability assessment. Thus,

we constructed a set of features based on the information provided by two widely used psycholinguistic databases: Celex and MRC, and a new database with age-of-acquisition norms for English words (Kuperman et al., 2012). We used Wordnet[12], a lexical database for English, for obtaining the semantic feature.

Morpho-Syntactic properties of Words

The Celex Lexical Database (Baayen, Piepenbrock & Gulikers, 1995) for English consists of information on the orthography, phonology, morphology, syntax and frequency for more than 50,000 English lemmas. The morphological properties of words in Celex include information about the derivational, inflectional and compositional features of the words along with information about their morphological origins and complexity. Syntactic properties of the words in Celex describe the various attributes of a word depending on its parts of speech. We used the proportion of occurrences per text of various morphological and syntactic properties of words as features (e.g., the ratio of transitive verbs, complex morphological words, and vocative nouns to the number of words that had Celex entries). Words in the document that are not included in the Celex database were ignored from this calculation. For the WeeBit corpus texts we analyzed, 40-50% of the lemmas were found in the Celex database.

In all, we used the 35 morphological and 49 syntactic properties that were expressed using character or numeric codes in the Celex database as features for our task and excluded word frequency statistics and properties which consisted of word strings. While more details about the morphological and syntactic properties of the lemmas can be found in the Celex user manual[13] with examples, the features that received higher weights in the model will be described while discussing our results. Table 3.5 and Table 3.6 list all the implemented morphological and syntactic features respectively. All counts in this list are divided by the number of words from a given text that had an entry in the Celex database. The lemmas for words were obtained by using the lemmatizer from MorphAdorner[14]. All the features from this group will be referred together as CELEX in this thesis.

[12] http://wordnet.princeton.edu/
[13] http://catalog.ldc.upenn.edu/docs/LDC96L14
[14] http://morphadorner.northwestern.edu/

Feature code	Feature Description
morphcomplex	morphologically complex words
morphmonomorphic	mono-morphic words
morphconversions	words involving conversions
morphcontractions	words with contracted form
morphirrelevant	words where morphology is irrelevant
morphobscure	words whose morphology is obscure
morphmayincluderoot	words whose morphology may include root
morphundetermined	words whose morphology is undetermined
foreignwords	foreign words
moreanalyses	words with more morphological analyses
nvaffcomp	Noun-Verb (NV) affix compounds
der	NV compounds analysed as derivation
comp	NV compounds analysed as compound
dercomp	NV comp. analysed as derivational compound
intrans	intransitive verbs
trans	transitive verbs
transintrans	transitive-intransitive verbs
unmarkedtrans	verbs with unmarked transitivity
sastem	words containing a stem
saaffix	words containing an affix
sasanda	words containing stem and affix
saflex	words with flectional form of a stem
sasandflex	words with stem and flectional form
alloblend	words with blend allomorphy
alloclip	words with clipping
alloderiv	words with derivational allomorphy
alloflex	words with flectional allomorphy
alloconv	words with conversion
subst	words with affix substitution
opacity	words with opacity

transderhash	words with derivational transformation
transderadd	derivational transformations, added letters
transderremov	derivational transformation, removed letters
infix	words with infixations
reversal	words with reversals

Table 3.5: Celex Morphological Features

Feature code	**Feature Description**
numnouns	nouns
numpropernouns	proper nouns
numadj	adjectives
numverb	verbs
numarticle	articles
numpron	pronouns
numadv	adverbs
numprep	prepositions
numconj	conjunctions
numinterj	interjections
numcountablen	countable nouns
numuncountablen	uncountable nouns
numsingularn	singular nouns
numpluraln	plural nouns
numgroupcount	countable group nouns
numgroupuncount	uncountable group nouns
numattrn	attributive nouns
numpostposN	post positive nouns
numvocN	vocative nouns
numExprN	nouns used with other words to make up a phrase
transV	transitive verbs

transcompV	verbs with object complement
intransV	intransitive verbs
ditransV	ditransitive verbs
linkingV	linking verbs
phrasalV	phrasal verbs
prepV	prepositional verbs
phrprepV	phrasal prepositional verbs
exprV	verbs used with other words to make up a phrase
ordAdj	ordinary adjectives
attrAdj	attributive adjectives
predAdj	predicative adjectives
postposAdj	post positional adjectives
exprAdj	adjectives used with other words to make up a phrase
ordAdv	ordinary adverbs
attrAdv	attributive adverbs
predAdv	predicative adverbs
postposAdv	post positional adverbs
combAdv	combinatoric adverbs
exprAdj	adverbs used with other words to make up a phrase
perPro	personal pronouns
demonPro	demonstrative pronouns
possPro	possessive pronouns
reflPro	reflexive pronouns
whPro	wh-pronouns
detPro	determinative pronouns
pronPro	pronominal pronouns
expPro	pronouns used with other words to make up a phrase
coordConj	coordinative conjunctions
subordConj	subordinative conjunctions

Table 3.6: Celex Syntactic Features

Psycholinguistic Features

The MRC Psycholinguistic Database (Wilson, 1988) is a machine-readable dictionary with around 1.5 million words along with their 26 linguistic and psychological attributes. It is a freely available online resource[15]. We used the measures of word familiarity, concreteness, imageability, meaningfulness and age of acquisition from this database as our features. Kuperman et al. (2012) compiled a database of age-of-acquisition ratings for over 50000 English words[16] through crowd sourcing. They compared their ratings with other existing age-of-acquisition norms that are also accessible through the database. We included all the Age of Acquisition (AoA) ratings from the database as features. More details on the norms and the procedure of obtaining the AoA ratings can be found in Kuperman et al. (2012). Table 3.7 lists all the Psycholinguistic features used in this thesis. All the psycholinguistic features, when used together, will be referred to as PSYCH in this thesis.

Feature Code	**Feature Description**
Features from MRC Psycholinguistic database	
familiarity	Average word familiarity rating
concreteness	Word concreteness rating
imagery	Word imagery rating
colMeaningful	Word meaningfulness rating according to Colorado norms
pavioMeaningful	Word meaningfulness rating according to Pavio norms
AoA_MRC	Avg. AoA of words
Features from Kuperman et al. (2012)	
AoA_Kup_Lem	Avg. AoA of lemmas
AoA_Kup	Avg. AoA of words
AoA_Bird_Lem	Avg. AoA of lemmas, Bird norm
AoA_Bristol_Lem	Avg. AoA of lemmas, Bristol norm

[15] `http://www.psych.rl.ac.uk`

[16] freely available at: `http://crr.ugent.be/archives/806)`

AoA_Cort_Lem	Avg. AoA of lemmas, Cortese & Khanna norm

Table 3.7: Psycholinguistic Features

Semantic Features

Finally, we used the average number of senses per word, calculated using the MIT Java Wordnet Interface[17] (Finlayson, 2014) as a semantic feature. We excluded auxiliary verbs for this calculation, as they tend to have multiple senses that do not necessarily contribute to reading difficulty.

Though the features based on lexical resources such as Celex, Wordnet, MRC database etc., are limited by the sizes of the respective databases, they capture a different type of information compared to the other feature categories we study. As we shall see in later sections, some of these features indeed received high weights in the regression model, confirming that a potential lack of coverage is not a problem that can invalidate these features in practice. One would assume this lack of coverage is most likely impact the higher reading levels given that those include less common words, which therefore are also covered less in the lexical resources. However, the features adapted from the SLA complexity literature should provide a good coverage of the properties distinguishing the more complex reading levels.

3.4 Experimental Setup

3.4.1 Modeling Method

We considered readability assessment primarily as regression, since regression helps us to identify reading levels on a numeric scale in a way that allows us to also identify the documents falling between levels. However with datasets where

[17] `http://projects.csail.mit.edu/jwi/`

there are fewer categories overall (e.g., BBC Subtitles), we modeled it as a text classification problem.

For regression, we considered only linear models since they are most readily interpretable and it is faster to build linear models. We report on the regression models using support vector regression in this chapter. We used the WEKA machine learning toolkit (Hall et al., 2009) for training and testing our models. For Support Vector Regression, we used the SMOReg (Sequential Minimal Optimization regression) implementation in WEKA with the default PolyKernel. The default exponent for PolyKernel in WEKA is 1, which makes it a linear kernel and thus will provide a human interpretable output. For classification, although we explored a wide range of learning algorithms, since the SMO classifier implementation in WEKA worked the best, we used that algorithm for our modeling process. In both classification and regression, when we used the same data set for training and testing, we used 10-fold Cross-Validation to test the internal validity of the model.

3.4.2 Evaluation Measures

For regression models, we report Pearson correlation coefficient (r) and Root Mean Square Error (RMSE) as our evaluation metrics. Pearson correlation coefficient measures the extent of linear relationship between two random variables. In readability assessment, a high correlation indicates that the texts at a higher difficulty level are more likely to receive a higher level prediction from the model and those at lower difficulty level would more likely receive a lower prediction. RMSE can be interpreted as the average deviation in grade levels between the predicted and the actual values. When comparing the performance of the regression model on other test-sets (i.e., during cross-corpus evaluations), we used Spearman's rank correlation coefficient (ρ) along with Pearson correlation, as the scales used in the various datasets are different. For classification models, we report classification accuracy as our evaluation measure.

3.5 Regression Model

We used the WeeBit corpus introduced in Section 3.2.1 to train our primary document level readability model. We used the entire feature set introduced above, which consists of a total of 151 features. Using 10 fold cross validation, the regression model achieved a Pearson correlation of 0.92 and an RMSE of 0.53. As a baseline comparison, we trained a model with only the traditional surface features (average sentence length and number of sentences per document). The model achieved a correlation of 0.6 and RMSE of 1.16. Clearly, the model with the full linguistic feature set performs much better than a model using only surface features. Although the SMOReg algorithm does not involve a feature selection procedure by default, we can infer the importance of the features in the model by looking at the weights assigned to them. The feature vectors are normalized by the SMOReg algorithm before training the model. Table 3.8 shows the five features with the highest positive and negative weights as assigned by the SMOReg model. Table 3.9 illustrates some of the features that were assigned very low weights by the model.

Feature	**Weight**	**Feature group (source)**
familiarity	+0.82	Word characteristics (Table 3.7)
AoA_Kup_Lem	+0.73	Word Characteristics (Table 3.7)
modvar	-0.61	Lexical Richness and POS (Table 3.3)
coordperc	+0.6	Syntactic Complexity (Table 3.4)
morphirrelevant	+0.56	Word Characteristics (Table 3.5)
depcperc	0.54	Syntactic Complexity (Table 3.4)
ver	-0.53	Lexical Richness and POS (Table 3.3)
pronouns	+0.51	Lexical Richness and POS (Table 3.3)
numcountablen	+0.47	Word Characteristics (Table 3.3)
nounvar	0.47	Lexical Richness and POS (Table 3.3)

Table 3.8: Top 10 Features with high weight in the WeeBit trained model

Among the top features we find lexical, syntactic, and word characteristic features, with age of acquisition and word familiarity being at the top, followed by the variability of the modifier use and several syntactic aspects, such as the use of coordinate phrases and dependent clauses per t-unit. The uninformative features

Feature	Weight	Feature group (source)
concreteness	+0.0001	Word Characteristics (Table 3.7)
predAdj	+0.0002	Word Characteristics (Table 3.6)
pavioMeaningful	+0.0017	Word Characteristics (Table 3.7)
numVPs	+0.0152	Syntactic Complexity (Table 3.4)
interj	-0.0024	Lexical Richness and POS (Table 3.3)
propernouns	-0.0038	Lexical Richness and POS (Table 3.3)
mlc	-0.0216	Syntactic Complexity (Table 3.4)
depcpert	-0.0509	Syntactic Complexity (Table 3.4)
expAdv	0	Word Characteristics (Table 3.6)
demonPro	0	Word Characteristics (Table 3.6)

Table 3.9: Top 10 Features with low weight in the WeeBit corpus trained model

also include word specific features such as concreteness and meaningfulness of a word and syntactic features such as the mean length of a clause. The heterogeneous nature of the features that are found to be useful for readability assessment supports our strategy to explore a rich linguistic feature basis on which to build readability models.

A model with so many features can be prone to over-fitting. Although performing a 10-fold cross-validation addresses this issue to some extent, establishing that the model performs well on cross-corpus evaluations would strengthen the claim that the model does not over-fit and would ensure that the model is generalizable. We therefore tested our model on a standard dataset, the Common Core Standards test set (Section 3.2.2).

3.6 Generalizability of Readability Model

When tested using the Common Core Standards test set, our model gave a Pearson correlation of 0.6, which is a drop from the correlation of 0.9 achieved during 10 fold cross-validation. However, since the scales and the related age groups in WeeBit are different from those in Common Core standards data, it is more appropriate to compare them in terms of Spearman's rank-correlation (ρ). The rank-correlation between our model's predictions and actual grades in the test-set was 0.69. A regression model trained with only surface features achieved a

Pearson correlation of 0.4 and a rank correlation of 0.5.

Nelson et al. (2012) compared the performances of six proprietary text difficulty metrics on five test sets. Since the Common Core standards dataset is a part of this study, it gives us a way to compare our system performance against seven proprietary systems. The systems compared in this study are:

1. Lexile (Metametrics, `http://www.lexile.com`)
2. ATOS (Renaissance Learning, `http://www.renlearn.com/atos`)
3. DRP analyzer (Questar Assessment Inc., `http://www.questarai.com/Products/DRPProgram`)
4. REAP (Carnegie Mellon University, `http://reap.cs.cmu.edu`)
5. Source Rater (Educational Testing Service, `https://texteval-pilot.ets.org/TextEvaluator`)
6. Pearson Reading Maturity Metric (Pearson Knowledge Technologies, `http://www.readingmaturity.com`)
7. Coh-Metrix (University of Memphis, `http://cohmetrix.memphis.edu`)

More details on the individual systems can be found in Nelson et al. (2012). Complementing this study, Flor et al. (2013) also used the grade level annotations of the Common Core standards test set to compare the Lexical Tightness measure they introduce, the Flesch-Kincaid Grade level formula, and the text length as a surface baseline. While Nelson et al. (2012) report their comparison in terms of Spearman's rank correlation (ρ), Flor et al. (2013) provide the Pearson correlation (r). To enable comparison with all of them, we report both of the measures for our models. Table 3.10 lists the performance of various systems on Common Core data as reported in the two papers and contrasts them with the results for our models.

As can be seen from the table, our readability model with all features performs on par with the best performing systems in the study and is outperformed only by

System	Spearman	Pearson
Our System	**0.69**	**0.61**
Nelson et al. (2012):		
REAP	0.54	–
ATOS	0.59	–
DRP	0.53	–
Lexile	0.50	–
Reading Maturity	**0.69**	–
SourceRater	**0.75**	–
Flor et al. (2013):		
Lexical Tightness	–	-0.44
Flesch-Kincaid	–	0.49
Text length	–	0.36

Table 3.10: Performance on CommonCore data

the SourceRater system developed by Educational Testing Service (ETS), which uses a combination of a cognitively oriented feature set and psychometric methods. In terms of Pearson correlation, our model performs better than other systems that reported the measure.

Since the Coh-Metrix performance was only reported graphically in Nelson et al. (2012), it is not included in this list. Among the various Coh-metrix dimensions used in the Nelson et al. (2012) study, a moderate negative ρ was observed for syntactic simplicity ($\sim$ -0.45), word concreteness ($\sim$ -0.4) and referential cohesion ($\sim$ -0.2) respectively.

3.7 Generalizability of Feature sets

It is clear from the above results that the readability model trained on the WeeBit corpus generalizes well across several standard datasets. Another aspect we wanted to investigate is the generalizability of the feature set used. In other words, building a model on WeeBit and testing it on other datasets establishes that the model (consisting of the features and their weights) is generalizable to a certain extent. However, how informative are the observations captured by the feature set in general and not specific to WeeBit texts? To answer this question, we trained and

tested regression models with the same feature set for different corpora using 10-fold cross-validation. Table 3.11 presents the performance of our feature set on the different corpora.

Corpus	Description	Pearson Correlation	RMSE
WeeBit	Section 3.2.1	0.9*	0.53
Common Core	Section 3.2.2	0.59*	2.69
TASA	Section 3.2.3	0.97*	1.77

Table 3.11: Using the same feature set to train multiple models (* $\rightarrow$ p$<$ 0.001)

The features were useful in building classification models for other training sets as well, showing that they are not specific to the characteristics of certain corpora. All the correlations were statistically significant. However, the performance varied between them. The model trained well with the TASA corpus, which was the largest among our data sets and also had a wide score range. This is despite the fact that we do not consider any features encoding the measures used in their formula (except sentence length). While models trained on WeeBit and TASA had a very high correlation and low RMSE, the model trained on Common Core corpus, had a lower correlation and higher RMSE than the others. This could be because of the fact that we are dealing with much smaller datasets that in addition make use of a larger scale range, i.e., a sparse data problem arising from few instances for a large set of possible values.

In our previous work, a subset of this feature set was also used successfully to train binary classification models of web-based datasets like Wikipedia-Simple Wikipedia, Time-TimeForKids, and a collection of normal news websites and those intended for children, resulting in $>$ 90% accuracies (Vajjala & Meurers, 2013). Hence, we can conclude from this experiment that, given enough training data, our feature set can be used to build a good model for a wide range of datasets annotated with reading-level judgments.

3.8 Genre Effects in Readability Models

Sheehan et al. (2008, 2010) studied the effect of text genre on readability assessment and established that the genre of a text influences readability assessment.

Flor et al. (2013) also showed that there was a difference in the performance of their readability model across different genres. Hence, to determine the genre dependence of our model, we studied the genre-wise performance of the WeeBit model on the Common Core Standards data. We chose this dataset as it includes genre annotations, while at the same time ensuring comparability with previous research. Table 3.12 presents the Pearson and Spearman correlations for the different genres of text in this dataset.

Genre	**# Docs**	r	ρ
Speech	13	0.41	0.35
Misc	44	0.61	0.69
Literature	56	0.44	0.51
Informative	55	**0.71**	**0.76**

Table 3.12: Model performance, by genre

Since the WeeBit dataset primarily consists of non-fiction articles on news events, it is not surprising that our model performs best for informational texts (ρ=0.76). This performance is at the level of the average performance of the best commercial system SourceRater on the overall Common Core Standards data set (cf. Table 3.10). The performance of the model on Speech texts was the worst compared to other categories, though. This would lead to a question: can we handle genre differences while using readability models? One way to approach the problem is to build genre specific readability models and follow a two-stage approach to readability assessment, as outlined in Sheehan et al. (2013). In stage-1, the text is assigned a genre by means of a genre classification model, and then, in stage-2, a reading level is assigned to the text using the genre specific readability model.

We explored this idea by constructing a genre specific readability model using our feature set, to identify age-specific TV programs. For this, we used the BBC-Subtitles, classified into three age-groups (cf. Section 3.2.4).

3.9 Genre Specific Readability Models

Reading, listening, and watching television programs are all ways to obtain information partly encoded in language. Just like books are written for different target groups, current TV programs target particular audiences, which differ in their interests and ability to understand language. For books and text in general, a wide range of readability measures have been developed to determine for which audience the information encoded in the language used is accessible. Different audiences are commonly distinguished in terms of the age or school level targeted by a given text.

While for TV programs the nature of the interaction between the audio-visual presentation and the language used is a relevant factor, in this thesis, we explored whether the language by itself is equally characteristic of the particular age groups targeted by a given TV program. We thus focused on the language content of the program as encoded in TV subtitles and explored the role of text complexity in predicting the intended age group of the different programs. We used the BBC Subtitles corpus introduced in Section 3.2.4 for this purpose.

Since there are only three age groups in this corpus, we treated this as a text classification problem and trained classification models with the SMO implementation in WEKA. From the distribution of the corpus in Table 3.2, it is clear that the three groups have large differences in terms of their sentence length. Thus, we first constructed a classification model with only sentence length as the feature. This yielded a classification accuracy of 71.4%, which we consider as our baseline (instead of a basic random baseline of 33%). We then constructed a model with all our features (cf. Section 3.3). This model achieves a classification accuracy of 95.9%, which is a 23.7% improvement over the sentence length baseline in terms of classification accuracy.

In order to understand what features contribute the most to classification accuracy, we applied feature selection on the entire set, using two algorithms available in WEKA, which differ in the way they select feature subsets:

- *InfoGainAttributeEval* evaluates the features individually based on their Information Gain (IG) with respect to the class.

- *CfsSubsetEval* Hall (1999) chooses a feature subset considering the correlations between features in addition to their predictive power.

Both feature selection algorithms use methods that are independent of the classification algorithm as such to select the feature subsets. Information Gain-based feature selection results in a ranked list of features, which are independent of each other. The Top-10 features according to this algorithm are listed in Table 3.13.

Feature Code	**Feature Group (Ref)**
AoA_Kup_Lem	Word Characteristics (Table 3.7)
numpp	Syntax (Table 3.4)
sasanda	Celex (Table 3.5
– avgparsetreeheight	Syntax (Table 3.4)
– numnp	Syntax (Table 3.4)
subst	Celex (Table 3.5
– numprep	Celex (Table 3.6
numuncountablen	Celex (Table 3.6
numclauses	Syntax (Table 3.4)
– senlen	Syntax (Table 3.4)

Table 3.13: Ranked list of Top-10 features using IG, for BBC Subtitles Corpus

As mentioned in the description, all Top-10 features encode different linguistic aspects of a text. While there are more syntactic features followed by Celex features in these Top-10 features, the most predictive feature is a psycholinguistic feature encoding the average age of acquisition of words. A classifier using only the Top-10 IG features achieves an accuracy of 84.5%.

Applying CfsSubsetEval to these Top-10 features set selects the six features not prefixed by a hyphen in the table, indicating that these features do not correlate much with each other. A classifier using only this subset of 6 features achieves an accuracy of 84.1%.

We also explored the use of CfsSubsetEval feature selection on the entire feature set instead of using only the Top 10 features. From the total of 152 features, CfsSubsetEval selected a set of 41 features. Building a classification model with only these features resulted in a classification accuracy of 93.9% which is only 2% less than the model including all the features. Table 3.14 shows the specific feature subset selected by the CfsSubsetEval method, including some examples

illustrating the morphological features. The method does not provide a ranked list, so the features here simply appear in the order in which they are included in the feature vector.

Feature Description
preposition phrases
t-units
co-ordinate phrases per t-unit
lexical words in total words
interjections
conjunctive phrases
word senses
verbs
verbs, past participle (VBN)
proper nouns
plural nouns
avg. corrected type-token ratio
avg. AoA acc. to ratings of Kuperman et al. (2012)
avg. AoA acc. to ratings of Cortese & Khanna (2008)
avg. word imageability rating (MRC)
avg. AoA according to MRC
morph. complex words (e.g., *sandbank*)
morph. conversion (e.g., *abandon*)
morph. irrelevant (e.g., *meow*)
morph. obscure (e.g., *dedicate*)
morph. may include root (e.g., *imprimatur*)
foreign words (e.g., *eureka*)
words with multiple analyses (e.g., *treasurer*)
noun verb affix compounds (e.g., *stockholder*)
lemmas with stem and affix (e.g., *abundant*=*abound*+*ant*)
flectional forms (e.g., *bagpipes*)
clipping allomorphy (e.g., *phone* vs. *telephone*)
deriv. allomorphy (e.g., *clarify–clarification*)

flectional allomorphy (e.g., verb *bear* $\mapsto$ adjective *born*)
conversion allomorphy (e.g., *halve–half*)
lemmas with affix substitution (e.g., *active=action+ive*)
words with reversion (e.g., *downpour*)
uncountable nouns
collective, countable nouns
collective, uncountable nouns
post positive nouns.
verb, expression (e.g., *bell the cat*)
adverb, expression (e.g., *run amok*)
reflexive pronouns
wh pronouns
determinative pronouns

Table 3.14: CfsSubsetEval feature subset

Table 3.15 summarizes the classification accuracies with the different feature subsets seen so far, with the feature count shown in parentheses.

Feature Subset (#)	**Accuracy**	**SD**
All Features (152)	95.9%	0.37
Cfs on all features (41)	93.9%	0.59
Top-10 IG features (10)	84.5%	0.70
Cfs on IG (6)	84.1%	0.55

Table 3.15: Accuracy with various feature subsets

We performed statistical significance tests between the feature subsets using the Paired T-tester (corrected), provided with WEKA and all the differences in accuracy were found to be statistically significant at $p < 0.001$. We also provide the Standard Deviation (SD) of the test set accuracy in the 10 folds of Cross Validation per dataset, to make it possible to compare these experiments with future research on this dataset in terms of statistical significance.

Table 3.16 presents the classification accuracies of individual features from the Top-10 features list (introduced in Table 3.13).

Feature Code	**Accuracy**
AoA_Kup_Lem	82.4%
numpp	74.0%
sasanda	77.7%
avgparsetreeheight	73.4%
numnp	73.0%
subst	74.3%
numprep	72.0%
numuncountablen	68.3%
numclauses	72.5%
senlen	71.4%

Table 3.16: Accuracies of Top-10 individual features

The table shows that all but one of the features individually achieves classification accuracies above 70%. The first feature (AoA_Kup_Lem) alone resulted in an accuracy of 82.4%, which is quite close to the accuracy obtained by all the Top-10 features together (84.5%).

To obtain a fuller picture of the impact of different feature groups, we also performed ablation tests removing some groups of features at a time. Table 3.17 shows the results of these tests along with the SD of the 10 fold CV. All the results that are statistically significant at $p < 0.001$ from the model with all features (95.9% accuracy, 0.37 SD) are indicated with a *.

Interestingly, removing the most predictive individual feature (AoA_Kup_Lem) from the feature set did not change the overall classification accuracy at all. Removing all of the AoA features or all of the psycholinguistic features also resulted in only a very small drop. The combination of the linguistic features, covering lexical and syntactic characteristics as well as the morphological, syntactic, orthographic, and phonological properties from Celex, thus seem to be equally characteristic of the texts targeting different age-groups as the psycholinguistic properties, even though the features are quite different in nature. In terms of separate groups of features, syntactic features alone performed the worst (77.5%) and lexical richness features the best (93.1%).

Features	**Acc.**	**SD**
All − AoA_Kup_Lem	95.9%	0.37
All − All AoA Features	95.6%	0.58
All − PSYCH	95.8%	0.31
All − CELEX	94.7%*	0.51
All − CELEX − PSYCH	93.6%*	0.66
All − CELEX − PSYCH − LEX (= SYN only)	77.5%*	0.99
LEX	93.1%*	0.70
CELEX	90.0%*	0.79
PSYCH	84.5%*	1.12

Table 3.17: Ablation test accuracies

To investigate which classes were mixed up by the classifier, consider Table 3.18 showing the confusion matrix for the model with all features and 10 fold CV.

classified as →	**CBeebies**	**CBBC**	**Adults**
CBeebies (0–6)	3619	156	1
CBBC (6–12)	214	3526	36
Adults (12+)	2	58	3716

Table 3.18: Confusion Matrix

We find that CBeebies is more often confused with the CBBC program for older children (156+214) and very rarely with the program for adults (1+2). The older children programs (CBBC) are more commonly confused with programs for adults (36+58) compared to CBeebies (1+2), which is expected given that the CBBC audience is closer in age to adults than the CBeebies audience.

Summing up, we can conclude from these experiments that the classification of transcripts into age groups is informed by a wide range of linguistic and psycholinguistic features. While for some practical tasks a few features may be enough to obtain a classification of sufficient accuracy, the more general take-home message is that authentic texts targeting specific age groups exhibit a broad range of linguistics characteristics that are indicative of the complexity of the language used.

3.9.1 Effect of text size and training data size

When we first introduced the properties of the BBC Subtitles corpus in Table 3.2, it appeared that sentence length and the overall text length could be important predictors of the target age-groups. However, the list of Top-10 features based on information gain was dominated by more linguistically oriented syntactic and psycholinguistic features.

Sentence length was only the tenth best feature by information gain and did not figure at all in the 43 features chosen by the CfsSubsetEval method selecting features that are highly correlated with the class prediction while having low correlation between themselves. As mentioned above, sentence length as an individual feature only achieved a classification accuracy of 71.4%.

The text length is not a part of any feature set we used, but considering the global corpus properties, we wanted to verify how well it would perform. Thus, we trained a model with only text length (#sentences per text) as a feature. This achieved a classification accuracy of only 56.7%.

The corpus consists of transcripts of whole TV programs and hence an individual transcript text typically is longer than the texts commonly used in readability classification experiments. This raises the question whether the high classification accuracies we obtained are the consequences of the larger text size (as measured by text length in words).

As a second issue, the training data size available for this data set for 10-fold cross-validation experiments is comparatively large, given the 3776 texts per level available in the overall corpus. We thus also wanted to study the impact of the training size on the classification accuracy achieved.

Pulling these threads together, we compared the classification accuracy against text length and training set size to better understand their impact. For this, we trained models with different text sizes (by considering the first 25%, 50%, 75% or 100% of the sentences from each text) and with different training set sizes (from 10% to 100%). Figure 3.1 presents the resulting classification accuracy in relation to training set size for the different text sizes. All models were trained with the full feature set, using 10-fold cross-validation as before.

As expected, both the training set size and the text size affect the classification

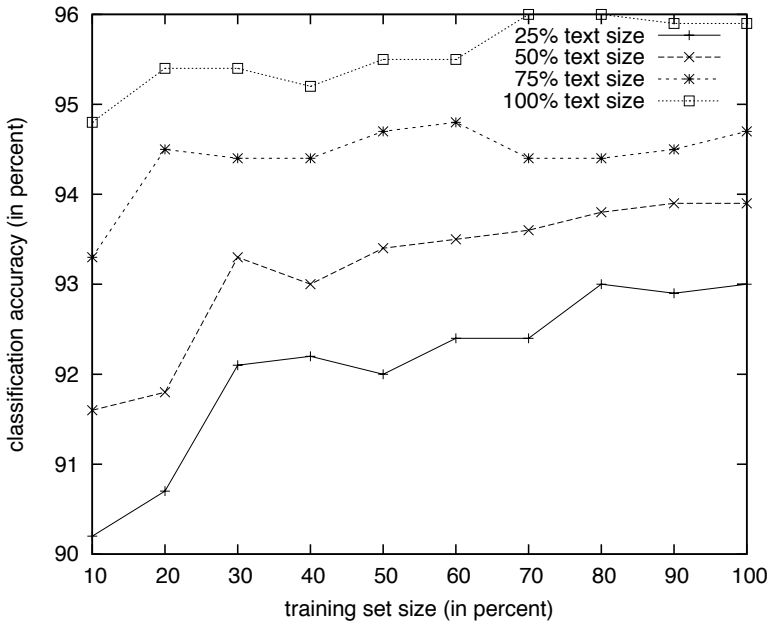

Figure 3.1: Classification accuracy for different text sizes and training set sizes

accuracy. However, the classification accuracy even for the smallest text and training set size is always above 90%, which means that the unusually large text and training size is not the main factor behind the very high accuracy rates. In all four cases of text size, there was a small effect of training set size on the classification accuracy. But the effect reduced as the text size increased. At 25% text size, for example, the classification accuracy ranged 90–93% (mean 92.1%, SD 0.9) as the amount of training set used increased from 10% to 100%. However, at 100% text size, the range was only 94.8–96% (mean 95.6%, SD 0.4).

Comparing the results in terms of text size alone, a larger text size resulted in a better classification accuracy in all cases, irrespective of the training set size. A longer text will simply provide more information for the various linguistic features, enabling the model to deliver better judgments about the text. However, despite the text length being reduced to one fourth of its size, the models built with our feature set always collect enough information to ensure a classification accuracy of at least 90%.

In the above experiments, we varied the text size from 10% to 100%. But since we are taking a certain percentage of data from each text, texts from CBBC

and Adults on average still are longer than CBEEBIES texts. While this reflects the fact that TV transcripts in real life are of different length, we also wanted to see what happens when we eliminate such length differences.

We thus trained classification models fixing the length of all documents to a concrete absolute length, starting from 100 words (rounded off to the nearest sentence boundary) increasing the text size until we achieve the best overall performance. Figure 3.2 displays the classification accuracy we obtained for the different (maximum) text sizes, for all features and feature subsets.

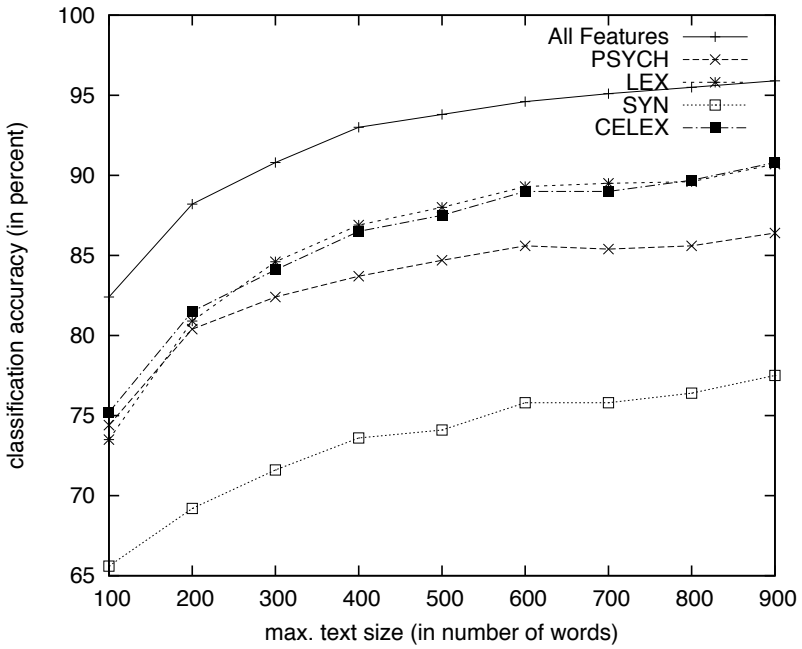

Figure 3.2: Classification accuracy for different absolute text sizes (in words)

The plot shows that the classification accuracy already reaches 80% accuracy for short texts, 100 words in length, for the model with all features. It rises to above 90% for texts that are 300 words long and reaches the best overall accuracy of almost 96% for texts which are 900 words in length. All the feature subsets follow the same trend too, with varying degrees of accuracy, which is always lower than the model with all features.

From this experiment, we can conclude that our feature set is generalizable to speech as well and we can build efficient genre specific readability models with

this approach.

3.10 Topic Differences

Apart from genre (spoken vs. informative vs. literary etc.,), another possible source of differences in readability scores could be the topic. Since the TASA corpus includes a topic annotation for most of its texts, we explored the utility of our feature set for building good topic specific models. Table 3.19 summarizes the results of the 10-fold cross-validation based regression model performance for different topics in terms of Pearson correlation and RMSE.

Topic	# Docs	DRP score range	r	RMSE
Health	∼1300	40–81	0.98	1.36
Science	∼5K	35–81	0.98	1.58
Language Arts	∼16K	28–110	0.95	1.62
Social Studies	∼ 10K	35–110	0.88	4.47
Business	∼ 1000	47–80	0.95	1.58
Miscellaneous	∼ 700	36–81	0.98	1.94
Home Economics	∼ 300	54–83	0.88	2.33

Table 3.19: Topic specific readability models with TASA corpus

All the models resulted in a high correlation and low RMSE. Compared to the other topics, however, the Social Studies model had a slightly lower correlation and a much higher RMSE. While understanding the reasons for specific topical differences needs further investigation, we can conclude from these results that the feature set is sufficiently general and informative across topics.

3.11 Chapter Summary

In this chapter, we explored the problem of automatic readability assessment by building multiple supervised learning models based on a rich feature set that models properties of the language studied in Second Language Acquisition and Psycholinguistics along with several linguistic features. We also established the generalizability of the feature set and models, by training and testing the models on

four existing readability annotated datasets. We studied the genre and topic dependence of models and built topic and genre specific models using existing, annotated corpora. We investigated the effect of training data size on classification accuracy. We also briefly studied the effect of text size on classification accuracy, which will be discussed in more detail in Chapter 5, where we look at the readability of single sentences.

While all the models trained well and achieved the second best reported result on the common core standards test set, we found out that the model is biased towards informational texts and performs poorly with spoken texts. This is not surprising, since the model was trained on informational texts. However, the features we used generalized well to spoken texts as well, and we achieved a 96% classification accuracy using a movie subtitles corpus. We also explored the effect of text size on prediction accuracy and found that while reducing the text sample size will result in a drop in accuracy, the drop is less for a model that considers all features instead of only certain categories of them. Finally, we briefly explored training topic specific readability models using the TASA corpus.

3.11.1 Outlook

Our experiments showed that we could build reliable models for assigning reading levels to various kinds of texts, which can be used in real life, and which have a strong linguistic grounding. The immediate next step is to develop a web-based or desktop based application that can be used by the non-technical users to select texts suiting their reading level.

The current approach does not take into account several aspects of the text (e.g., discourse). Future directions include creating experiments where we can feasibly combine these aspects towards building a holistic approach to readability assessment. The role of genre and domain specific nature of texts also needs further consideration. Finally, we did not take the user or task into account while building these readability models. Approaches that can integrate these aspects into current readability models need to be explored in future.

Chapter 4

Understanding the effect of text complexity on readers: An Eye-tracking Study

Abstract

In this chapter, we explore the effect of text complexity on online processing and offline performance of the readers. For this purpose, we conducted an eye-tracking experiment where the participants were asked to read texts belonging to two reading levels. They were also asked to answer recall and comprehension questions for us to be able to analyze their performance outcomes. In addition, we also explored the effect of language proficiency on both the online and offline processes. Our results show that text complexity was a significant predictor for three of the six eye tracking measures we studied while reader's L2 proficiency had a significant effect on two of them. In terms of the outcome measures, while proficiency affects both recall and comprehension, text complexity is correlated only with recall performance.

The work described in this chapter was supported by LEAD Intramural Research Grant, project number: 19110507 (2013-2015). Other members of the project are: Alexander Eitel, Detmar Meurers and Katharina Scheiter.

4.1 Introduction

Automatic readability models primarily rely on gold-standard data with grade levels assigned based on the judgments of teachers and other language experts. However, it is important to ensure that the model predictions for a given text reflect the actual comprehension difficulties readers may face with the text. Cunningham & Mesmer (2014) recently argued that the difficulty scores assigned to gold-standard texts like Common Core Standards exemplars should be based on the students' reading performance with the texts rather than expert judgments. One way to validate the usage of the models of text complexity created based on expert judgments is by comparing the model predictions with the reading and comprehension performance of readers for selected texts. On a related note, though it is clear that some texts can be difficult to read for certain target audience, it is not clear if simplifying these texts will result in better recall and comprehension for the target audience. A number of studies in the fields of education and psychology explored these aspects in the past.

For example, Evans (1972) compared the effect of unsimplified versus grammatically simplified versions of five prose selections on 12th grade students, using multiple choice questions and cloze tests. These experiments showed that the students performed significantly better with simplified versions. Walmsley et al. (1981) investigated the effect of text simplification on the reading comprehension of elderly (60+) readers. Their results showed that while simplifying by readability formulae had no effect on comprehension, subjective rewriting resulted in an improvement in some cases, the reading ability of the readers was a significant predictor in the process. Green & Olsen (1988) studied reader preferences for and comprehension of original and adapted fiction with 58 second grade students. Using two original children's books and their adapted versions re-written by a publisher, they showed that while children preferred original texts over the adapted versions, there was no significant performance difference between the two versions in terms of comprehension.

In a slightly different context, Charrow (1988) used three versions of a car manufacturer's recall letter (original, simplified based on formulae, rewritten based on guidelines) and tested their effect on 56 participants who were buyers of cars.

