Automatic Readability Assessment

by

Lijun Feng

A dissertation submitted to the Graduate Faculty in Computer Science
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy,
The City University of New York

2010

This manuscript has been read and accepted for the Graduate Faculty in Computer Science in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Matt Huenerfauth

Date | Chair of Examining Committee

Theodore Brown

Date | Executive Officer

Noémie Elhadad	Columbia University
Heng Ji	Queens College, CUNY
Andrew Rosenberg	Queens College, CUNY
Virginia Teller	Hunter College, CUNY

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Automatic Readability Assessment

by

Lijun Feng

Adviser: Professor Matt Huenerfauth

We describe the development of an automatic tool to assess the readability of text documents. Our readability assessment tool predicts elementary school grade levels of texts with high accuracy. The tool is developed using supervised machine learning techniques on text corpora annotated with grade levels and other indicators of reading difficulty. Various independent variables or features are extracted from texts and used for automatic classification. We systematically explore different feature inventories and evaluate the grade-level prediction of the resulting classifiers. Our evaluation comprises well-known features at various linguistic levels from the existing literature, such as those based on language modeling, part-of-speech, syntactic parse trees, and shallow text properties, including classic readability formulas like the Flesch-Kincaid Grade Level formula. We focus in particular on discourse features, including three novel feature sets based on the density of entities, lexical chains, and coreferential inference, as well as features derived from entity grids. We evaluate and compare these different feature sets in terms

of accuracy and mean squared error by cross-validation. Generalization to different corpora or domains is assessed in two ways. First, using two corpora of texts and their manually simplified versions, we evaluate how well our readability assessment tool can discriminate between original and simplified texts. Second, we measure the correlation between grade levels predicted by our tool, expert ratings of text difficulty, and estimated latent difficulty derived from experiments involving adult participants with mild intellectual disabilities. The applications of this work include selection of reading material tailored to varying proficiency levels, ranking of documents by reading difficulty, and automatic document summarization and text simplification.

To 蕤佳

and my parents 徐世凤 (09.11.1941 – 17.01.2010) and 冯立恒

Acknowledgments

My deep gratitude goes first to three people who shaped this thesis: my adviser Matt Huenerfauth, Noémie Elhadad, and Martin Jansche. Without any of them, I could not have come close to finishing this thesis.

As my adviser, Matt has supported me in the past years in every way that one could wish: he has guided my study wholeheartedly with patience and enthusiasm, given me complete freedom to explore my research interests, sought every opportunity to allocate resources to support and facilitate my study, and offered me his time and friendship generously with utter understanding whenever I needed them. I'm truly grateful to have such a wonderful mentor.

I am fortunate to know Noémie, who had an equally important role in my graduate study. She inspired my interest in NLP through the projects she assigned in her 2006 Statistical NLP course. She introduced me to Matt, and I was fortunate to be involved in a collaborative project between them. Their framework for an envisioned text simplification system for adults with intellectual disabilities laid the groundwork for the evolvement of my thesis. Throughout my study, she generously set time aside to meet with me regularly and guided my work together with Matt. I am truly thankful to her invaluable mentoring, as well as her warm friendship.

Equally important to my academic progress is my best friend and husband Martin. Aside from his unconditional love, kindness, patience and encouragement, with which he has been supporting me through my whole PhD years, he also provided critical technical guidance, advice and help for my thesis study. As a knowledgeable NLP expert, his comments and candid critiques often helped me stay focused on what still needs to be improved. I thank him for sacrificing his valuable sleeping time to validate my code, fix/maintain my computer, help me catch mysterious bugs and have discussions with me on my endless questions, for listening to my frustrations and cheering me up with his bright humor, and for putting up with my daily complaints during my entire thesis writing.

I would like to thank my thesis committee Virginia Teller, Heng Ji and Andrew Rosenberg for reading the thesis and giving me valuable feedback. I thank Virginia for always being there for me – from my survey exam to my thesis defense – asking insightful questions and pushing me to address the bigger picture. I appreciate Heng's kind gestures in making resources available for my study and constantly watching out future opportunities for my research. I thank Andrew for having many discussions with me, teaching me statistical methods and giving me useful suggestions on improving my presentation skills.