They used multiple choice questions and opinion questionnaire to study the differences between text versions and found that while the simplified version based on readability formulae did not result in better comprehension, the guidelines based version did. Smith (1988) studied the effect of linguistic complexity of instructions on the performance of children in playing games and showed that the comprehension also depended on the task that the children were asked to perform and not on text complexity alone. Britton & Gülgöz (1991) used the Kintsch's reading comprehension model (Kintsch & van Dijk, 1978) to revise a 1000 word expository text and showed that the free recall of this principled revision increased compared to that of the original version.

More recently, Crossley et al. (2014) used a moving windows self-paced reading task to study the effect of text simplification on text comprehension and reading time of second language English learners. This task uses a moving window where the sentence is not seen at once but as parts, like words or phrases. Participants usually see the next word/phrase by pressing a spacebar or some such button. So, the context of the sentence or text is not seen together with the current word/phrase in this task. Nine texts (each written in three versions) from onestopenglish.com were used in this experiment. Comprehension was assessed through yes/no questions and the participants also took an English language proficiency test. This experiment showed that while text complexity significantly correlated with the reading time (explaining $\sim$12% of the variance), its effect was no longer significant if the participant's English proficiency was taken into account. In terms of comprehension, while text complexity was significant, less proficient readers have benefited more from reading the simplified versions than highly proficient readers. That is, the effect of text complexity on comprehension depended on language proficiency.

In cognitive psychology research, studying eye movement patterns is a standard method for understanding the cognitive processes involved in reading and comprehension (e.g., Just & Carpenter, 1980; Rayner, 1998; Jr, Staub & Rayner, 2007). Eye tracking, though time and cost consuming, provides a more natural way to study the reading processes compared to reading time studies. In addition, it allows us to study the processes like re-reading of the text by the users. We can also get multiple measures of processing compared to self-paced reading which

gives us only one measure (reading time). These measures will also provide us more insights into how readers respond to different kinds of reading difficulties.

Eye movements in reading research are typically studied in terms of fixations, saccades and regressions. Fixations refer to the relatively stationary positions of the eye at specific areas of the text and saccades refer to the rapid eye movements between two fixations. Regressions refer to the cases where the reader revisits and fixates on parts that were already read. Reader's comprehension difficulties were shown to manifest in longer fixations, shorter saccades and more regressions in previous research (c.f. Rayner (1998) for a review). Rayner et al. (2006) explicitly studied how text's difficulty level affects eye movement measures in reading. They used a collection of 32 text passages and asked 32 students to rate them on a scale of 1–10. The passage difficulty from these ratings ranged from 2.8 (relatively easy) to 6.6 (moderately difficult). They found out that the text difficulty rating correlated strongly with average fixation duration, number of fixations and total time. Readers' performance with comprehension questions correlated negatively with difficulty, indicating that the readers had difficulties with more difficult passages, but this correlation was statistically insignificant. To our knowledge, this is the only other existing study that used eye tracking as a method to study the effect of text difficulty on the cognitive processing of the readers.

4.1.1 Our Study

In this study, we performed an eye-tracking experiment to understand the cognitive correlates of text complexity and the effect of text complexity on reader performance. We also studied the effect of reader's language proficiency on eye movements and performance outcomes.

Our study differs from Rayner et al. (2006) in terms of the materials used, from Crossley et al. (2014) in terms of the experimental methods and from both the studies in terms of the additional variables studied. While Rayner et al. (2006) used a set of unrelated text passages, we used parallel versions of texts in easy and difficult versions for the task, similar to Crossley et al. (2014). This enables us to specifically study the effect of simplifying a text on reading performance. While Crossley et al. (2014) used texts from the same source as we did, they did a

self-paced reading time study without using eye-tracking. In terms of evaluation, Rayner et al. (2006) evaluated the reader's performance by asking four multiple choice comprehension questions per text and Crossley et al. (2014) used four yes/no questions per text to assess comprehension. In our study, we developed a set of eight recall and six comprehension questions per text, such that they can be answered by reading any version of the text. As in Crossley et al. (2014), we also study the effect of the reader's language proficiency on both reading performance and eye-tracking measures. Finally, in this study, we compare three different notions of text complexity - an expert rating, a psycholinguistic measure and the output of a computational linguistic model, unlike the previous two studies, that employ only a single notion of text complexity based on expert (Crossley et al., 2014) and non-expert (Rayner et al., 2006) judgments.

The rest of the chapter describes our study in detail and is organized as follows: Section 4.2 lists the primary research questions in this study and our hypotheses about them. Section 4.3 describes the experimental procedure including the variables studied. Section 4.4 explains our data analysis methods. Section 4.5 discusses the results from the analysis and Section 4.6 concludes the chapter with pointers to future work.

4.2 Research Questions and Hypotheses

In this study, we explored the effect of text complexity and reader's English proficiency by means of eye-tracking measures and two measures of reading performance - recall and comprehension. We compared three notions of text complexity in this study: expert annotated complexity as given by onestopenglish[2] writers, the score assigned by our readability model from Chapter 3 and a psycholinguistic measure, Surprisal. The primary research questions of this experiment are:

1. Does text complexity affect online-processing? - Based on existing research, we expect that text complexity will result in an increase in the number of fixations, average fixation duration and revisits (more revisits ⇒ more regressions).

[2] http://www.onestopenglish.com/

2. Does text complexity affect the offline performance outcomes of the readers? Based on existing research, we expect that an increase in text complexity will result in a decrease in both recall and comprehension scores.

3. Does language proficiency influence any of the online or offline measures? We hypothesize that there will be an interaction between text complexity and a reader's language proficiency. Based on previous research, it can be expected that the impact of text complexity on reading performance is more pronounced for less proficient readers than highly proficient readers.

4. Are there differences between the different notions of text complexity in terms of explaining the online and offline measures? - Since the different notions encode different features of text complexity, we could hypothesize that there are differences.

5. Can effects of reading difficulty on online processing behavior explain differences in learning outcomes? - We hypothesize that there is a possibility that the online processing variables mediate the effect of text complexity or reader proficiency on the eventual performance of the reader in terms of their learning outcomes.

We studied the first question by assessing the participants' eye movements while reading and the second question by asking recall and comprehension questions about the texts read by them. For the third question, we assessed the performance of the participants in an English proficiency test. For the fourth question, we compared three different notions of text complexity - expert assigned labels, a computational model of linguistic complexity, and a psycholinguistic measure. For the last question, we employed mediation analysis.

4.3 Experimental Method

4.3.1 Participants

We ran the experiment using 49 native speakers of German (33 female, 16 male) chosen from a population of university students from various disciplines (average

age: 24.3, range: $[19, 32]$). All students belonged to the University of Tübingen and were recruited by a group email. They were asked to do German and English proficiency tests before the experiment. For assessing German, we used the LGVT 6-12 (Schneider et al., 2007), a standard German reading speed and comprehension test. English proficiency was assessed through a c-test used at the University of Tübingen, which will be explained in the next section. We also collected other information from the participants, such as the years of exposure to English and other languages known.

4.3.2 Texts

Simplified texts primarily come in two forms: adapted or abridged versions of original texts. Manual simplification can be performed either with the aim of modifying specific words and syntactic structures in a uniform manner (for example, according to a readability formula) or by rewriting the texts following the intuition of the authors who typically have an idea about the linguistic capabilities of the target audience.

In the present study, we used materials that were intuitively simplified by writers at onestopenglish.com, a website run by MacMillan publishing group for second language learners. Each week, onestopenglish.com chooses one news article from *The Guardian*, an internationally renowned British weekly journal and rewrites it into three versions: beginner, intermediate and advanced. Texts from this website have been used by other researchers in the past to study the linguistic differences between simplified and unsimplified versions (e.g., Allen, 2009a; Crossley et al., 2012) and the effect of simplification on readers (Crossley et al., 2014). We chose *four texts, each having two versions* (advanced and beginner), for our experiments. We chose the texts through manual inspection in such a way that they differed linguistically in form but without loss in meaning. Since we did an eye-tracking study, we restricted the length of the text given to the participants to $\sim 250 - 300$ words and made sure that the text lengths do not differ much between versions of the same text. Table 4.1 shows the number of words for each text, in both versions.

Text_Version	Num. words
1_Difficult	296
1_Easy	298
2_Difficult	286
2_Easy	234
3_Difficult	248
3_Easy	230
4_Difficult	312
4_Easy	306

Table 4.1: Number of words in the texts used for the experiment

Recall and comprehension questions were prepared after reading both text versions, ensuring that each version contains same answers to the questions. There were eight recall questions and six comprehension questions per text. While the recall questions primarily dealt with the factual information that is directly available in the text, comprehension questions were yes/no questions that needed drawing inferences. All the texts we used and the associated recall and comprehension questions can be seen in Appendix B.

4.3.3 Experimental Procedure

We used the iView X^{TM} Hi-Speed eye-tracker from SensoMotoric Instruments (SMI) for running our eye-tracking experiment and used the software SMI BeGaze[3] with the Reading package for analyzing the data from this experiment. The participants were randomly assigned to four experimental conditions, which differed in the ordering of the texts used, and the reading level (referred to as text order in this chapter). We followed a latin square design which ensured that each participant read each of the four texts, two in easy and two in difficult versions. No participant read the same text in two different versions. All participants read a trial text before starting to read the actual texts and answered recall and comprehension questions about the text. The trial text was written to suit the intermediate reading

[3] `http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/begaze-analysis-software.html`

level. This was done to ensure that the participants understood the task. The texts were shown in two pages on the eye-tracker since they were about 300-350 words long. The eye-tracking experiment was conducted by a trained research assistant with experience in using the eye-tracking equipment.

The step-by-step procedure of the experiment was as follows:

1. The participants provided their demographic details and were briefed about the experiment. They then took the LGVT 6-12 test for German.

2. After calibrating the eye-tracker, the participants read a trial text and answered the recall and comprehension questions on paper.

3. Then, the participants read all the four texts that were assigned to their condition one by one, each time answering the questions on paper and getting back to the eye-tracker.

4. After finishing the reading, the participants did the English proficiency test and answered a few questions about the experiment.

The entire experimental process per participant lasted about 60-90 minutes.[4]

4.3.4 Dependent Variables

Eye-tracking measures

We explored 6 eye-tracking measures for this study.

1. **Fixation count**: This refers to the average number of fixations per sentence in a text. Previous eye-tracking research established that the reader's comprehension difficulties are reflected in the eye movements through increased fixations (Rayner, 1998). Fixation count also had a strong correlation with text difficulty in Rayner et al. (2006).

[4]The text collection, questions creation, eye-tracking experiment and the data collection was performed in collaboration with Alexander Eitel and Katharina Scheiter at the Knowledge Media Research Center, Tübingen. The online interface for the English c-test was created by Magdalena Wolska.

2. **Average Fixation Duration**: This refers to the average duration (in milliseconds) of fixations in a text. It is expected that the difficulties in processing the text are reflected in eye movements through longer fixations (e.g., Just & Carpenter, 1980; Rayner, 1998). This measure also correlated strongly with text difficulty in the Rayner et al. (2006) study.

3. **First fixation duration**: This refers to the duration of the first fixation in a sentence in milli seconds (ms). This is one of the measures used for studying lexical activation while reading and is used as a word frequency measure in reading research (Rayner, 1998).

4. **First pass duration**: This is the sum of the fixation durations during the first pass through a sentence (ms). This is indicative of the comprehension processes in reading research.

5. **Second pass duration**: This is the sum of the fixation durations during the second pass through a sentence (ms). This can be indicative of re-reading of a sentence which may indicate a reading difficulty.

6. **Revisits**: Average revisits is understood as a measure of further visits to a an area of interest and is defined as: (number of glances/subjects with atleast one visit)-1.

Performance Outcome Measures

We explored two aspects of performance outcome, recall and comprehension. They were measured as follows[5]:

1. **Recall**: The recall questions asked factual information from the text, that can be reproduced verbatim by the readers.

2. **Comprehension**: The comprehension questions tested the understanding of the readers by asking yes/no questions which require drawing inferences from the text.

[5] As mentioned earlier, example questions can be seen in Appendix B.

4.3.5 Independent Variables

We used three independent variables in this study - text complexity, participant's English proficiency and the order in which the participants read texts.

Text Complexity

The texts we considered for this study encode text complexity as a binary variable having labels -"beginner" and "advanced". These labels are assigned by the authors at onestopenglish.com, who are experienced in preparing simplified texts for English language learners. We consider this as the primary notion of text complexity in our analysis. However, as mentioned earlier, we also consider two other notions of text complexity. The first one is our own readability model (cf. Chapter 3), which is based on a range of linguistic features that model the word level and syntactic properties of text.

The other notion of complexity comes from a psycholinguistic measure - surprisal, which is a measure of expected cognitive load during sentence processing, based on information theory. It is expected to be indicative of the word level difficulty in context. A higher surprisal value implies higher processing difficulty. Surprisal values for texts can be computationally estimated by some of the freely available software. We took surprisal scores as measured by Roark parser (Roark et al., 2009) for the texts used in this study. We took the average of the "total surprisal" of all sentences from the parser to get an estimate of the surprisal for each text.

Both these measures consistently rated the "easy" document as being at a lower reading level than the "difficult" document. This means that the "easy" document is more readable according to both the measures. All the three notions of text complexity we used in this study are summarized in Table 4.3.5.

Measure	Source	Description
Binary Complexity	onestopenglish.com	has two values - easy, difficulty
Surprisal	(Roark et al., 2009)	Not restricted to a specific value. Higher values of surprisal indicate difficult text
VM (cf. Chapter 3)	Vajjala & Meurers (2014b)	outputs a score on a scale of 1-6.

Table 4.2: Text Complexity Measures Used in this Study

Table 4.3 shows the Surprisal scores and the reading level from VM for the texts we used. As mentioned earlier, both the approaches assigned lower scores to the easier versions than difficult versions.

Text_Version	**VM**	**Surprisal**
1_Difficult	5.19	207.5
1_Easy	3.9	147.2
2_Difficult	4.2	193.2
2_Easy	3.1	112.3
3_Difficult	4.1	165.4
3_Easy	3.0	124.6
4_Difficult	5.4	181.9
4_Easy	4.8	144.4

Table 4.3: Text Complexity Scores for the Texts Used

English Proficiency

We considered the L2 proficiency of the reader as one of the predictor. For assessing English proficiency of the participants, we used an online cloze test (Taylor, 1953) from the placement tests used by the University of Tübingen. This test consists of five text passages where word-endings of some of the words are missing and test takers should fill in the blanks. The test taker's proficiency is based on the number of correctly filled in answers. The exact test we used is attached in

Appendix C. The average score for the English proficiency test among our participants was 72.6 (range: $[21, 112]$, sd: 20.24) where a score of 100 or above is considered extremely proficient.

Text Order

In addition to these two variables, since the users read four texts, we considered the order in which a text was given to the user (which depended on the experimental condition) as a third independent variable.

4.4 Data Analysis Methods

We studied the relationships between the dependent and independent variables using linear regression and Generalized Additive Mixed Models (GAMMs). While linear models are easy to interpret, additive models allow the flexibility to model complex, non-linear interactions between variables. In GAMMs, the response variable is modeled as being dependent on the smooth functions of predictor variables. GAMMs have been used successfully in modeling experimental data in psycholinguistic research in the recent past (e.g., Wieling et al., 2014).

While additive models in general allow us to explicitly model non-linear relationships between the variables, they do not take the random effects nature of the data into account. Mixed-effects models make a distinction between fixed and random factors among the independent variables. Fixed effects are the variables with a limited range of options where it is possible to exhaust all the possible values (e.g., the reading difficulty of a text is measured on some scale with boundaries). Random effects refer to those that have values sampled from a population and it is not possible to exhaustively cover all the levels in an experiment e.g., participants who did an experiment. We cannot exhaustively enlist all the participants in the entire population group for our experiment. So the variation due to participants can be considered a random effect. Thus, GAMMs allow us to have the benefits of both additive and mixed effects models. We constructed our GAMMs by using thin-plate regression splines for smoothed components, as implemented in the

mgcv package in R (Wood, 2006)[6].

4.4.1 Fixed and Random Effects in GAMMs

Fixed Effects: we consider text complexity, participant's English proficiency and the order in which the four texts appeared in a given condition (1-4) as fixed effects.

Random Effects: There are two likely random-effect factors that may cause a systemic variation.

1. Participants: We run our experiments with a limited number of participants, who form only a sample of a much larger population who are the target readers of difficult texts. Hence, we consider participant as a random factor. Although proficiency of the participant is modeled as a fixed effect, this factor may account for other possible unknown variations between them.

2. Texts: We used four texts, written in two versions in our experiment. The texts represent only a sample among a large set of texts that the learners may read and does not exhaustively cover all the possible options. Hence, we consider this as a random effect.

Both the random effect factors are considered as random intercepts in constructing the GAMM models reported in the next section, as considering them as random slopes did not result in any significant performance difference compared to the model with random intercepts.

4.4.2 Model Comparison

We evaluated the trade off between the complexity of a GAMM and the simplicity of a linear model by comparing them in terms of the variance explained (R^2). Comparison between linear models and GAMMs was performed based on the Akaike Information Criterion (AIC). An AIC difference of >2 is considered as a threshold for choosing the model with lower AIC as a better model.

[6] `http://cran.r-project.org/web/packages/mgcv/`

4.4.3 Mediation Analysis

Finally, we used mediation analysis to understand if the effects of text difficulty on online processing can explain the performance outcomes of the participants (Question 4 from Section 4.1). Mediation analysis is the process of studying the relationship between the dependent and independent variables by means of a third "mediator" variable. In mediation models, it is generally hypothesized that the independent variable influences the mediator, which in turn influences the dependent variable. It is usually used to understand the underlying mechanism behind a known relationship. We performed this analysis using the mediation package in R (Tingley et al., 2014)[7].

4.5 Results

For linear models, we consider one of the measures of text complexity, the participant's language proficiency and text order as the independent variables. We considered the interaction between proficiency and text complexity in this model.

For GAMM models, we consider one of the measures of text complexity, participant's language proficiency and the order in which a participant read the text as the fixed effects. For our experiment, since they read four texts, the text order is a bounded (factor) variable with values: 1, 2, 3, 4. The interaction between proficiency and complexity is specified using the "by=" option in the smooth function for proficiency. We considered the participant and text variation as random effects. For both linear models and GAMMs, while we constructed the models using all the three measures of complexity (onestopenglish, Surprisal and VM), a detailed discussion of the results was done for a model with a binary notion of complexity only.

For the sake of uniformity in comparisons between variables, we considered a non-transformed version of the data, without removing any outliers, for all the models reported in this chapter. We also left the model setup in terms of the fixed and random effects the same for all the dependent variables. Thus, the GAMM summaries discussed here refer to models of the form:

[7] http://cran.r-project.org/web/packages/mediation/

model = gam(dependent_variable ∼ Difficulty + s(Proficiency) + TextId +
s(Proficiency, by=Difficulty) + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

More detailed discussion of the effect of model comparisons due to data transformations, removal of outliers/missing data for individual variables and a comparison of model fit can be seen in Appendix A. A study of the three-way interaction between text order, proficiency and text complexity is also presented in the appendix.

4.5.1 Online Processing Variables

Fixation Count

In linear models, binary text complexity had a significant effect ($p < 0.001$) on fixation count, where the fixation counts increased for difficult texts compared to easy texts. The effect of proficiency was significant ($p < 0.05$) and the fixation counts decreased with an increase in reader proficiency. Text order had a significant effect ($p<0.05$) and the fixations increased for texts read at a later point of time during the experiment. Both the other notions of text complexity Surprisal ($p<0.001$) and VM ($p<0.01$) too had a significant effect on fixation count, with the direction being the same as binary text complexity, i.e., increased values of surprisal and VM resulted in increased fixation counts. In all the cases, the interaction between proficiency and complexity was not significant in the linear models. All the linear models explained 10-12% of the variance.

In comparison, the GAMM models with all the three notions of text complexity explained $\sim 65\%$ of the variance. All the three notions of complexity were significant predictors of fixation count ($p<0.001$). Proficiency and the interaction of proficiency with complexity were not significant but the random effect due to participants and texts were both significant ($p<0.001$ and $p<0.05$ respectively). Text order too had a significant effect on Fixation Count ($p<0.001$), with the number of fixations increasing with the text id. Table 4.4 summarizes the GAMM model with binary complexity, in terms of the parametric coefficients and the significant smooth terms.

As seen in Table 4.4, fixation counts decrease by 2.35, as one moves from

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	10.5682	0.8467	12.48	< 0.001
DifficultyEasy	-2.3575	0.4115	-5.73	< 0.001
TextOrder	0.8464	0.2051	4.13	<0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
Participant	Yes	39.40	6.09	< 0.001
Text	Yes	2.13	2.45	0.02
Variance Explained (R^2 adj): 65.2%				

Table 4.4: Summary of the GAMM model for Fixation Count

difficult to easy text. Further, fixation counts seem to increase with the number of texts the participant read i.e., the participants appear to fixate more on the texts they read later, compared to those that they read in the beginning. With respect to our hypotheses, we conclude that the fixation count is strongly influenced by text complexity.

Average Fixation Duration

In linear models, only Proficiency had a significant effect ($p<0.001$) on average fixation duration, with higher proficiencies resulting in lower fixation durations. The interaction between proficiency and complexity (in all three notions) was not significant. All the linear models explained about 16-17% of the variance.

The effect of text complexity on average fixation duration was not significant in the GAMM models as well, for all the three notions of complexity. Proficiency ($p< 0.05$) and Text Order ($p<0.001$) were significant. Participants had more average fixation duration for the later texts than those that they read in the beginning. Random effects due to participant and text variation were both significant ($p<0.001$ for both). The interaction between text complexity and proficiency was not significant. The GAMM model explained 74% of the variance. Table 4.5 summarizes the GAMM model with binary complexity, in terms of the parametric coefficients and the significance of smooth terms.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	146.863	10.1432	14.479	< 0.001
DifficultyEasy	0.323	4.0968	0.079	0.937
TextOrder	9.0981	2.1122	4.307	<0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
Proficiency	No	2.031	3.121	0.044
Participant	Yes	39.64	7.43	< 0.001
Text	Yes	2.63	5.80	< 0.001
Variance Explained (R^2 adj): 74%				

Table 4.5: Summary of the GAMM model for Average Fixation Duration

The model summary quantitatively shows the effect of text order on Average Fixation Duration and indicates the non-linear nature of proficiency (edf >1). To conclude, while we originally hypothesized that average fixation duration is affected by text complexity, our results show that the it is affected only by the language proficiency of the participant and not text complexity.

First Fixation Duration

Neither difficulty nor proficiency had any significant effect on first pass duration, for linear models. With GAMMs, although none of them had any significant effect, the random variation due to participants was significant ($p < 0.05$).

First Pass Duration

Neither difficulty nor proficiency had any significant effect on first pass duration, for linear models. However, with GAMMs, while text complexity was not a significant predictor, the effect of proficiency ($p < 0.01$) and the interaction effects between proficiency and text complexity were both significant. The random effect due to variation among participants was also significant ($p<0.001$) in this case. There were no differences between the different notions of text complexity and all of them were not significant predictors of first pass duration. The GAMM model with binary complexity explained 51.2% of the variance. Table 4.6 sum-

marizes this model in terms of the parametric coefficients and the significance of smooth terms.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	691.23	69.84	9.897	< 0.001
DifficultyEasy	-27.00	43.12	-0.626	0.532
TextOrder	17.3	19.20	0.901	0.369
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
Proficiency	No	0.79	9.598	0.0054
Proficiency:DifficultyDifficult	No	0.74	8.7	0.01
Proficiency:DifficultyEasy	No	6.63	2.75	0.009
Participant	Yes	34.8	3.356	< 0.001
Variance Explained (R^2 adj): 51.2%				

Table 4.6: Summary of the GAMM model for First Pass Duration

By the evidence from this GAMM model, we can conclude that the first pass duration is affected by proficiency and its interaction with text complexity.

Second Pass Duration

In linear models, binary text complexity had a significant effect ($p < 0.001$) on second pass duration, with increased durations for difficult texts compared to easy texts. The effect of proficiency was also significant ($p < 0.01$) and the second pass durations decreased with an increase in proficiency. Surprisal ($p<0.001$) and VM ($p<0.01$) too had significant effect on second pass duration, with the direction being the same as binary text complexity i.e., increased values of surprisal and VM resulted in increased second pass reading time. In all the cases, the interaction between proficiency and complexity was not significant. The models explained 11-13% of the variance with linear regression.

In GAMM models, while the effect of text difficulty was significant in all three notions of complexity ($p < 0.001$), the effect or proficiency and the interaction between proficiency and text difficulty were not significant. Compared to the 13% of variance explained by the linear model, all the three GAMM models

explained ∼63% of the variance, with much lower AIC. Table 4.7 summarizes the coefficients of the GAMM model with binary text complexity.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	1747.16	167.35	10.44	< 0.001
DifficultyEasy	-597.07	94.85	-6.295	< 0.001
TextOrder	115.43	42.41	2.722	0.007
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
Participant	Yes	38.54	4.45	< 0.001
Variance Explained (R^2 adj): 63%				

Table 4.7: Summary of the GAMM model for Second Pass Duration

As it can be observed from the model summary in Table 4.7, the second pass duration reduces (by 597 ms) as one moves from difficult texts to easy texts. In this model, the proficiency of the participant does not seem to have any effect while the random variation due to participants is significant. This difference between the linear model and GAMM for proficiency may be attributed to the fact that the random effects take into account the differences due to proficiency along with other causes of variation between participants. In terms of the research hypotheses, we can conclude based on the evidence from this GAMM model that the text difficulty is a significant predictor of second pass reading time.

Revisits

In linear models, Proficiency had a significant effect on the number of Revisits ($p<0.01$) with the number of revisits decreasing with increasing proficiency. The binary notion of complexity and Surprisal were both significant ($p<0.05$) and the number of revisits increased with increased difficulty, as expected. There was no interaction between proficiency and text difficulty in any of the cases and the best model, with binary complexity, explained ∼10% of the variance.

For Revisits, GAMM models showed a significant effect of text difficulty and text order ($p < 0.001$). Proficiency did not have a significant effect on the number of revisits. But both the random effects due to participants ($p < 0.001$) and texts

($p < 0.01$) are significant. There were no differences among the three notions of complexity and all of them explained ∼75% of the variance while maintaining a low AIC compared to the linear model. The model with binary complexity explained 74.9% of the variance. Table 4.8 shows the model summary for this model.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	2.583	0.3888	6.643	< 0.001
DifficultyEasy	-0.7618	0.1522	-5.005	< 0.001
TextOrder	0.4674	0.0782	5.97	<0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
Participant	Yes	41.99	10.36	< 0.001
Text	Yes	2.59	4.862	0.001
Variance Explained (R^2 adj): 74.9%				

Table 4.8: Summary of the GAMM model for Revisits

In this model, as in the case of other variables like fixation count, the number of revisits decreases between difficult to easy texts. The participants also have slightly higher revisits for later texts than earlier ones. To summarize, considering the initial hypotheses, revisits are significantly affected by text complexity but not proficiency and there was no interaction between proficiency and text difficulty.

4.5.2 Outcome Variables

Recall Measure

In linear models, Proficiency had a significant effect on the number of correctly answered recall questions ($p<0.001$) with the increase in proficiency resulting in increased scores. The binary notion of complexity ($p<0.05$), Surprisal ($p<0.05$) and VM ($p<0.05$) were all significant and the recall scores decreased with increased difficulty, as expected. There was no interaction between proficiency and text difficulty in any of the cases and the best model, with binary complexity, explained 36.9% of the variance.

With GAMMs, there was a significant effect of all the fixed and random effect variables, along with an interaction effect between Proficiency and Text Complexity. There were no significant differences between different notions of text complexity and all the three models explained $\sim$ 55–57% of the variance. Table 4.9 shows the model summary for the model with binary complexity.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	3.453	0.371	9.29	< 0.001
DifficultyEasy	0.648	0.1971	3.285	0.001
TextOrder	0.297	0.099	2.989	0.003
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
Proficiency	No	0.667	59.884	< 0.001
Proficiency:DifficultyDifficult	No	1.028	13.318	< 0.001
Proficiency:DifficultyEasy	No	0.667	6.096	0.04
Participant	Yes	27.63	1.505	< 0.001
Text	Yes	2.39	4.022	0.002
Variance Explained (R^2 adj): 55.6%				

Table 4.9: Summary of the GAMM model for Recall Scores

The effect of text complexity on recall reflects what was expected in the hypotheses, that the recall scores increase with the reduction of text difficulty. As in earlier cases, text order had a significant effect on recall scores. However, the direction of the effect is unexpected compared to what we observed for the eye-tracking measures. For eye-tracking measures, the increase of a measure with text order indicates that the participants are taking longer to read the later texts compared to first text. However, for recall scores, this would mean that the users are performing (slightly) better as they read more texts. This could mean that the users are getting familiar with the task (of answering recall questions).

It is interesting to note that both proficiency and its interaction with difficulty are significant predictors of recall scores. Figure 4.1 demonstrates the interaction between text complexity and language proficiency for recall performance in the form of a contour plot, where the lighter colors indicate higher recall scores.

In this figure, we can notice that the relation between proficiency and recall scores is nearly linear although it is slightly different for difficult texts compared to easy texts. On an average, less proficiency users have more difficulties with

Figure 4.1: Interaction between Proficiency and Text complexity for Recall

difficult texts compared to easy texts, as it can be seen by the difference in recall scores between difficult and easy texts. However, as proficiency increases, this difference decreases, indicating that the high proficiency readers relatively less affected by text difficulty. So, we can conclude that both text complexity and participant's language proficiency have a significant effect on the recall performance of the participants and there is a strong interaction between them.

Comprehension Measure

With linear models, comprehension performance was significantly affected only by the proficiency of the user ($p<0.001$) and there was no interaction with text complexity. The linear model explained 18.2% of the variance.

The GAMM models also did not show any significant effect of complexity (in all three notions) and proficiency was a significant predictor of comprehension score ($p < 0.001$). The random variation due to texts was also significant ($p < 0.01$). This model explained 27.6% of the variance and is summarized in Table 4.10.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	3.951	0.279	14.12	< 0.001
DifficultyEasy	0.039	0.154	0.255	0.799
TextOrder	0.108	0.077	1.401	0.163
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
Proficiency	No	1.313	10.051	< 0.001
Text	Yes	2.39	4.351	0.001
Variance Explained (R^2 adj): 27.6%				

Table 4.10: Summary of the GAMM model for Comprehension Scores

Though we hypothesized that comprehension scores are affected by text complexity, we can now conclude that it depends only on the proficiency of the participant and not on the reading level of the text, without any significant interaction between them.

4.5.3 Mediation Analysis

The experiments discussed above demonstrate that some of the eye-tracking measures are impacted by text complexity and all except first fixation duration were affected by the language proficiency of the reader. We also found that one of the outcome variables, recall, was influenced by both text complexity and readers' language proficiency while only the latter affected the comprehension scores.

Given this background, we explored whether the effect of text complexity and proficiency on online processing can be used to explain the differences in the learning outcomes of the participants. So, we did mediation analysis considering the eye-tracking measures as mediator variables and the outcome measures as dependent variables. We explored two paths of mediation models with text complexity and language proficiency as the independent variables respectively.

To perform the mediation analyses, we need to ensure that the relationship between the eye-tracking measures and outcome measure is statistically significant. As only average fixation duration showed a significant effect, we did not perform the analyses with other eye-tracking measures. There was no mediation effect of average fixation duration on the outcome variables when both proficiency and difficulty are considered as the independent variables.

4.6 Discussion

The results from the above analyses confirm that the eye-movement patterns of the readers are sensitive to the complexity of the text they are reading, as was seen by increased fixation counts, second pass reading time and revisits with increased text complexity. Average fixation duration and first pass reading time were affected by language proficiency and not text complexity. There was an interaction between proficiency and text complexity only for first pass reading time. While some of these observations concur with previous research, some of them do not.

The Rayner et al. (2006) study concluded that the difficulty rating for the text significantly positively correlated with average fixation duration and number of fixations among the measures we studied. In our study, while text difficulty was a significant predictor of the number of fixations, the effect was not significant for average fixation duration. One explanation for these contrasting results can be attributed to the operationalization of ”text complexity” in both the studies. While the Rayner et al. (2006) study consisted of 32 texts rated independently by 32 students on a scale of 1–10, our study used three definitions of text complexity - which were all non significant predictors of average fixation duration. Further, we considered four pairs of texts written in two versions each based on text complexity whereas Rayner et al. (2006) used texts that do not have this property.

However, it is interesting to note that reader's language proficiency, which was not considered in their study showed a significant effect in our study. With respect to the relation between comprehension scores and text difficulty, like in Rayner et al. (2006) study, our results also showed no statistically significant relation with text difficulty.

Crossley et al. (2014) performed a self-paced reading time study and the conclusions drawn differ in terms of the effect of text complexity on comprehension, compared to our study. Crossley et al. (2014) showed a significant effect of text complexity on comprehension, but it depended on the language proficiency of the participant. In our study, only readers' language proficiency was a significant predictor of comprehension scores, and the model explained 27.6% of variance compared to 16% of the variance explained by their model with both proficiency and complexity as predictors. One possible reason for these differences could be that we did not use same texts for the experiments. This possibility gains weight when we consider the fact that random variation due to texts was significant ($p < 0.01$, Table 4.10)[8].

An important distinction between the present study and others in terms of the analysis lies in the use of text order as an additional fixed effect predictor and studying the presence of participant and the text as random effects. Our results showed that the text order had a significant effect for four of the six eye-tracking measures and with the recall scores. The random effect due to participant variation was significant for all the eye-tracking measures and recall scores. The random effect due to text variation with three of the six eye-tracking measures and both the outcome measures. This shows the value of using mixed effects modeling for analyzing the data in this study. Further, the interaction between proficiency and complexity for some of the variables was not captured by the linear models. But the GAM models showed significant interaction between the two variables for first pass reading time and recall scores. Finally, the variance explained by the GAM models was consistently higher than those of linear models with the best model (for revisits) explaining a variance of 75%. This is an important factor in selecting a model, if it has to be useful to make future predictions. These three observations

[8] As mentioned earlier, the texts we used, in both versions, and the questions we prepared can be shared for research use. An example text is provided with all these details in Appendix A.

from the results justify the choice of GAMMs as an appropriate tool to analyze the data from this experiment and demonstrate the role of potential variation among the dependent variables due to study design and differences between participants and text materials.

In addition to the results reported here, three-way interaction between text complexity, proficiency and text order, which was not considered in our hypotheses, turned out to be significant for all the measures except average fixation duration and comprehension score. For most of the cases, removal of outliers and transforming the data resulted in significantly better models. A discussion of the 3-way interactions and model comparisons can be seen in Appendix A.

4.6.1 Conclusions

To summarize, in this chapter, we described an experiment to evaluate the effect of reading easy versus difficult to read texts on the cognitive processing and performance outcomes of non-native English speakers. We used a collection of manually simplified texts, each in two versions, compiled from an external source onestopenglish.com. We did an eye-tracking study and also obtained scores for recall and comprehension questions on the texts read by the participants, in both versions. We analyzed the data using linear models and using GAMMs with fixed and random effects factors. We also investigated if there is a possible mediation effect of the processing measures on performance outcomes. The primary conclusions from this experiment are summarized as follows:

1. Among the processing measures we studied:

    - Fixation count, second pass duration and the revisits counts were significantly affected by text complexity.
    - Average fixation duration and first pass reading time were affected only by the readers' language proficiency.
    - First fixation duration was not significantly affected by either proficiency or text complexity.
    - There was a significant interaction between text complexity and language proficiency for only first pass reading time.

2. Among the outcome measures,

    - Text complexity was a significant predictor only for recall scores and not the comprehension scores while participant's language proficiency was a significant predictor for both.
    - There was an interaction between text complexity and language proficiency for recall scores.

3. There seemed to be no differences between the three notions of text complexity we used (human, linguistic and psycholinguistic) in terms of the amount of variance explained by them.

4. The order in which the participants read the texts was a significant predictor of four eye-tracking measures (fixation count, average fixation duration, second pass reading time and revisits) and recall score.

5. The random effect of participant variation was significant for all the eye-tracking measures and recall scores. The random effect of text variation was significant for three eye-tracking measures (fixation count, average fixation duration, revisits) and for both the performance outcome measures.

6. There was no mediation of the processing measures on the differences in the performance outcomes.

4.6.2 Outlook

Our results show that text simplification can be effective in improving some of the performance outcomes of the readers, which calls for the construction of efficient automatic text simplification systems for second language learners. While our experiments were performed using manually simplified texts, it would be interesting to check the output of automatic text simplification in a similar experimental setup. Identifying specific linguistic variables that correlate with the processing and performance measures is an interesting direction to pursue, from a linguistic perspective. Another challenging problem is to understand the relation between the nature of the simplification performed (e.g., lexical, syntactic or semantic)

and the dependent variables. Finally, the fact that there were no significant differences between the different notions of complexity is an interesting result and would merit further study. From a modeling perspective, developing models with better fit for the data is a next step[9].

[9] As mentioned earlier, a detailed discussion on constructing and analyzing various models for individual variables and a study of the three way interactions between text order, language proficiency and text complexity in the data can be seen in Appendix C.

Chapter 5

Readability Analysis of Sentences: Motivation, Methods and Applications

Abstract

In this chapter, we study the problem of assessing readability at the sentence level. We first explore the use of the document level readability model from Chapter 3 directly on sentences and later explore sentence level readability as a pairwise ranking approach. Using multiple in-corpus and cross-corpus evaluations, we establish that sentence level readability is better assessed by considering it as a ranking problem. In the process of investigating the problem, we also created a new sentence level readability corpus, which contains each sentence written in three versions based on the reading level of the learners. This corpus can serve as a useful resource for future research on this topic. Finally, we briefly explored the idea of using sentence level readability model as a means to provide more fine-grained readability judgments at the document level.

Some of the experiments described in this chapter are reported in Vajjala & Meurers (2014a), Vajjala & Meurers (under review).

5.1 Introduction

Text Simplification is the process of simplifying the linguistic form of a text without losing its meaning. In an educational context, the purpose of text simplification generally is to adapt the text complexity to facilitate comprehension by the target audience, such as language learners or students with disabilities. In such contexts, it is important to have a method to evaluate the degree of simplification performed. Further, in order to automate the process of text simplification, or to assist the manual creation of simplified text, it is useful to have an approach that chooses the possible targets for text simplification in a text, rather than simplifying everything possible. Readability assessment at a sentence level is very useful for these two tasks. We describe its use for the first task (i.e., evaluating simplification) in this chapter.

Sentence level readability assessment is a recent area of research and has been studied in the context of Automatic Text Simplification (Napoles & Dredze, 2010; Dell'Orletta et al., 2014), in Computer Assisted Language Learning applications for selecting appropriate sentences for language learning exercises (Segler, 2007; Pilán et al., 2014) and indirectly, for machine translation (Stymne et al., 2013).

Napoles & Dredze (2010) compared English Wikipedia and Simple Wikipedia at document and sentence level in terms of readability. In the absence of aligned pairs of sentences between Wikipedia and Simple Wikipedia, they started with the assumption that all sentences in Simple Wikipedia are simple and vice versa. With this assumption, they achieved a 80% binary classification accuracy for sentences. Dell'Orletta et al. (2014) studied sentence level readability classification for Italian text simplification. In their approach, they start with an assumption that all sentences in a corpus of "easy to read" newspaper texts are easy but all sentences in "difficult" texts need not be difficult. They created a corpus of 1745 manually annotated "difficult" sentences by sampling sentences from the "difficult" texts of the corpus through an experiment with two annotators. Considering several linguistic features, they report sentence level classification accuracy of $\sim$85%. It has to be noted however that both the approaches lack the presence of a parallel, sentence aligned unsimplified-simplified corpus and in the case of Dell'Orletta et al. (2014) the texts are not aligned either.