I would like to thank the following people and organizations for their various roles in my personal and academic life: Wei Chen, Paola Garcia, Allen Harper, Tiziana Ligorio, Pengfei Lu, Karthik Natarajan, Josh Waxman and Henry Wiyanto for helping conduct user studies; Tsan Huang, Janice and Martin for annotating reading difficulty of LocalNews2008 for me with their expertise knowledge; Shravan Vasishth for teaching me statistical meth-

ods; Noémie Elhadad and Regina Barzilay, the Weekly Reader Corporation and LiteracyNet for making their corpora available for my research; the CUNY Research Foundation for funding the 2008 user study; Ted Brown for securing funding and teaching positions for me through my study at CUNY; professors Robert Haralick, Amotz Bar-Noy, Noson S. Yanofsky, Neng-Fa Zhou and Dexter D. Scott for inspiring my interest in Computer Science and encouraging me to pursue a higher degree in this field; Lina Garcia, our assistant program officer, for her indispensable mastery of time and space.

Last year around this time, while preparing for my thesis proposal, I was struggling with deep pain when I heard from home that my mom was seriously ill. This pain has never ceased since then. Mom, had I known that's our last visit, I would have put off everything to spend more time with you. Nothing could be compared with your role in my life. I thank you and Dad for bringing me up as I am. I love you. Whenever I see the moon light outside the window ever so bright and cool, I miss you.

Without the love, support and comfort from my whole family, I could not have gained strength so soon to continue and finish my thesis work. I thank my Dad, my brother and three sisters, together with their families, for always being strong and there for me. I am thankful for the frequent phone calls from home, and cherish the late night conversations with and messages from my beloved sisters.

Inge and Walter, your yearly visits always bring me joy and warmth, I am deeply in debt to your love.

Flora, you are such a 乖乖 – you make me happy, you make me smile. And Martin, you too.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Readability is commonly defined as a measure of ease with which a written text can be understood. What makes a text easy or hard to read has been the central topic of readability research for the past 80 years and continues to attract considerable interest. Historically, researchers have approached this problem with the assumption that the readability of a text could be measured by a simple function of a few text properties that are objective and easy to determine (Miller and Kintsch, 1980). Many traditional metrics that claim to measure text readability often relied on a limited set of superficial text features, such as sentence length, number of syllables per word, word frequency, etc. These metrics are easy to compute, but they have been proven to be highly unreliable (Collins-Thompson and Callan, 2004; Feng et al., 2009; Petersen and Ostendorf, 2009; Si and Callan, 2001).

With the advancement of Natural Language Processing (NLP) technology, text readability has received increased attention (Barzilay and Lapata, 2008; Collins-Thompson and Callan, 2004; Heilman et al., 2007, 2008; Petersen and Ostendorf, 2009; Pitler and Nenkova, 2008; Schwarm and Ostendorf,

2005; Si and Callan, 2001). Language models and parsers have been used to explore more complex lexical features and syntactic constructs in aiding readability study. Compared with the earliest readability scores, the number of independent variables explored by recent statistical methods has grown considerably. However, like their earlier predecessors, these modern approaches are still mostly limited to statistical analysis of a text at lexical and syntactical level.

There is evidence that readability formulas do measure factors, such as sentence length and complexity of vocabulary, that may reflect readability (Bormuth, 1966). But it is clear that the limited number of independent variables explored by traditional and recent approaches capture only a fraction of all text properties that actually contribute to text comprehensibility. Aside from vocabulary complexity and sentence structure, many important factors, such as the structure of the text, the definition of discourse topic, discourse cohesion and coherence, the purpose of the author and so on, play a central role in determining reading difficulties of a text (Freeman, 1978; Gourlay, 1978; Kintsch and Vipond, 1979) . However, up until recently, readability research has made little progress beyond lexical and syntactic analysis. It is partly because lexical and syntactic features are easier to define and measure with existing techniques, while factors such as discourse topic and discourse coherence require much more complex semantic analysis, and thus remain as challenging problems.

It is commonly agreed that reading ease is not determined by intrinsic text properties alone; rather, it results from an interaction of complex language comprehension processes between the reader and the text (Davison and Kantor, 1982; Gray and Leary, 1935; Miller and Kintsch, 1980). In

addition to innate text properties, factors on the readers' side, such as the readers' literacy skills, prior knowledge, motivation and interest to read also influence the readability or comprehensibility of a text directly. Traditional and recent approaches to readability have rarely addressed factors from the readers' side. Many readability models were created without any particular group of readers in mind.

Research development in cognitive science concerning reading has greatly advanced our understanding of cognitive processes underlying text comprehension. Established major theories have commonly agreed that the goal of reading is to construct a coherent memory representation of a text by the reader. This definition of reading points out a few important aspects for readability research: a) the constructive nature of reading indicates that reading ease results equally, if not more, from the reader's active effort to comprehend a text as the text itself. b) Constructing a coherent memory representation of a text involves complex language comprehension activities, most of them occur at high level discourse comprehension rather than word identification and sentence processing. c) Various memory systems, in particular working memory, have great influence on the quality of these language comprehension activities.