Zhu et al. (2010) created a parallel corpus of sentences in unsimplified and simplified versions from Wikipedia-Simple Wikipedia for English, which is publicly available. Since in this corpus, there is clear binary distinction between the reading levels of the sentence, we explored sentence level readability starting with this corpus, so that we would not run into the problem about the quality of the reading level annotation of the training corpus.

We explored three learning methods for this task:

1. Consider sentence level readability as a binary classification task (simple versus hard) and build machine learning models with the Wikipedia-Simple Wikipedia sentence aligned corpus (Section 5.2).

2. Consider sentence level readability as a continuum rather than binary and use the document level readability model to assign reading levels to sentences (Section 5.3).

3. Rank the sentences based on their reading level through pair-wise ranking (Section 5.4).

For all the approaches, we used the feature set described in Chapter 3.

While our primary motivation for exploring readability at the sentence level is to apply it for text simplification, another way to look at it is as a more fine-grained modeling of readability. The existing readability corpora have a global readability score per text. However, a difficult text may contain parts that are of varying levels of difficulty with more percentage of difficult parts. Similarly, an easy text can contain some sentences that are easy and some that are difficult. To develop a model that takes this aspect into account, we used the sentence level readability model to get an estimate about textual readability. In other words, we developed an approach to get a global (textual) estimate of readability through the distribution of local (sentence) readability estimates. This experiment about estimating textual readability through sentence level models is described in Section 5.5. Finally, Section 5.6 concludes this chapter with some pointers to future research directions.

5.2 Sentential Readability as Binary Classification

In this approach, our aim is to build a binary classifier that classifies sentences into easy or hard categories. For this purpose, we would ideally need a corpus with labels for individual sentences. So, we used the Wikipedia-Simple Wikipedia sentence aligned corpus.

5.2.1 Wikipedia-Simple Wikipedia corpus

Simple Wikipedia targets students, children, adult language learners and people with reading difficulties[2] so a corpus of Wiki-Simple Wikipedia sentence pairs is suitable for our purpose. We use the sentence aligned corpus created by Zhu et al. (2010). They used a collection of ∼65k parallel articles from Wikipedia and Simple Wikipedia to create a sentence aligned corpus consisting of ∼100k pairs. We used this corpus after removing the sentence pairs that remained unchanged in both versions. This sub-corpus consists of 80,912 sentence pairs.

For each of the pairs in the Wikipedia-Simple Wikipedia Sentence Aligned Corpus introduced above, we labeled the sentence from Wikipedia as *hard* and that from Simple English Wikipedia as *simple*. The corpus thus consisted of single sentences, each labeled either *simple* or *hard*. On this basis, we constructed a binary classification model. Our document level readability model does not include discourse features, so all the features can also be computed for individual sentences. We built a binary sentence level classification model using several classification algorithms implemented in WEKA. However, to maintain compatibility with the experiments reported in the previous chapter, we report results using Sequential Minimal Optimization (SMO) algorithm in this chapter.

A binary classification approach to determine whether a given sentence is *simple* or *hard* was disappointing, reaching only 66% accuracy in a 10-fold cross-validation setting, for a balanced dataset. Experiments with other algorithms did not yield any better results. To study how the classification performance is impacted by the size of the training data, we experimented with different training set sizes. Figure 5.1 shows the classification accuracy with different data sizes.

[2] `http://simple.wikipedia.org/wiki/Simple_English_Wikipedia`

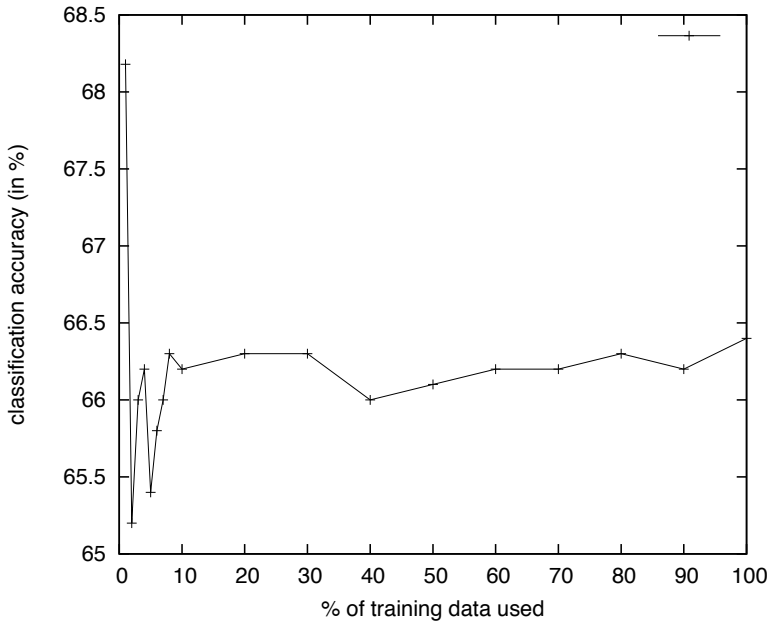

Figure 5.1: Training size vs. classification accuracy

The graph shows that beyond 10% of the training data, more training data did not result in significant differences in classification accuracy. Even at 10%, the training set contained around 10K instances per category, so the variability of any of the patterns distinguished by our features is sufficiently represented.

We also explored whether feature selection could be useful. A subset of features chosen by removing correlated features using the CfsSubsetEval method in WEKA did not improve the results, yielding an accuracy of 65.8%. A simple baseline based on the sentence length as single feature resulted in an accuracy of 60.5%, underscoring the limited value of the rich feature set in this binary classification setup.

For the sake of a direct comparison with the document level model, we also explored modeling the task as a regression on a 1–2 scale. In comparison to the document level model, which as discussed in Chapter 3 had a correlation of 0.9, the sentence level model achieves only a correlation of 0.4. A direct comparison is also possible when we train the document level model as a five-class classifier

with SMO. This model achieved a classification accuracy of ∼90% on the documents, compared to the 66% accuracy of the sentence level model classifying sentences. So under each of these perspectives, the sentence level models trained using sentence level readability data are much less successful than the document level models on the document task.

But does this indicate that it is not possible to accurately identify the reading level distinctions between simplified and unsimplified versions at the sentence level? Is there not enough information available when considering a single sentence? To answer this question, we explored the hypothesis that there is a distribution of sentences belonging to various reading levels in both easy and difficult to read texts, and hence, we cannot consider all sentences in "easy to read" texts as easy and "difficult" texts as difficult.

5.3 Using Document Level model on sentences

To verify this hypothesis, we applied the document level readability model from Chapter 3 to the Wikipedia-Simple Wikipedia corpus sentence pairs. On one hand, this will help us verify our hypothesis. On the other hand, it will also answer our question about the accuracy of the features in identifying distinctions in the reading levels of sentences. Figure 5.2 shows the distribution of Wikipedia and Simple Wikipedia sentences according to the predictions of our document level readability model trained on the WeeBit corpus (Chap 3). As we are using a regression model, the values sometimes go beyond the training corpus' scale of 1–5. For ease of comparison, we rounded off the reading levels to the five level scale, i.e., 1 means 1 or below, and 5 means 5 or above.

As seen in the Figure 5.2, our model could identify sentences belonging to all the reading levels on its scale. It also determines that a high percentage of the Simple Wikipedia sentences belong to lower reading levels, with over 45% at the lowest reading level; yet there also are some Simple Wikipedia sentences which are assigned the highest readability level. In contrast, the regular Wikipedia sentences are evenly distributed across all reading levels. From this distribution, we hypothesized that the nature of the simplification is relative and not absolute. That is, while for each pair of the Wikipedia-Simple Wikipedia sentence aligned cor-

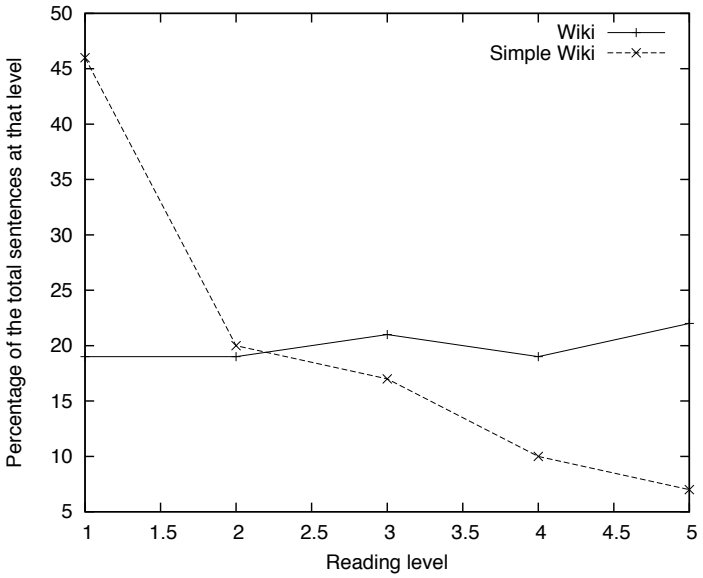

Figure 5.2: Reading levels of Wikipedia and Simple Wikipedia sentences

pus we used, the Wiki sentence was harder than the Simple Wikipedia sentence, it does not necessarily mean that each of the Wikipedia sentences is harder than each of the Simple Wikipedia sentences. For example, consider two (hard, easy) sentence pairs with the levels $(2, 1)$ and $(5, 3)$ respectively.

In that case, the low accuracy of the binary classifier may thus simply result from the inappropriate assumption of an absolute, binary classification viewing each of the sentences originating from Simple Wikipedia as simple and each from the regular Wiki as hard. The confusion matrices of the binary classification too suggested some support for this hypothesis, as more *simple* sentences were classified as *hard* compared to the other way around. This can occur when a *simple* sentence is simpler than its *hard* version, but could actually be simplified further – and as such may still be harder than another unsimplified sentence. The hypothesis thus amounts to saying that the two-class classification model mistakenly turned the relative difference between the sentence pairs into a global classification of individual sentences, independent of the pairs they occur in.

We used the readability scores assigned to the sentences by the document level readability model, to determine for how many pairs the relative reading levels of

the sentences are identified correctly by the model. In other words, we calculated the percentage of pairs (S, N) in which the reading level of a simplified sentence (S) is identified as less than, equal to, or greater than the unsimplified (normal) version of the sentence (N), i.e., $S < N$, $S = N$, and $S > N$. Where simplification split a sentence into multiple sentences, we computed S as the average reading level of the split sentences.

Given the regression model setup, we can consider how big the difference between two reading levels determined by the model should be in order for us to interpret it as a categorical difference in reading level. Let us call this discriminating reading level difference the *d-level*. For example, with $d = 0.3$, a sentence pair determined to be at levels $(3.4, 3.2)$ would be considered a case of $S = N$, whereas $(3.4, 3.7)$ would be an instance of $S < N$. The d-value can be understood as a measure of how fine grained the model is in identifying reading level differences between sentences. If we consider the percentage of samples identified as $S <= N$ as an accuracy measure, Figure 5.3 shows the accuracy for different d-values.

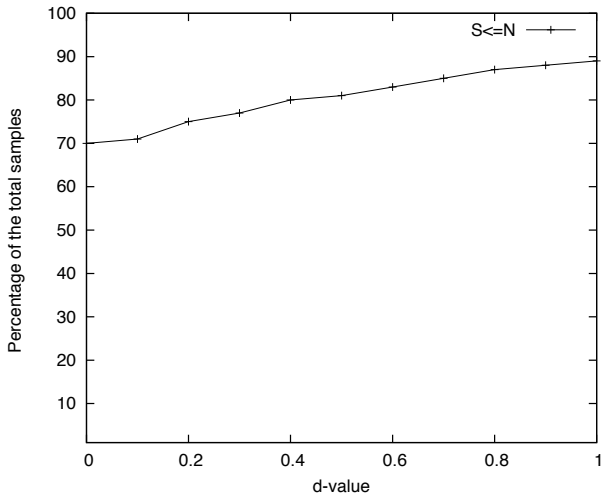

Figure 5.3: Accurately identified $S <= N$

We can observe that the percentage of instances that the model correctly identifies as $S <= N$ steadily increases from 70% to 90% as *d* increases. While the value of

d in theory can be anything, values beyond 1 are uninteresting in the context of this study. At $d=1$, most of the sentence pairs already belong to $S=N$, so increasing d beyond 1 would defeat the purpose of identifying reading level differences. The higher the d-value, the more of the simplified and unsimplified pairs are lumped together as indistinguishable. Spelling out the different cases from Figure 5.3, the number of pairs identified correctly ($S<N$), equated ($S=N$), and misclassified ($S>N$) as a function of the d-value is shown in Figure 5.4.

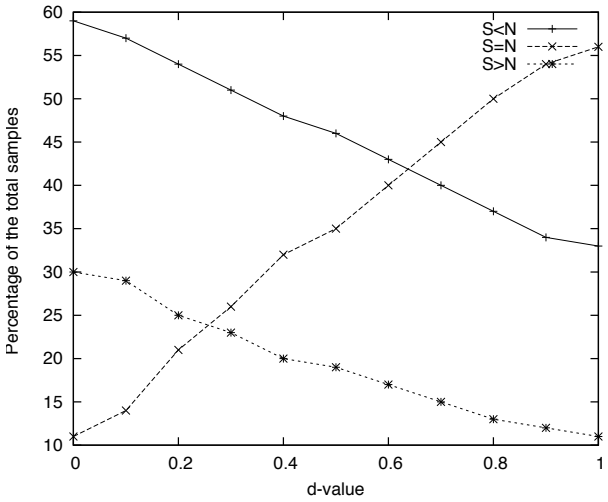

Figure 5.4: Model accuracy by d-value

At $d=0.4$, around 50% of the pairs are correctly classified, 20% are misclassified, and 30% equated. At $d=0.7$, the rate of pairs for which no distinction can be determined already rises above 50%. For d-values between 0.3 and 0.6, the percentage of correctly identified pairs exceeds the percentage of equated pairs, which in turn exceeds the percentage of misclassified pairs.

5.3.1 Influence of reading level on accuracy

We saw in Figure 5.2 that the Wikipedia sentences are uniformly distributed across the reading levels, and for each of these sentences, a human simplified version is included in the corpus. Even sentences identified by our readability model as belonging to the lower reading levels thus were further simplified. This leads us

to investigate whether the reading level of the unsimplified sentence influences the ability of our model to correctly identify the simplification relationship. To investigate this, we separately analyzed pairs where the unsimplified sentences had a higher reading level and those where it had a lower reading level, taking the middle of the scale (2.5) as the cut-off point. Figure 5.5 shows the accuracies obtained when distinguishing unsimplified sentences of two levels.

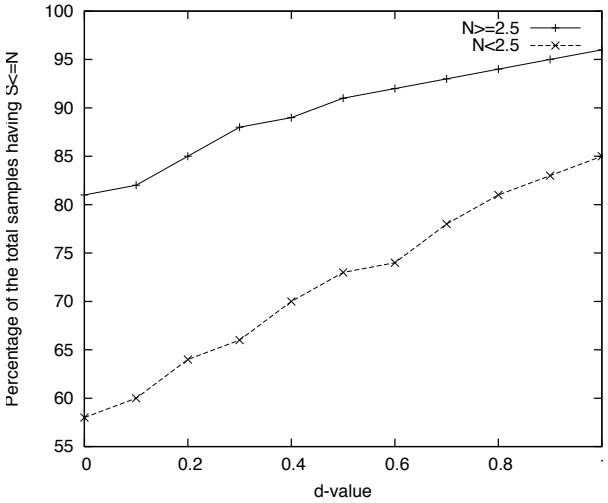

Figure 5.5: Accuracy ($S <= N$) for different N types

For the pairs where the reading level of the unsimplified version is high, the accuracy of the readability model is high (80–95%). Presumably the complex sentences for which the model performs best offer more syntactic and lexical material informing the features used. In the other case, the accuracy drops to 65–75% (for $0.3 <= d <= 0.6$). Figure 5.6 and Figure 5.7 graphically demonstrate the difference, by splitting Figure 5.5 in to three cases again ($S < N$, $S = N$, $S > N$).

While the pairs with a high-level unsimplified sentence in Figure 5.6 follow the pattern in Figure 5.4, the results in Figure 5.7 for the pairs with an unsimplified sentence at a low readability level establish that the model essentially is incapable to identify readability differences.

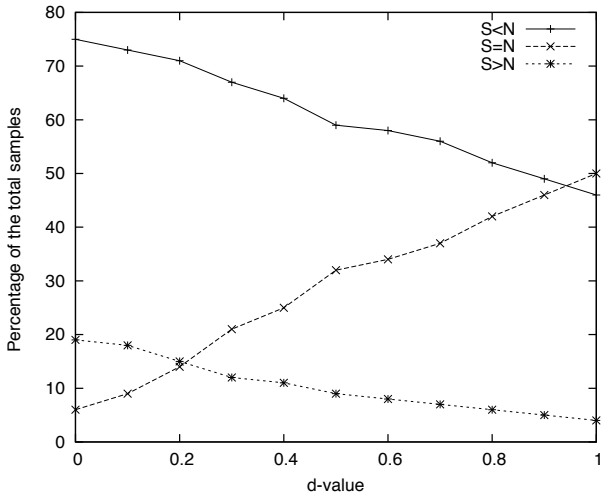

Figure 5.6: Results for $N >= 2.5$

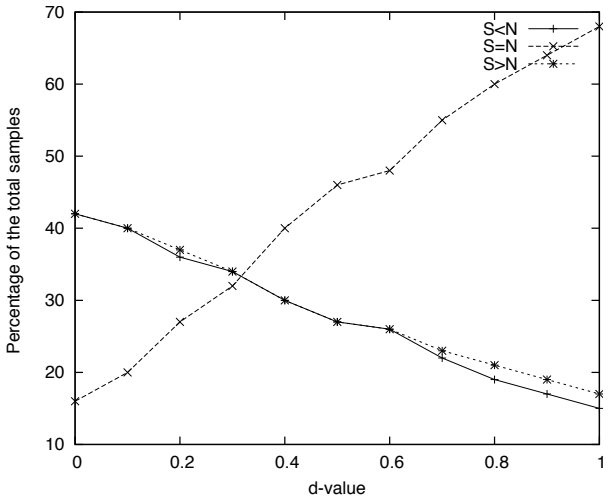

Figure 5.7: Results for $N < 2.5$

Summing up, the experiments discussed in this section show that a document level readability model trained on the WeeBit corpus can provide insightful perspectives on the nature of simplification at the sentence level. The results emphasize the relative nature of readability and the need for more features capable of identifying characteristics that can distinguish sentences at lower levels.

5.4 Comparison through Pair-wise ranking

We extended the notion of the relative nature of simplification further and developed a pair-wise ranking based approach which directly captures the idea of relative levels of readability. This significantly outperformed the previous approach to compare sentential reading levels. We modeled sentential complexity as a pair-wise ranking problem. Pair-wise ranking is one of the methods to rank data instances based on some parameter. "Learning to rank" (Li, 2014) methods are typically used in Information Retrieval (IR) for ranking search results based on their relevance. The objective of these methods is to learn to rank a set of data instances. In IR, it is used to compare a pair of documents in terms of their relevance to a given query. In our case, given a pair of sentences where one is the simplified version of the other, the aim of the ranker is to predict which one of them is simpler than the other.

Thus, the learning problem for us is to compare versions of a sentence and rank them based on their reading difficulty, while trying to minimize inversion of ranks in the process. That is, the aim of this approach is to put the difficult sentence at a higher rank than its easier version as much as possible. For example, for a given sentence s_i, let us say there are three versions - x_{i1}, x_{i2}, x_{i3}. Now, if x_{i1} is more complex than x_{i2} and x_{i3}, then, $x_{i1} > x_{i2}$ and $x_{i1} > x_{i3}$ become pairs. These preference pairs are viewed as instances for the classifier in a pair-wise ranking problem. It has to be noted that in this setup, each simple sentence is assumed only to be simpler than its unsimplified version. No assumptions are made about its simplicity or difficulty with respect to other sentences.

5.4.1 Algorithms

We first compared three pair-wise ranking algorithms for our initial models and chose the best performing algorithm for the rest of our experiments. The three algorithms we explored are:

- **RankSVM** (Herbrich, Graepel & Obermayer, 2000; Joachims, 2002): uses a Support Vector Machine for learning to perform pair-wise classification. It is one of the earliest and most commonly applied ranking algorithms. It

is also popular for NLP tasks and was used for ranking children's literature texts based on their reading level in the past (Ma et al., 2012a).

- **RankNet:** (Burges, Shaked, Renshaw, Lazier, Deeds, Hamilton & Hullender, 2005): is a pair-wise ranking algorithm that is a modified version of the traditional back-propagation based neural network, applied to ranking problems. Thus, instead of updating the parameters iteratively for each instance, the update is done for each preference pair. It is known to perform well for practical use and was successfully used in a real-life search engine to rank search results and was not applied for readability assessment before.

- **RankBoost** (Freund, Iyer, Schapire, & Singer, 2003) is an algorithm that uses boosting for pair-wise ranking. It uses a linear combination of several weak rankers to produce the final ranking. The algorithm is typically applied in collaborative filtering problems and to our knowledge, was not used in this context before.

Apart from the learning methods, the algorithms also differ in terms of their loss functions. RankSVM, RankNet and RankBoost have hinge loss, exponential loss and logistic loss functions respectively. We used publicly available implementations of these algorithms for training our models - SVM^{rank} Joachims (2006)[3] for Ranking SVM and RankLib[4] software for RankNet and RankBoost.

Evaluation: Since our learning goal is to minimize the number of wrongly ranked pairs, we measure the efficiency of the approach in terms of the percentage of correctly ordered pairs i.e., the percentage of pairs in which the difficult version gets a higher rank than its simplified counterpart. We refer to this as accuracy in the rest of this section.

Since the learning process depends on the pair-wise constraints generated per sentence instance, the number of simplified versions of a sentence in the corpus too influences the learning process. That is, if a sentence s_i has two versions S_1 and S_2 where S_1 is the difficult version, there is only one pair-wise constraint in

[3] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html
[4] http://sourceforge.net/p/lemur/wiki/RankLib/

this case - $S_1 > S_2$. However, if the sentence s_i had three rewritten versions instead of 2, $\{S_1, S_2, S_3\}$ in the decreasing order of difficulty, then the constraints generated are: $S_1 > S_2, S_1 > S_3, S_2 > S_3$. Thus, there are more constraints to learn from, per instance, for the rankers. To understand the effect of the levels of simplification on the pair-wise ranking accuracy, we need a corpus of parallel texts with more than two levels of simplification. For this, we created a sentence simplification corpus from OneStopEnglish.com website.

5.4.2 OneStopEnglish corpus

OneStopEnglish (OSE) is an English teachers' resource website published by the Macmillan Education Group. They publish Weekly News Lessons[5] which consist of news articles sourced from the newspaper *The Guardian*. The articles are rewritten by teaching experts in a way targeting English language learners at three reading levels (elementary, intermediate, advanced). We acquired permission from OSE to use the articles for research purposes and downloaded the weekly lessons from September 2012–March 2014, which resulted in a collection of 76 article triplets. Each article is included with an elementary, an intermediate, and an advanced version so that overall the corpus contains 228 articles.

Corpus pre-processing: The weekly lessons are pdf files consisting of a pre-test about the topic of the article, the re-written news article, and exercises related to the article. We first parsed the pdfs using iTextPDF[6] to extract the article text, excluding everything else. Since our aim is to compare different versions of a sentence, we took each article triplet and sentence-aligned two at a time using TF-IDF and cosine similarity, following previous research on monolingual sentence alignment (Nelken & Shieber, 2006; Zhu et al., 2010).

We created two versions of the corpus:

OSE3: For the sentences which exist in all three versions of an article (elementary, intermediate, advanced), we obtain a triplet of sentences. We selected all

[5] http://www.onestopenglish.com/skills/news-lessons/weekly-topical-news-lessons

[6] http://itextpdf.com

triplets for which each pair of sentences was above a minimum similarity threshold of 0.7 (based on manual qualitative analysis using different thresholds). Overall, we identified 837 sentence triplets and call this corpus resource OSE3[7]. An example of a sentence that was rewritten across the three levels is shown below:

Adv: *In Beijing, mourners and admirers made their way to lay flowers and light candles at the Apple Store.*

Int: *In Beijing, mourners and admirers came to lay flowers and light candles at the Apple Store.*

Ele: *In Beijing, people went to the Apple Store with flowers and candles.*

OSE2 We also compiled an additional two-level corpus consisting of pairs of sentences from this data. For this, we extracted the 4575 pairs of sentences that were above the minimum similarity threshold. After removing the sentences that remained unchanged in both versions, we were left with 3113 sentence pairs, where one is the simplified version of the other. We will refer to this corpus as OSE2.

Sentence labels: To apply ranking, we need to have a numeric score for the feature vectors of sentences. So, we gave a score of 2 to the more difficult version and 1 to its simplified version in the sentence pair. In the case of a triplet, we gave the sentences scores of 3,2,1 for advanced, intermediate and elementary levels respectively. We used the same notation for the WIKIPEDIA-SIMPLE WIKIPEDIA sentence level corpus as well, and used that too in these experiments. In the case of sentences that were split into two in the simplified version, we scored both the simple sentences as 1 so that pair-wise constraints will not be generated between them. Since pair-wise ranking only considers relative ranks, the ranking procedure is not dependent on the absolute reading levels of sentences (which is fortunate since no such readability level gold-standard annotation exists for the sentences in the two corpora used).

[7]The sentence-aligned corpus is available for research use.

Baseline: As some text simplification approaches used Flesch-Kincaid Grade Level (FKGL) as a readability measure for text simplification (e.g.,Woodsend & Lapata (2011a)) we consider it as our baseline in this approach.

5.4.3 Comparison with Regression

We started with an experiment directly comparing the ranking approach with the results reported in Section 5.3. So we trained a ranking model on the entire Wiki dataset with a 10-fold cross validation (CV) setup, using SVM^{Rank}. The ranking model achieved an accuracy of 82.7%. The standard deviation between the ten folds was 8.4%. This high level of variability may indicate that the nature of what constitutes simplifications in the Simple Wikipedia varies significantly, as may be expected for a collaborative editing setup –a potentially interesting issue to explore in the future.

However the 82.7% accuracy achieved by the ranking model is a significant increase over the results achieved in Section 5.3, where we used a document level regression model directly on sentences. While that model predicted the order correctly in 59% of the cases, in 11% of the cases, it gave the same score to both sentences in the pair. So assuming that the 11% are randomly assigned to easy or difficult, we get an accuracy of 64.5% for this model (59+5.5). In comparison, using FKGL instead of training a model achieved a ranking accuracy of 72.3%. While the ranking approach clearly outperforms the regression setup and also achieves a 10% improvement over the FKGL baseline for this task, we were interested in further exploring the problem in terms of a comparison between ranking approaches and the generalizability of the ranking approach in cross-corpus and multi-level simplification scenarios.

5.4.4 Comparison between ranking algorithms and corpora

We continued with training more ranking models using WIKI and OSE2 corpora. To make the results comparable for these two corpora, for each of the training sets WIKI-TRAIN and OSE2-TRAIN we selected 2000 sentence pairs, and the remaining part was used as the test set (WIKI-TEST: 78912 pairs, OSE2-TEST:

1113 pairs). We first compared the three ranking algorithms mentioned in Section 5.4.1 in terms of the percentage of correctly ranked pairs. Table 5.1 shows the performance of three ranking algorithms using two training sets with within and cross-corpus evaluation setups. The base-lines for WIKI-TEST and OSE2-TEST, obtained using FKGL formula are 69% and 69.6% respectively.

Test set	SVM^{rank}	RankNet	RankBoost
Training: WIKI-TRAIN			
WIKI-TEST	**81.8%**	72.5%	76.4%
OSE2-TEST	74.6%	59.1%	70.2%
Training: OSE2-TRAIN			
WIKI-TEST	77.5%	73.8%	74.8%
OSE2-TEST	**81.5%**	69%	75.5%

Table 5.1: Performance of ranking algorithms

The table shows that SVM^{rank} performed the best among the ranking algorithms we tried. In the following, we therefore only report results with this algorithm. The second observation is that there always was a drop in performance for cross-corpus evaluation. The drop is smaller for the model trained on the OSE2 corpus, which suggests that the OSE2 corpus covers a more representative, broader range of simplifications. Taking that idea further, we explored improving cross-corpus performance using two methods enriching the training data.

5.4.5 Improving cross-corpus performance

First, we combined the two training sets to create a new, hybrid training set WIKI-OSE2-TRAIN, which should increase the representativeness and range of the simplifications included in the training data. Second, we used the three level corpus OSE3 to train the ranker to simultaneously considering a broader range of the simplifications (given that a ranker will learn a single set of weights for ranking the three pairs in a set for OSE3, instead of three sets of weights for ranking each pair independently). Since the OSE3 corpus had only 837 sentence triplets in total, we assign 750 of them to the training set (OSE3-TRAIN) and the remaining ones as the test set OSE3-TEST. The baseline performance for OSE3-TEST using FKGL was 71.3%. Table 5.2 shows the results for the three test sets for models

trained on the combined WIKI-OSE2-TRAIN and the OSE3-TRAIN training sets.

Test set	Training Set	
	WIKI-OSE2-TRAIN	OSE3-TRAIN
WIKI-TEST	81.3%	78.6%
OSE2-TEST	80.7%	82.4%
OSE3-TEST	79.7%	79.7%

Table 5.2: Performance with WIKI-OSE2-TRAIN and OSE3 corpora

As expected, the accuracy for the combined, more varied training set results in a comparable performance across the three tests sets. The results for the OSE3-TRAIN training set providing the ranker triples over which to learn the weights are less clear. Since one combines multiple data sources and the other has multiple levels of text simplification in it, these datasets are perhaps capturing more diverse simplification options than the other two datasets taken independently. The performance of both the models with OSE3-TEST test set differs in terms of individual instances (as a brief manual inspection showed), but the overall accuracy did not change for both the models. The fact that all results, cross-corpus and same-corpus are considerably close together supports the assumption that reliable sentence level readability ranking models, which generalize across very different data sets, can be built.

5.4.6 Influence of Training set size

Considering the relatively low cross-corpus accuracy on OSE2-TEST of the model trained on WIKI-TRAIN (74.6%) as seen in Table 5.1, we wondered whether increasing the amount of training data would improve the results. We therefore trained on increasingly larger portions of the Wikipedia-Simple Wikipedia data set up to the full set of 80k pairs and tested it with both OSE2-TEST and OSE3-TEST datasets. Figure 5.8 shows the accuracy on the two test-sets for increasing training set size.

The accuracy curve is essentially flat, with the model on the largest training set reaching 76.3%, less than two percent above the result for the model using only 2k pairs for training. The Wikipedia-Simple Wikipedia data set thus does not seem to offer the machine learner the kind of variety of simplifications needed

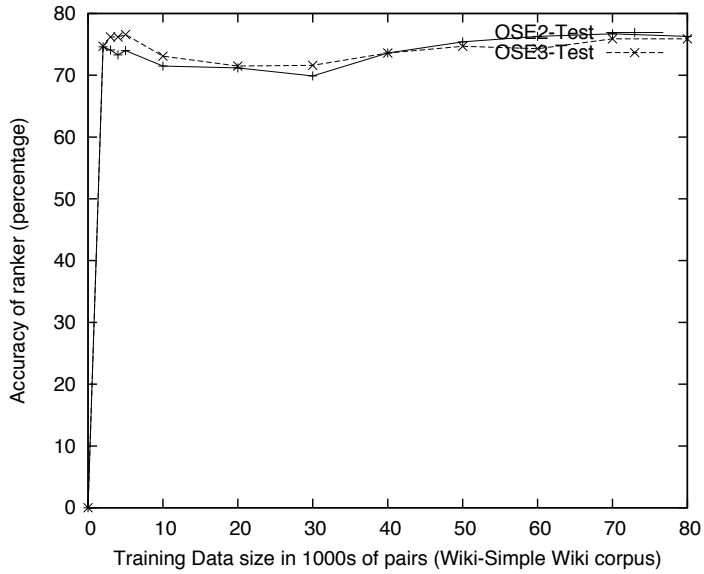

Figure 5.8: Training set size vs. accuracy

to generalize better to the OSE2-TEST and OSE3-TEST sets. The results for any training set size remain below those of the model trained on the combined WIKI-OSE2-TRAIN data set.

From these experiments, we can conclude that a ranking approach is useful for comparing simplified and unsimplified sentences in terms of readability. The approach performs with an accuracy of over 80% in several train-test setups. These results support that such an approach can meaningfully be used for evaluating text simplification.

5.4.7 Feature Selection

Since the above experiments established the validity of the approach, we now turn to feature selection. Apart from giving us an understanding about how much can we achieve with how less, this also gives us an opportunity to understand the linguistic properties of simplification better. In feature selection, we first investigated the contribution of different feature groups to ranking accuracy. Figure 5.9 shows the performance of the ranking models trained using four feature groups

and a model trained with all the features, using WIKIOSE-TRAIN dataset, which generalized well across test-sets.

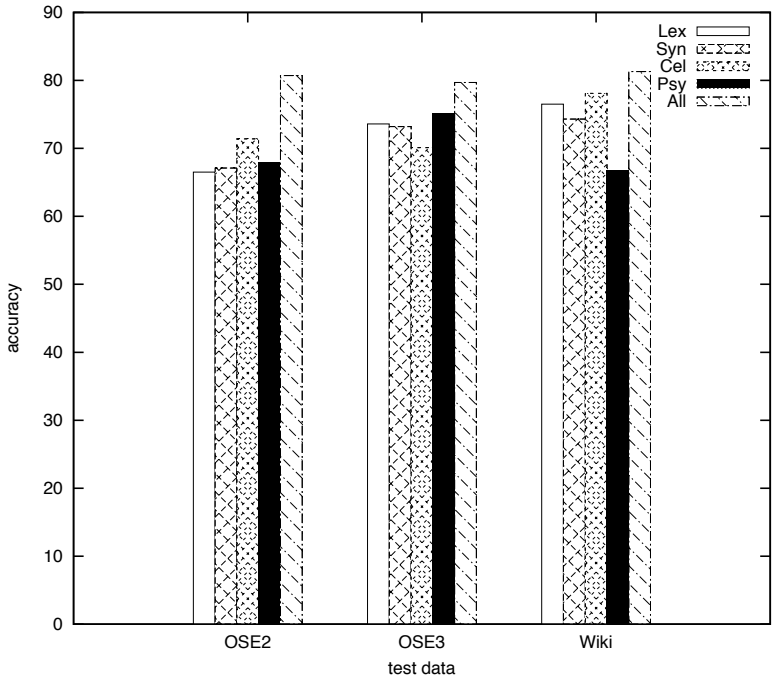

Figure 5.9: Performance of different feature groups

The performance of these feature groups seem to vary with the test-sets used. For example, CEL features seemed to perform poorly compared to other groups for OSE3-TEST whereas PSY features performed poorly for Wiki. However, in general, the model with all the features performed better than models with sub-groups of features in all the cases. From this observation, we can conclude that modeling multiple dimensions of readability could be more useful than choosing a single aspect for a generalizable approach to evaluate text simplification.

Impact of individual features To understand the linguistic nature of simplifications, it is also useful to understand which individual features are more informative by themselves. Hence, we trained single feature ranking models and ranked the features based on their performance on the test sets. Tables 5.3 and 5.4 show the list of single features that achieved $> 60\%$ accuracy for two training sets -

WIKI-TRAIN and OSE2-TRAIN, using within corpus test-set evaluation[8].

feature	feature group	accuracy
numsubtrees	SYN	72.1%
cttr	LEX	70.4%
senlen	SYN	69.7%
AoA-kup-Lem	PSY	64.8%
numconstituents	SYN	63.3%
mlt	SYN	63.2%

Table 5.3: Accuracy of single feature models for WIKI-TRAIN/WIKI-TEST

feature	feature group	accuracy
AoA-kup-Lem	PSY	72.8%
cttr	LEX	66.7%
numsubtrees	SYN	64.4%
mlc	SYN	63.2%
imagery	PSY	63.2%
familiarity	PSY	63.2%
colMeaningful	PSY	63.2%
concreteness	PSY	61.7%

Table 5.4: Accuracy of single feature models for OSE2-TRAIN/OSE2-TEST

While the WIKI model had only 6 features that individually performed with an accuracy above 60% (4 SYN, 1 LEX, 1 PSY), OSE2 had 8 features (5 PSY, 2 SYN, 1 LEX) features. There were 3 and 5 CEL features respectively that performed with more than 50% accuracy in both datasets respectively. Thus, we can notice a combination of word-level and syntactic features in this list. While WIKI model seems to be influenced more by the syntactic features, word level psycholinguistic feature seems to play more important role in the OSE2 model, when considered as single features. One possible reason for avg. sentence length being more predictive in WIKI model and not OSE2 model could be that the Wikipedia dataset consisted of a lot of deletions ($\sim$ 45% of the sentences had major deletions) compared to OSE dataset, where sentences were mostly rewritten

[8]We experimented with a variety of age of acquisition norms and lexical diversity measures, but only report one of each above. Interestingly, the accuracies obtained using the various AoA norms substantially differed, between 37% and 72.8%, also due to coverage.

or paraphrased instead of deleting the content. Hence, sentence length as a single feature for OSE2-TEST data achieved a accuracy of only 57.5%, compared to the 69.7% shown for WIKI-TEST data in Figure 5.3.

The number of psycholinguistically motivated features (age of acquisition, concreteness, meaningfulness, imagery) in the OSE2 model is interesting and would merit a more detailed study. Information about the role of these features could also be useful for lexical simplification approaches like that of Jauhar & Specia (2012), who used some of the features from the MRC psycholinguistic database to rank word substitutes for lexical simplification.

5.4.8 Simplification at different levels

We also explored, whether the nature of the simplification differs between advanced sentences being simplified compared to intermediate sentences being (further) simplified. We split the OSE3-TRAIN and OSE3-TEST datasets into two pairs of datasets ADV-INTER-TRAIN, ADV-INTER-TEST and INTER-ELE-TRAIN, INTER-ELE-TEST respectively for this purpose. Table 5.5 shows the differences in the performance of the ranking approach between the two levels of simplification. Overall results showing simplification is somewhat different at these two different levels.

Train:	ADV-INTER	INTER-ELE
ADV-INTER-TEST	73.6%	74.7%
INTER-ELE-TEST	81.6%	80.5%

Table 5.5: Simplification at different levels

The performance on INTER-ELE test-set was better when tested with both the models. To understand the reason, we explored the nature of the simplification involved at these two different levels by testing the predictive power of individual features. While only AoA features achieved an accuracy of above 60% ADV-INTER model, Table 5.6 shows the list of features that individually achieved an accuracy of >60% for intermediate to beginner level. In general, the predictive power of features seem to be higher for INTER-ELE model compared to ADV-INTER model. This may perhaps mean that the amount of simplification is more

in the former transition than the latter.

feature	feature group	accuracy
AoA-Kup-Lem	PSY	77%
imagery	PSY	67.8%
CTTR	LEX	67.8%
meaningfulness	PSY	66.7%
concreteness	PSY	65.5%
familiarity	PSY	64.4%
MLC	SYN	64.4%
#sub trees	SYN	64.4%
num. senses	LEX	64.4%

Table 5.6: Single feature ranking models for INTER-ELE simplification

It can be argued that the syntactic features seen in these tables (e.g., avg. length of a clause/t-unit, num. subtrees etc.,) are correlated with text length. However, since simplification can also involve sentence rewrites that do not affect the sentence length as such (e.g., paraphrasing, reordering), the degree of simplification is more reflected through the use of specific syntactic structures than sentence length alone, as the results clearly indicate.

5.4.9 Error Analysis

To understand if there is a systematic pattern in the errors made by the ranker, we did a manual analysis of errors. For this, we took the results of training with OSE3-TRAIN data and testing with the OSE3-TEST. Since this is the smallest test set (87 triplets), and had only 53 misclassified pairs in total (79.7% accuracy), we chose this dataset for a quick manual analysis of errors. The following are four example sentence pairs/triplets from the test set. While the first two were ranked correctly by the ranker, the last two illustrate the cases where the ranker failed.

- Example 1:
  - *adv: He warned that it was too early to use oxytocin as a treatment for the social difficulties caused by autism and cautioned against buying oxytocin from suppliers online.*

  - *int: He warned that it was too early to use oxytocin as a treatment for the social difficulties caused by autism and said people should not buy oxytocin online.*
  - *ele: He said that it was too early to use oxytocin as a treatment for the social difficulties caused by autism and said people should not buy oxytocin online.*

- Example 2:

  - *int: DNA taken from the wisdom tooth of a European hunter-gatherer has given scientists a glimpse of modern humans before the rise of farming.*
  - *ele: Scientists have taken DNA from the tooth of a European hunter-gatherer and have found out what modern humans looked like before they started farming.*

- Example 3:

  - *adv: Its inventor, Bob Propst, said in 1997, "the cubiclizing of people in modern corporations is monolithic insanity."*
  - *int: Its inventor, Bob Propst, said, in 1997, "the use of cubicles in modern corporations is crazy."*
  - *ele: The inventor, Bob Propst, said, in 1997, "the use of cubicles in modern companies is crazy."*

- Example 4:

  - *adv: A special "auditor" declares him 96.9% "made in France" and Montebourg visits to present him with a medal.*
  - *int: A special "auditor" declares him 96.9% "made in France" and Montebourg visited to present him with a medal.*

In Example 1, the transformation from *adv* to *int* is primarily paraphrasing ("and cautioned against buying oxytocin" vs "and said people should not buy oxytocin") where was the transformation from *int* to *ele* is that of a simple lexical

substitution (”He warned” vs ”He said”). However, in Example 2, there was a significant re-ordering of the sentence with some paraphrasing (”before the rise of farming” vs ”before they started farming”). In both these cases, our model identified the changes as simplification (according to the original writers of the text) and ranked the different versions correctly in terms of their reading complexity. This may lead us to a conclusion that the model identifies paraphrases and lexical substitutions efficiently at multiple levels.

However, the model is not as effective with the sentence triplet in Example 3. One would assume that the *adv* version of the sentence is more difficult than the other two versions of the sentence. The pair-wise ranking from our model was: $adv < int$; $int > ele$; $adv > ele$. Though the model identified a simple lexical substitution between *int* and *ele* correctly, it failed to identify the transformation from ”the cubiclizing of people” to ”the use of cubicles” or ”monolithic insanity” to ”crazy” as simplification. This could possibly be because the parse structure as such did not alter much despite the rephrasing and because neither ”cubiclizing” nor its lemmatized version ”cubiclize” existed in the psycholinguistic databases we used. Including frequencies of word usage may perhaps be useful in such cases. In Example 4, where the only change between the sentence version is a tense difference (visits vs visited), the model failed to identify the rank-order correctly, which could perhaps because of the fact that the feature vector would not have changed much in this case. Whether a change in tense could be considered simplification is another issue we would not discuss in this paper. While we did not find any systematic failure in our models yet, based on the current results and observations, we could conclude that this approach can efficiently identify simplifications that are both lexical and syntactic in nature.

5.5 Sentences to Documents - Local to Global Readability Estimates

Now that we have an approach to rank sentences based on their readability, we explored the possibility of moving back from sentences to documents. The reading level assigned to a text is a global estimate for that text, on an average. However,

it is possible for a text to have parts with varying degrees of complexity. Having a model that can compare sentences or paragraphs in terms of their reading difficulty may enable us to develop models for more fine grained assessment of the reading level of a text. Since we now have models that can rank short texts based on their reading level, it is possible for us to estimate readability locally between sentences. We hypothesized that employing these models to estimate document level readability would provide as a more fine grained global readability estimates based on the ranking of sentence level estimates.

Applying the ranking model on the sentences in a document results in a prediction generated per sentence by the model. While the predictions by themselves don't mean anything, sorting them by their absolute value will give us a way to compare sentences within a text in terms of their reading level. Thus, a large difference between the predicted values of two sentences would mean that they are widely separated in terms of their difficulty level while a small difference would mean the model cannot identify the rank order difference with more confidence. This aspect of the ranking models allows us to develop an estimate of document readability based on the distribution of the ranker predictions over the sentences in the document. One way to draw an inference is to look at the skewness of the distribution around median. If the document is skewed towards the right, it could mean that there are more difficult parts (sentences in our case) in the document and if its left-skewed, it may mean that the document has more easy to read sentences and a few difficult sentences. Comparing the rank correlation between actual grade levels of the text and the skewness value per text provides us a way to assess the utility of this approach for new unseen, text data.

To verify this hypothesis, we used sentence level readability ranking models developed in the previous sections to rank Common Core Standards texts (Refer Chapter 3) based on their reading level. We compared multiple measures of skewness to get a skew estimate per text. The measures are obtained from Apache Commons Mathematics library[9]. We used the following measures in this study:

[9] `http://commons.apache.org/proper/commons-math/`

1. Skewness, which is measured as:

$$\frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(x_i - mean)^3}{std^3} \tag{5.1}$$

   where n is the number of values (in our case, sentences in a text), x_i is the prediction of the ranker for i^{th} sentence, mean is the mean of the predictions and std is the standard deviation.

2. Kurtosis, which is measured as:

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(x_i - mean)^4}{std^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{5.2}$$

3. Galton Skewness, given by:

$$\frac{Q1 + Q3 - 2 * Q2}{Q3 - Q1} \tag{5.3}$$

   where Q1, Q2,Q3 are the lower, median and upper quartile respectively.

4. Range: The difference between maximum and minimum value of the predictions for a text.

5. Stdev: The standard deviation of the distribution of predictions for a text.

We used the WIKI-TRAIN, OSE2-TRAIN and OSE3-TRAIN models to rank the sentences within each text from Common Core Standards corpus. We then compared the rank-correlations between all the skewness measures and the actual grade levels. The measures: skewness, kurtosis and galton did not result in a statistically significant correlation with actual grade levels. Hence, we report only the results with Range and Stdev for all the three models, in Table 5.7.

The rank-correlations obtained for both the measures are in the range of 0.46-0.52, which is similar to what was obtained by Lexile scale on this data set as reported in (Nelson et al., 2012). These correlations are much lesser than what were obtained for the other reported results on this dataset, including that of our

Model	ρ for Range	ρ for Stdev
WIKI-TRAIN	0.52	0.49
OSE2-TRAIN	0.51	0.47
OSE3-TRAIN	0.50	0.46

Table 5.7: Rank-correlations for the Common Core Standards dataset

approach from Chapter 3. However, this is an interesting problem to explore further, considering the fact that the predictions of this model are based on a training data which has more fine grained readability estimates than any other dataset used so far.

5.6 Conclusions

To summarize, we studied approaches compare sentences and their simplified versions in terms of their reading level. We showed that a pairwise ranking approach performs better than classification or regression for this task. It identifies the order in terms of their reading level correctly with an accuracy of over 80%, the best accuracy we are aware of, using parallel sentence aligned data. We performed within corpus and cross-corpus evaluations with two very different sentence aligned corpora and showed that the approach generalizes well across corpora. In this process, we created a new resource of sentence-aligned simplified texts based on OneStopEnglish texts rewritten by experts for language learners into three reading levels. This approach and the corpus could be useful for the evaluation of text-simplification systems for language learners in real life educational settings.

We also studied the role of individual features and groups of features in predicting the ranking order between simplified and unsimplified versions of the sentences. We found out that the psycholinguistic features like Age-of-acquisition seem to be more predictive as individual features. However, using all the features results in a much better model that performs well in cross-corpus settings too.

We compared the differences in text simplification process between multiple levels. Our results show that more simplification operations happen between intermediate to elementary level compared to advanced to intermediate levels, in terms of the features we studied. A short error analysis did not reveal any sys-

tematic error patterns in our approach yet, and deserves further study. Though we primarily studied this problem to compare versions of the same sentence, the approach is equally applicable in choosing the target sentences for simplification of a text.

Finally, we briefly explored the possibility of using the sentence level readability estimates to detect the reading level of the document, based on the skewness of the ranker predictions. This approach gave a rank-correlation of 0.5 when the range of predictions (maximum value – minimum value) was considered as a measure of skewness, on Common Core Standards dataset. While this approach did not result in an improvement over previously reported results, the initial experiments appear promising and they could be a step towards getting fine grained readability estimates for a text.

5.6.1 Outlook

Understanding which differences between accuracies are statistically significant and exploring feature selection in further detail in terms of selecting the best features for the ranker while removing the correlated ones Geng et al. (2007) are the immediate directions one could pursue.

Apart from that, it would be interesting to apply the approach to evaluate the output of real automatic text simplification systems and compare their performance in terms of readability. Going beyond complexity, in the long term it could be interesting to extend the approach to a full framework for evaluating automatic text simplification systems by integrating aspects of fluency and grammaticality.

Chapter 6

Text Simplification as Machine Translation: Role of training corpora and language models

Abstract

In this chapter, we describe an approach to generate simplified texts with Statistical Machine Translation (SMT). We applied a Phrase Based Machine Translation (PBMT) approach as implemented in Moses SMT toolkit (Hoang et al., 2007) for this purpose. We experimented with focused training datasets and language models. In our experiments, we considered Automatic Text Simplification (ATS) as a machine translation task from English into Simple English. Our approach currently only handles word replacements and paraphrases as it is trained using a focused set of sentences that consist of only these operations. We show that this smaller focused model performs better than a model trained on a large, noisy Wikipedia based training set for the same simplification operations. We also performed a cross-corpus evaluation of the simplification approach and our results lead us to the conclusion that for text simplification, purely phrase-based machine translation does not generalize to data from another source.

6.1 Introduction

One of our goals in this thesis was to explore the possibility of automatic text simplification as a means to provide comprehensible texts for language learners. Most of the reported automatic text simplification approaches primarily rely on Wikipedia-Simple Wikipedia sentence aligned corpus as the training resource. However, as we noticed in the previous chapter and as was observed by Amancio & Specia (2014) as well, a majority of the transformations in Wikipedia were deletions, where content from the source got removed in the target version. Since simplification does not necessarily mean removal of content, it is desirable to have an approach that will simplify the form without the loss of content. With this goal in mind, we aimed at performing lexical simplification and paraphrasing using a machine translation approach.

With this aim, we deal with two issues related to automatic text simplification based on machine translation.

1. the role of training data and language models on the quality of machine translation for handling lexical simplifications and paraphrases

2. the cross-corpus generalizability of the approach using a new sentence level simplification corpus

We hypothesize that training the machine translation system with data that contains only certain kinds of simplifications would give us a model that learns how to do those transformations correctly. Thus, instead of using a larger, noisy training corpus, which would result in a model that handles more transformations, but performs poorly, we used a smaller, focused training data. Our results confirmed our hypothesis and the translation model trained with specific transformations achieved better performance than a model trained on the entire data consisting of a lot of noise, for lexical simplification and paraphrase operations. Automatic Text Simplification has been approached as a machine translation problem in the recent past (e.g., Zhu et al., 2010; Coster & Kauchak, 2011a; Wubben et al., 2012). While most of the approaches experimented with the machine translation process, some explored the role of language models (Kauchak, 2013) and some others explored ways to re-rank the translated strings for text simplification

(Wubben et al., 2012). To our knowledge, the role of training data on simplified text generation has not been explored earlier in this strand of research. Our experiments showed that the focused training data approach resulted in better simplification in terms of handling lexical simplifications and paraphrases.

As we now have access to another sentence aligned simplification corpus (onestopenglish corpus described in the previous chapter), we investigated the cross-corpus generalizability of the above mentioned approach. For this experiment, we compared combinations of training data and language models using Wiki and OSE corpus sentence aligned data. Our results show that all models generally performed poorly with cross-corpus evaluation.

The rest of this chapter is organized as follows: Section 6.2 describes the corpora used in these experiments for building translation and language models. Section 6.3 describes the methods used in this chapter - for training and testing the machine translation models. Section 6.4 describes the experiments we performed and analyses the results obtained. Finally, Section 6.6 summarizes the conclusions of our experiments with some pointers to extending the work.

6.2 Corpora

6.2.1 Training and Development Data

We used two pairs of training and development corpora in this chapter.

1. WIKIALL consists of 90,000 sentence pairs of simplifications from Zhu et al. (2010) as training data and 8000 sentence pairs as development data. This is a noisy corpus consisting of all kinds of sentential transformations with a large amount of content deletion.

2. WIKIOSE is a dataset consisting of two parts:

    - A subset of WIKIALL which includes only sentence pairs that had transformations involving lexical simplification or paraphrasing. We obtained this subset by comparing the normal and simplified versions in terms of the difference in text length and cosine similarity. In addition, we eliminated sentences that got split during simplification. This

subset consisted of ∼39K pairs of sentences. We used 35K for training the translation system, 2K as the development set and the remaining as the test set.

- A subset of the OSE2 dataset explained in the previous chapter, consisting of 2300 sentences involving only paraphrases and lexical simplifications. Of these 2300, 800 were included in the training set, 750 sentence pairs in the development set and the remaining were used as test set.

6.2.2 Test Data

Several of the previous approaches used the test set from Zhu et al. (2010), consisting of 100 complex sentences and the corresponding simplified version consisting of 131 sentences (e.g., Woodsend & Lapata, 2011a; Wubben et al., 2012). Since our translation models are designed to only handle specific simplifications, a test set consisting of all transformations, especially sentence splits, would not be suitable for our task. So, we created two test sets - WIKI-TEST and OSE-TEST consisting of 2000 and 750 pairs of sentences each, where the simplifications performed are only word/phrase replacements and paraphrases. The sentence pairs used in test sets were not used in any of the training and development datasets[1].

6.2.3 Language Models

We used four language models in this approach:

1. LM0: Language model built with only sentences belonging to Simple Wikipedia from the WIKIALL corpus

2. SW: Language model built using the entire Simple Wikipedia in plain text, as crawled in August 2014.

3. OSEB: Language model built using only text from the One Stop English corpus, beginner level texts.

[1] All training, development and test sets used in these experiments can be shared for research use.

4. SW-OSEB: Language model built by combining the texts used in SW and SW-OSEB.

6.3 Methods

We followed a Phrase Based Machine Translation (PBMT) approach to develop simplification models. PBMT is a form of Statistical Machine Translation (SMT) where the translation units are phrases instead of words. In this model, the aim of the SMT system is to segment the source sentence into phrases and generate target translations for the phrases. The phrasal alignment and translation are based on statistical probabilities and hence rely on having large amounts of parallel training data. The PBMT system typically consists of two models: a translation model and a language model. In our context, while the translation model is concerned with generating possible simplifications for an unsimplified sentence, the language model is one of the factors in deciding the likelihood score of the generated translations.

To train the PBMT system, we used the Moses[2] toolkit (Hoang et al., 2007), which is one of the most commonly used tools for building machine translation system prototypes. We did not alter the PBMT pipeline in terms of the external tools used or the training option configurations. Tri-gram Language models with Kneser-Ney smoothing were trained using IRSTLM[3], which is integrated in Moses. GIZA++(Och & Ney, 2003)[4] was used for word alignment. Tuning was performed by Minimum Error Rate Tuning (MERT), the default in Moses.

6.3.1 Evaluation

We used the BLEU (Bilingual Evaluation Understudy) metric (Papineni et al., 2002) for evaluating and comparing the machine translation outputs from our systems, as in other related work. BLEU metric scores the quality of the machine translated text in terms of its closeness to reference translations and is known to

[2] http://www.statmt.org/moses/
[3] https://hlt.fbk.eu/technologies/irstlm
[4] http://www.statmt.org/moses/giza/GIZA++.html

achieve a high correlation with human judgments of translation quality. The reference translations in our case are the human simplified simple Wikipedia sentences. Mathematically, BLEU is calculated as the geometric mean of the n-gram precisions of machine translated output with respect to corresponding gold standard translations created by human translators, multiplied with a penalty for sentences that are shorter than the reference translation.

We consider the BLEU score that we get when we return an unsimplified text as the baseline for this task. That is, assuming that no simplification has been performed by the system, our baseline merely calculates the BLEU scores between unsimplified sentence and its human simplified version as they appear in the original corpus. A comparison of the baseline with the actual BLEU score obtained from the system output will give us an estimate of how much simplification was actually done by the system, instead of looking at the standalone BLEU score. We used the Multi-BLEU script in Moses toolkit to get the BLEU scores between sentence versions. This program gives a score of 100 for a perfect match with the human simplified version. To estimate the statistical significance of BLEU scores, we used MultEval[5] (Clark et al., 2011).

6.4 Experiments and Results

We first compared the effect of using a focused training corpus on the quality of the translated output. Table 6.1 and Table 6.2 show a comparison of BLEU scores for both the training sets used for WIKI-TEST and OSE-TEST test sets respectively. For these models, we built a language model consisting of only Simple Wikipedia sentences from the training corpus (LM0). In both the tables, baseline refers to the BLEU score between the original unsimplified version from Wikipedia and human simplified version from Simple Wikipedia, as explained earlier. *With Simplification* BLEU score refers to the score between the human simplified text and the machine generated text.

Clearly, using a focused training corpus was useful for WIKI-TEST, where the BLEU score had a huge increase from 73.51 to 95.71. Compared to the baseline

[5] `https://github.com/jhclark/multeval`

Training Data	BLEU - WIKI-TEST	
	Baseline	With Simplification
WIKIALL	71.78	73.51
WIKIOSE	71.78	95.71

Table 6.1: BLEU comparison on WIKI-TEST data for models using different training data

Training Data	BLEU - OSE-TEST	
	Baseline	With Simplification
WIKIALL	78.41	73.87
WIKIOSE	78.41	70.69

Table 6.2: BLEU comparison on OSE-TEST data for models using different training data

BLEU score (71.78), the WIKIALL model had a very small increase compared to the WIKIOSE model, which had a huge improvement. Both the models were significantly better than their respective baseline models ($p < 0.001$).

However, this improved accuracy did not transfer to a cross-corpus evaluation setup. The score for the simplified version of OSE-TEST was worse than the baseline BLEU for both models. This implies that the simplification system performed even worse than just returning the text without performing any simplification. Between them, the WIKIALL model received a slightly higher BLEU score than the WIKIOSE model. On the one hand, this may indicate that the nature of simplification is different between both the corpora and that the OSE corpus was under-represented in the training process. On the other hand, having a larger language model may result in improved simplifications, if we assume that the nature of the simplifications in both the corpora are similar.

We compared the performance of three other language models along with the language model used for the above comparison, using the WIKIOSE dataset as the training set for re-training a translation model. Table 6.3 and Table 6.4 show the comparison of BLEU scores for the two test sets, with all the language models.

As the results show, changing the language models did not seem to result in a lot of improvement in the BLEU score. The BLEU for OSE-TEST showed a slight increase with both SW and OSEB language models compared to LM0,

Language Model	BLEU - WIKI-TEST	
	Without Simplification	With Simplification
LM0	71.78	95.71
SW	71.78	95.66
OSEB	71.78	95.63
SW-OSEB	71.78	95.68

Table 6.3: BLEU comparison on WIKI-TEST data using different language models for translation

Language Model	BLEU - OSE-TEST	
	Without Simplification	With Simplification
LM0	78.41	70.69
SW	78.41	71.51
OSEB	78.41	70.81
SW-OSEB	78.41	70.67

Table 6.4: BLEU comparison on OSE-TEST data using different language models for translation

though the score was still less than returning the same text without performing any simplification. This indicates that using a larger language model or a language model focused on the target domain is not useful in improving the cross-corpus translation performance. The results also indicate that the language model cannot alter the translation output drastically, as it is clear from the case of the WIKI-TEST test set achieving a BLEU score of 95.63 even when trained on a small, out of corpus language model (OSEB).

6.5 Output Examples and Analysis

Finally, we compared the output produced by our approach with the outputs of other reported systems. We used the translation model trained on WIKIOSE and the language model SW for this comparison, as this seemed to perform slightly better with cross-corpus performance. Table 6.5 shows an example Wikipedia sentence, its simple Wikipedia version and translations by various systems. The outputs for all the systems are as reported in Wubben et al. (2012). We manually verified that the example sentence was not seen in any of our training and

development sets.

Reference	Sentence
Actual sentence	the judge ordered that chapman should receive psychiatric treatment in prison and sentenced him to twenty years to life, slightly less than the maximum possible of twenty-five years to life.
Simple Wikipedia version	he was sentenced to twenty-five years to life in prison in 1981.
Word-substitution baseline	the judge ordered that chapman should **have** psychiatric treatment in prison and sentenced him to twenty years to life, slightly less than the maximum possible of twenty-five years to life.
Zhu et al. (2010)	the judge ordered that chapman should **get** psychiatric treatment. *in prison and sentenced him to twenty years to life, less maximum possible of twenty-five years to life.*
Woodsend & Lapata (2011a)	the judge ordered that chapman should **will get** psychiatric treatment in prison. *he sentenced him to twenty years to life to life.*
Wubben et al. (2012)	the judge ordered that chapman should **get** psychiatric treatment in prison and sentenced him to twenty years to life, **a little bit** less than the **highest** possible to twenty-five years to life.
Our Approach	the judge ordered that chapman should **get** psychiatric treatment in prison and sentenced him to twenty years to life, slightly less than the maximum possible of twenty-five years to life .

Table 6.5: Example Output - Comparison with other approaches

We can notice that in the Simple Wikipedia version, a significant proportion of the sentence got deleted, resulting in a loss of information. The word substitution baseline changed ”receive” to ”have”. The systems by Zhu et al. (2010) and Woodsend & Lapata (2011a) attempted to split the sentence into two, but resulted in an ungrammatical output. Wubben et al. (2012) performed three word

replacement operations and our approach replaced ”receive” with ”get”.

Table 6.6 shows three more examples of our approach's output with examples from both the test sets:

Reference	Sentence
Example 1 (from WIKI-TEST)	
Original	In economics, hyperinflation is inflation that is very high or ”out of control” , a condition in which prices increase rapidly as a currency loses its value.
Simple Wiki	In economics, hyperinflation is inflation that is ”out of control,” when prices increase very fast as money loses its value.
Our model output	In economics, hyperinflation is inflation that is very high or ”out of control”, a condition in which prices increase **very fast** as a currency loses its value.
Example 2 (from OSE-TEST)	
Original	These are the once seemingly sci-fi questions that can now be experimentally tackled in the lab.
OSE-Simplified	These are the seemingly sci-fi questions that can now be experimentally tackled in the lab.
Our model output	These are the seemingly sci-fi questions that can now be **tackled experimentally** in the lab.
Example 3 (from OSE-TEST)	
Original	He secured a full-time job in administration and worked as a DJ.
OSE-Simplified	After leaving college, he got a full-time job in administration and worked as a DJ.
Our model output	He **had obtained a job** in administration and worked as a DJ.

Table 6.6: A few more Example Outputs

As it can be seen from the example translations, the model is capable of performing lexical simplifications and a small amount of re-ordering (experimentally tackled versus tackled experimentally in Example 2), even in a cross-corpus sce-

nario in some cases. However, as the BLEU scores from Tables 1–4 show, in most of the cases, the model failed to generalize to the new corpus.

Since in a practical application scenario the model has to be able to translate new, unseen data, more work is needed to make the model robust for cross-corpus performance. While acquiring more training data is one direction to pursue, another direction could be to explore syntactic simplification. Further, extracting lexical simplification and paraphrasing rules from phrase table entries and using them in a rule-based system is another option that could be useful.

6.6 Conclusions

In this chapter, we explored automatic text simplification as phrase based machine translation, with the aim of simplifying in form while retaining the meaning. Hence, we considered only a training subset consisting of only a restricted set of simplification operations instead of using the entire data comprising of a lot of deletion decisions. This resulted in a better simplification, in terms of the BLEU scores. We used an additional test set apart from the Wikipedia-Simple Wikipedia data that was used for this task so far and our results showed that the phrase based machine translation approach did not transfer to the new corpus. We explored the utility of different language models for improving the cross corpus performance and our experiments showed that language models did not improve the BLEU scores. Finally we compared the output of our system with other existing systems, and briefly analyzed some of the simplifications performed by our approach. The results showed that the system captures lexical simplifications and small paraphrase operations efficiently.

6.6.1 Outlook

As a next step, we would explore the evaluation of simplification by applying the sentence level readability model from the previous chapter, on a small set of new, unseen data and compare it with human judgments. We are currently working on an approach that identifies split points in a difficult sentence based on typed dependency representation of sentences. Adding these split rules can enhance the

capabilities of the current approach by handling the splitting of longer sentences. We could also explore means to combine the phrase table mappings generated by the phrase-based machine translation approach with a set of manually created rules that handle other simplification operations involving larger amounts of rewriting and sentence splitting. Finally, viewing text simplification as syntactic monolingual machine translation may result in more generalizability to a new corpus, which needs to be explored in the future.

Part II

Linguistic Complexity in other Educational Contexts

Overview

Apart from readability assessment and text simplification, the analysis of linguistic complexity can be useful in other applications related to the educational context. Some of them include: assessing student writing (e.g., Lu, 2010, 2012), analysis of textbooks (e.g., Fitzgerald et al., 2015), and developing diagnostic tools for content authors (e.g., Agrawal et al., 2012). Considering readability issues in developing personalized education and adaptive learning tools is also gaining attention (e.g., AMPLE project at IBM[6]).

In this thesis, we studied the usefulness of linguistic complexity features for two educational applications:

1. **L2 Writing Analysis**: We used the linguistic complexity measures developed in Chapter 3 to classify L2 English writing into pre-defined proficiency levels.

2. **Analyzing German textbooks**: We used a collection of text complexity features developed for German by Hancke (2013) to analyze the German Geography textbooks across grades, school types and publishers and studied the differences among them in terms of the linguistic features.

The two chapters that follow will describe both the applications in detail.

[6] http://researcher.watson.ibm.com/researcher/view_group.php?id=4975

Chapter 7

Assessing L2 Writing with Text Complexity Measures

Abstract

In this chapter, we explore the usefulness of our readability features for analyzing L2 English writing. Using learner datasets annotated with proficiency level, we built classification and regression models for automatic proficiency prediction with our feature set. While the results with some datasets are promising, we did not get a uniform performance across all the datasets, which may indicate both the differences between the datasets we used in this experiment and the insufficiency of the features to capture relevant aspects of proficiency. We briefly compare the most predictive features for proficiency classification based on learner L1 background and the results indicate some differences in terms of most predictive syntactic features.

7.1 Introduction

Automatic scoring of learner essays is one of the most popular educational applications of Natural Language Processing (NLP). Producing a free form text such as an essay is a part of several language assessment tests in high-stakes and low-stakes scenarios. The aim of automated assessment is to analyze such texts based

on the given examination scale and criteria. It has obvious advantages in terms of reducing the amount of manual work involved in assessing student scripts and complimenting the human examiner judgments. Automated essay scoring is already being used along with human grading in several online exams like Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT). It can also be useful in a placement test that one may take at a language teaching institute before starting to learn a language at a certain level or serve as a guiding tool for language learners in self-assessment. It can also be used as a writing assistance tool for language learners (Burstein et al., 2003). Apart from this, automated approaches can also enable us to identify distinctive features at a proficiency level, thereby providing us with insights about the process of language acquisition.

Automated Assessment of student essays has been an active area of research for over four decades now, and commercial assessment systems are being used to evaluate writing in computer-based tests. From superficial measures like word length and sentence length to sophisticated natural language processing techniques, a wide range of factors such as grammatical correctness, error rate, language quality and proficiency were considered for this task (e.g., Williamson, 2009; Burstein & Chodorow, 2010; Yannakoudakis et al., 2011; Crossley et al., 2011b). Dikli (2006) provides an overview of the working of various Automated Essay Scoring systems.

While this research is primarily focused on English language, with the creation of learner corpora in various European languages, automatic approaches for classifying learner essays into various proficiency levels began to emerge. Approaches for morphologically rich languages also made use of language specific morphological features, which were not explored before in the case of English. Ostling et al. (2013) reported on a proficiency classification approach for Swedish, modeling with features like word length, sentence length, POS tag densities and corpus based entropy features. Hancke & Meurers (2013) and Hancke (2013) described a proficiency classification approach for German based on European CEFR standards using a broad range of lexical, syntactic and morphological features. Vajjala & Lõo (2013) and Vajjala & Lõo (2014) describe a proficiency classification approach for Estonian learner texts based on the CEFR scale, using a collection of part-of-speech and morphological features.

The primary purpose of these approaches is to predict the proficiency of L2 learners based on their written production. There are also studies that performed a qualitative analysis of criterial features between proficiency levels in Second Language Acquisition (SLA) literature. Kyle & Crossley (2014) used a range of lexical sophistication indices and showed that the measures explain 47.5% of the variance in holistic scores of lexical proficiency of second language English learners. Characteristics like lexical richness, syntactic complexity, error patterns of learners and other characteristics too were studied in the recent past in SLA research community for English (e.g., Tono, 2000; Lu, 2010, 2012; Vyatkina, 2012). Although this strand of research is primarily focused on English, recent research has started to focus on other languages as well (Gyllstad et al., 2014).

SLA research often characterizes L2 proficiency in terms of three dimensions: *Complexity, Accuracy* and *Fluency* (e.g., Housen & Kuiken, 2009; Norris & Ortega, 2009). Hence, we hypothesized that the complexity features of the texts produced by English learners could be useful predictors for assessing the L2 writing proficiency of learners. Our feature set contains some of the lexical richness and syntactic complexity features which were also originally used to analyze L2 English writing in Second Language Acquisition (SLA) research (e.g. Lu, 2010, 2012). These features turned out to be useful for readability classification of properly written English texts intended for both L1 (WeeBit) and L2 (OneStopEnglish) learners. Now, turning back to where we started, we used our text complexity feature set described in Chapter 3, to study their impact in performing proficiency classification of L2 English. We evaluate the effectiveness of our approach by training supervised machine models on several publicly accessible L2 English learner corpora.

The experiments and results are described in the following sections of the chapter: Section 7.2 describes all the corpora we used in the experiments reported in this chapter. Section 7.3 describes the basic experimental setup and evaluation methods. Section 7.4 describes the experiments for assessing proficiency and their results. Finally, Section 7.5 concludes the Chapter with an overview of future directions.

7.2 Corpora

We used four publicly accessible, proficiency annotated corpora for our task.

7.2.1 The FCE corpus

Yannakoudakis et al. (2011) released a corpus of First Certificate of English (FCE) responses from the Cambridge Learner Corpus (CLC). The CLC is a collection of texts produced by takers of English as Second or Other Language (ESOL) exams, which consists of scripts produced by learners and their associated scores. Along with the release of this dataset, they also experimented with various feature groups, ranging from n-gram models to phrase structure rules and error rates. The system predictions with these features achieved a correlation of 0.75 with actual scores. We used the publicly accessible version of this corpus[1]. It contains 1238 ESOL examination scripts written by 1238 distinct learners at upper-intermediate level, scored on a scale of 1–40. Following their guidelines, we consider 1141 texts from the year 2000 as the training set and 97 scripts from the year 2001 as the test set in our experiments. We will refer to this corpus as FCE for the rest of this chapter and to the training and test sets as FCE-TRAIN and FCE-TEST respectively.

7.2.2 BuiD corpus

The British University in Dubai (BUiD) Arab Learner corpus (Randall & Groom, 2009) is a collection of 1865 English learner texts written by students with an Arabic L1 background from the last year of secondary school and the first year of university. The texts were scored and assigned to six proficiency levels according to the Common Educational Proficiency Assessment (CEPA) examination standard in United Arab Emirates. Each level consists of about 250-300 texts. Table 7.1 shows the distribution of texts across CEPA levels in the corpus and descriptive statistics about the texts. We will refer to this corpus as BUID for the rest of this chapter.

[1] http://ilexir.co.uk/applications/clc-fce-dataset/

Proficiency Level	Num. Texts	Num. Sentences per text	Avg. Sentence Length
cepa1	283	2.6	22.6
cepa2	293	5.5	22.8
cepa3	251	7.6	24
cepa4	297	9.8	24.7
cepa5	299	12.7	18.5
cepa6	252	12.3	19.0

Table 7.1: BUiD Arab Learner Corpus

7.2.3 ICNALE Corpus

The International Corpus Network of Asian Learners of English (ICNALE) corpus (Ishikawa, 2011) consists of 5600 essays written by college students in ten countries and areas in Asia as well as by English native speakers. The learner essays are assigned to four proficiency levels following the CEFR guidelines (A2, B1, B2, B2+). Table 7.2 shows the distribution of texts in the corpus and basic statistics about it. We will refer to this corpus as ICNALE for the rest of this chapter. To our knowledge, this corpus was not used to perform proficiency classification before.

Proficiency Level	Num. Texts	Num. Sentences per text	Avg. Sentence Length
A2	960	15.1	18
B1_1	1904	14.9	16
B1_2	1872	13.8	17.7
B2	464	13.6	18.3

Table 7.2: The ICNALE Corpus

7.2.4 TOEFL11 Corpus

The TOEFL11 corpus (Blanchard et al., 2013) consists of essays written by English learners with 11 native language (L1) backgrounds. It consists 1100 essays per native language and has proficiency annotations belonging to three categories: low, medium, high. We used the version consisting of the training and development sets used for the Native Language Identification shared task, 2013. The

corpus is now available from Linguistic Data Consortium[2]. Table 7.3 shows a description of the corpus across proficiency levels. We will refer to this corpus as TOEFL11 for the rest of this chapter. Although the corpus has been extensively used for Native Language Identification, to our knowledge, it was not used for proficiency classification before.

Proficiency Level	Num. Texts	Num. Sentences per text	Avg. Sentence Length
Low	1201	10	32.8
Medium	5964	15.3	25.2
High	3834	17.8	21.4

Table 7.3: The TOEFL11 Corpus

7.3 Experimental Setup

We used the same feature set from Chapter 3 for this task, for all the datasets. We explored both classification and regression approaches for this task, based on the continuous or discrete nature of the notion of proficiency used in the datasets. When we considered the task as classification, we trained classification models on the entire dataset as well as a subset in which all the categories are represented equally. This was done to compare the performance in the two setups and ensure that there is no bias towards majority classes in the model that has more training data for one class. Where there is a bias towards the majority class in such cases, we report the confusion matrices from both balanced and unbalanced models.

For training the machine learning models, we used WEKA toolkit (Hall et al., 2009), as in earlier chapters. For classification, we will report the results with SMO algorithm and for regression, using SMOReg, as in previous chapters. In terms of the evaluation, we follow the same procedure as in earlier chapters: classification accuracy for classification and correlation and root mean square error (RMSE) for regression.

[2] https://catalog.ldc.upenn.edu/LDC2014T06

7.4 Experiments and Results

7.4.1 With FCE

We first trained a regression model with FCE-TRAIN as the training set and FCE-TEST as the test set. The model achieved a correlation of 0.54 and an RMSE of 5.3. For a 10-fold CV setup using the entire dataset, we got a correlation of 0.51 and an RMSE of 4.8. While both the correlations are significant ($p<0.001$), the numbers leave us with a lot of scope for improvement. In a previous work on this dataset, Yannakoudakis et al. (2011) reported the highest correlation of 0.75, with a large feature set consisting of word and POS n-grams, phrase structure rules, and several error rate features. Error rate features and n-gram features played a significant role in their model performance, both of which are not considered in our model. The significant positive correlation of our model predictions with actual scores that we achieved so far seem to indicate that complexity features contribute to proficiency scoring. They could improve the overall prediction performance when considered together with other groups of features that are known to work for this task.

7.4.2 With BUID

Keeping the experimental setup the same as the previous experiment, we trained a regression model with BUID dataset, using a 10 fold CV. The SMOReg model achieved a correlation of 0.9 and an RMSE of 0.81. Since the number of sentences per text varied a lot between CEPA levels in this corpus (Table 7.1), we trained a model with only number of sentences as a feature, which achieved a correlation of 0.63 and an RMSE of 1.3. However, including this feature to our overall feature set did not result in any improvement in the correlation and increased the RMSE by 0.02 (0.83 vs 0.81). Though we are not aware of any other work that made use of this corpus for proficiency classification, the obtained numbers show that the model trained well.

Since the features resulted in a good regression model, we explored feature selection for this dataset. We used two feature selection approaches: CfsSubsetEval (Hall, 1999) and ReliefFAttributeEval (Robnik-Sikonja & Kononenko, 1997)

methods implemented in WEKA. CfsSubsetEval selects a group of attributes such that there is less degree of redundancy between them. ReliefFAttributeEval selects an attribute by comparing its value for a sampled instance with other instances belonging to the same and different classes. Both the algorithms work on both discrete and continuous classes.

Reducing the feature set using CfsSubsetEval method resulted in a subset consisting of 33 features, which resulted in the same performance as the model with all features. Considering top 30 features from the ReliefFAttributeEval ranked feature set too resulted in a performance similar to the model with all the features. Table 7.4 lists all the features selected by CfsSubsetEval and Table 7.5 lists the top 30 features returned by ReliefFAttributeEval algorithms respectively. The exact description of all the features can be found in Chapter 3 (Section 3.3). While ReliefFAttributeEval provides a ranked list, CfsSubsetEval does not provide a ranked list of features. So, the features appear in the order in which they appear in the dataset in Table 7.4 and as ranked list in decreasing order in Table 7.5.

nouns	proper nouns	pronouns
conj	advvar	adjvar
modvar	nounvar	numfuncwords
numvb	numvbd	numvbg
numvbp	ttr	cttr
morphcomplex	morphcontractions	morphirrelevant
moreanalyses	intrans	trans
sastem	sasanda	saflex
alloblend	alloderiv	transderhash
numgroupuncount	numExprN	perPro
pronPro	subordConj	AoA_kup_lem

Table 7.4: Best features for BUID corpus, using CfsSubsetEval method

Interestingly, the list of features using both methods consists entirely of word-level features like POS tags, morphological properties, type-token ratio and psycholinguistic features. Neither the surface features like sentence length nor the syntactic features are seen in this list. One reason could be that the learner sentences are perhaps not well formed enough for the parsers to correctly parse the sentences.

cttr	nounvar	lexicals
ttr	predAdj	numVbd
numattrn	numsingularn	modvar
numnouns	saflex	transderhash
numvocN	attrAdj	pavioMeaning
numVerb	pronPro	advvar
numcountablen	trans	ordAdj
adjvar	detPro	pronouns
numconj	numgroupcount	numtunits
numpron	numprep	numuncountablen

Table 7.5: Best features for BUID corpus, using ReliefFAttributeEval method

We also explored this as a classification task instead of regression. Training a SMO classifier for 6 categories resulted in an exact accuracy of 64.8% and an adjacent accuracy (where the prediction is within one level of difference from the actual value) of 96.9%. Table 7.6 shows the confusion matrix for the model.

classified as$->$	cepa1	cepa2	cepa3	cepa4	cepa5	cepa6
cepa1	208	58	13	3	0	1
cepa2	46	188	50	7	2	0
cepa3	6	70	116	53	4	2
cepa4	0	2	34	184	72	5
cepa5	0	0	0	57	200	42
cepa6	0	0	0	6	57	189

Table 7.6: Confusion Matrix for BUiD dataset

Clearly, the confusion between adjacent classes is much larger compared to classes separated by more levels. This shows that the feature set is able to capture the continuous nature of proficiency in this dataset. Two conclusions that can be drawn from this set of experiments with this corpus are:

1. Our approach resulted in a good model for proficiency scoring, achieving a correlation of 0.9 when trained as regression and an adjacent accuracy of 96.9% when trained as classification.