The following thesis presents research on developing an automatic text readability assessment tool, focusing on advancing readability related features to various discourse levels while taking user characteristics into account. The primary goal of the thesis is to advance our understanding of, and quantify, what makes a text easy or difficult to read, in particular for readers with mild intellectual disabilities (MID). In addition to enriching features studied by previous research, such as those based on language

modeling, part-of-speech, syntactic parse trees, and shallow text properties, we focus in particular on novel discourse features based on density of entities, lexical chains, coreferential inference, as well as those derived from well-known entity grids. In designing discourse feature representations, we take working memory, which plays a fundamental role in the process of discourse comprehension activities, and the limitation of which is especially characteristic to adults with intellectual disabilities, into account. We use supervised machine learning techniques to build classifiers with these features and evaluate and compare their effectiveness in detecting and predicting reading difficulty of texts assigned with elementary grade levels. In order to assess how well our readability assessment tool generalize to texts from different domain, we manually created corpora consisting of original and simplified texts adapted specifically for adults with mild intellectual disabilities. We evaluate how well our tool can differentiate between original and simplified texts. We compare the correlations of predictions by our tool with independent measure of text difficulty rated by experts and estimated latent difficulty derived from experiments involving adult participants with mild intellectual disabilities. The ultimate goal of this study is not simply to model and understand readability issues, but also to aid in the development of automatic language processing tools that can rewrite texts to be more readable. We envision that this work can be useful in a variety of applications, including selection of reading material tailored to varying proficiency levels, ranking of documents by reading difficulty, and automatic document summarization and text simplification.

This thesis is organized in the following way. Chapter 2 discusses the motivations and background of the proposed study. Chapter 3 surveys

relevant literature and previous work with critiques. Chapter 4 presents research guidelines and methods of the study and describes our research hypothesis concerning particular reading difficulties arising from various level of discourse comprehension. Chapter 5 describes the characteristics of various corpora and how they were used in the study. Chapter 6 presents techniques used to extract features at various linguistic levels and describes implementation of features in detail.

Chapter 7 presents research on building and evaluating an automatic readability assessment tool on corpora annotated with grade levels. We use a set of efficient and robust machine learning techniques to address the task of automatic text readability prediction. This chapter consists of three major contributions. Firstly, in addition to refining and improving previously explored features, we propose and evaluate four subsets of discourse features, three of them are novel and have not been studied before in readability research. Secondly, we conduct thorough experiments and analysis to assess and compare feature effectiveness at various linguistic levels, which has not been done in the field before. Thirdly, we experiment with various combinations of features at various linguistic levels to improve and optimize model performance. Our best model achieves 74% accuracy, outperforming the current state of art (accuracy 63%) by nearly 11%.

In Chapter 8 we further evaluate our automatic text readability assessment tool on unseen data. We focus on investigating how well models built with grade levels generalize to unseen texts and what are their limitations. It is challenging to evaluate model performance on unseen data, because the reading difficulty of the texts contained within is unknown. We address this problem by two separate approaches: we have experts annotate text

difficulty of the same set of data; we design a reading experiment based on the same set of data, recruit adult readers with ID to read assigned texts and answer simple comprehension questions. We then use a hierarchical latent trait model to infer text difficult based on test participants' reading ability (this is based on joint work, see Section 8.4 for details). These two alternative measures of text difficulty allow us to evaluate model performances and investigate relations between grade level predictions, expert ratings and inferred text difficulty for adults with ID.

Chapter 9 summarizes major observations, conclusions and contributions of this thesis, and proposes directions for future work.

Chapter 2

Motivations & Background

To better understand our approach and the motivations of our research, we outline in section 2.1 dominant theoretical framework concerning cognitive processes underlying reading comprehension. In section 2.2, we describe particular reading difficulties that are often characteristic of adult readers with ID. A broad setting of possible applications of our research are discussed in section 2.3.

2.1 Understanding Reading Comprehension

Research on text comprehension in the past 40 years has greatly advanced our understanding of cognitive processes underlying reading comprehension. Several influential theories have been developed concerning the encoding, representation, retrieval, and application of linguistic and other types of knowledge crucial to successful text understanding (Lorch and van den Brock, 1997). Despite their differences, the theoretical frameworks have generally shared a consensus view that the goal of reading is to con-

struct a coherent memory representation of a text (Anderson and Bower, 1973; Collins and Loftus, 1975; Frederiksen, 1975; Gernsbacher, 1990, 1997; Haviland and Clark, 1974; Kintsch and van Dijk, 1978; Schank, 1975). The significance of viewing text comprehension as memory construction is that it defines comprehension in terms of the coherence of the representation the reader constructs (Lorch and van den Brock, 1997) and the relation between the reader and the text. How well a text is comprehensible to the reader depends on to what extent the reader's representation captures the local and global coherence relations intended by the author (Lorch and van den Brock, 1997).