2. Syntactic features do not seem to be playing any role in predicting the L2

proficiency of the writer, in this dataset.

7.4.3 With ICNALE

Since the ICNALE corpus had four proficiency levels, we explored both classification and regression for this task. As the dataset is imbalanced, we considered only 464 instances per category (number of instances in the category with least representation, B2) to train our models. When modeled as classification, the model achieved an exact classification accuracy of 44% and an adjacent accuracy of 80.3%. When modeled as regression, we got a correlation of 0.5 and an RMSE of 0.9. Since the models did not achieve a good performance, we did not explore feature selection in detail. However, an Information Gain based ranked list of features for this model too was dominated by non-syntactic features. One reason for the poor performance of the model could be that the distinctions between the levels in this model (A2, B11, B12, B2) are perhaps more fine grained than those in the BUiD dataset. However, more study is needed to understand the nature of the relation between the features used and this dataset.

7.4.4 With TOEFL11

As the TOEFL11 corpus consists of three proficiency levels - low, medium and high, we considered it as a three class classification problem. Considering the entire corpus (Table 7.3) resulted in a skew towards the majority class (medium), as can be seen from the example confusion matrix shown in Table 7.7.

classified as$->$	low	medium	high
low	527	670	4
medium	79	5004	881
high	0	1409	2425

Table 7.7: Confusion Matrix for TOEFL11, unbalanced dataset

Hence, we trained additional models with a balanced training set consisting of 1201 instances per class. This resulted in a 2% decrease in classification accuracy, but the per-class accuracies seemed comparable now, without a skew to the

medium class. Table 7.8 shows the confusion matrix for this data, which resulted in accuracy of 70.5%.

classified as− >	low	medium	high
low	931	249	21
medium	170	722	309
high	5	309	887

Table 7.8: Confusion Matrix for TOEFL11, balanced dataset

Feature selection using CfsSubsetEval for this model resulted in a subset of 45 features, which together resulted in an accuracy of 69.3%. Choosing top-20 features based on Information Gain method resulted in a classification accuracy of 69.6%. Table 7.9 shows the top-20 features obtained using Information Gain.

Rank	Feature	Rank	Feature
1	cttr	11	advvar
2	lexicals	12	AoA_Kup_Lem
3	nounvar	13	numsubtrees
4	verbvar	14	mlc
5	modvar	15	numconstituents
6	opacity	16	numcountablen
7	concreteness	17	ttr
8	familiarity	18	imagery
9	colMeaningful	19	moreanalyses
10	adjvar	20	transderadd

Table 7.9: Top-20 features, for TOEFL11 corpus

Unlike the BUID data where there were no syntactic features in the list of most predictive features, in this dataset, syntactic features figure in the ranks between 10–20. While choosing only top-10 features resulted in a classification accuracy of 61.7%, adding the next 10 features resulted in an 8% improvement in accu-

racy. Thus, syntactic features seem to contribute to the overall accuracy for this dataset. We are not aware of any previous work on proficiency classification for this dataset, but we can consider 70% accuracy as a good start, considering the limited aspects of proficiency addressed by our feature set.

Most Predictive Features, by L1

Now we compared the most predictive features between the L2 writings of learners belonging to specific L1s. We took a sample of 4 L1s (Arabic, Italian, Korean, Telugu), belonging to four language families, to explore the most predictive features for each subset. Table 7.10 shows the top-10 features ranked by their Information Gain, for 4 subsets of the TOEFL11 data, representing four native languages.

Rank	Arabic	Italian	Chinese	Telugu
1	cttr	cttr	cttr	cttr
2	lexicals	familiarity	ttr	adjvar
3	AoA_Kup_Lem	colMeaningful	numprep	nounvar
4	familiarity	concreteness	modvar	lexicals
5	nounvar	AoA_Kup_Lem	numpp	numtunits
6	colMeaningful	lexicals	advvar	modvar
7	sastem	imagery	lexicals	verbvar
8	verbvar	numvbp	nounvar	numnp
9	sasanda	nounvar	numcountablen	advvar
10	morphcomplex	numvbg	verbvar	numclauses

Table 7.10: Top-10 features, by native language of the learners

As we observe in this table, while most of the word-level features are seen commonly for all L1s, there are one and three parse tree based features respectively for Chinese and Telugu data. Although we do not have any hypotheses yet on why this is the case, it could be an interesting direction to explore in the future.

7.5 Conclusions

In this chapter, we explored the usefulness of our text complexity features, for the task of assessing L2 learner English writing. We collected proficiency level annotated datasets from standardized criterion created by various sources and used our feature set to construct classification and regression models. While our approach effectively modeled proficiency with two of the datasets (BUiD and TOEFL11), it was not successful with ICNALE data and had a significant but less than the previously reported scores with FCE data. The models with these datasets have correlations ranging from 0.5 to 0.9 and classification accuracies ranging from 40-75%. This inconsistency across datasets may indicate differences between the datasets in terms of the notion of proficiency, which needs to be investigated in better detail in future. It could also indicate the insufficiency of the feature set in capturing the notion of L2 proficiency. This is possible since we do not consider features that were known to work for L2 proficiency classification, like spelling and grammar errors, for example. However, it is interesting that the models we constructed captured the notion of proficiency to some extent, although they relied on a feature set that does not consider these typically used predictors of L2 proficiency. Using this feature set in conjunction with learner errors and other features typically used in essay scoring systems could result in improved prediction accuracies.

We briefly compared the most predictive features for four native languages (Arabic, Italian, Chinese, Telugu) in the TOEFL11 corpus, by building L1 specific proficiency classification models. While the lexical richness features feature prominently in all the four languages, syntactic features seem to be useful for predicting Telugu native speakers' English proficiency. More detailed analysis is needed in this direction, to analyze the differences between L2 texts produced by learners of different L1s and the influence of native language on proficiency. Performing mediation analyses studying the influence of native language on specific syntactic constructs or word usage, which in turn affect the proficiency can reveal more about the relationship between native language and L2 proficiency.

Chapter 8

Analyzing Reading Demands in German Textbooks

Abstract

In this chapter, we describe an application of text complexity features for analyzing the reading demands in German schoolbooks. We used a corpus of Geography textbooks from grades five to ten, for two types of German schools - Gymnasium and Hauptschule. This corpus consists of books from four publishers. We compared the text snippets from these books in terms of the linguistic features described in Hancke (2013). We built text classifiers to predict the grade level/school type of these texts. The prediction accuracy for school-wise classification (binary) is 74.5% and for grade wise classification (three classes), it is 53.3%. In addition to performing text classification, we also studied a subset of individual features that show differences between categories. Our results lead us to a conclusion that while there exist differences between grades, schools and publishers, the predictive models perform better for school wise classification and significant differences exist between publishers at the school level.

The work described in this chapter was supported by LEAD Intramural Research Grant, project number: 19110506 (2013-2014). Other members of the project are: Karin Berendes, Doreen Bryant and Detmar Meurers

8.1 Introduction

Textbooks are the primary reading material for students at school. Hence, the form of language used in textbooks would play an important role in comprehension and language development of the students. Recently, educational standards like Common core[2] in the United States proposed that the students should have a "staircase of increasing text complexity" with grade level. This would imply that the reading materials for students, including the textbooks they read, should increase in their linguistic complexity with grade level.

Textbooks at various levels of instruction from school to college were compared for text complexity in the past in terms of traditional readability formulae (e.g., Aukerman, 1965; Flory et al., 1992) and other features encoding linguistic properties of texts (e.g., Lively & Pressey, 1923; Patty & Painter, 1931; Thorndike & Lorge, 1944; Dufty, Graesser, Louwerse & McNamara, 2006; Pyburn & Pazicni, 2014). Longitudinal analyses of textbooks from several decades were conducted using a range of linguistic features encoding text complexity (e.g., Davidson, 2005; Gamson et al., 2013; Lu et al., 2014), to study the trend of these features at different grade levels over a period of time. More recently, Fitzgerald et al. (2015) studied the text characteristics related to text complexity in primary grade texts and found that nine textual characteristics were important for complexity analysis at that level. They also concluded that the interplay between text characteristics is important to understand text complexity.

On the other hand, textbooks and other such instructional material have been used as the training corpus to develop computational readability models in the recent past (e.g., Heilman, Collins-Thompson, Callan & Eskenazi, 2007; Sato, Matsuyoshi & Kondoh, 2008; Francois & Watrin, 2011; Jiang, Sun, Gu & Chen, 2014; Pilán, Vajjala & Volodina, 2015), for English as well as other languages. While it is practical to consider textbooks as a training corpus for building readability models, the question: are textbook writers explicitly taking readability issues into consideration? demands further study. One way to understand this, apart from doing qualitative research through interviews with publishers and writers, is to quantitatively study the linguistic characteristics of these texts, as was done in

[2] http://www.corestandards.org/

earlier research mentioned in the previous paragraph.

In a related context, Britton et al. (1989), discussed about the lack of explicit awareness for the writers about the rewrites they have performed to improve text comprehension. Relating this to the textbook writers, a method to analyze and quantify the linguistic complexity of a text is also useful to develop a writing assistance tool for textbook writers, to make them aware of the language properties that may make a text difficult to comprehend. This is one of the motivations for this work.

In this chapter, we describe a first approach to analyze German textbooks in terms of differences between grades and school types using a set of existing text complexity features (Hancke, 2013). To this end, we compiled a corpus of German textbooks. In Germany, students can study in different types of high schools based on their performance in the primary school. The possible difference between school-types in terms of the linguistic complexity of textbook content was not explored in previous research. One reason for this could be that the research so far focused on the textbooks used in the US, where such a difference does not exist. In this study, we analyzed textbooks belonging to two school types in Germany - Gymnasium and Hauptschule. Earlier studies on textbook analysis did not look into the issue of possible differences between publishers. Our experiments showed that publisher is an important factor affecting the prediction of grade levels and school types for our dataset. This could be a potential factor to consider for English textbooks as well. Thus, this chapter contributes both to general problem of analyzing textbooks for linguistic complexity and to doing this analysis specifically for the German language.

To summarize, the primary research questions studied in this chapter are:

1. Can we automatically predict the grade level of a text using linguistic complexity features?

2. Can we automatically predict the school type?

3. What role does the publisher play in both the prediction tasks?

The rest of this chapter is organized as follows: Section 8.2 describes the corpus of German textbooks used in our experiments. Section 8.3 explain the ex-

perimental setup, features and analysis methods used in this research. Section 8.4 describe the text classification experiments and the results obtained and Section 8.5 describes the distribution of some of the features between categories of texts by grade, school and publisher. Section 8.6 concludes the chapter with pointers to future research.

8.2 Corpus

We used is a collection of Geography textbooks used in Baden-Württemberg state of Germany, in two types of German high schools - Gymnasium and Hauptschule as our corpus. While Gymnasium is the high school that typically leads to university education, Hauptschule trains the students for vocational education. It consisted of textbooks belonging to grades five to ten and produced by four publishers. In total, 35 books were scanned and edited using the OmniPage[3] Optical Character Recognition software.

Although the corpus consists of expository texts, instructions, exercises, definitions etc., we used only the expository texts from this corpus, to avoid genre influence on the modeling process. We chose textbooks belonging to only one subject in order to avoid the subject level differences. The subset of the corpus we used in the experiments reported in this paper consists of 2928 texts in total. Since some of the textbooks are intended to be used for two grades, we combined the grades into three groups - 5and6, 7and8, 9and10 (containing 1097, 958 and 873 texts respectively in total). Table 8.1 describes the corpus used in this chapter.

8.3 Features

We used a subset of the features described in Hancke (2013) for this task. Hancke (2013) studied the problem of proficiency classification of L2 German learners and implemented several lexical, semantic, syntactic and morphological properties of the texts, error rates, specific parse tree rules and tense patterns, and language model features for this task. Some of these features were also used

[3] https://en.wikipedia.org/wiki/OmniPage

Grade level	# texts in Gymnasium	# texts in Hauptschule	Total per publisher
Publisher A			
5and6	245	156	
7and8	146	223	1044
9and10	119	155	
Publisher B			
5and6	116	127	
7and8	147	70	627
9and10	108	59	
Publisher C			
5and6	202	136	
7and8	150	58	920
9and10	234	140	
Publisher D			
5and6	0	115	
7and8	0	164	337
9and10	0	58	
Total per school	1467	1461	2928

Table 8.1: The Reading Demands Corpus

for a German readability classification task earlier (Hancke et al., 2012b). We used all the measures of lexical diversity and variation, word frequency features from DlexDB database (Heister et al., 2011), lexical relatedness features from GermaNet (Henrich & Hinrichs, 2010), and syntactic and morphological features from that feature set for this task. In addition, we used a rule based Propositional Idea Density feature, based on the CPIDR system (Brown et al., 2007; Covington, 2007), implemented for German (Schulz, 2012). In all, our feature set consisted of 136 features. The implementation details of all the features are discussed in Hancke (2013), Chapter 5.

8.4 Classification Experiments

For text classification experiments, we used WEKA, as in earlier chapters. We report here results with SMO classification algorithm. All the reported results are obtained after 10-fold Cross Validation, and with balanced training sets, where all categories are equally represented. We compare the performance of features based on classification accuracy, as in earlier chapters.

8.4.1 Question 1: Grade-wise classification

Since this is a three-class classification problem with a balanced data set, we have a random baseline of 33%. We started with classification models using only traditional features (word length, num. syllables per word, sentence length), with and without considering school and publisher as features, which we consider as our baselines. A model with all the features reached the highest accuracy of 53.3%, which is a 7% improvement over the baseline with traditional features. Table 8.2 shows a summary of classification accuracies with various sub-groups of features.

Feature set	**# Features**	**Accuracy**
Random Baseline	0	33%
Without school, publisher as features		
Word length + Sentence Length + num. Syllables	3	46.4%
Syntactic Features	47	43.2%
Lexical Features	46	47.9%
Morphological Features	42	48.7%
Propositional Idea Density (PID)	1	35.2%
All	136	51.8%
With school, publisher as features		
Word length + Sentence Length + num. Syllables	5	46.3%
Syntactic Features	49	44.6%
Lexical Features	48	49.6%
Morphological Features	44	50.2%
Propositional Idea Density (PID)	3	39.7%
All	138	**53.3%**

Table 8.2: Grade Wise Classification

It is clear from this table that considering school and publisher as features while performing grade-wise classification seem to result in slightly better classification models for all the feature sets. However, an accuracy of 53.3% leaves

a lot of scope for improvement. To verify if the low performance is because of combining texts from Gymnasium and Hauptschule together, we trained separate grade-wise classification models for Gymnasium and Hauptschule texts. Table 8.3 and Table 8.4 show the summary of classification accuracies, with various feature sets, including publisher as an additional feature in all cases. Since we considered balanced training datasets, we had 443 instances per category for Gymnasium texts and 412 instances per category for Hauptschule texts.

Feature set	**# Features**	**Accuracy**
Random Baseline	0	33%
Word length + Sentence Length + num. Syllables	4	47.2%
Syntactic Features	48	46.2%
Lexical Features	47	51%
Morphological Features	43	50.3%
Propositional Idea Density (PID)	2	42.2%
All	137	**56.4%**

Table 8.3: Grade Wise Classification, Gymnasium texts

Feature set	**# Features**	**Accuracy**
Random Baseline	0	33%
Word length + Sentence Length + num. Syllables	4	45.6%
Syntactic Features	48	47.9%
Lexical Features	47	50.2%
Morphological Features	43	51.5%
Propositional Idea Density (PID)	2	43.9%
All	137	**54%**

Table 8.4: Grade Wise Classification, Hauptschule texts

As the tables indicate, performing grade-wise classifications separately for school types did not result in any drastic improvements in the results. There was a 3% improvement in classification accuracy for Gymnasium texts and less than 1% improvement for Hauptschule texts, compared to the combined dataset when the entire feature set was used. There are also small amounts of differences between feature groups.

While this may indicate no difference between the datasets in terms of grade wise distinctions between the features we used, performing a cross corpus evaluation between them may provide us more information regarding the differences between them. Hence, we performed two cross-corpus tests, using Gymnasium texts trained model to assess Hauptschule texts and vice-versa. Table 8.5 and Table 8.6 show the confusion matrices for these comparisons, for the models trained with all the features respectively.

classified as $->$	5and6	7and8	9and10
5and6	277	104	31
7and8	237	115	60
9and10	210	103	99

Table 8.5: Classifying Hauptschule Texts with Gymnasium Texts model

classified as $->$	5and6	7and8	9and10
5and6	107	126	210
7and8	93	105	245
9and10	72	47	324

Table 8.6: Classifying Gymnasium Texts with Hauptschule Texts model

It is interesting to observe from these two tables that most of the Hauptschule texts get classified in to the lowest grade level (5and6) followed by the middle level (7and8), when classified using the model built with Gymnasium texts. On the contrary, most of Gymnasium texts get classified in to the highest grade level (9and10) when classified using the model built with Hauptschule texts. Though the classification accuracies for the original models used for this cross corpus

evaluation are not very high, it is clear from both the confusion matrices that there are differences between the schools for grade-wise classification.

We did not explore grade-wise classification for specific publishers in detail due to the lack of sufficient training data for all the publishers, school types and grade combinations. Comparing in cases where a balanced training set consisted of more than 100 instances per category, Publisher A and Publisher C achieved classification accuracy of 55.2% and 58.4% respectively for Gymnasium texts. This may indicate a small amount of grade-wise differences between the publishers in terms of the feature set used.

Apart from a three-way classification between grades, we explored binary classification between grades to understand which of the grade-pairs are more easily distinguishable by our feature set. Table 8.7 shows the performance of binary grade wise classifications for both Gymnasium and Hauptschule texts.

Description	data: Gymnasium texts	data: Hauptschule texts
5and6 vs 7and8	64.5%	64.8%
7and8 vs 9and10	71.1%	68.2%
5and6 vs 9and10	76.6%	74.7%

Table 8.7: Binary Classification between grades

The binary classification accuracies are at least 10% higher than three-way classification. However, it is interesting to note that the classification accuracies for distinguishing between 5and6 and 9and10 are at least 5% higher than those for other classifications, for both Gymnasium and Hauptschule texts. This means that there is an ordinal trend between the classes, for the features used, which makes the prediction for the middle-class difficult.

To conclude the grade-wise classification experiments, our current models reached an accuracy of 56% for three-class classification and up to 76.6% for binary classification. While this leaves us with a lot of scope for improvement in terms of feature sets and modeling, this could also raise questions about the assumption that the textbooks possess clearly distinguishable differences between features that are assumed to contribute to text complexity, which needs to be explored in future.

8.4.2 Question 2: School wise classification

We next explored the second question about predicting the correct school type for a given text. This is a binary classification problem with two classes - Gymnasium (Gym) and Hauptschule (HS). Table 8.8 shows the classification accuracies for a balanced training set consisting of 1461 texts per category, using different subsets as features, with and without considering grade and publisher as features.

Feature set	**# Features**	**Accuracy**
Random Baseline	0	50%
Without grade, publisher as features		
Word length + Sentence Length + num. Syllables	3	61.5%
Syntactic Features	47	63.1%
Lexical Features	46	68.3%
Morphological Features	42	60.4%
Propositional Idea Density (PID)	1	51.1%
All	136	69.8%
With grade, publisher as features		
Word length + Sentence Length + num. Syllables	5	66.8%
Syntactic Features	49	69.1%
Lexical Features	48	71.7%
Morphological Features	44	69.3%
PID	3	62.4%
All	138	**74.5%**

Table 8.8: School Wise Classification

The school-wise classification model seems to be more affected by the addition of grade and publisher as features. For example, the accuracy of morphological features and PID increased by 8.9% and 11.3% respectively with the addition of grade and publisher as features. For grade-wise classifications the differences in accuracies between adding and not adding school and publisher as

features was only around 2-3% for any feature set. This may indicate that there are more differences between grades and publishers, when comparing Gymnasium and Hauptschule texts. So, we trained separate Gym versus HS binary classification models per grade first and per publisher next, to observe if there is a large variation between classification accuracies of the models. Table 8.9 shows the Gym versus HS classification by grade, with and without publisher as a feature.

Grade level	**Accuracy**
With publisher as a feature	
5and6	75.5%
7and8	77.7%
9and10	77.5%
Without publisher as a feature	
5and6	73.1%
7and8	67.3%
9and10	72.9%

Table 8.9: Gym vs HS classification, by grade level

We can observe from the figure that there is not a lot of performance difference between grades when we consider publisher as a feature. However, when we remove publisher as a feature, we see a slight drop in performance for 5and6, where as the drop is relatively high for 9and10 (4.6%) and higher (10.4%) for 7and8. This clearly indicates that there are large differences among publishers, in distinguishing between Gymnasium and Hauptschule texts, especially at the grade level 7and8.

Table 8.10 shows the comparison between three publishers, for Gym vs HS classification, with and without considering grade as a feature. Since the fourth publisher does not have texts for Gymnasium, we did not consider the publisher for this experiment. We considered equal number of instances per category while training the classification models, like in earlier experiments. So, for each classifier considered here, the random baseline is 50%.

Publisher	Accuracy with grade as feature	Accuracy without grade as feature
A	64%	63.9%
B	79.9%	79.9%
C	76%	75%

Table 8.10: Publisher differences for Gym versus HS classification

As the table shows, there are clear differences among publishers between the two school types. While a classifier trained with texts from publisher A achieved a classification accuracy of 63.9%, both publisher B and publisher C texts achieved accuracies that are more than 10% higher than publisher A.

Thus, for our second question about school wise differences, our experiments lead us to the conclusion that the differences between school types are strongly dependent on the information about the textbook publisher and that there are significant differences between publishers in terms of Gym vs HS classification.

8.5 Experiments with Individual Features

To understand the nature of the differences between grades/schools/publishers more specifically, we compared them in terms of individual features. We used Information Gain feature selection from WEKA, to choose the most predictive features for a given classification scenario, and studied the difference in the feature distribution between groups.

8.5.1 Differences among Grades

Table 8.11 shows the deviance explained by individual features, for grade-wise differences, considering schools and publishers separately, along with the statistical significance of individual features. The features listed in the table were chosen based on Information Gain, considering all the data, and after manually removing correlated features. We chose 5 features each from LEX, SYN and MORPH feature groups respectively.

Feature	Publisher A				Publisher B				Publisher C				Publisher D		All			
	Gym		HS		Gym		HS		Gym		HS		HS		Gym		HS	
	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.
Lexical-POS Features																		
LEX_avgWordLength	**	2.1%	***	5%	***	8.4%	***	7.7%	***	8.2%	***	7.9%	***	7.2%	***	5%	***	6.5%
LEX_adjectiveVariation	***	5.75%	n.s.	0.05%	***	5.6%	n.s.	1.5%	***	6.6%	**	2.9%	n.s.	0.9%	***	6.4%	**	0.7%
LEX_adverbVariation	**	2%	n.s.	0.8%	n.s.	0.4%	*	2.2%	n.s.	0.5%	n.s.	0.4%	***	6.3%	***	1%	***	1.4%
LEX_modifierVariation	**	1.4%	n.s.	0.4%	***	3%	n.s.	0%	***	3.4%	n.s.	1%	**	1.5%	***	2.6%	n.s.	0%
LEX_verbTokenRatio	n.s.	0.1%	n.s.	0%	**	2.3%	n.s.	0.2%	***	3.1%	***	3.6%	***	3.4%	***	1.4%	**	0.7%
Syntactic Features																		
SYN_avgNumNonTerminalsPerSentence	*	1%	***	3.4%	***	3.6%	**	3.6%	***	5%	***	4%	*	1.45%	***	2.8%	***	3%
SYN_avgSentenceLength	**	1.4%	***	3.7%	***	3.7%	***	4.3%	***	5%	***	4.9%	*	1.75%	***	3.1%	***	3.5%
SYN_avgParseTreeHeight	***	2.3%	***	2.4%	**	2.6%	**	4.15%	***	2.25%	***	6.6%	n.s.	0.8%	***	2.4%	***	3.3%
SYN_averagePPLengthInWords	***	4.1%	*	0.95%	***	3.8%	n.s.	0.6%	***	2.2%	**	2.6%	n.s.	0.9%	***	3.1%	***	1%
SYN_averageNPFrequency	**	1.75%	**	1.35%	**	2.6%	**	3.2%	***	4.3%	***	2.05%	*	1.8%	***	3.1%	***	1.9%
Morphological Features																		
MORPH_derivedNounsToNounsRatio	***	7.3%	***	5.4%	***	8.95%	**	4.4%	***	11%	***	8.1%	***	8.8%	***	8.8%	***	7%
MORPH_ungT	***	7.2%	***	5.7%	***	8%	**	3.5%	***	8.6%	***	9.2%	***	8%	***	7.1%	***	7.05%
MORPH_ionT	n.s.	0.15%	***	3.4%	*	1.7%	*	1.9%	***	3.9%	**	3%	***	6.5%	***	1.1%	***	3.8%
MORPH_entT	n.s.	0.45%	*	1%	***	5.6%	***	6%	***	2.1%	n.s.	0.05%	**	2.4%	***	2.4%	***	1.2%
MORPH_keitT	***	2.9%	n.s.	0.1%	*	1.55%	*	1.8%	**	1.2%	n.s.	0%	n.s.	0.04%	***	1.9%	n.s.	0%

Table 8.11: Grade Wise differences with single features - by school and by publisher

The differences between publishers can be seen more clearly here, in terms of some of the individual features. For example, for feature LEX_adjectiveVariation, all the publishers show significant differences among grade levels Gymnasium texts and more than 5% of the deviance is explained by the single feature for this school. However, only one publisher (C) shows significant differences between grades for Hauptschule texts with respect to this feature, explaining only 2.9% of the deviance. On the other hand, very small amount of deviance is explained by another feature LEX_adverbVariation, for publishers A, B and C, while it explains 6.3% of deviance for Publisher D, which has only Hauptschule texts.

In syntactic features, only one feature SYN_avgParseTreeHeight explained more than 5% of the deviance for Publisher C, for Hauptschule texts, while for the same publisher, it explained only 2.2% for the Gymnasium texts. More deviance is explained by morphological features MORPH_derivedNounsToNounsRatio and MORPH_ungT for all the publishers with some publishers showing larger differences between Gymnasium and Hauptschule, in terms of the deviance explained.

To conclude, this grade wise differences table shows that the some publishers show grade level differences explaining up to 11% of deviance for some of the features, and there are both between publisher and between school type differences for several of the features.

8.5.2 Differences between Schools

In this section, we explored the differences between school types, at each grade level, for each publisher. As before, the features were chosen based on Information Gain and after manually removing correlated features. We chose 5 features each from LEX, SYN and MORPH feature groups respectively.

Table 8.12 shows the differences between schools at each grade, for publisher A, along with the overall deviance explained by a feature considering texts from all grades together.

Feature	Publisher A							
	5and6		7and8		9and10		All	
	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.
LEX_textLengthBaseline	***	2.7%	n.s.	0%	n.s.	0%	*	0.5%

LEX_mtld	*	1.1%	n.s.	0.1%	n.s.	0.1%	n.s.	0%
LEX_squaredVerbVariation	***	2.9%	n.s.	0%	n.s	0.5	*	0.4%
LEX_adjectiveVariation	n.s	0%	n.s.	0.5%	***	8.55%	*	0.3%
LEX_avgWordLength	n.s	0.6%	n.s.	0%	***	3.8%	*	0.4%
SYN_avgNumNonTerminalsPerSentence	*	0.9%	***	2.6%	n.s.	0%	**	0.5
SYN_avgSentenceLength	n.s.	0.6%	**	1.7%	n.s.	0.1%	*	0.4
SYN_longestDependency	**	2%	n.s.	0	n.s.	0	n.s.	0.2%
SYN_averageNPFrequency	n.s.	0	*	1.2%	n.s.	0.2%	n.s.	0.2%
SYN_averagePPFrequency	*	0.9%	*	1.2%	*	0.7%	**	0.8%
MORPH_averageCompoundDepth	n.s	0%	n.s	0.1%	n.s.	0%	n.s.	0.1%
MORPH_compoundNounsToNounsRatio	n.s	0%	***	5.1%	n.s.	0.7%	***	1.3%
MORPH_derivedNounsToNounsRatio	n.s	0.4%	n.s.	0.3%	**	2.6%	*	0.4%
MORPH_ungT	***	3.05%	n.s.	0.7%	***	7.2%	***	2.1%
MORPH_aturT	n.s	0.6%	n.s.	0%	***	6.95%	n.s.	0%

Table 8.12: School wise differences at each grade, for Publisher A

While most of the features chosen for this publisher show little differences at all the grades in terms of deviance explained, some of the features, explain higher deviance for between school differences, at specific grade levels. For example, LEX_adjectiveVariation explains 8.55% of the deviance at grades 9and10 but less than 1% of deviance at other grade levels. Similarly, MORPH_aturT and MORPH_compoundNounsToNounsRatio explain 6.95% and 5.1% of deviance at 9and10 and 7and8 respectively, while explaining less than 1% deviance at other grade levels. However, on the whole, Publisher A appears to make less distinctions between school types, at all levels, in terms of the features we considered.

Table 8.13 shows the differences between schools at each grade, for publisher A, along with the overall deviance explained by a feature considering texts from all grades together.

Feature	Publisher B							
	5and6		7and8		9and10		All	
	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.
LEX_textLengthBaseline	***	23.1%	n.s.	0.6%	n.s	1.35%	***	5.7%
LEX_mtld	***	4.4%	*	1.8%	n.s.	0.1%	***	2.1%
LEX_squaredVerbVariation	***	19.2%	n.s.	0.2%	n.s	0	***	3.45%
LEX_adjectiveVariation	n.s.	1.2%	***	5.7%	***	6.3%	***	4.9%
LEX_avgWordLength	**	2.5%	n.s.	1.4%	**	4.6%	***	3%
SYN_avgNumNonTerminalsPerSentence	***	14.4%	***	37.7%	***	10.2%	***	20.3

SYN_avgSentenceLength	***	16%	***	30.6%	***	10.2%	***	19%
SYN_longestDependency	***	7.3%	***	8.1%	**	3.8%	***	7.1%
SYN_averageNPFrequency	***	5.9%	***	17.7%	**	4.3	***	9.8%
SYN_averagePPFrequency	***	9%	***	14.1%	***	6.2%	***	9.9%
MORPH_averageCompoundDepth	n.s	1.4%	n.s.	0.1%	n.s.	0%	n.s	0.2%
MORPH_compoundNounsToNounsRatio	n.s	0.9%	n.s.	0%	n.s.	0.4%	***	0.3%
MORPH_derivedNounsToNounsRatio	*	1.7%	***	12.4%	**	5.15%	***	6.3%
MORPH_ungT	n.s	0.7%	***	10.5%	*	2.8%	***	4.1%
MORPH_aturT	n.s	0.3%	n.s.	0.1%	*	3.3%	n.s.	0.5%

Table 8.13: School wise differences at each grade, for Publisher B

Compared to Publisher A, Publisher B has several features that explain more deviance, and show differences at each grade level, between school types. For example, LEX_textLengthBaseline and LEX_adjectiveVariation explain 23.1% and 19.2% of deviance respectively for between school differences, at grade 5and6, while not showing any significant differences between schools at other grade levels. SYN_avgSentenceLength and SYN_avgNumNonTerminalsPerSentence explain more than 10% of deviance at each grade level while explaining 37.7% and 30.6% respectively in the grade 7and8. Similarly, more than 10% of deviance is explained at grade 7and8 for two morphological features that encode the use of compound nouns and derived nouns: MORPH_compoundNounsToNounsRatio and MORPH_derivedNounsToNounsRatio. This shows that this publisher makes distinctions between schools for several features, and sometimes the distinctions are clearly seen between grades too, with more differences seen at grade 7and8.

Table 8.14 shows the differences between schools at each grade, for publisher A, along with the overall deviance explained by a feature considering texts from all grades together.

Feature	Publisher C							
	5and6		7and8		9and10		All	
	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.	Sig.	Dev.
LEX_textLengthBaseline	***	15.6%	n.s.	0%	**	1.7%	***	3.9%
LEX_mtld	***	12.6%	n.s.	1%	***	4.4%	***	6.3%
LEX_squaredVerbVariation	***	8.5%	n.s.	0.1%	n.s	0.2	***	1.6%
LEX_adjectiveVariation	***	6.1%	***	9.6%	***	7.9	***	7.1%
LEX_avgWordLength	***	4.95%	**	3.9%	***	4.4%	***	4.05%
SYN_avgNumNonTerminalsPerSentence	***	11.4%	***	16.3%	***	10.5%	***	11.2%

SYN_avgSentenceLength	***	12.9%	***	16%	***	8.1%	***	10.2%
SYN_longestDependency	***	11.7%	*	2.9%	***	5.1%	***	6.9%
SYN_averageNPFrequency	***	6.4%	***	8.4%	***	9.6%	***	8%
SYN_averagePPFrequency	***	8.2%	***	8%	***	4.6%	***	6.15%
MORPH_averageCompoundDepth	*	1.1%	n.s.	0%	**	2.2%	***	1.2%
MORPH_compoundNounsToNounsRatio	n.s	0.6%	*	2.9%	***	6.1%	***	2.9%
MORPH_derivedNounsToNounsRatio	n.s	0%	n.s.	1.5%	n.s.	0.2%	n.s.	0.3%
MORPH_ungT	*	1%	n.s.	0.7%	n.s.	0.3%	*	0.5%
MORPH_aturT	n.s	0.1%	n.s.	0.3%	n.s.	0%	n.s	0.05%

Table 8.14: School wise differences at each grade, for Publisher C

Publisher C makes more distinctions between school types at the lowest grade for three lexical richness features - LEX_textLengthBaseline, LEX_mtld and LEX_squaredVerbVariation. At other grades, these features do not seem to explain any larger deviance for between schools comparison. All the 5 SYN features considered here indicate significant differences between schools, at all grade levels. Among the MORPH features, only MORPH_compoundNounsToNounsRatio explains 6.1% of the deviance, at the highest grade level 9and10. These results indicate that the publisher C makes more distinctions between LEX features at lower levels, SYN features at all levels and MORPH features do not play an important role for this publisher. To summarize, these tables show that the publishers display large differences between school types than grade levels for some of the individual features.

8.6 Conclusions

In this chapter, we analyzed German geography textbooks in terms of linguistic complexity features. We built prediction models for classifying texts by their grades and school types. The features we included in the analysis encoded lexical, POS, syntactic, morphological properties of language and propositional idea density.

In terms of the grade-wise classification, we achieved a maximum prediction accuracy of 53.3% (56.4% for Gymnasium texts alone and 54% for Hauptschule texts alone), for a three-way classification into grades: 5and6, 7and8 and 9and10.

The ability to separate between 5and6 and 9and10 was >10% higher than the classification accuracy for 5and6 and 7and8 and >5% higher than 7and8 vs 9and10 classification. These results show that there are differences between grades in terms of the features used but it is not sufficient to explain all the variance between grades.

For school-wise classification, we achieved a binary classification accuracy of 74.5% for Gymnasium versus Hauptschule classification. While distinguishing between the schools at individual grade levels was slightly larger (75.5% for 5and6, 77.7% for 7and8 and 77.5% for 9and10 respectively), the classification results for this task were strongly influenced by the publisher of the text. Our models classified texts from Publisher A into Gymnasium or Hauptschule correctly with 64% accuracy. But for the same task, with texts from Publisher B and Publisher C, we achieved classification accuracies of 79.9% and 76%, which are significantly better than the accuracy with Publisher A. This clearly shows the differences between publishers in terms of writing texts for Gymnasium and Hauptschule.

Finally, we explored the specific contributions of some of the features, in terms of the variance explained by them for distinguishing between grades and schools. We selected 15 features (5 LEX, 5 SYN, 5 MORPH) each for grade and school-wise differences and studied the differences between publishers at each grade level and school type. Our results indicated that the publishers make more distinctions between school types than between grade levels.

While the low prediction accuracies for grade-wise classification may indicate that the features are not capturing the existing differences in linguistic complexity between grades, it has to be noted that some of these features are known to correlate with text complexity in English (e.g., adjective variation) and some of the described features were successfully used in the past for German readability classification in Hancke et al. (2012b) for classifying between texts written for children and adults and for proficiency classification of L2 German learners. The school wise classification models and the experiments with individual features also showed that the features explain more variation between school types than the grade levels. All these facts may lead us to a conclusion that linguistic complexity is perhaps implicitly considered while preparing texts, at a school level rather than grade level, and the factors considered differ from publisher to

publisher.

8.6.1 Outlook

Adding features that encode other properties of the text like cohesion and coherence will be the next immediate direction to pursue. It would also be interesting to do a comparative study with textbooks from other languages, using a comparable feature set that encodes lexical, syntactic and morphological aspects of the language.

Part III

Conclusions

Chapter 9

Conclusions and Future Work

Readability assessment and text simplification are useful in providing relevant, comprehensible texts for language learners. In this thesis, we proposed computational approaches for both these tasks based on linguistic modeling. We studied the problem of automatic readability assessment of texts and explored its usefulness at the sentence level to compare the degree of simplification between manually simplified sentences. In the context of automatic text simplification, we investigated the role of training data and language models in modeling text simplification as statistical machine translation.

To understand the effect of text complexity on readers and the cognitive correlates of linguistic complexity, we performed an eye-tracking study with L2 English readers where they read texts manually written in two versions differing in text complexity. This experiment showed that the eye-tracking measures and the reading outcomes are influenced by both text complexity and language proficiency.

Beyond readability, we used the feature set we developed for readability assessment to another educational application - assessing the language proficiency of L2 English writing. Finally, we applied an existing German text complexity feature set from Hancke (2013) for analyzing the Geography textbooks used in German schools at different grades and school types.

The rest of this chapter will describe the conclusions and contributions of this thesis in better detail: Section 9.1 presents chapter wise conclusions from the

thesis. Section 9.2 lists the specific contributions of this thesis in terms of research results and the resources created. Section 9.3 discusses the limitations of this thesis and Section 9.4 presents some ideas to overcome these limitations, pointing to future research directions.

9.1 Summary

9.1.1 Readability Assessment of Texts

In Chapter 3, we modeled text complexity based on a rich feature set comprising of 150 features that consisted of measures of text complexity derived from Second Language Acquisition (SLA) and psycholinguistics research and several part of speech tag and syntactic parse tree based features. During this process, we compiled a new corpus of texts annotated by their grade-level, called WeeBit, by combining two online sources of graded texts - WeeklyReader and BBC-BiteSize. We followed a supervised machine learning approach and modeled readability assessment as regression, for developing a readability model based on this corpus and our feature set. Apart from testing the internal validity of the model by means of cross-validation, we also established its generalizability, by testing on multiple existing readability annotated sets. With cross-validation, our model achieved an average correlation of 0.9. For predicting the grade level of texts in the common core standards exemplars data set, which has become a standard test set in the contemporary readability assessment work, our readability model achieved a rank correlation of 0.69, which is the second best reported result after ETS' SourceRater so far (as reported in Nelson et al. (2012)).

We established the generalizability of the feature set and studied the genre and topic dependence of our models. Since the primary dataset we used consisted of texts of informative nature like news articles, our model predictions worked well for texts of that genre and performed poorly with spoken text like transcribed speech. To overcome this issue, we explored the possibility of creating topic and genre specific models using existing, annotated corpora. Our genre specific readability model trained on speech texts from a corpus of television subtitles achieved 96% classification accuracy for classifying texts into three age groups.

We also studied the effect of text size on classification accuracy for this corpus. Our results showed that while reducing the text sample size resulted in a drop in accuracy, the drop is less for a model that considers all features instead of only certain categories of them. Finally, we briefly investigated topic specific readability models, constructed based on the TASA corpus and all topic based models achieved correlations of above 0.8.

9.1.2 Effect of Text Complexity on Readers

In Chapter 4, we described an eye-tracking experiment for studying the effect of text complexity on the online processing of the users and their performance outcomes. Along with text complexity, we studied the influence the readers' English language proficiency on these measures. We analyzed the data using linear regression and Generalized Additive Mixed Models and investigated if there is a possible mediation effect of the processing measures on performance outcomes.

Among the eye tracking measures we studied, fixation count, second pass duration and number of revisits were significantly affected by text complexity, decreasing with an increase in text complexity. Reader's language proficiency was a significant predictor for average fixation duration and first pass duration. There was an interaction between text complexity and language proficiency in the case of first pass duration. In terms of the outcome measures, text complexity was a significant predictor only for recall scores while the reader's language proficiency was a strong predictor for both recall and comprehension scores. There was an interaction between proficiency and text complexity for recall scores.

The order in which the subjects read the texts had a significant effect for four of the six eye-tracking measures and the recall scores in the outcome measures. The random effect due to participant variation was significant for seven of the eight dependent variables studied, which indicates the importance of considering it in the modeling process. The random effect due to text variation was significant for fixation count, average fixation duration and revisits among the eye-tracking measures and for recall scores in the outcome measures. Finally, there was no mediation of the processing measures on the differences in the performance outcomes.

9.1.3 Readability at the sentence level

In Chapter 5, we explored the problem of assessing the readability of sentences. We showed that a pairwise ranking approach performs better than classification or regression in distinguishing between sentences based on their reading level. Our approach ranks the sentences in terms of their reading level correctly with an accuracy of over 80%, the best accuracy for this task we are aware of. We performed in-corpus and cross-corpus evaluations to establish the generalizability of the approach. During this process, we created a resource of sentence-aligned simplified texts based on onestopenglish.com texts created by experts for English as second language learners. The corpus consists of parallel, simplified versions of sentences belonging to three reading levels. This approach and the corpus will be useful for the evaluation of text-simplification systems for language learners in real life educational settings. Though we primarily studied this problem to compare versions of the same sentence, the approach is equally applicable in choosing the target sentences for simplification of a text after ranking them based on their complexity.

We briefly explored the idea of applying the sentence level, ranking based readability model to estimate text level readability. We hypothesized that this approach to go from local (sentence level) readability estimates to global (document level) estimates will enable us to get the distribution of linguistic complexity within a document. An initial testing of the model with common core standards exemplar texts achieved a rank correlation of 0.51, which is comparable to some of the existing commercial and academic systems (Lexile = 0.5, DRP = 0.53, REAP = 0.54) on this dataset. This is a direction that needs to be explored in detail in the future, as it can potentially result in more fine-grained estimates of readability, since it is trained on parallel, sentence level simplified data instead of full texts which potentially have sentences of varying difficulty.

9.1.4 Automatic Text Simplification

In Chapter 6, we developed an approach to perform automatic text simplification, which handles lexical simplification and paraphrasing, by modeling it as phrase based machine translation problem. We aimed at simplifying in form while re-

taining the meaning, without deleting content. Hence, we considered a training subset consisting of only a restricted set of simplification operations instead of using the entire simplification dataset comprising of a lot of deletion decisions. This resulted in a better simplification, in terms of the BLEU scores, increasing the score from 71.8 to 95.7 for the Wikipedia data.

We performed cross-corpus evaluation of this approach using the OneStopEnglish sentences and our results showed that the phrase based machine translation approach did not transfer to the new corpus. To our knowledge, this is the first cross-corpus evaluation of an automatic text simplification approach. Finally, We explored the utility of different language models for improving the cross corpus performance and our experiments showed that changing the language model did not improve the BLEU scores for a cross-corpus setup.

9.1.5 Readability Features for L2 Proficiency Classification

In Chapter 7, we used the readability features we developed in Chapter 3, for the task of assessing L2 learner English writing. We collected four publicly accessible L2 learner essay data sets annotated with proficiency levels based on some standardized criterion. All the datasets were created by various external sources. We constructed classification and regression based proficiency assessment models on these data sets.

This approach performed accurate proficiency classification with two of the datasets (BUiD and TOEFL11), achieved significant but less than previously reported results for the third dataset (FCE) and was not very successful with the fourth data set (ICNALE). The differences in results may either indicate the differences among the datasets in terms of the notion of proficiency or the insufficiency of the complexity features to generalize, which needs to be explored in future. It should be noted that our proficiency assessment models relied on a feature set that does not consider typically used features to model learner data like n-gram models and error patterns of learners. We also did not attempt to pre-process the learner text for spelling/grammar errors. So, using this feature set in conjunction with learner error features and other features typically used in essay scoring systems could result in improved prediction accuracies for proficiency classification.

9.1.6 Analyzing Linguistic Complexity of German School books

In Chapter 8, we applied the readability analysis approach we followed to another language, German. We analyzed German geography textbooks using the linguistic complexity features developed by Hancke (2013) for German proficiency classification. These features encode lexical, POS, syntactic and morphological properties of the language. We built prediction models for classifying texts by their grades and school types.

We achieved a maximum prediction accuracy of 53.3%, for a three way grade-wise classification (5and6, 7and8 and 9and10). For classifying between Gymnasium and Hauptschule texts, we achieved a classification accuracy of 74.5%. However, this accuracy differed based on the publisher, ranging from 64% to 79.9% among three publishers. We also explored the specific contributions of some of the features for distinguishing between grades and schools, in terms of the variance explained. We selected 15 features (5 lexical, 5 syntactic, 5 morphological) each for grade and school-wise differences and studied the differences between publishers at each grade level and school type. Our results indicated that the publishers make more distinctions between school types than between grade levels. This study leads to a conclusion that the linguistic complexity is perhaps implicitly considered while preparing texts, at a school level rather than grade level, and the factors considered differ from publisher to publisher.

9.2 Contributions

The specific contributions of this thesis, in terms of the research and the resources created are listed below:

9.2.1 Research

1. In this thesis, we showed that an integrated approach combining parse tree and POS based features with those derived from research in SLA and Psycholinguistics can be useful for accurately estimating the readability of a text in terms of a given grading rubric. Methodologically speaking, we es-

tablished the validity of the approach by performing multiple cross-corpus evaluations, which was not performed as extensively in the past research on this topic.

2. We conducted an eye-tracking study to understand the cognitive correlates of linguistic complexity and how text complexity and language proficiency affect the performance outcomes of the readers. To our knowledge, this is the first approach to study the relationship of text complexity and reader's language proficiency with eye-tracking and performance outcome variables, all considered together.

3. We explored readability modeling at the sentence level and studied its utility for assessing the degree of text simplification and briefly studied its use for a fine-grained estimation of document level readability. We studied sentence readability as a pair-wise ranking problem, which resulted in more than 80% accuracy for the task. To our knowledge, sentence level readability was not modeled as pair-wise ranking before. We also performed cross-corpus evaluations to ensure the generalizability of the approach. Previous research on this topic did not look into validating the results through cross-corpus comparisons.

4. In modeling automatic text simplification as machine translation, we studied the role of training corpora and language models for improving the quality of simplification performed and also evaluated the generalizability of the approach by means of cross-corpus evaluation. While the use of language models was explored before, the other two aspects were not studied much in the past research on automatic simplification.

5. We showed that the feature set we developed for readability assessment can be useful for L2 proficiency classification as well, by studying four externally created L2 English writing datasets.

6. We proposed an approach to evaluate the content of German school textbooks in terms of their linguistic complexity, as a step towards developing better materials for students, by taking their grade level into account.

9.2.2 Resources

In this thesis, we created the following corpus resources[1]:

1. the WeeBit corpus, consisting of texts annotated with 5 levels based on age-group, for first language English learners, with more than 600 texts per level.

2. the OneStopEnglish corpora: OSE2 consisting of ∼3000 pairs of parallel sentences belonging to two reading levels and OSE3, consisting of ∼800 triplets of parallel sentences belonging to three reading levels.

9.3 Limitations

In this thesis, readability is modeled only considering the features based on word-level part-of-speech, morphological and psycholinguistic information and syntactic parse trees into account. Other useful features such as frequency of occurrence of words/phrases, idiomatic expressions, n-gram patterns of words or POS tags are not taken into account in this feature set. The discourse structure of the text was not considered at all.

Further, the approach described in this thesis modeled readability considering the text's linguistic properties alone. This approach ignores the role of the user-based variables like language proficiency, topical interest, motivational levels and other factors. The interaction of a user with a text is also based on the context and the task in which reading and comprehension of the text are needed. Reading a text for answering specific questions about it could be different from reading it for getting an overview about a topic, for example. The current version of our model does not account for these differences between users and for different tasks.

With regard to text simplification, the model considers a very limited set of simplification operations, whereas text simplification in reality can involve many more operations, including deletion. Finally, the model seems to work well with

[1]The code for extracting all the relevant features for texts can be shared for research use. We are currently working on a web-interface that provides a reading level estimate and details about the individual features, for various text documents and urls, based on our readability model.

wikipedia texts, on which it is trained but fails to work on new texts from other sources like onestopenglish.com.

9.4 Outlook

As mentioned earlier, our current readability assessment approach does not take discourse aspects and frequency of word usage or syntactic patterns into account. Future enhancements to the feature set can include these aspects. Approaches that integrate user modeling and task specificity into readability models could be useful in suggesting appropriate reading materials to the learners, considering their proficiency, topical interest, motivation levels and other background information.

Our eye-tracking experiment showed that text complexity has an effect on some of the online-processing and outcome variables while learner proficiency plays a more important role with outcome variables. Using the data from this experiment as a basis to combine a user model (i.e., proficiency score), task specificity (i.e., recall/comprehension performance) with text complexity may be useful in moving towards a real-life application to provide appropriate reading materials for learners, that takes the text, the user and their interaction into account. Identifying specific linguistic variables that correlate with the online processing and performance measures and understanding the relation between the nature of the simplification performed (e.g., lexical, syntactic or semantic) and the dependent variables are interesting problems to pursue.

We studied sentence level readability by comparing human simplifications. The next step would be to apply this approach to the output of automatic text simplification systems. In the long term, it would be interesting to extend this to include fluency and grammaticality aspects of texts to develop a full framework of evaluating automatic text simplification.

In automatic text simplification, we followed a conservative approach to handle only a small subset of possible simplifications. Adding other transformations like sentence splitting could be an immediate extension to this part. Since phrase based machine translation does not seem to generalize to new data, it would be useful to explore the usefulness of syntactic translation for text simplification. Extracting rules from phrase-structure trees, or combining rules with statistical

approaches need to be explored in future to develop a practical, usable text simplification system that can either assist writers to prepare simplified texts or generate grammatical, fluent and readable text for language learners.

Zusammenfassung

Sowohl in der kindlichen Entwicklung als auch im Erwachsenenalter spielt das Lesen eine zentrale Rolle für den Lernprozess und Wissenserwerb. Allerdings sind nicht alle Texte jedem potentiellen Leser zugänglich. So können beispielsweise Leseschwierigkeiten auftreten, wenn die Sprachkenntnis eines Lesers nicht der linguistischen Komplexität des Texts entspricht. In solchen Fällen kann eine Vereinfachung des Textes hinsichtlich der linguistischen Form unter Beibehaltung des Inhalts zur Verständlichkeit für den Leser beitragen. Die vorliegende Dissertation behandelt die Bewertung der Lesbarkeit und der Vereinfachung von Texten aus computerlinguistischer Perspektive.

Die Ergebnisse der Arbeit zeigen, dass sich die Lesbarkeit eines Textes durch einen integrativen Ansatz recht genau vorhersagen lässt, wenn sowohl linguistische Features als auch Erkenntnisse aus dem Zweitspracherwerb und der Psycholinguistik anhand gegebener Bewertungskriterien berüksichtigt werden. Aus methodischer Sicht wird die Gültigkeit des Ansatzes durch mehrere Evaluierungen über verschiedene Korpora hinweg nachgewiesen. Die vorliegende Arbeit modelliert ferner die Textlesbarkeit auf der Satzebene und untersucht die Nützlichkeit solcher Modelle für den Grad der Textvereinfachung. Dabei wird auch die Verwendung der Satzlesbarkeit für eine detaillierte Einschätzung der Lesbarkeit auf der Dokumentebene kurz untersucht. Weiterhin werden korpusübergreifende Evaluierungen durchgeführt, um die Generalisierbarkeit des Ansatzes zu gewährleisten. In einer Eye-Tracking-Studie wurde darüberhinaus gezeigt, dass sowohl die Textkomplexität als auch die Sprachkenntnisse des Lesers die Echtzeitverarbeitung und Leistungsergebnisse beeinflussen. In dieser Arbeit wird die automatische Textvereinfachung als maschinelles Übersetzungsproblem behandelt. Hierfür wird untersucht, inwiefern Language Models und die verwendeten Trainingskorpora die Textvereinfachung qualitiativ optimieren können. Hierbei wurde wiederum die Generalisierbarkeit durch korpusübergreifende Evaluierung sichergestellt.

Über die Bewertung der Lesbarkeit hinausgehend wird gezeigt, dass die entwickelten Features auch für die Klassifikation von Texten in verschiedene Zweitsprachniveaus nützlich sein können. Dies wird durch den Vergleich von vier ex-

tern erstellten, geschriebenen L2-Datensätzen demonstriert. Schließlich wird ein Ansatz vorgeschlagen, in dem der Inhalt deutscher Schulbücher hinsichtlich ihrer linguistischen Komplexität evaluiert wird, um künftig dazu beizutragen, bessere Lernmaterialien zu entwickeln, die auch die Klassenstufe der Schüler berücksichtigen.

Bibliography

Abrahamsson, E., T. Forni, M. Skeppstedt & M. Kvist (2014). 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 (PITR)*. ACL, pp. 57–65.

Agrawal, R., S. Chakraborty, S. Gollapudi, A. Kannan & K. Kenthapudi (2012). Empowering Authors to Diagnose Comprehension Burden in Textbooks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 967–975.

Aitchison, J. (2011). *The Articulate Mammal: An Introduction to Psycholinguistics*. Routledge Classics.

Allen, D. (2009a). A study of the role of relative clauses in the simplification of news texts for learners of English. *System* 37(4), 58–599.

Allen, D. (2009b). Using a corpus of simplified news texts to investigate features of the intuitive approach to simplification. In *Proceedings of the Corpus Linguistics Conference 2009*.

Aluisio, S., L. Specia, C. Gasperin & C. Scarton (2010). Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*. pp. 1–9.

Aluísio, S. M., L. Specia, T. A. Pardo, E. G. Maziero & R. P. Fortes (2008). Towards Brazilian Portuguese automatic text simplification systems. In *Pro-*

ceeding of the eighth ACM symposium on Document engineering. New York, NY, USA, DocEng '08, pp. 240–248.

Amancio, M. & L. Specia (2014). An Analysis of Crowdsourced Text Simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. ACL, pp. 123–130.

Amoia, M. & M. Romanelli (2012). SB: mmSystem - Using Decompositional Semantics for Lexical Simplification. In *In Proceedings of First Joint Conference on Lexical and Computational Semantics (SEM)*. Association for Computational Linguistics, pp. 482–486.

Anderson, J. (1983). Lix and Rix: Variations on a Little-Known Readability Index. *Journal of Reading* 26(6), 490–496.

Anderson, R. C. & A. Davison (1988). *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, Lawrence Erlbaum Associates, chap. 2 Conceptual and Empirical Basis of Readability Formulae, pp. 23–53.

Angrosh, M. & A. Siddharthan (2014). Text simplification using synchronous dependency grammars: Generalising automatically harvested rules. In *Proceedings of the 8th International Natural Language Generation Conference*. Philadelphia, Pennsylvania: Assoc. for Computational Linguistics, pp. 16–25.

Aranzabe, M. J., A. D. de Ilarraza & I. Gonzalez-Dios (2012a). First Approach to Automatic Text Simplification in Basque. In *Proceedings of the First worshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*. pp. 1–8.

Aranzabe, M. J., A. D. de Ilarraza & I. Gonzalez-Dios (2012b). Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. In *SEPLN Journal*. pp. 61–68.

Aukerman, R. C. (1965). Readability of Secondary School Literature Textbooks: A First Report. *The English Journal* 54(6), 533–540.

Baayen, R. H., R. Piepenbrock & L. Gulikers (1995). The CELEX Lexical Databases.

Bach, N., Q. Gao, S. Vogel & A. Waibel (2011). TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, pp. 474–482.

Barlacchi, G. & S. Tonelli (2013). ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In *14th International Conference on Computational Linguistics and Intelligent Text Processing, (CICLing)*. Springer, pp. 476–487.

Barzilay, R. & N. Elhadad (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, EMNLP '03, pp. 25–32.

Begeny, J. C. & D. J. Greene (2013). Can Readability Formulas be successfully used to gauge the difficulty of reading materials? *Psychology in the Schools* 51(2), 1520–6807.

Beinborn, L., T. Zesch & I. Gurevych (2012). Towards fine-grained readability measures for self-directed language learning. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*. pp. 11–19.

Bendersky, M., W. B. Croft & Y. Diao (2011). Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*. New York, NY, USA: ACM, WSDM '11, pp. 95–104.

Bennöhr, J. (2005). A Web-based Personalised Textfinder for Language Learners. Master's thesis, School of Informatics, University of Edinburgh.

Biran, O., S. Brody & N. Elhadad (2011). Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 496–501.

Blake, C., J. Kampov, A. K. Orphanides, D. West & C. Lown (2007). UNC-CH at DUC 2007: Query Expansion, Lexical Simplification and Sentence Selection Strategies for Multi-Document Summarization. In *Proceedings of Document Understanding Conference*. URL `http://duc.nist.gov/pubs/2007papers/unc-ch.blake.final.pdf`.

Blanchard, D., J. Tetreault, D. Higgins, A. Cahill & M. Chodorow (2013). *TOEFL11: A Corpus of Non-Native English*. Tech. rep., Educational Testing Service.

Bormuth, J. R. (1966). Readability: A New Approach. *Reading Research Quarterly* 1(3), 79–132.

Boston, M. F., J. T. Hale, U. Patil, R. Kliegl & S. Vasishth (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research* 2(1), 1–12.

Boston, M. F., J. T. Hale, S. Vasishth & R. Kliegl (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes* 26(3), 301–349.

Bott, S., L. Rello, B. Drndarevic & H. Saggion (2012). Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *In Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. pp. 357–374.

Bott, S. & H. Saggion (2011a). Spanish Text Simplification: An Exploratory Study. In *27th Conference of the Spanish Society for Natural Language Processing*. pp. 87–95.

Bott, S. & H. Saggion (2011b). An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *ACL Workshop on Monolingual Text-to-Text Generation*. pp. 20–26.

Box, G. E. P. & D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society* 26(2), 211–252.

Britton, B. K., L. V. Dusen, S. Gulgöz & S. M. Glynn (1989). Instructional Texts Rewritten by Five Expert Teams: Revisions and Retention Improvements. *Journal of Educational Psychology* 81, 226–239.

Britton, B. K. & S. Gülgöz (1991). Using Kintsch's Computational Model to Improve Instructional Text: Effects of Repairing Inference Calls on Recall and Cognitive Structures. *Journal of Educational Psychology* 83, 329–345.

Broda, B., M. Ogrodniczuk, B. Nitoń & W. Gruszczyński (2014). Measuring Readability of Polish Texts: Baseline Experiments. In *Proceedings of The 9th edition of the Language Resources and Evaluation Conference (LREC)*. URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/427_Paper.pdf`.

Brooke, J., V. Tsang, D. Jacob, F. Shein & G. Hirst (2012). Building Readability Lexicons with Unannotated Corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Montréal, Canada: Association for Computational Linguistics, pp. 33–39.

Brouwers, L., D. Bernhard, A.-L. Ligozat & T. Francois (2014). Syntactic Sentence Simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. ACL, pp. 47–56.

Brown, C., T. Snodgrass, M. A. Covington, R. Herman & S. J. Kemper (2007). Measuring propositional idea density through part-of-speech tagging. poster presented at Linguistic Society of America Annual Meeting, Anaheim, California.

Bruce, B. & A. Rubin (1988). *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, Lawrence Erlbaum Associates, chap. 1 Readability Formuals: Matching Tool and Task, pp. 5–22.

Bulté, B. & A. Housen (2014). Conceptualizing and measuring short-term changes in {L2} writing complexity. *Journal of Second Language Writing* 26(0), 42 – 65. Comparing perspectives on {L2} writing: Multiple analyses of a common corpus.

Burges, C., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton & G. Hullender (2005). Learning to rank using gradient descent. In *Proceedings of the International Conference on Machine Learning*. pp. 89–96.

Burstein, J. & M. Chodorow (2010). *Progress and New Directions in Technology for Automated Essay Evaluation*, Oxford University Press, chap. 36, pp. 487–497. 2nd ed.

Burstein, J., M. Chodorow & C. Leacock (2003). Criterion: Online Essay Evaluation: An Application for Automated Evaluation of Student Essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-03)*. Acapulco, Mexico, pp. 3–10.

Candido, Jr., A., E. Maziero, C. Gasperin, T. A. S. Pardo, L. Specia & S. M. Aluisio (2009). Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg, PA, USA, EdAppsNLP '09, pp. 34–42.

Canning, Y. & J. Tait (1999). Syntactic Simplification of Newspaper Text for Aphasic Readers. In *Proceedings of SIGIR-99 Workshop on Customised Information Delivery*. pp. 6–11.

Canning, Y., J. Tait, J. Archibald & R. Crawley (2000). Cohesive Generation of Syntactically Simplified Newspaper Text. In *Third International Workshop on Text, Speech and Dialogue, TSD 2000, Brno, Czech Republic, September 13-16, 2000*. Springer, pp. 145–150.

Carroll, J., G. Minnen, Y. Canning, S. Devlin & J. Tait (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*. Madison, Wisconsin: Association for the Advancement of Artificial Intelligence (AAAI), pp. 7–10.

Carroll, J., G. Minnen, D. Pearce, Y. Canning, S. Devlin & J. Tait (1999). Simplifying Text for Language-Impaired Readers. In *Proceedings of the 9th Confer-*

ence of the European Chapter of the Association for Computational Linguistics (EACL). pp. 269–270.

Caseli, H., T. Pereira, L. Specia, T. Pardo, C. Gasperin & S. Aluísio (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics. Research in Computing Science (CICLING 2009 Proceedings)* 41, 59–70.

CCSSO (2010a). Common Core State Standards. National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.

CCSSO (2010b). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects. Appendix B: Text Exemplars and Sample Performance Tasks*. Tech. rep., National Governors Association Center for Best Practices, Council of Chief State School Officers.

Chae, J. & A. Nenkova (2009). Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the ACL*. pp. 139–147.

Chall, J. S. & E. Dale (1995). *Readability Revisted: The New Dale-Chall Readability Formula*. Brookline Books.

Chandrasekar, R., C. Doran & B. Srinivas (1996). Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*. pp. 1041–1044.

Chandrasekar, R. & B. Srinivas (1996). *Automatic Induction of Rules for Text Simplification*. Tech. Rep. IRCS Report 96–30, Upenn, NSF Science and Technology Center for Research in Cognitive Science.

Charrow, V. (1988). *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, Lawrence Erlbaum Associates, chap. 4. Readability vs Comprehensibility: A Case-study in improving a real document, pp. 85–114.

Chen, Y.-T., Y.-H. Chen & Y.-C. Cheng (2013). Assessing Chinese Readability using Term Frequency and Lexical Chain. *Computational Linguistics and Chinese Language Processing* 18(2), 1–18.

Chung, J.-W., H.-J. Min, J. Kim & J. C. Park (2013). Enhancing readability of web documents by text augmentation for deaf people. In *WIMS '13 Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*. pp. 30:1–30:10.

Clark, J. H., C. Dyer, A. Lavie & N. A. Smith (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, pp. 176–181.

Clercq, O. D., V. Hoste, B. Desmet, P. V. Oosten, M. D. Cock & L. Macken (2014). Using the crowd for readability prediction. *Natural Language Engineering* -, 1–33.

Coleman, M. & T. L. Liau (1975). A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology* 60, 283–284.

Collins-Thompson, K. & J. Callan (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*. Boston, USA, pp. 193–200.

Collins-Thompson, K. & J. Callan (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology* 56(13), 1448–1462.

Cortese, M. J. & M. M. Khanna (2008). Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods* 43, 791–794.

Coster, W. & D. Kauchak (2011a). Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. Portland, Oregon: Association for Computational Linguistics, pp. 1–9.

Coster, W. & D. Kauchak (2011b). Simple English Wikipedia: A New Text Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 665–669.

Covington, M. A. (2007). *CPIDR 3 User Manual*. Computer Analysis of Speech for Psychological Research (CASPR) Research Report 2007-03, The University of Georgia, Artificial Intelligence Center, Athens, GA.

Coxhead, A. (2000). A New Academic Word List. *Teachers of English to Speakers of Other Languages* 34(2), 213–238.

Crossley, S. A., D. Allen & D. S. McNamara (2012). Text simplification and comprehensible input: A case for an intuitive approach. In *Language Teaching Research*. vol. 16, pp. 89–108.

Crossley, S. A., D. B. Allen & D. McNamara (2011a). Text Readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language* 23(1), pp. 84–101.

Crossley, S. A., D. F. Dufty, P. M. McCarthy & D. S. McNamara (2007). Toward a new readability: A mixed model approach. In D. S. McNamara & G. Trafton (eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society*. Cognitive Science Society, pp. 197–202.

Crossley, S. A., J. Greenfield & D. S. McNamara (2008). *Assessing Text Readability Using Cognitively Based Indices*, Teachers of English to Speakers of Other Languages, Inc. 700 South Washington Street Suite 200, Alexandria, VA 22314, pp. 475–493.

Crossley, S. A., T. Salsbury & D. S. McNamara (2011b). Predicting the proficiency level of language learners using lexical indices. In *Language Testing*. pp. 1–21.

Crossley, S. A., H. S. Yang & D. S. McNamara (2014). What's so simple about simplified texts? A computational and psycholinguistic investigation of text

comprehension and text processing. *Reading in a Foreign Language* 26(1), 92–113.

Cunningham, J. W. & H. A. Mesmer (2014). Quantitative Measurement of Text Difficulty: What's the Use? *The Elementary School Journal* 115, 255–269.

Daelemans, W., A. Höthker & E. T. K. Sang (2004). Automatic Sentence Simplification for Subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. pp. 1045–1048.

Dale, E. & J. S. Chall (1948a). A Formula for Predicting Readability. *Educational research bulletin; organ of the College of Education* 27(1), 11–28.

Dale, E. & J. S. Chall (1948b). A Formula for Predicting Readability: Instructions. *Educational research bulletin; organ of the College of Education* 27(2), 37–54.

Dale, E. & R. W. Tyler (1934). A Study of the Factors Influencing the Difficulty of Reading Materials for Adults of Limited Reading Ability. *The Library Quarterly* 4, 384–412.

Damay, J. J. S., G. J. D. Lojico, K. A. L. Lu, D. B. Tarantan & E. C. Ong (2006). SIMTEXT: Text Simplification of Medical Literature. In *3rd National Natural Language Processing Symposium - Building Language Tools and Resources*.

Daowadung, P. & Y.-H. Chen (2012). Stop Word in Readability Assessment of Thai Text. In *12th IEEE International Conference on Advanced Learning Technologies*. pp. 497–499.

Davidson, R. A. (2005). Analysis of the complexity of writing used in accounting textbooks over the past 100 years. *Accounting Education* 14(1), 53–74.

De Belder, J. & M.-F. Moens (2012). A dataset for the evaluation of lexical simplification. *Lecture Notes in Computer Science* 7182, 426–437.

Dell'Orletta, F., S. Montemagni & G. Venturi (2011). READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the*

2nd Workshop on Speech and Language Processing for Assistive Technologies. pp. 73–83.

Dell'Orletta, F., S. Montemagni & G. Venturi (2012). Genre–oriented Readability Assessment: a Case Study. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education (SLP-TED)*. pp. 91–98.

Dell'Orletta, F., S. Montemagni & G. Venturi (2013). Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An Exploratory Study on Italian Fictional Prose. In *Proceedings of Recent Advances in Natural Language Processing*. pp. 189–197.

Dell'Orletta, F., M. Wieling, A. Cimino, G. Venturi & S. Montemagni (2014). Assessing the Readability of Sentences: Which Corpora and Features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA9)*. Baltimore, Maryland, USA: ACL, pp. 163–173.

Devlin, S. & G. Unthank (2006). Helping aphasic people process online information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. New York, NY, USA: ACM, Assets '06, pp. 225–226.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment (JTLA)* 5(1), 4–35.

dos Santos Marujo, L. C. (2009). REAP.PT (Reap-Portuguese). Master's thesis, Universidade Tecnica de Lisboa.

Drndarevic, B. & H. Saggion (2012a). Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. *The Spanish Society for Natural Language Processing (SEPLN)* 49, 13–20.

Drndarevic, B. & H. Saggion (2012b). Towards Automatic Lexical Simplification in Spanish: An Empirical Study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Montréal, Canada: Association for Computational Linguistics, pp. 8–16.

Drndarevic, B., S. Stajner & H. Saggion (2012). Reporting Simply: A Lexical Simplification Strategy for Enhancing Text Accessibility. In *Proceedings of "Easy to read on the web" online symposium*.

DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa, California: Impact Information.

DuBay, W. H. (2006). *The Classic Readability Studies*. Costa Mesa, California: Impact Information.

Dufty, D. F., A. C. Graesser, M. M. Louwerse & D. S. McNamara (2006). Assigning grade levels to textbooks: Is it just readability? In *Proceedings of the 28th Annual Meetings of the Cognitive Science Society*. Mahwah, NJ: Erlbaum, pp. 1251–1256.

Ellimoottil, C., A. Polcari, A. Kadlec & G. Gupta (2012). Readability of Websites Containing Information About Prostate Cancer Treatment Options. *Readability of Websites Containing Information About Prostate Cancer Treatment Options* 188, 2171–2176.

Evans, R., C. Orasan & I. Dornescu (2014). An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. ACL, pp. 131–140.

Evans, R. J. (2011). Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing* 26(4), 371–388.

Evans, R. V. (1972). The Effect of Transformational Simplification on the Reading Comprehension of Selected High School Students. In *Journal of Literacy Research*. pp. 273–281.

Falkenjack, J. & A. Jonsson (2014). Classifying easy-to-read texts without parsing. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. ACL, pp. 114–122.

Falkenjack, J., K. H. Mühlenbock & A. Jönsson (2013). Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*. pp. 27–40.

Farr, J. N., J. J. Jenkins & D. G. Paterson (1951). Simplification of Flesch Reading Ease Formula. *Journal of applied psychology* 35(35), 333–337.

Feblowitz, D. & D. Kauchak (2013). Sentence Simplification as Tree Transduction. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. pp. 1–10.

Feng, L. (2008). *Text Simplification: A Survey*. Tech. rep., CUNY.

Feng, L. (2010). Automatic Readability Assessment. Ph.D. thesis, City University of New York (CUNY).

Feng, L., N. Elhadad & M. Huenerfauth (2009). Cognitively Motivated Features for Readability Assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, pp. 229–237.

Feng, L., M. Jansche, M. Huenerfauth & N. Elhadad (2010). A Comparison of Features for Automatic Readability Assessment. In *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.*. pp. 276–284.

Ferraro, G., H. Suominen & J. Nualart (2014). Segmentation of patent claims for improving their readability. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. ACL, pp. 66–73.

Finlayson, M. A. (2014). Java Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation. In *Proceedings of the 7th Global Wordnet Conference*. pp. 78–85.

Fitzgerald, J., J. Elmore, H. Koons, E. H. Hiebert, K. Bowen, E. E. Sanford-Moore & A. J. Stenner (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology* 107, 4–29.

Flesch, R. F. (1943). Marks of Readable Style: A Study in Adult Education. Ph.D. thesis, Columbia University.

Flesch, R. F. (1948). A New Readability Yardstick. *Journal of Applied Psychology* 32(3), 221–233.

Flor, M. & B. B. Klebanov (2014). Associative Lexical Cohesion as a Factor in Text Complexity. *International Journal of Applied Linguistics: Special issue on Recent Advances in Automatic Readability Assessment and Text Simplification* 165:2, 223–258.

Flor, M., B. B. Klebanov & K. M. Sheehan (2013). Lexical Tightness and Text Complexity. In *Proceedings of the Second Workshop on Natural Language Processing for Improving Textual Accessibility*. pp. 29–38.

Flory, S. M., T. J. P. Jr. & M. F. Tassin (1992). Measuring readability: A comparison of accounting textbooks. *Journal of Accounting Education* 10, 151–161.

Forsyth, J. (2014a). Automatic Readability Prediction for Modern Standard Arabic. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*. ELRA, pp. 9–15.

Forsyth, J. (2014b). Automatic Readability Prediction for Modern Standard Arabic. Master's thesis, Brigham Young University.

François, T. & E. Miltsakaki (2012). 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*. Montréal, Canada: Association for Computational Linguistics, pp. 49–57.

François, T. L. (2009). Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. Stroudsburg, PA, USA: Association for Computational Linguistics, EACL '09, pp. 19–27.

Franchina, V. & R. Vacca (1986). Adaptation of Flesch readability index on a bilingual text written by the same author both in Italian and English languages. In *Linguaggi*, Linguaggi, pp. 47–49.

Francois, T. & C. Fairon (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. pp. 466–477.

Francois, T. & P. Watrin (2011). On the Contribution of MWE-based Features to a Readability Formula for French as a Foreign Language. In *Proceedings of Recent Advances in Natural Language Processing*. pp. 441–447.

François, T., N. Gala, P. Watrin & C. Fairon (2014). FLELex: a graded lexical resource for French foreign learners. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. pp. 3766–3773.

Freund, Y., R. Iyer, R. Schapire, & Y. Singer (2003). An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research* 4, 933–969.

Futagi, Y., I. W. Kostin & K. M. Sheehan (2007). Reading Level Assessment for Literacy and Expository Texts. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*. The Cognitive Science Society, p. 1853.

Gamson, D. A., X. Lu & S. A. Eckert (2013). Challenging the Research Base of the Common Core State Standards: A Historical Reanalysis of Text Complexity. *Educational Researcher* 42(7), 381–391.

Gasperin, C., E. Maziero, L. Specia, P. T.S.P. & S. Aluisio (2009). Natural language processing for social inclusion: a text simplification architecture for different literacy levels. In *XXXVI Seminário Integrado de Software e Hardware (SEMISH-2009)*. Bento Gonçalves, Brazil, pp. 387–401.

Geng, X., T.-Y. Liu, T. Qin & H. Li (2007). Feature Selection for Ranking. In *Proceedings of SIGIR Conference*. pp. 548–552.

Gonzalez-Dios, I., M. J. Aranzabe & A. Díaz de Ilarraza (2014a). Making Biographical Data in Wikipedia Readable: A Pattern-based Multilingual Approach. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 11–20.

Gonzalez-Dios, I., M. J. Aranzabe, A. Díaz de Ilarraza & H. Salaberri (2014b). Simple or Complex? Assessing the readability of Basque Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 334–344.

Grabar, N., T. Hamon & D. Amiot (2014). Automatic diagnosis of understanding of medical words. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. ACL, pp. 11–20.

Graesser, A. C., a. Z. C. Danielle S. McNamara, M. Conley, H. Li & J. Pennebaker (2014). Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *The Elementary School Journal* 115, 210–229.

Graesser, A. C., D. S. McNamara & J. M. Kulikowich (2012). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher* 40(5), 223–234.

Granowsky, A. & M. Botel (1974). Background for a New Syntactic Complexity Formula. *The Reading Teacher* 28 (1), 31–35.

Gray, W. S. & B. E. Leary (1935). *What makes a book readable: With special reference to adults of limited reading ability, an initial study*. Chicago, Illinois, USA: The University of Chicago Press.

Green, G. M. & M. S. Olsen (1988). *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, Lawrence Erlbaum Associates, chap. 5. Preferences for and Comprehension of Original and Readability Adapted Materials, pp. 115–140.

Green, M. J. (2014). 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 (PITR)*. ACL, pp. 38–46.

Gunning, R. (1968). *The Technique of Clear Writing*. New York: McGraw-Hill Book Company, 2nd ed.

Gyllstad, H., J. Grandfeldt, P. Bernardini & M. Källkvist (2014). Linguistic correlates to communicative proficiency levels of the CEFR: The case of syntactic complexity in written L2 English, L3 French and L4 Italian. *EUROSLA Yearbook* 14(1), 1–30.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann & I. H. Witten (2009). The WEKA Data Mining Software: An Update. In *The SIGKDD Explorations*. vol. 11, pp. 10–18.

Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, The University of Waikato, Hamilton, NewZealand.

Hancke, J. (2013). Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language. Master's thesis, International Studies in Computational Linguistics. Seminar für Sprachwissenschaft, Universität Tübingen.

Hancke, J. & D. Meurers (2013). Exploring CEFR classification for German based on rich linguistic modeling. In *Learner Corpus Research 2013, Book of Abstracts*. Bergen, Norway.

Hancke, J., D. Meurers & S. Vajjala (2012a). Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 1063–1080.

Hancke, J., D. Meurers & S. Vajjala (2012b). Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbai, India, pp. 1063–1080.

Hansberry, D., N. Agarwal, S. Gonzales & S. Baker (2014). Are We Effectively Informing Patients? A Quantitative Analysis of On-line Patient Education Resources from the American Society of Neuroradiology. *American Journal of Neuroradiology* 35(7), 1270–1275.

Hara, T., C. Chen, Y. Kano & A. Aizawa (2013). Modeling Comma Placement in Chinese Text for Better Readability using Linguistic Features and Gaze Information. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. p. 49–58.

Heilman, M., K. Collins-Thompson, J. Callan & M. Eskenazi (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-07)*. Rochester, New York, pp. 460–467.

Heilman, M., K. Collins-Thompson & M. Eskenazi (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications at ACL-08*. Columbus, Ohio, pp. 71–79.

Heilman, M. & N. Smith (2010). Extracting Simplified Statements for Factual Question Generation. In *Proceedings of the Third Workshop on Question Generation*. pp. 11–20.

Heimann Mühlenbock, K. (2013). I see what you mean:Assessing readability for specific target groups. Ph.D. thesis, University of Gothenburg.

Heister, J., K.-M. Würzner, J. Bubenzer, E. Pohl, T. Hanneforth, A. Geyken & R. Kliegl (2011). dlexDB - eine lexikalische Datenbank fr̈ die psychologische und linguistische Forschung. *Psychologische Rundschau* 62, 10–20.

Henrich, V. & E. Hinrichs (2010). GernEdiT - The GermaNet Editing Tool. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), System Demonstrations*. pp. 19–24.

Herbrich, R., T. Graepel & K. Obermayer (2000). *Large margin rank boundaries for ordinal regression*, MIT Press, Cambridge, MA, pp. 115–132.

Hoang, H., A. Birch et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45nd Annual Meeting of the Association for Computational Linguistics (ACL '07)*. pp. 177–180.

Horn, C., C. Manduca & D. Kauchak (2014). Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL Short Papers)*. Assoc. for Computational Linguistics, pp. 458–463.

Housen, A. & F. Kuiken (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* 30(4), 461–473. URL `http://applij.oxfordjournals.org/content/30/4/461.full.pdf`.

Hruby, G. G. & U. Goswami (2011). Neuroscience and Reading: A Review for Reading Education Researchers. *Reading Research Quarterly* 46 (2), 156–172.

Huenerfauth, M., L. Feng & N. Elhadad (2009). Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. New York, NY, USA: ACM, Assets '09, pp. 3–10.