Reading is a complex cognitive process that requires the smooth coordination of various cognitive and language abilities as well as memory systems. Many theories have been developed since the early 1970s to understand the cognitive processes which a reader goes through to process a text and construct a memory representation. One of the most important contributions made by the development of theories is *knowledge representation*, which provides semantic networks as a metaphor for text structure. A reader's task in the early stage of reading involves word identification and sentence processing, with the goal of extracting meaning from basic component units of the text. These basic meaning units, often referred to as propositions, are then stored in memory systems to form the building blocks for the memory construction. According to semantic network theory, these meaning units are not placed in memory randomly; rather, they are organized and structured (Gernsbacher, 1990, 1997; Harm and Seidenberg, 1999; Kintsch and van Dijk, 1978; Schank, 1975; Stanovich, 1985) by various relations, in particular inferential relations, such as referential, causal, spa-

tial, temporal, instrumental, predictive or elaborative inferences (Lorch and van den Brock, 1997).

The framework of semantic network structure implies that much of the comprehension process involves identifying and representing relations among propositions. Reading comprehension is often viewed as an on-line process: Information is processed sequentially at sentence to sentence level. The reader uses the incrementally available information on-line to direct the comprehension processes. In the initial stage of reading, conceptual information processed from the first few sentences is organized and structured to form the foundation of a semantic network, which serves as the base of the reader's memory representation. A coherent memory representation is constructed and maintained by the reader's ability to identify the relations among concepts and propositions and to connect them to the existing semantic network when processing each new sentence. In many cases, relations among propositions are not stated explicitly. Rather, they are established by referential devices, such as the use of anaphora and ellipses, or carefully structured by implicit causal or other types of inferential relations. The reader has to solve references to establish entities in a text and make appropriate inferences by applying previously acquired knowledge to fill in implicit relations.

The perspective of reading comprehension as memory construction also emphasizes the central role memory systems, in particular working memory, play in various language comprehension activities. Working memory is the cognitive system responsible for temporary storage and simultaneous manipulation of information (Merrill et al., 2003; Stanovich, 1985). During the on-line processes of reading comprehension, working memory provides

access to processed text information for search and retrieval of relevant referents necessary for comprehension. To facilitate the construction of coherent memory representations, working memory also provides mechanisms to enhance relevant information and suppress contextually irrelevant information (Gernsbacher, 1990, 1997; Merrill et al., 2003).

The greatest impact that working memory has on language comprehension is the storage and processing function it provides during memory construction of discourse. However, the resources working memory can provide and coordinate for comprehension processes is not unlimited. A widely accepted measure of working memory capacity, called working memory span, was developed by Daneman and Carpenter (1980). They had individuals read or listen to a series of unrelated sentences. After the whole set of sentences was presented, subjects were asked to recall the last word of each sentence. This task requires that test participants use both processing and storage components of the working memory. The working memory span is defined as the number of sentences that individuals could process and still recall the last word of each sentence. Daneman and Carpenter's method has been widely used by many groups of researchers. Numerous empirical results have shown that individual differences in working memory capacity is highly correlated with variation in both overall reading ability and specific reading skills (Daneman and Carpenter, 1980; Daneman and Merickle, 1996; Daneman and Tardif, 1987; Dixon et al., 1988; Just and Carpenter, 1992; King and Just, 1991; Masson and Miller, 1983). Low working memory capacity has been shown to be related to a reduction in the speed and accuracy with which sentences can be processed (King and Just, 1991). The decline in comprehension performance for poor language

comprehenders was generally assumed to be related to added processing requirements associated with language comprehension activities, thus fewer resources of working memory capacity were left for storage and retrieval functions (Merrill et al., 2003). In general, the relation between working memory capacity and language comprehension can be stated as follows: when increasing demands are placed on the reader's working memory capacity, or when the reader must use more of the available capacity for processing activities, comprehension performance is likely to suffer (Bilsky, 1985; Conners, 2003; Merrill et al., 2003; Stanovich, 1985).