Inui, K., A. Fujita, T. Takahashi, R. Iida & T. Iwakura (2003). Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the Second International Workshop on Paraphrasing, held at ACL 2003*. pp. 9–16.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE projects. In G. Weir, S. Ishikawa & K. Poonpon (eds.), *Corpora and language technologies in teaching, learning and research*, Glasgow, UK: University of Strathclyde Publishing, pp. 3–11.

Islam, Z. & A. Mehler (2013). Automatic Readability Classification of Crowd-Sourced Data based on Linguistic and Information-Theoretic Features. *Computación y Sistemas (CICLING 2013 Proceedings* 17(2), 113–123.

Islam, Z., A. Mehler & R. Rahman (2012). Text Readability Classification of Textbooks of a Low-Resource Language. In *26th Pacific Asia Conference on Language,Information and Computation*. pp. 545–553.

Islam, Z., M. R. Rahman & A. Mehler (2014). *Readability Classification of Bangla Texts*, Springer Berlin Heidelberg, vol. 8404 of *Lecture Notes in Computer Science*, pp. 507–518.

Jakobsen, G. (1971). *Dansk LIX 70*. Landsforeningen af læsepædagoger. København: Dragør.

Jameel, S., W. Lam & X. Qian (2012). Ranking Text Documents Based on Conceptual Difficulty Using Term Embedding and Sequential Discourse Cohesion. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*. vol. 1, pp. 145–152.

Jauhar, S. K. & L. Specia (2012). UOW-SHEF: SimpLex – Lexical Simplicity Ranking based on Contextual and Psycholinguistic Features. In *In proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM)*. pp. 477–481.

Jiang, Z., G. Sun, Q. Gu & D. Chen (2014). An Ordinal Multi-class Classification Method for Readability Assessment of Chinese Documents. In *Knowledge Science, Engineering and Management*, Springer International Publishing, vol. 8793 of *Lecture Notes in Computer Science*, pp. 61–72.

Joachims, T. (2002). Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, pp. 217–226.

Joachims, T. (2006). Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 133–142.

Johannsen, A., H. Martínez, S. Klerke & A. Søgaard (2012). EMNLP@CPH: Is frequency all there is to simplicity? In *First Joint Conference on Lexical and Computational Semantics (*SEM)*. pp. 408–412.

Jonnalagadda, S. & G. Gonzalez (2009). Sentence Simplification Aids Protein-Protein Interaction Extraction. In *Proceedings of The 3rd International Symposium on Languages in Biology and Medicine, Jeju Island, South Korea, November 8-10, 2009*.

Jonnalagadda, S. & G. Gonzalez (2010). BioSimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. In *AMIA Annual Symposium Proceedings*. pp. 351–356.

Jonnalagadda, S., L. Tari, J. Hakenberg, C. Baral & G. Gonzalez (2009). Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. In *Proceedings of the NAACL-HLT 2009, Boulder, USA, June*. pp. 177–180.

Jr, C. C., A. Staub & K. Rayner (2007). *Eye movement research: A window on mind and brain*, Oxford:Elsevier Ltd., chap. Eye movements in reading words and sentences, pp. 341–372.

Junior, A. C., A. Copestake, L. Specia & S. M. Aluísio (2011). Towards an on-demand Simple Portuguese Wikipedia. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*. pp. 137–147.

Just, M. & P. Carpenter (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review* 87, 329–355.

Kandel, L. & M. A. Moles (1958). Application de l'indice de Flesch a la langue francaise. *Cahiers Etudes de Radio-Television* 19, 253–274.

Kandula, S., D. Curtis & Q. Zeng-Treitler (2010). A semantic and syntactic text simplification tool for health content. In *In Proceedings of AMIA Annual Symposium*. pp. 366–370.

Kandula, S. & Q. Zeng-Treitler (2008). Creating a Gold Standard for the Readability Measurement of Health Texts. In *AMIA 2008 Symposium Proceedings*. pp. 353–357.

Kane, L., J. Carthy & J. Dunnion (2006). Readability Applied to Information Retrieval. In *Proceedings of European Conference on Information Retrieval (ECIR)*. Springer, pp. 523–526.

Kanungo, T. & D. Orr (2009). Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, WSDM '09, pp. 202–211.

Karpov, N., J. Baranova & F. Vitugin (2014). Single-sentence Readability Prediction in Russian. In *Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST)*. pp. 91–100.

Kate, R. J., X. Luo, S. Patwardhan, M. Franz, R. Florian, R. J. Mooney, S. Roukos & C. Welty (2010). Learning to Predict Readability using Diverse Linguistic Features. In *23rd International Conference on Computational Linguistics (COLING 2010)*.

Kauchak, D. (2013). Improving Text Simplification Language Modeling Using Unsimplified Text Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pp. 1537–1546.

Kauchak, D., O. Mouradi, C. Pentoney & G. Leroy (2014). Text Simplification Tools: Using Machine Learning to Discover Features that Identify Difficult Text. In *47th Hawaii International Conference on System Sciences (HICSS)*. pp. 2616–2625.

Keskisärkkä, R. (2012). Automatic Text Simplification via Synonym Replacement. Master's thesis, Linköping University.

Keskisärkkä, R. & A. Jönsson (2012). Automatic Text Simplification via Synonym Replacement. In *In Proceedings of The Fourth Swedish Language Technology Conference*. pp. 46–47.

Keuleers, E., P. Lacey, K. Rastle & M. Brysbaert (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods* 44, 287–304.

Kidwell, P., G. Lebanon & K. Collins-Thompson (2011). Statistical Estimation of Word Acquisition with Application to Readability Prediction. In *Journal of the American Statistical Association. 106(493):21-30*.

Kim, A.-Y., H.-J. Song, S.-B. Park & S.-J. Lee (2014). Device-Dependent Readability for Improved Text Understanding. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 1396–1404.

Kim, J. Y., K. Collins-Thompson, P. N. Bennett & S. T. Dumais (2012). Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of the fifth ACM international conference on Web search and data mining*. New York, NY, USA: ACM, WSDM '12, pp. 213–222.

Kincaid, J. P., R. P. J. Fishburne, R. L. Rogers & B. S. Chissom (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

Kintsch, W. & T. A. van Dijk (1978). Toward a Model of Text Comprehension and Productions. *Pschological Review* 85(5), 363–394.

Klaper, D., S. Ebling & M. Volk (2013). Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. pp. 11–19.

Klare, G. (1952). Measures of the Readability of Written Communication: An Evaluation. *The Journal of Educational Psychology* 43(7), 385–399.

Klare, G. R. (1969). Automation of the Flesch Reading Ease Readability Formula, with Various Options. *Reading Research Quarterly* 4(4), 550–559.

Klare, G. R. (1974). Assessing Readability. *Reading Research Quarterly* 10(1), 62–102.

Klebanov, B. B., K. Knight & D. Marcu (2004). Text Simplification for Information-Seeking Applications. In *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*. Springer Verlag, pp. 735–747.

Klerke, S. (2012). Automatic Text Simplification in Danish: Sampling a restricted space of rewrites to optimize readability using lexical substitutions and dependency analyses. Master's thesis, University of Copenhagen.

Klerke, S. & A. Søgaard (2012). DSim: Danish parallel corpus for text simplification. In *In Proceedings of Language Resources and Evaluation Conference (LREC), 2012*. pp. 4015–4018.

Klerke, S. & A. Søgaard (2013). Simple, readable sub-sentences. In *Proceedings of the ACL Student Research Workshop*. pp. 142–149.

Kuperman, V., H. Stadthagen-Gonzalez & M. Brysbaert (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4), 978–990.

Kyle, K. & S. A. Crossley (2014). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly* .

Lal, P. & S. Rüger (2002). Extract-based Summarization with Simplification. In *In Proceedings of Document Understanding Conference (DUC) 2002*. URL `http://duc.nist.gov/pubs/2002papers/imperial_rueger.pdf`.

Landauer, T. & D. Way (2012). Improving Text Complexity Measurement through the Reading Maturity Metric. Annual meeting of the National Council on Measurement in Education.

Larsson, P. (2006). Classification into Readability Levels Implementation and Evaluation. Master's thesis, Uppsala Universitet, Department of Linguistics and Philology, Language Technology Programme.

Lau, T. P. (2006). Chinese Readability Analysis and Its Applications on the Internet. Master's thesis, CUHK, Hongkong.

Lau, T. P. & I. King (2006). Bilingual Web Page and Site Readability Assessment. In *In Proceedings of the World Wide Web Conference (WWW)*. pp. 993–994.

Lavalley, R. & K. Berkling (2014). Data Exploration of Sentence Structures and Embellishments in German texts: Comparing Children's Writing vs Literature. In *Proceedings of the 12th edition of the KONVENS conference Vol. 1*. URL `http://hildok.bsz-bw.de/frontdoor/index/index/docId/268`.

Lenzner, T. (2013). Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty? *Sociological Methods and Research* 43(4), 677–698.

Leroy, G. & D. Kauchak (2013). The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association* E1, 169–172.

Levy, R. & G. Andrew (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*. Genoa, Italy, pp. 2231–2234.

Li, H. (2014). *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan and Claypool Publishers.

Ligozat, A.-L., C. Grouin, A. Garcia-Fernandez & D. Bernhard. (2012). ANNLOR: A Naive Notation system for Lexical Outputs Ranking. In *In English Lexical Simplification Proceedings of the 6th International Workshop on Semantic Evaluation*. pp. 487–492.

Liu, X., W. B. Croft, P. Oh & D. Hart (2004). Automatic recognition of reading levels from user queries. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, SIGIR '04, pp. 548–549.

Lively, B. A. & S. L. Pressey (1923). A Method for Measuring the "Vocabulary Burden" of Textbooks. *Educational Adminstration and Supervision* 9(7), 389–398.

Lorge, I. (1944). Predicting Readability. *Teachers College Record* 45, 404–419.

Louis, A. & A. Nenkova (2013). What Makes Writing Great? First Experiments on Article Quality Prediction in the Science Journalism Domain. *Transactions of the Association for Computational Linguistics* 1, 341–352.

Lu, L. & N. Parameswaran (2009). Sentence Simplification Based Ontology Mapping. In *Proceedings of the Twenty-Second International FLAIRS Conference (2009)*.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Languages Journal* pp. 190–208.

Lu, X., D. A. Gamson & S. A. Eckert (2014). Lexical difficulty and diversity in American elementary school reading textbooks: Changes over the past century. *International Journal of Corpus Linguistics* 19(1), 94–117.

Lucisano, P. & M. E. Piemontese (1988). Gulpease: Una formula per la predizione della diffi- colta' dei testi in lingua italiana. *Scuola e Citta* 3, 57–68.

Ma, Y., E. Fosler-Lussier & R. Lofthus (2012a). Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg, PA, USA: Association for Computational Linguistics, NAACL HLT '12, pp. 548–552.

Ma, Y., R. Singh, E. Fosler-Lussier & R. Lofthus (2012b). 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*. Stroudsburg, PA, USA: Association for Computational Linguistics, PITR '12, pp. 58–64.

Marujo, L., J. Lopes, N. Mamede, I. Trancoso, J. Pino, M. Eskenazi, J. Baptista & C. Viana (2009). Porting REAP to European Portuguese. In *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*. pp. 69–72.

McCall, W. A. & L. M. Crabbs (1926). *Standard Test Lessons in Reading*. 5. Teachers College, Columbia University, Bureau of Publications.

Mccarthy, D., F. E. Sussex & R. Navigli (2007). SemEval-2007 Task 10: English lexical substitution task. In *In Proceedings of the 4th workshop on Semantic Evaluations (SemEval-2007*. pp. 48–53.

McLaughlin, G. H. (1969). SMOG Grading – a New Readability Formula. *Journal of Reading* 12(8), 639–646.

Medero, J. & M. Ostendorf (2011). Identifying Targets for Syntactic Simplification. In *ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2011)*. pp. 69–72.

Medero, J. & M. Ostendorf (2013). Atypical Prosodic Structure as an Indicator of Reading Level and Text Difficulty. In *Proceedings of NAACL-HLT 2013*. Assoc. for Computational Linguistics, pp. 715–720.

Mesmer, H. A., J. W. Cunningham & E. H. Hiebert (2012). Toward a Theoretical Model of Text Complexity for the Early Grades: Learning From the Past, Anticipating the Future. *Reading Research Quarterly* 47(3), 235–258.

Miltsakaki, E. (2009). Matching readers' preferences and reading skills with appropriate web texts. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*. Stroudsburg, PA, USA: Association for Computational Linguistics, EACL '09, pp. 49–52.

Miltsakaki, E. & A. Troutt (2007). Read-X: Automatic Evaluation of Reading Difficulty of Web Text. In T. Bastiaens & S. Carliner (eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*. Quebec City, Canada: AACE, pp. 7280–7286.

Miltsakaki, E. & A. Troutt (2008). Real Time Web Text Classification and Analysis of Reading Difficulty. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio: Association for Computational Linguistics, pp. 89–97.

Miwa, M., R. Sætre, Y. Miyao & J. Tsujii (2010). Entity-Focused Sentence Simplification for Relation Extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. pp. 788–796.

Mothe, J. & L. Tanguy (2005). Linguistic features to predict query difficulty. In *ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications*. pp. 7–10.

Nakatani, M., A. Jatowt & K. Tanaka (2010). Adaptive Ranking of Search Results by Considering User's Comprehension. In *Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2010)*. ACM Press, Suwon, Korea, pp. 182–192.

Napoles, C. & M. Dredze (2010). Learning simple Wikipedia: a cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*. Stroudsburg, PA, USA: Association for Computational Linguistics, CL&W '10, pp. 42–50.

Nelken, R. & S. M. Shieber (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. In *In 11th Conference of the European Chapter of the Association of Computational Linguistics*. Assoc. for Computational Linguistics, pp. 161–168.

Nelson, J., C. Perfetti, D. Liben & M. Liben (2012). *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance*. Tech. rep., The Council of Chief State School Officers.

Newbold, N., H. McLaughlin & L. Gillam (2010). Rank by Readability: Document Weighting for Information Retrieval. In H. Cunningham, A. Hanbury & S. Rüger (eds.), *Advances in Multidisciplinary Retrieval*, Springer Berlin / Heidelberg, vol. 6107 of *Lecture Notes in Computer Science*, pp. 20–30.

Nietzio, A., B. Scheer & C. Bühler (2012). How Long Is a Short Sentence? – A Linguistic Approach to Definition and Validation of Rules for Easy-to-Read Material. In *13th International Conference on Computers Helping People with Special Needs (ICCHP)*. pp. 369–376.

Nishikawa, H., T. Makino & Y. Matsuo (2013). A Pilot Study of Readability Prediction with Reading Time. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. pp. 78–84.

Norris, J. M. & L. Ortega (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics* 30(4), 555–578. URL `http://applij.oxfordjournals.org/content/30/4/555.full.pdf`.

Och, F. J. & H. Ney (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1), 19–51.

Ostling, R., A. Smolentzov, B. Tyrefors Hinnerich & E. Höglin (2013). Automated Essay Scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia: Association for Computational Linguistics, pp. 42–47.

Ott, N. (2009). Information Retrieval for Language Learning: An Exploration of Text Difficulty Measures. Master's thesis, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen, Germany.

Ott, N. & D. Meurers (2010). Information Retrieval for Education: Making Search Engines Language Aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications* 3(1–2), 9–30.

Ozasa, T., G. R. S. Weir & M. Fukui (2007). Measuring Readability for Japanese Learners of English. In *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*. Pan-Pacific Association of Applied Linguistics, pp. 122–125.

Papineni, K., S. Roukos, T. Ward & W.-J. Zhu (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *40th Annual meeting of the Association for Computational Linguistics*. pp. 311–318.

Patty, W. & W. Painter (1931). A Technique for Measuring the Vocabulary Burden of Textbooks. *Journal of Educational Research* 24, 127–134.

Pearson, P. D. & E. H. Hiebert (2014). The State of the Field: Qualitative Analyses of Text Complexity. *The Elementary School Journal* 115, 161–183.

Pera, M. S. & Y.-K. Ng (2012). BReK12: A Book Recommender for K-12 Users. In *In Proceedings of SIGIR*. ACM, pp. 1037–1038.

Petersen, S. E. (2007). Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education. Ph.D. thesis, University of Washington.

Petersen, S. E. & M. Ostendorf (2006). Assessing the Reading Level of Web Pages. In *Ninth International Conference on Spoken Language Processing (Interspeech-ICSLP)*. Pittsburgh, Pennsylvania.

Petersen, S. E. & M. Ostendorf (2007). Text Simplification for Language Learners: A Corpus Analysis. In *Speech and Language Technology for Education (SLaTE)*. pp. 69–72.

Petersen, S. E. & M. Ostendorf (2009). A machine learning approach to reading level assessment. *Computer Speech and Language* 23, 86–106.

Petrov, S. & D. Klein (2007). Improved Inference for Unlexicalized Parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York, pp. 404–411.

Phani, S., S. Lahiri & A. Biswas (2014). Inter-rater Agreement Study on Readability Assessment in Bengali. In *International Conference On Natural Language Processing And Cognitive Computing (ICONACC)-2014*. Manipur University, Imphal.

Pilán, I., S. Vajjala & E. Volodina (2015). A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. In *Proceedings of CICLING 2015- Research in Computing Science Journal Issue (to appear)*.

Pilán, I., E. Volodina & R. Johansson (2013). Automatic Selection of Suitable Sentences for Language Learning Exercises. In L. Bradley & S. Thouësny (eds.), *20 Years of EUROCALL: Learning from the Past, Looking to the Future. Proceedings of the 2013 EUROCALL Conference*. pp. 218–225.

Pilán, I., E. Volodina & R. Johansson (2014). Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA9)*. Baltimore, Maryland, USA: ACL, pp. 174–184.

Pitler, E., A. Louis & A. Nenkova (2010). Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, ACL '10, pp. 544–554.

Pitler, E. & A. Nenkova (2008). Revisiting readability: a unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, EMNLP '08, pp. 186–195.

Pringle, M., B. Natesh & K. Konieczny (2013). Patient information leaflet on mastoid surgery risks: assessment of readability and patient understanding. *The Journal of Laryngology and Otology* 127(11), 1078–1083.

Pyburn, D. T. & S. Pazicni (2014). Applying the Multilevel Framework of Discourse Comprehension To Evaluate the Text Characteristics of General Chemistry Textbooks. *Journal of Chemical Education* 91(6), 773–783.

Quarteroni, S. & S. Manandhar (2006). Incorporating User Models in Question Answering to Improve Readability. In *Proceedings of the Workshop KRAQ'06 on Knowledge and Reasoning for Language Processing*. Trento, Italy, pp. 50–57.

Qumsiyeh, R. & Y.-K. Ng (2011). ReadAid: A Robust and Fully-Automated Readability Assessment Tool. In *In Proceedings of the IEEE 23rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, pp. 539–546.

Randall, M. & N. Groom (2009). The BUiD Arab Learner Corpus: a resource for studying the acquisition of L2 English spelling. In *Proceedings of the Corpus Linguistics Conference (CL)*. Liverpool, UK.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124, 372–422.

Rayner, K., K. H. Chace, T. J. Slattery & J. Ashby (2006). Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading* 10(3), 241–255.

Rello, L., H. Saggion & R. Baeza-Yates (2014). Keyword Highlighting Improves Comprehension for People with Dyslexia. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. pp. 30–37.

Rello, L., H. Saggion, R. Baeza-Yates & E. Graells (2012). Graphical Schemes May Improve Readability but Not Understandability for People with Dyslexia. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Montréal, Canada: Association for Computational Linguistics, pp. 25–32.

Roark, B., A. Bachrach, C. Cardenas & C. Pallier (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 324–333.

Robnik-Sikonja, M. & I. Kononenko (1997). An adaptation of Relief for attribute estimation in regression. In *In proceedings of the Fourteenth International Conference on Machine Learning*. pp. 296–304.

Salama, A., K. Oflazer & S. Hagan (2013). Typesetting for Improved Readability using Lexical and Syntactic Information. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. pp. 719–724.

Salesky, E. & W. Shen (2014). Exploiting Morphological, Grammatical, and Semantic Correlates for Improved Text Difficulty Assessment. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA9)*. Assoc. for Computational Linguistics, pp. 155–162.

Sato, S. (2014). Text Readability and Word Distribution in Japanese. In *Proceedings of The 9th edition of the Language Resources and Evaluation Conference (LREC)*.

Sato, S., S. Matsuyoshi & Y. Kondoh (2008). Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *LREC'08*.

Scarton, C. & S. M. Aluísio (2010). Coh-Metrix-Port: a readability assessment tool for texts in Brazilian Portuguese. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*.

Schneider, W., M. Schlagmüller & M. Ennemoser (2007). *LGVT 6-12: Lesegeschwindigkeits-und-verständnistest für die Klassen 6-12*. Hogrefe.

Schulz, S. (2012). *Thinking in propositions - Propositional idea density as a cross-language complexity measure*. Tech. rep., Eberhard Karls University of Tübingen.

Schwarm, S. & M. Ostendorf (2005). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*. Ann Arbor, Michigan, pp. 523–530.

Segler, T. M. (2007). Investigating the Selection of Example Sentences for Unknown Target Words in ICALL Reading Texts for L2 German. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Seretan, V. (2012). Acquisition of Syntactic Simplification Rules for French. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Shardlow, M. (2012). Bayesian Lexical Simplification. 2012.

Shardlow, M. (2013a). A Comparison of Techniques to Automatically Identify Complex Words. In *Proceedings of the ACL Student Research Workshop*. pp. 103–109.

Shardlow, M. (2013b). The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. pp. 69–77.

Shardlow, M. (2014). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing* pp. 58–70.

Sheehan, K. M., M. Flor & D. Napolitano (2013). A Two-Stage Approach for Generating Unbiased Estimates of Text Complexity. In *Proceedings of the Second Workshop on Natural Language Processing for Improving Textual Accessibility*. pp. 49–58.

Sheehan, K. M., I. Kostin & Y. Futagi (2008). When Do Standard Approaches for Measuring Vocabulary Difficulty, Syntactic Complexity and Referential Cohesion Yield Biased Estimates of Text Difficulty? In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*.

Sheehan, K. M., I. Kostin, Y. Futagi & M. Flor (2010). *Generating Automated Text Complexity Classifications That Are Aligned With Targeted Text Complexity Standards*. Tech. rep., ETS.

Sheehan, K. M., I. Kostin, D. Napolitano & M. Flor (2014). The TextEvaluator Tool: Helping Teachers and Test Developers Select Texts for Use in Instruction and Assessment. *The Elementary School Journal* 115, 184–209.

Shen, W., J. Williams, T. Marius & E. Salesky (2013). A Language-Independent Approach to Automatic Text Difficulty Assessment for Second-Language Learners. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*. pp. 30–38.

Si, L. & J. Callan (2001). A Statistical Model for Scientific Readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*. ACM, pp. 574–576.

Siddharthan, A. (2002a). An Architecture for a Text Simplification System. In *In Proceedings of the Language Engineering Conference 2002 (LEC 2002)*. pp. 64–71.

Siddharthan, A. (2002b). Resolving Attachment and Clause Boundary Ambiguities for Simplifying Relative Clause Constructs. In *Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computational Linguistics (ACL 2002)*. pp. 60–65.

Siddharthan, A. (2003). Preserving Discourse Structure when Simplifying Text. In *Proceedings of the European Natural Language Generation Workshop (ENLG)*. pp. 103–110.

Siddharthan, A. (2004). *Syntactic simplification and text cohesion*. Tech. Rep. UCAM-CL-TR-597, University of Cambridge Computer Laboratory.

Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Research on Language and Computation* 4(1), 77–109.

Siddharthan, A. (2011). Text Simplification using Typed Dependencies: A Comparison of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*. pp. 2–11.

Siddharthan, A. (2014). A Survey of Research on Text Simplification. *International Journal of Applied Linguistics: Special issue on Recent Advances in Automatic Readability Assessment and Text Simplification* 165:2, 259–298.

Siddharthan, A. & A. Copestake (2002). Generating Anaphora for Simplifying Text. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*. pp. 199–204.

Siddharthan, A. & N. Katsos (2010). Reformulating Discourse Connectives for Non-Expert Readers. In *In Proceedings of Human Language Technologies: the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pp. 1002–1010.

Siddharthan, A. & N. Katsos (2012). Offline Sentence Processing Measures for testing Readability with Users. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Montréal, Canada: Association for Computational Linguistics, pp. 17–24.

Siddharthan, A. & A. Mandya (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: ACL, pp. 722–731.

Sinha, M., T. Dasgupta & A. Basu (2014). Text Readability in Hindi: A Comparative Study of Feature Performances Using Support Vectors. In *Proceedings of The 11th International Conference on Natural Language Processing (ICON)*.

Sinha, M., S. Sharma, T. Dasgupta & A. Basu (2012). New Readability Measures for Bangla and Hindi Texts. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING):Posters*. pp. 1141–1150.

Sinha, R. (2012). UNT-SIMPRANK: Systems for Lexical Simplification Ranking. In *In Proceedings of First Joint Conference on Lexical and Computational Semantics (SEM)*. pp. 493–496.

Sitbon, L. & P. Bellot (2008). A readability measure for an information retrieval process adapted to dyslexics. In *In Second international workshop on Adaptive Information Retrieval (AIR)*. pp. 52–57.

Sjöholm, J. (2012). Probability as readability:A new machine learning approach to readability assessment for written Swedish. Master's thesis, Linköpings universitet.

Smith, C. S. (1988). *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, Lawrence Erlbaum Associates, chap. Chapter 10: Factors of Linguistic Complexity and Performance, pp. 247–279.

Specia, L. (2010). Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR'10)*. pp. 30–39.

Stajner, S., B. Drndarevic & H. Saggion (2013). Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computación y Systemas* 17(2), 251–262.

Stajner, S., R. Mitkov & H. Saggion (2014). 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 (PITR)*. Gothenburg, Sweden: ACL, pp. 1–10.

Stajner, S. & H. Saggion (2013). Readability Indices for Automatic Evaluation of Text Simplification Systems: A Feasibility Study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan: Asian Federation of Natural Language Processing, pp. 374–382.

Stenner, A. J. (1996). Measuring reading comprehension with the Lexile framework. In *Fourth North American Conference on Adolescent/Adult Literacy*.

Stymne, S., J. Tiedemann, C. Hardmeier & J. Nivre (2013). Statistical Machine Translation with Readability Constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. pp. 375–386.

Sung, Y.-T., J.-L. Chen, J.-H. Cha, H.-C. Tseng, T.-H. Chang & K.-E. Chang (2014). Constructing and validating readability models: the method of integrat-

ing multilevel linguistic features with machine learning. *Behavior Research Methods* 47(2), 1–15.

Takahashi, T., T. Iwakura, R. Iida, A. Fujita & K. Inui (2001). KURA: A Transfer-Based Lexico-Structural Paraphrasing Engine. In *Proc. of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS) Workshop on Automatic Paraphrasing: Theories and Applications*. pp. 37–46.

Tan, C., E. Gabrilovich & B. Pang (2012). To Each His Own: Personalized Content Selection based on Text Comprehensibility. In *In Proceedings of WSDM*. pp. 233–242.

Tanaka-Ishii, K., S. Tezuka & H. Terada (2010). Sorting texts by readability. *Comput. Linguist.* 36(2), 203–227.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly* 30, 415–433.

Temnikova, I. (2012). Text Complexity and Text Simplification in the Crisis Management domain. Ph.D. thesis, University of Wolverhampton, UK.

Tham, T. M. (1987). Linguistic variables as predictors of Chinese text readability. Master's thesis, National University of Singapore.

Thomas, S. R. & S. Anderson (2012). WordNet-based lexical simplification of a document. In *Proceedings of KONVENS 2012*. pp. 80–88.

Thorndike, E. L. (1921). Word Knowledge in the Elementary School. *Teachers College Record* 28(5), 334–370.

Thorndike, E. L. & I. Lorge (1944). *The Teacher's Word Book of 30,000 Words*. New York: Bureau of Publications, Teachers College, Columbia Unversity.

Tingley, D., T. Yamamoto, K. Hirose, L. Keele & K. Imai (2014). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software* 59(5), 1–38.

Todirascu, A., T. François, N. Gala, C. Fairon, A.-L. Ligozat & D. Bernhard. (2013). Coherence and Cohesion for the Assessment of Text Readability. In *Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Science*. pp. 11–19.

Tonelli, S., K. Tran Manh & E. Pianta (2012). Making Readability Indices Readable. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Montréal, Canada: Association for Computational Linguistics, pp. 40–48.

Tono, Y. (2000). A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora. In *PALC'99: Practical Applications in Language Corpora*. pp. 323–340.

Toutanova, K., D. Klein, C. Manning & Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*. Edmonton, Canada, pp. 252–259.

Tur, G., D. Hakkani-Tür, L. Heck & S. Parthasarathy (2011). Sentence simplification for spoken language understanding. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*. pp. 5628–5631.

Uitdenbogerd, A. L. (2005). Readability of French as a Foreign Language and its Uses. In *Proceedings of the 10th Australasian Document Computing Symposium*. pp. 19–25.

Vajjala, S. & K. Lõo (2014). Automatic CEFR Level Prediction for Estonian Learner Text. *NEALT Proceedings Series Vol. 22* pp. 113–128.

Vajjala, S. & K. Lõo (2013). Role of Morpho-syntactic features in Estonian Proficiency Classification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8), Association for Computational Linguistics*.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *In Proceedings*

of the 7th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 163—-173.

Vajjala, S. & D. Meurers (2013). On The Applicability of Readability Models to Web Texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. pp. 59–67.

Vajjala, S. & D. Meurers (2014a). Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL, Gothenburg, Sweden: Association for Computational Linguistics, pp. 288–297.

Vajjala, S. & D. Meurers (2014b). Exploring Measures of "Readability" for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In *Proceedings of the Third Workshop on Predicting and Improving Text Readability for Target Reader Populations*. Gothenburg, Sweden: ACL, pp. 21–29.

Vajjala, S. & D. Meurers (2014c). Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification* 165(2), Thomas François and Delphine Bernhard.

Van Heuven, W. J., P. Mandera, E. Keuleers & M. Brysbaert (2014). Subtlex-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology* pp. 1–15.

van Oosten, P. & V. Hoste (2011). Readability annotation: replacing the expert by the crowd. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg, PA, USA: Association for Computational Linguistics, IUNLPBEA '11, pp. 120–129.

Van Oosten, P., V. Hoste & D. Tanghe (2011). A posteriori agreement as a quality measure for readability prediction systems. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*. Berlin, Heidelberg: Springer-Verlag, CICLing'11, pp. 424–435.

van Oosten, P., D. Tanghe & V. Hoste (2010). Towards an Improved Methodology for Automated Readability Prediction. In *LREC'10*.

Vogel, M. & C. Washburne (1928). An Objective Method of Determining Grade Placement of Children's Reading Material. *The Elementary School Journal* 28, 373–381.

Vor der Brück, T. & S. Hartrumpf (2007). A semantically oriented readability checker for German. In Z. Vetulani (ed.), *Proceedings of the 3rd Language & Technology Conference*. Poznań, Poland: Wydawnictwo Poznańskie, pp. 270–274.

Vor der Brück, T., S. Hartrumpf & H. Helbig (2008a). A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators. *Informatica* 32(4), 429–435.

Vor der Brück, T., H. Helbig & J. Leveling (2008b). *The readability checker DeLite*. Tech. Rep. Technical Report 345-5/2008, Fakultät für Mathematik und Informatik, FernUniversität in Hagen.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal* 96(4), 576–598.

Walker, A., A. Siddharthan & A. Starkey (2011). Investigation into Human Preference between Common and Unambiguous Lexical Substitutions. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG)*. Association for Computational Linguistics, pp. 176–180.

Walmsley, S. A., K. M. Scott & R. Lehrer (1981). Effects of Document Simplification on the Reading Comprehension of the Elderly. *Journal of Literacy Research* 13(3), 237–248.

Wan, X., H. Li & J. Xiao (2010). EUSUM: extracting easy-to-understand english summaries for non-native readers. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, SIGIR '10, pp. 491–498.

Weir, G. R. S. & N. K. Anagnostou (2008). Collocation frequency as a readability factor. In *Proceedings of the 13th Conference of Pan-Pacific Association of Applied Linguistics*. Pan-Pacific Association of Applied Linguistics.

West, M. (1953). *A General Service List of English Words*. London: Longmans.

Wieling, M., S. Montemagni, J. Nerbonne & R. H. Baayen (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language* 90(3), 669–692.

Williams, S. & E. Reiter (2008). Generating basic skills reports for low-skilled readers*. *Nat. Lang. Eng.* 14(4), 495–525.

Williamson, D. M. (2009). A Framework for Implementing Automated Scoring. In *The annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME)*.

Wilson, M. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers* 20(1), 6–11.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Woodsend, K. & M. Lapata (2011a). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Assoc. for Computational Linguistics, pp. 409–420.

Woodsend, K. & M. Lapata (2011b). WikiSimple: Automatic Simplification of Wikipedia Articles. In *In Proceedings of the 25th National Conference on Artificial Intelligence*. pp. 927–933.

Wubben, S., A. van den Bosch & E. Krahmer (2012). Sentence Simplification by Monolingual Machine Translation. In *Proceedings of ACL 2012*. pp. 1015–1024.

Yan, X., D. Song & X. Li (2006). Concept-based document readability in domain specific information retrieval. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, pp. 540–549.

Yannakoudakis, H., T. Briscoe & B. Medlock (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, HLT '11, pp. 180–189. Corpus available: `http://ilexir.co.uk/applications/clc-fce-dataset`.

Yatskar, M., B. Pang, C. Danescu-Niculescu-Mizil & L. Lee (2010). For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the NAACL*. pp. 365–368.

Yuka, T., O. Yoshihiko & Y. Hisao (1988). A computer readability formula of Japanese texts for machine scoring. In *Proceedings of the 12th conference on Computational linguistics - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, COLING '88, pp. 649–654.

Zhao, J. & M.-Y. Kan (2010). Domain-specific iterative readability computation. In *Proceedings of the 10th annual joint conference on Digital libraries*. New York, NY, USA: ACM, JCDL '10, pp. 205–214.

Zhu, Z., D. Bernhard & I. Gurevych (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China*. pp. 1353–1361.

Appendices

Appendix A

Additional GAM models and Analysis for the Eye-tracking Data

Abstract

In this appendix, we describe additional modeling experiments for the eye-tracking data used in Chapter 4. As mentioned earlier, the experiments reported in the chapter primarily maintained a uniformity in the model settings across all the dependent variables studied, for easy comparison between them. Here, we report experiments that consider the effect of performing specific transformations based on the variable distribution and removing outliers for individual models. We also explore and compare a diverse set of models to understand the nature of the interactions and the effectiveness of tensor smooths in improving the model fit. These experiments reveal some new insights into the data that were not seen in the model settings followed in Chapter 4. Apart from observing three way interactions that were not observed in the earlier modeling, in many cases, these experiments also result in a better model fit.

A.1 Methods

As in Chapter 4, the modeling experiments are performed by Generalized Additive Mixed Modeling (GAMM) as implemented in the mgcv package in R[1]. Model evaluation and selection were done by using CompareML function from *Interpreting Time Series and Autocorrelated Data Using GAMMs* (itsadug) package in R[2]. Data transformations and outlier removal are performed based on the manual inspection of QQplots of the dependent variables[3]. In the following sections, we describe the models we explored and present the model summary for the best model, for each of the dependent variables. In all the experiments, we consider only the binary notion of complexity.

A.2 Fixation Count

In chapter 4, the reported model for fixation count considered text complexity, proficiency, text order, and the interaction between proficiency and complexity as fixed effects and the subject and text variation as random effects. As mentioned earlier, we did not perform any data transformation nor did we remove any outliers from the data. We now report some of the more complex models, which achieve significantly better performance over this model. Since fixation count had a skew in the distribution, we first performed a log transform to achieve normality and then removed three outlier instances, which seemed like missing data. We then trained gam model with the same parameters as in Chapter 4.

Specifically, the default model from Chapter 4 looks like below, and the variance explained is 65.3%.

fixation0 = gam(Fixation.Count $\sim$ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

The model with the log transformed variable looks like below, and the variance explained by this model is 73.8%.

[1] http://cran.r-project.org/package=mgcv

[2] `http://cran.r-project.org/package=itsadug`

[3] The code and data can be shared for replication of the results.

fixation1 = gam(LogFixationCount $\sim$ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

The log transformed model was significantly better than the default model ($p < 0.001$) as reported by the CompareML method. Thus, transforming the data and removing the outliers resulted in a superior model.

We then explored the a three way interaction between the three fixed effect variables (difficulty, proficiency, text order). Since proficiency was not a significant predictor, we considered it only with interaction terms in this model. Although TextOrder as a main effect was a significant predictor ($p < 0.001$), removing it did not result in any difference in the model performance. So, we considered TextOrder too only in the interaction term. The final interaction model looks like below and explained 73.7% of the variance:

fixation2 = bam(LogFixationCount $\sim$ Difficulty + s(Proficiency, TextOrder, by = Difficulty, k=4) + s(Participant, bs=”re”) + s(Text, bs=”re”), data=fcdata)

Although all the interactions were significant ($p < 0.001$) in this model *fixation2*, it is not significantly different from *fixation1* with a very small difference in fREML values. To build a model based on these interactions, that can perform better than the model without one, we used tensor product smooths instead of the normal smooths for the interaction. That is, the model now is trained with the following setting:

fixation3 = gam(LogFixationCount $\sim$ Difficulty + te(Proficiency, TextOrder, by = Difficulty, k=4) + s(Participant, bs=”re”) + s(Text, bs=”re”), data=fcdata)

This model explains 76.5% of the variance, which is a statistically significant improvement from *fixation1*. The Q-Q plot of the residuals for this model displayed an outlier, removing which resulted in the model explaining 78.9% of the variance, which is statistically significant compared to the model performance

with the outlier included. The summary of this final model is shown in Table A.1.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	2.478	0.0481	51.51	< 0.001
DifficultyEasy	-0.178	0.023	-7.61	< 0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
te(Proficiency,TextOrder):DifficultyDifficult	No	8.095	4.273	< 0.001
te(Proficiency,TextOrder):DifficultyEasy	No	4.544	7.549	< 0.001
Participant	Yes	41.86	11.020	< 0.001
Text	Yes	2.154	3.015	0.007
Variance Explained (R^2 adj): 78.9%				

Table A.1: Best Performing Model for Fixation Count

This model is a better fit for the data than the model discussed in Chapter 4 for fixation count (cf. Table 4.4), which explained 65.2% of the variance. In addition, while the interaction between proficiency and text complexity was not significant in that model, clearly, a three-way effect of proficiency and text order with difficulty was significant, as we see in Table A.1. The interaction can be visually observed in Figure A.1. Note that the fixation count values refer to the transformed value and not the original value.

From this figure, it is clear that the low proficiency readers make higher number of fixations when they read difficult texts in general compared to high proficiency readers (green color indicates lower values). However, the number of fixations also increase depending on when they read the texts. The fixation counts are clearly lower when they read difficult texts at the end of the experiment compared to the beginning. This effect is less pronounced in high proficiency readers although they seem to experience more fixations around third text. In comparison, for easy texts, readers of all proficiencies have less fixations than the difficult texts, in the texts they read at the beginning. However, as they read more texts, low proficiency readers seem to have more fixations while high proficiency readers are relatively less affected. This model shows a clear interaction between proficiency and difficulty, which was not observed in our baseline model from Chapter 4. In addition, it also indicates that the effect of a possible fatigue from reading more texts also differs with reader proficiency.

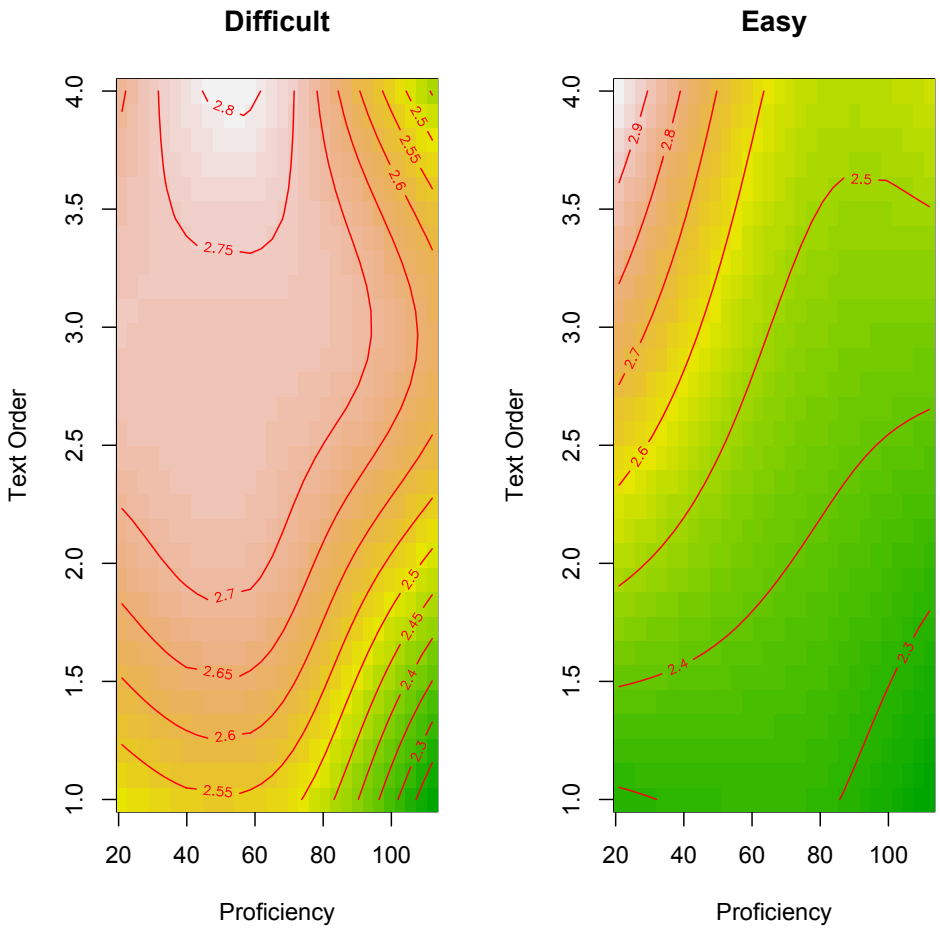

Figure A.1: 3-Way Interaction Visualization for Fixation Count

A.3 Average Fixation Duration

The baseline performance of the GAMM model for Average Fixation Duration (AFD) explained 74% of the variance (cf. Table 4.5). The experiments described below explore the role of data transformation and three-way interactions in improving the model fit. AFD values had some missing data (zero values or near zero values), which were removed first. To eliminate the skew in the distribution and transform the data to make the Q-Q plot approximate a normal distribution,

we used the BoxCox transformation (Box & Cox, 1964) as implemented in MASS package[4] in R. This resulted in transforming the AFD as $AFD^{2.6}$.

The baseline model, from Chapter 4, looked like below, explained 74% of the variance.

afd0 = gam(afd $\sim$ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder + s(Participant, bs = "re") + s(Text, bs = "re"), data=dat)

The model with the transformed variable, which looks like below, performed poorer than the baseline model, explaining 54% of the variance.

afd1 = gam($afd^{2.6}$ $\sim$ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder + s(Subject, bs = "re") + s(Text, bs = "re"), data=dat)

The model encoding a three-way interaction between Proficiency, Text Order and Difficulty too resulted in explaining 56.2% of the variance. Applying tensor smooth for the three way interaction did not result in any statistically significant improvement in the model fit. Hence, we can conclude from the existing evidence that data transformation and employing complex models was not useful in improving the model fit for average fixation duration.

A.4 First Fixation Duration

The baseline First Fixation Duration (FFD) from Chapter 4 explained 13% of the variance, with none of the fixed effects were significant. Only the random effect due to participant variation was significant ($p < 0.05$). This baseline model structure looked as follows:

ffd0 = gam(ffd $\sim$ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder + s(Participant, bs = "re") + s(Text, bs = "re"), data=dat)

The Q-Q plot for FFD distribution showed two outliers at both the extreme ends of the FFD value range. Removing the two outliers and re-training the default model resulted in a model that explained 47% of variance, which is significantly better than the default model. Removing four outliers based on the inspection of the Q-Q plot of this model and retraining with the same setting as the default model resulted in a significantly better model ($p < 0.001$), that explained 58.4%

[4] http://cran.r-project.org/package=MASS

of the variance. As in the default model, only the random effect due to participant variation was significant

Finally, we built a model removing all the main effects (since they are not significant) and including a three-way interaction with tensor smooths. This model structure looks like below:

ffd1 = gam(ffd $\sim$ te(Proficiency, TextOrder, by=Difficulty, k=4) + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

This model showed a significant interaction effect and explained 62.6% of the variance, which is significantly better than the previous model ($p < 0.003$). The summary of this model can be seen in Table A.2.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	194.41	4.11	47.33	< 0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
te(Proficiency,TextOrder):DifficultyDifficult	No	12.18	2.848	0.0011
Participant	Yes	38.05	4.978	< 0.001
Variance Explained (R^2 adj): 62.6%				

Table A.2: Best Performing Model for First Fixation Duration

The model summary from Table A.2 shows that there is a significant interaction between Proficiency and Text Order for Difficult texts. Figure A.2 shows this interaction.

From the figure, we can infer that the low proficiency readers, who face a difficult text in the beginning have a longer first fixation duration (140-170 ms from the wiggly surfaces) compared to the high proficiency readers (100-130 ms). However, the interactions seem to be more complex to interpret in the medium proficiency levels and may merit further study. To summarize, excluding the outliers and employing tensor smooth resulted in a much better model fit for first fixation duration, explaining 62.6% of the variance (compared to 13% variance for the model from Chapter 4). This model also showed a significant interaction between proficiency, text order and text complexity.

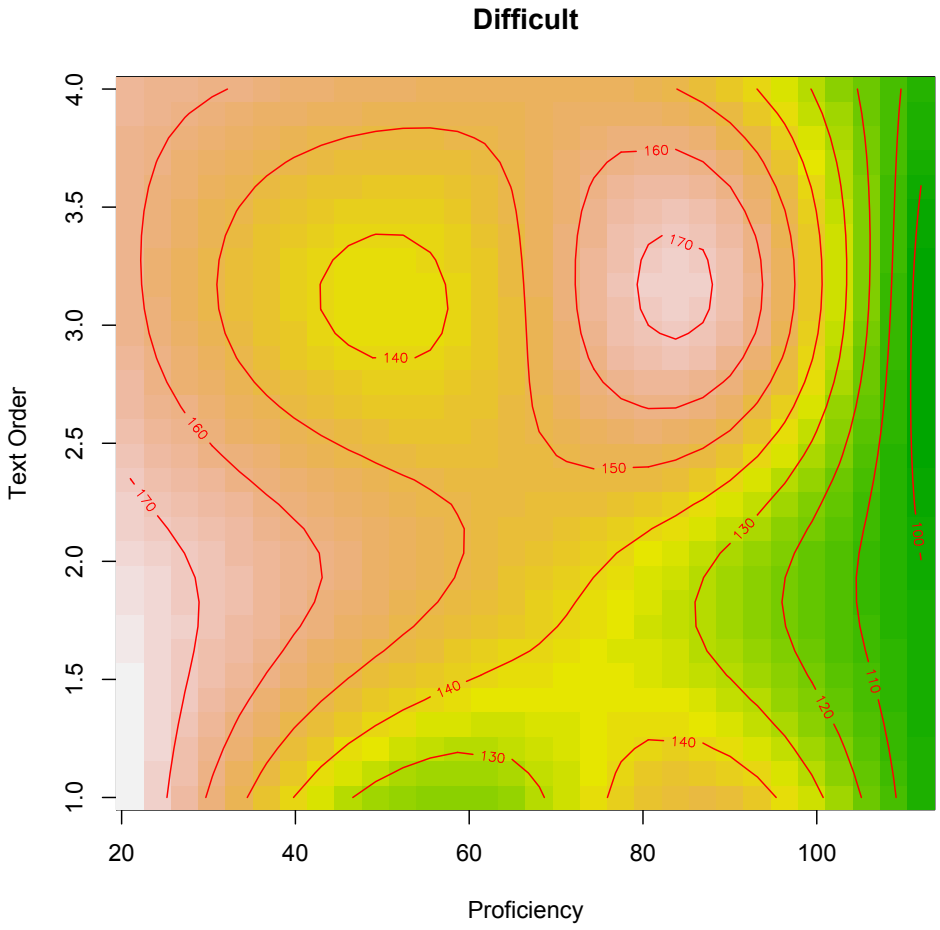

Figure A.2: 3-way interaction for First Fixation Duration

A.5 First Pass Duration

The First Pass Duration (FPD) model from Chapter 4 resulted in explaining 51.2% of the variance (cf. Table 4.6). This model, which can be considered a baseline for these experiments, looks as below:

fpd0 = gam(fpd ∼ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

The following experiments describe exploratory analysis to improve the model fit.

Manual inspection of the Q-Q plot of FPD values resulted in a log-transformation and removal of 5 observations (out of 192). Now, we re-trained the model with the same settings as the baseline model. This explained 52.6% of the variance. But the improvement in performance was not statistically significant.

We now added a three-way interaction, specified as follows:

fpd1 = gam(logfpd $\sim$ Difficulty + s(Proficiency, TextOrder, by=Difficulty) + TextOrder + s(Participant, bs = "re") + s(Text, bs = "re"), data=dat)

This model explained 52.4% of the variance and enhancing this model with tensor smooths resulted in a model which explained 54.4% of the variance, which was also significantly better ($p < 0.001$) than the baseline model. The summary of this final model is shown in Table A.3.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	5.55	0.07	79.43	< 0.001
DifficultyEasy	-0.095	0.043	-2.176	0.03
TextOrder	0.374	0.018	20.613	< 0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
te(Proficiency,TextOrder):DifficultyDifficult	No	4.41	34.64	< 0.001
te(Proficiency,TextOrder):DifficultyEasy	No	2.582	61.321	< 0.001
Participant	Yes	38.13	4.237	< 0.001
Variance Explained (R^2 adj): 54.4%				

Table A.3: Best Performing Model for First Pass Duration

It is interesting to note that the three-way interaction between the fixed effects turned out to be a significant predictor in this case too. Figure A.3 shows a visualization of this interaction (for log-transformed FPD).

We can notice from the figure that high proficiency readers have low FPD when they read easy texts, irrespective of the text order. However, while reading difficult texts, the FPD is longer at the beginning compared to reading difficult texts towards the end. On the contrary, low proficiency readers experience more FPD when they read difficult texts at a later stage than in the beginning. They also on an average have more FPD than high proficiency readers for both easy and difficult texts. While the interaction seems to be more linear for easy texts,

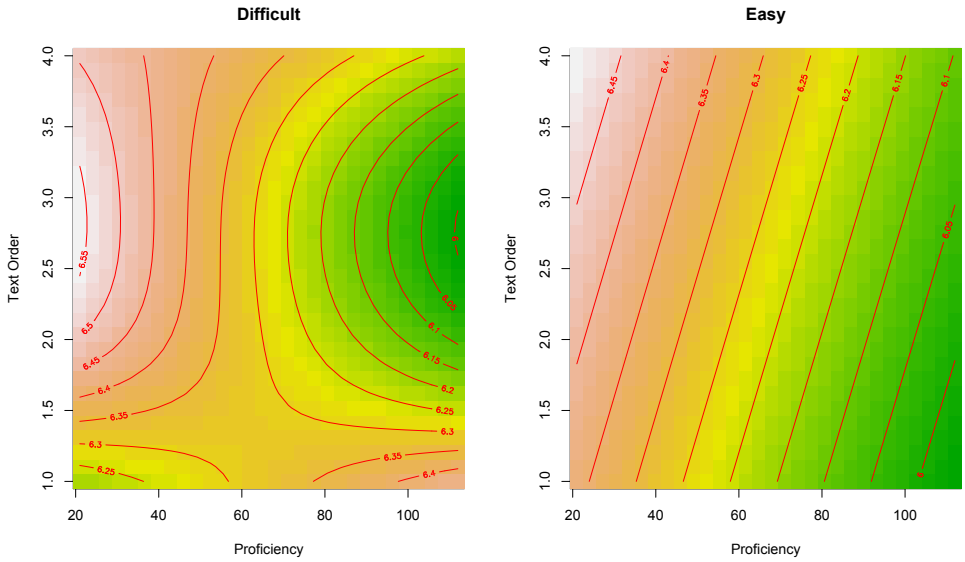

Figure A.3: Interaction between Proficiency and Text Order for First Pass Duration

we see more wiggly surfaces for difficult texts. The significance and interactions between variables also changed compared to the baseline model from Chapter 4.

A.6 Second Pass Duration

The Second Pass Reading Time (SPD) model from Chapter 4 resulted in explaining 63% of the variance (cf. Table 4.7). This model, which can be considered a baseline for these experiments, looks as below:

spd0 = gam(spd ∼ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

The following experiments describe a further analysis of the data and modeling methods to improve the model fit for the data. Firstly, removing the missing data and performing a log transform on the data resulted in a model with 189 observations (out of the total 192). Re-training the model now with this new data

and the log transformed variable resulted in an R^2 value of 59.8%. The difference between this model and the baseline model was not statistically significant. We now added a three-way interaction, specified as follows:

spd1 = gam(logspd ∼ Difficulty + s(Proficiency, TextOrder, by=Difficulty) + TextOrder + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

This model resulted in an R^2 of 61.9% and is a statistically better model ($p < 0.001$). Enhancing this model with tensor smooths resulted in a significant improvement ($p<0.001$), resulted in an R^2 value of 67.4%, where all the effects except that of the random effect due to text variation were significant. This model is summarized in Table A.4.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	6.11	0.09	70.43	< 0.001
DifficultyEasy	-0.296	0.046	-6.395	< 0.001
TextOrder	0.521	0.026	19.878	< 0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
te(Proficiency,TextOrder):DifficultyDifficult	No	7.785	22.296	< 0.001
te(Proficiency,TextOrder):DifficultyEasy	No	5.32	29.646	< 0.001
Participant	Yes	38.43	5.456	< 0.001
Variance Explained (R^2 adj): 67.4%				

Table A.4: Best Performing Model for Second Pass Duration

Figure A.4 shows a visualization of this interaction (for log-transformed SPD).

There is a clear difference between Difficult vs Easy texts in the nature of the interactions. While low proficiency readers had more SPD with difficult texts irrespective of the text order, the high proficiency readers experienced this after the first text. In the easy condition, while readers of all proficiencies had lower SPD for texts they read in the beginning, low proficiency readers experienced longer SPD as they read more texts. This effect was not seen with the increase in proficiency. Thus, we can conclude that there is a clear three-way interaction between proficiency, text order and text complexity, for second pass reading time.

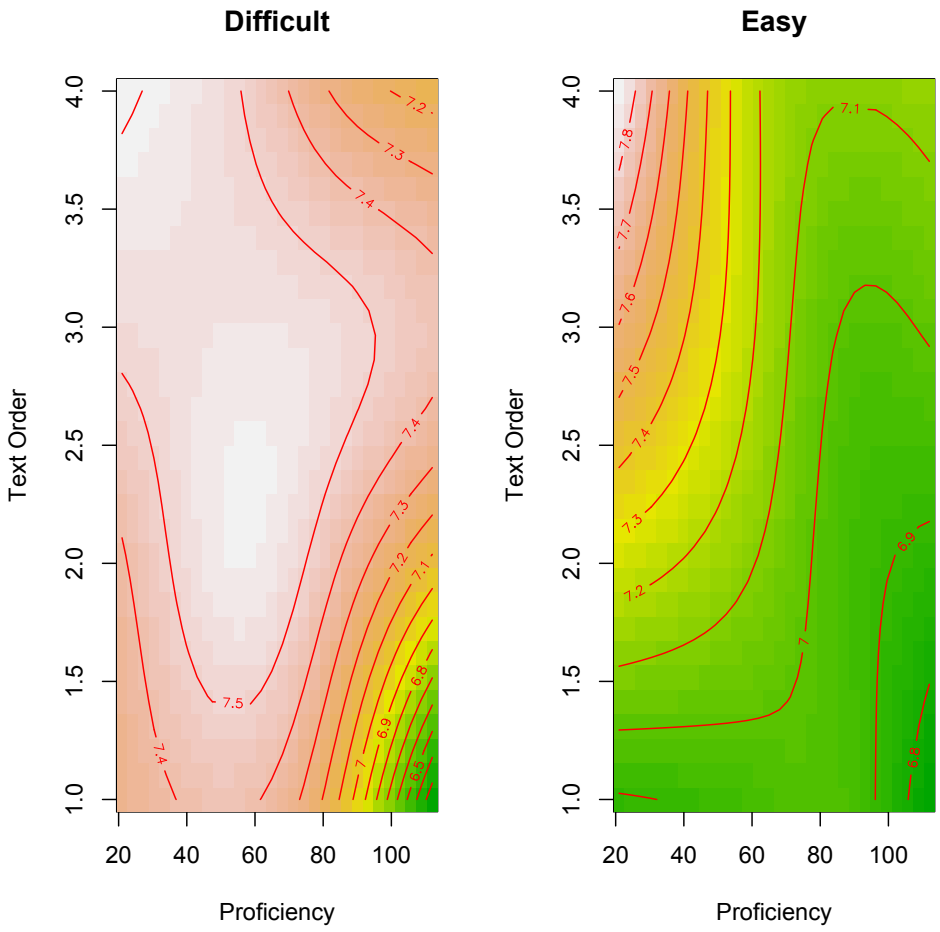

Figure A.4: Interaction between Proficiency and Text Order for Second Pass Duration

A.7 Revisits

The Revisits model from Chapter 4 resulted in explaining 74.9% of the variance (cf. Table 4.8). This model, which can be considered a baseline for these experiments, looks as below:

rev0 = gam(sprt ∼ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

Considering a three-way interaction for the above model, which looks like below, resulted in a statistically insignificant difference in model fit (R^2 = 75.2%).

rev1 = gam(sprt ∼ Difficulty + s(Proficiency, TextOrder, by=Difficulty) + TextOrder + s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

Using tensor smooth for rev1 resulted in a significantly better model ($p < 0.001$), which explained 77.4% of the variance. Removing zero valued data (potential missing data) did not result in a better model, so we can consider this as our final model, which is summarized below in Table A.5.

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	2.157	0.32	6.696	< 0.001
DifficultyEasy	-0.7056	0.146	-4.83	< 0.001
TextOrder	-0.616	0.083	7.383	< 0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
te(Proficiency,TextOrder):DifficultyDifficult	No	9.570	2.161	0.019
Participant	Yes	41.64	10.312	< 0.001
Text	Yes	1.992	2.435	0.013
Variance Explained (R^2 adj): 77.4%				

Table A.5: Best Performing Model for Revisits

A.8 Recall Score

As reported in Chapter 4 (cf. Table 4.9), the baseline model for recall scores explained a variance of 55.6%. Adding tensor smooths and three-way interactions to this model resulted in a significantly better model ($p < 0.05$) which explained 58.9% of the variance. This model summary is shown in Table A.6.

A.9 Comprehension Score

The baseline model for comprehension scores from Chapter 4 (cf. Table 4.10) explained 27.6% of variance and was of the form:

compre0 = gam(compre ∼ Difficulty + s(Proficiency) + s(Proficiency, by=Difficulty) + TextOrder

Parametric Coefficients				
Variable	Estimate	Std. Error	t value	p-value
Intercept	3.006	0.321	9.347	< 0.001
DifficultyEasy	0.679	0.192	3.527	< 0.001
TextOrder	0.467	0.089	5.202	< 0.001
Significant Smooth Terms				
Variable	is Random Effect?	Est. deg. of freedom	F	p-value
s(Proficiency)	No	0.9887	51.29	$p < 0.001$
te(Proficiency,TextOrder):DifficultyDifficult	No	5.78	3.194	0.006
Participant	Yes	29.272	1.806	< 0.001
Text	Yes	2.009	3.817	0.0014
Variance Explained (R^2 adj): 58.9%				

Table A.6: Best Performing Model for Recall

+ s(Participant, bs = ”re”) + s(Text, bs = ”re”), data=dat)

Removing all the insignificant terms from this model (Difficulty, TextOrder, Participant as random effect) resulted in a model with a lower R^2 value (22.3%), but there was no statistically significant difference between this model and the baseline model. The three-way interaction was also not significant for comprehension scores. Hence, stripping off the baseline model, the final model has the following structure:

compre1 = gam(compre $\sim$ s(Proficiency) + s(Text, bs=”re”), data=dat)

A.10 Discussion

To conclude, this additional analysis involving transformations, three-way interactions and tensor smooths improved in the model fit for some of the variables and did not result in a better model than the one in Chapter 4 for others. However, the three-way interaction was significant in all the cases except average fixation duration and comprehension score. This gains more importance if we consider the fact that the interaction between proficiency and difficulty was not significant for most of these variables (except first pass duration and recall scores). Among all the measures studied, the models for comprehension scores resulted in the worst fit for the data. Considering other possible factors influencing comprehension in future studies may result in models that can predict a reader's comprehension with better accuracy.

Appendix B

Texts and Questions used for the Eye-tracking Experiment

The four texts (in two versions each) and their respective questions that were used in the eye-tracking experiment described in Chapter 4 are shown below. These texts are taken from onestopenglish.com. The original articles are longer than the texts used here. Since this is an eye-tracking experiment, we only took the first 4,5 paragraphs from each text, so that the texts fit into two screens. We prepared recall and comprehension questions for each text by ourselves.

B.1 Text 1

B.1.1 Difficult Version

Amsterdam still looks liberal to tourists, who were recently assured by the Labour Mayor that the city's marijuana-selling coffee shops would stay open despite a new national law tackling drug tourism. But the Dutch capital may lose its reputation for tolerance over plans to dispatch nuisance neighbours to "scum villages" made from shipping containers.

The Mayor, Eberhard van der Laan, insists his controversial new £810,000 policy to tackle antisocial behaviour is to protect victims of abuse and homophobia from harassment. The camps where antisocial tenants will be rehoused for three to six months have been called "scum villages" because the policy echoes proposals

from Geert Wilders, the far-right populist, who last year demanded that ”repeat offenders” be ”sent to a village for scum”.

But Bartho Boer, a spokesman for the Mayor, denies that the plans are illiberal. ”We want to defend the liberal values of Amsterdam,” he says. ”We want everyone to be who he and she is - whether they are gay and lesbian or stand up to violence and are then victims of harassment. We as a society want to defend them.” According to Boer, the villages are not for ”the regular nuisance between two neighbours where one has the stereo too loud on Saturday night” but ”people who are extremely violent and intimidating, and in a clear situation where a victim is being repeatedly harassed”.

Those deemed guilty of causing ”extreme havoc” will be evicted and placed in temporary homes of a ”basic” nature, including converted shipping containers in industrial areas of the city. ”We call it a living container,” says Boer. Housing antisocial tenants in these units, which have showers and kitchens and have been used as student accommodation, will ensure that they are not ”rewarded” by being relocated to better accommodation.

B.1.2 Easy Version

To tourists, Amsterdam still seems very liberal. Recently the city’s Mayor told them that the coffee shops that sell marijuana would stay open, although there is a new national law to stop drug tourism. But the Dutch capital has a plan to send antisocial neighbours to ”scum villages” made from shipping containers, and so maybe now people won’t think it is a liberal city any more.

The Mayor, Eberhard van der Laan, says his new plan to solve the problem of antisocial behaviour will cost £810,000. The plan hopes to protect victims of abuse and homophobia. The camps, where antisocial families will live for three to six months, have been called ”scum villages” because last year Geert Wilders, the far-right politician, said that offenders should go to ”a village for scum”.

Bartho Boer, a spokesman for the Mayor, says that the plans are not illiberal. ”We want to defend the liberal values of Amsterdam,” he says. ”We want everyone to be who he and she is - whether they are gay and lesbian or try to stop violence and are then victims of harassment. We want to defend them.” According to Boer,

the villages are not for ”a problem neighbour who has the stereo too loud on Saturday night” but ”people who are very violent and in a clear situation where a victim is harassed again and again”.

People found guilty of violent harassment will be evicted from their homes and put in temporary homes, including shipping containers in industrial areas of the city. ”We call it a living container,” says Boer. The containers have showers and kitchens and have been used as student accommodation. They are going to use the containers because they want to show that if people are antisocial they do not get better accommodation.

B.1.3 Recall Questions

1. What is the new national law about?
2. What is Amsterdam known for?
3. What are scum villages made from?
4. Who is going to live in the scum villages?
5. For how long will tenants be rehoused?
6. Who is Bartho Boer?
7. Where are the temporary houses placed?
8. Who has lived in the temporary houses before?

B.1.4 Comprehension Questions (Yes/No answers)

1. The mayor of Amsterdam wants the city to remain liberal.
2. The plan is to rehouse gays and lesbians to safer areas.
3. Geert Wilders proposed a new law that suggests sending repeat offenders to scum villages.
4. The mayor of Amsterdam echoed proposals that Geert Wilders be sent ”to a village for scum”.

5. Marijuana-buying tourists are unaffected by the mayor's proposal.

6. Student housing containers including showers and kitchens are not considered to be attractive accommodation.

B.2 Text 2

B.2.1 Difficult Version

For almost 125 years, the secrecy surrounding the recipe for Coca-Cola has been one of the world's great marketing ploys. As the story goes, the fizzy drink's famous "Merchandise 7X" flavourings have remained unchanged since they were concocted in 1886. Today, the recipe is entrusted only to two Coke executives, neither of whom can travel on the same plane for fear the secret goes down with them.

Now, one of America's most celebrated radio broadcasters claims to have discovered the Coke secret. Ira Glass, presenter of the public radio institution This American Life, says he has tracked down a copy of the recipe, the original of which is still supposedly held in a burglar-proof vault at the Sun Trust Bank in Atlanta, Georgia.

The formula was created by John Pemberton, an Atlanta chemist and former Confederate army officer who crafted cough medicines and other concoctions in his spare time. In 1887, he sold the recipe to a businessman, Asa Griggs, who immediately placed it for safekeeping in the then Georgia Trust Bank.

Glass came across a recipe that he believes is the secret formula in a back issue of Pemberton's local paper, the Atlanta Journal-Constitution, while he was researching an entirely different story. Tucked away on an inside page of the 8 February 1979 edition, he stumbled on an article that claimed to have uncovered the closely guarded 7X formula.

The column was based on information found in an old leather-bound notebook that belonged to Pemberton's best friend and fellow Atlanta chemist, RR Evans. Glass was intrigued and, after some digging, found that the notebook had been handed down the generations until it reached a chemist in Georgia called Everett

Beal, whose widow still possesses it.

B.2.2 Easy Version

The recipe for Coca-Cola has been a secret for almost 125 years. This has been an important part of Coca-Cola's marketing plans. According to the story, the famous seven flavourings used in the fizzy drink have not changed since Coca-Cola was first made in 1886. Today, people say, only two Coke executives know the recipe. They cannot travel together on the same plane in case there is a crash and the secret dies with them.

Now, one of America's most famous radio broadcasters says he has discovered the Coke secret. Ira Glass, of the programme This American Life, says he has found a copy of the recipe. People believe the original recipe is kept in a bank in Atlanta.

John Pemberton, an Atlanta chemist, first created the recipe for Coca-Cola. In 1887, he sold the recipe to a businessman who immediately placed it in a local bank so it would be safe.

Glass found a recipe that he believes is the secret formula in an old copy of a local newspaper while he was researching a different story. On an inside page he found an article about the secret formula with seven flavourings.

The recipe came from an old notebook that belonged to Pemberton's best friend, RR Evans. Glass did some research and found the notebook had been passed from generation to generation until it reached a chemist in Georgia called Everett Beal.

B.2.3 Recall Questions

1. For how long has the Coca-Cola recipe been a secret?
2. Who is the Coca-Cola recipe entrusted to today?
3. Who claims to have a copy of the recipe?
4. Where is the original recipe of Coca-Cola supposedly held in?
5. What is the occupation of John Pemberton?

6. In what kind of paper did Glass come across the presumptive formula?
7. Where was the information found the column was based on?
8. How did Everett Beal got to possess the notebook?

B.2.4 Comprehension Questions

1. The two coke executives always travel together to keep the recipe safe.
2. Keeping the Coke recipe a secret was part of a marketing plan.
3. An old notebook with the Coke recipe is held in a bank in Georgia.
4. The Coke secret was discovered by a columnist from a local newspaper.
5. John Pemberton created the formula and placed the original recipe in a bank.
6. The article uncovering the seven flavourings was placed on an inside page of a local paper.

B.3 Text 3

B.3.1 Difficult Version

Tigers are more numerous in Nepal than at any time since the 1970s, a new census has revealed, giving conservationists hope that the big cats, whose numbers have been dropping across south Asia for 100 years, can be saved.

The number of wild royal bengal tigers in Nepal has increased to 198 – a 63.6% rise in five years – the government survey showed. "This is very encouraging," said Maheshwar Dhakal, an ecologist with Nepal's Department of National Parks and Wildlife Conservation.

The census is based on the examination of pictures from more than 500 cameras placed in five protected areas and three wildlife corridors. More than 250 conservationists and wildlife experts worked on the survey, which cost about £ 250,000. Dhakal said that a parallel survey was conducted in India and the results from both countries will be published later in 2013. "It will take a few more

months for India, which now has 1,300 big cats in several huge protected areas, to finalize the data," he added. Nepal has pledged to double the population of tigers by the year 2022 from 121 in 2009 when the last systematic tiger count took place.

Increasing prosperity in Asia has pushed up prices for tiger skins and the body parts used in traditional Chinese medicines. International gangs pay poor local Nepali significant sums to kill the cats. The skin and bones are handed to middlemen, who pass easily through the porous border to India, where the major dealers are based.

B.3.2 Easy Version

According to a new survey, there are more tigers in Nepal than at any time since the 1970s. The number of big cats has been decreasing in south Asia for 100 years, but conservationists now hope that we can save them.

The number of wild royal bengal tigers in Nepal has increased to 198 – a 63.6% increase in five years – the survey showed. "This is very good news," said Maheshwar Dhakal, an ecologist with Nepal's Department of National Parks and Wildlife Conservation.

The survey looked at pictures from more than 500 cameras in five protected areas and three wildlife corridors. More than 250 conservationists and wildlife experts worked on the survey, which cost about £ 250,000. Dhakal said that there was a similar survey in India and the results from both countries will be published later in 2013. "It will take a few more months for India, which now has 1,300 big cats in several huge protected areas, to finish the survey," he added. Nepal says it will double the population of tigers by the year 2022 from 121 in 2009 to 242.

Some rich people want tiger skins. Tiger body parts are used in traditional Chinese medicine. International gangs pay poor local Nepali people large amounts of money to kill the cats. The skin and bones are taken through the border to India, where the big dealers are.

B.3.3 Recall Questions

1. What information is the statement based on that tigers are more numerous in Nepal than at any time since the 1970s?

2. For how long has the number of tigers been decreasing across south Asia?
3. What kind of tigers showed a 63.3% rise of population in five years in Nepal?
4. Who is Maheshwar Dhakal from Nepal's Department of National Parks and Wildlife Conservation?
5. Besides protected areas, where are the cameras from the census placed ?
6. Which country also does a survey on their big cats?
7. Who pays poor local Nepali to kill the cats?
8. Where are the major dealers of tiger parts located?

B.3.4 Comprehension Questions (Yes/No answers)

1. The number of tigers in Nepal is increasing because they are moving in from India.
2. Pictures from more than 500 cameras were evaluated to catch people hunting the tigers.
3. More than 250 conservationists and wildlife experts work on the Indian survey.
4. In Nepal the population size of tigers is smaller than in India.
5. If interest in traditional chinese medicine increases, tiger protection will become even more important.
6. To evaluate the effort of doubling the tiger population until 2022, the cameras will again be needed.

B.4 Text 4

B.4.1 Difficult Version

Happiness is found by living in the now, particularly if the now involves having sex, according to a major study into mental wellbeing. But the study also found that people spend nearly half their time (46.7%) thinking about something other than what they are actually doing.

The benefits of living in the moment are extolled by many philosophical and religious traditions, but until now there has been scant scientific evidence to support the advice. Psychologists at Harvard University collected information on the daily activities, thoughts and feelings of 2,250 volunteers to find out how often they were focused on what they were doing, and what made them most happy. They found that people were happiest when having sex, exercising or in conversation, and least happy when working, resting or using a home computer. And although subjects' minds were wandering nearly half of the time, this consistently made them less happy.

The team concluded that reminiscing, thinking ahead and daydreaming tend to make people more miserable, even when they are thinking about something pleasant. Even the most engaging tasks failed to hold people's full attention. Volunteers admitted to thinking about something else at least 30% of the time while performing these tasks, except when they were having sex, when people typically had their mind on the job around 90% of the time.

"Human beings have this unique ability to focus on things that aren't happening right now. That allows them to reflect on the past and learn from it; it allows them to anticipate and plan for the future; and it allows them to imagine things that might never occur," said Matthew Killingsworth, a doctoral student in psychology and lead author of the study. "At the same time, it seems that human beings often use this ability in ways that are not productive and, furthermore, can be destructive to our happiness," he added.

B.4.2 Easy Version

According to a new study into mental wellbeing, people are happiest if they live for the present moment, particularly if this involves having sex. But the study also found that people spend almost half their time (46.7%) thinking about other things and not about what they are actually doing.

Many philosophical and religious traditions tell us that we should live for today. However, until now there has not been much scientific evidence to support this idea. Psychologists at Harvard University collected information on the daily activities, thoughts and feelings of 2,250 volunteers to find out how often they concentrated on what they were doing, and what made them most happy. They found that people were happiest when having sex, exercising or having a conversation. People were least happy when working, resting or using a home computer. They also found that people were not concentrating nearly half of the time and that this made them less happy.

The researchers found that thinking about the past, thinking ahead and daydreaming make people more miserable, even when they are thinking about something pleasant. Even the most interesting tasks did not make people concentrate all the time. Participants in the study said they were thinking about something else at least 30% of the time while they did these tasks, except when they were having sex, when they were concentrating around 90% of the time.

"Humans are the only creatures that can think about things that aren't happening right now. They can think about the past and learn from it; they can think about the future and plan for it; and they can also imagine things that might never happen," said Matthew Killingsworth, the main researcher. "At the same time, human beings often use this ability in ways that are not productive, and it can also make us unhappy," he added.

B.4.3 Recall Questions

1. What was the study about?

2. How much time do people spend thinking about something other than WHAT they're actually doing?

3. Besides thoughts and feelings, what kind of information did Psychologists at Harvard University collect?

4. How many volunteers participated in the study?

5. When were people happiest besides WHEN having sex or a conversation?

6. When were people least happy besides when working or using a home computer?

7. even FOR THE most interesting tasks, What did THE TASKS fail to do?

8. Who is Matthew Killingsworth?

B.4.4 Comprehension Questions (Yes/No answers)

1. The findings from the study confirm many Philosophical and religious traditions.

2. The study implies that people who feel miserable at a given moment can improve their well-being by day dreaming.

3. The aim of the study was to test the effectiveness of trainings intended to make people happier.

4. Most people are able to stay concentrated while having sex.

5. Researchers interviewed psychologists at Harvard University to find out how daily activities relate to mental well-being.

6. According to the study, people are unhappy almost half of their time.

Appendix C

C-Test for English Proficiency, used in the Eye-tracking experiment

More than 3.2m homes and businesses across the north-east US have been left without power after a freak snowstorm killed at least eight people and disrupted transport across the region. From Maryland t[ ] Maine, offi[ ] said i[ ] would ta[ ] days t[ ] restore elect[ ], even tho[ ] the snow[ ] ended o[ ] Sunday. T[ ] storm sma[ ] record tot[ ] for Oct[ ] and wors[ ] as i[ ] moved no[ ]. Communities i[ ] western Massachusetts we[ ] among t[ ] hardest h[ ]. The wea[ ] was bla[ ] for a[ ] least s[ ] deaths, a[ ] states of emergency were declared in New Jersey, Connecticut, Massachusetts and parts of New York. Roads and railways became blocked and flights cancelled, with passengers on a JetBlue flight stuck on a plane in Hartford, Connecticut, for more than seven hours on Saturday.

Airports are wasting billions of pounds on unnecessary security checks for travellers who pose no threat to planes, according to the airline industry's global body, amid growing support for an airport screening regime that gives preferential treatment to low-risk passengers. The Intern[ ] Air Tran[ ] Association, wh[ ] members inc[ ] British Air[ ], Virgin Atla[ ] and mo[ ] than 200 glo[ ] airlines, sa[ ] main term[ ] were strug[ ] to co[ ] with moun[ ] layers o[ ] safety regul[ ] that n[ ] cost t[ ] financially trou[ ] industry $7.4bn (?4.6bn) a ye[ ] to impl[ ]. Tony Tyler, dire[ ] general o[ ] IATA, sa[ ]: "We sp[ ] a hu[ ] amount of resource on screening people who quite frankly do not need it. "We need to find a better way of doing it.

I've never met anyone who didn't like pizza. This econ[ ] migrant fr[ ] impoverished Nap[ ] is t[ ] epitome o[ ] the Amer[ ] dream: popul[ ] by t[ ] Italian comm[ ], adapted t[ ] suit n[ ] world tas[ ] and th[ ] exported aro[ ] the wo[ ], it's t[ ] ultimate immi[ ] success st[ ]. Of cou[ ], Italians ca[ ] take a[ ] the cre[ ] for wh[ ] is qu[ ] simply t[ ] world's best snack. As the Oxford Companion to Food points out, the linguistic link between pizza and pitta is surely no coincidence ? topped breads have been popular around the Mediterranean since classical times, and Etruscans were baking schiacciata in the Tuscan region over 2,000 years ago.

Natural light is just one potential factor in a child's eyesight. The ti[ ] children sp[ ] outdoors co[ ] be lin[ ] to a red[ ] risk o[ ] being sh[ ]-sighted, rese[ ] suggests. A[ ] analysis o[ ] eight prev[ ] studies b[ ] Cambridge Unive[ ] researchers fo[ ] that f[ ] each addit[ ] hour sp[ ] outside p[ ] week, t[ ] risk fr[ ] myopia fe[ ] by 2%. Expo[ ] to amb[ ] light a[ ] looking a[ ] distant objects could be key factors, they said. The studies involved more than 10,000 children and adolescents.

Jack London called it "the call of the wild", but for us it's the call of the world ? and we are responding while we can still walk. For ag[ ] we ha[ ] envied sch[ ] leavers th[ ] gap ye[ ], their backp[ ] wanderings wit[ ] needing t[ ] rush ho[ ] after a f[ ] weeks t[ ] pay hom[ ] to t[ ] grindstone. T[ ] wild id[ ] has be[ ] in gest[ ] for we[ ] over a ye[ ], in t[ ] run-u[ ] to my wife's retir[ ]. I a[ ] a free[ ] writer and have for years been able to drop everything at the drop of a hat, but Vivienne is a psychotherapist, and the idea of abandoning patients for long periods was out of the question. Now she has done the brave deed and retired, and there's no question but that next Saturday we fly off to our first port of call: Cape Town.