Working memory capacity has been demonstrated to be a good predictor of general language comprehension ability because of its high correlation observed in empirical studies with general measures of language comprehension. In addition, research has also shown that the comprehensibility of a text can be well predicted by an analysis of the demands it places on the reader's working memory (Britton and Gulgoz, 1991; Miller and Kintsch, 1980). Based on Kintsch's theory that reading is to construct a coherent memory representation of a text (Kintsch and van Dijk, 1978), Miller and Kintsch (1980) implemented a computational prose processing model to simulate certain aspects of comprehension processes that a reader must go through in order to construct a coherent memory representation. Their assumption was that at those points in the comprehension process at which the model has difficulty locating and maintaining coherent relations, human readers should experience similar difficulties (Miller and Kintsch, 1980). Two types of these particular difficulties involve reinstatements and inferences. The model assumes certain constraints placed on the readers' working memory capacity and asserts that only a fraction of already read

text can be held in working memory. If a segment of text is read which is not related to the current contents of working memory, long-term memory has to be searched for relevant text that has already been processed. If the search is successful, that part of the text is reinstated in working memory to maintain the coherence of the text. If the search is not successful, the reader has to make appropriate inferences based on previously acquired knowledge (Miller and Kintsch, 1980). In their study, 20 short texts of varying readability were selected and propositions contained in each text were hand-annotated. The model simulates the text processing task of the reader by operating on the propositions with pre-defined rules. Each text was then read by 120 students and reading time and recall were assessed. As expected, the experiment results show high correlation of model predictor variables, such as inferences and reinstatements, with subjects' reading time, recall and text readability (defined as reading time per proposition recalled in terms of correlations). Subsequent studies following a similar approach by Britton et al. (1990) and Britton and Gulgoz (1991) further corroborated that inferences are major sources of text locations where reading difficulties occur and inserting inferences makes instructional texts more readable.

Our research on text readability benefits from established major theories and important empirical findings concerning reading by viewing readability as a result of the interaction between the text and the reader's prose processing ability. These theories and findings support our emphasis on the following elements of our approach to readability:

- Text readability is not determined by intrinsic text properties alone. Rather, reading ease or difficulty results from the interaction of the reader and the text.

- The goal of reading is to construct a coherent memory representation of a text. Word identification and sentence parsing are part of basic comprehension processes that occur at the low level of text comprehension. Much of reading difficulties arise from higher level of discourse comprehension, which involves mostly evaluating and identifying relations among conceptual information, solving references to establish entities in a text and making various types of inferences to fill in missing information.

- Working memory has great impact on various language comprehension activities, because it provides temporary storage and simultaneous manipulation of information and coordinates resources that are necessary for comprehension processes during reading.

- Working memory capacity constantly places constraints on readers' attempt to understand a text. Individual differences in working memory capacity account for some of the variation in comprehension performance.

- Text comprehensibility can be well predicted by an analysis of the demands it makes of readers' working memory.

2.2 Adults with Intellectual Disabilities and Literacy Challenges

According to the 2006 American Community Survey (U.S. Census Bureau, 2006), about 5% of the civilian non-institutionalized population, approximately 13.5 million people age 16 or above in the United States, have intellectual disabilities (ID), with intelligence test scores of 70 or below. Among this group of people, about 85% are in the category of mild intellectual disabilities (MID) (IQ range 50–75) (Drew and Hardman, 2004). We will use the term "intellectual disabilities" (ID) or "mild intellectual disabilities" (MID) henceforth. People with ID face many challenges in their daily lives; one of these challenges lies in the area of reading literacy. Proficient reading skills are crucial to a successful life in modern society.

However, low literacy is prevalent in individuals with ID. Research on language comprehension of people with ID has consistently found that individuals with ID have reading skills below their mental age and lag far behind their peers without ID (Dunn, 1954; Jones et al., 2006; Katims, 2000; Merrill, 1924; Samuels, 2002; Sheperd, 1967). In particular, this "reading lag" begins as early as a mental age of 8 or 9, and increases thereafter (Conners, 2003; Jenkinson, 1989; Merrill, 1924). Many studies are consistent with this finding. A study conducted by Jones et al. (2006) assessing the reading comprehension of adults with mild intellectual disabilities (MID) reported that the average reading skills of subjects (mean age: 46 years 10 months; standard deviation: 13 years 9 month) were below that of average 7-year-old readers without disabilities.

Several factors contribute to the lower literacy skills of adults with ID.

Above all, the limitation of their cognitive functioning due to various degree of impairments affects their reading comprehension directly. Moreover, research has shown extreme limitations of working memory associated with intellectual disabilities (Hale and Borkowski, 1991; Pulsifer, 1996). As discussed in section 2.1, working memory has a direct influence on language comprehension processes because of its temporary storage function and simultaneous manipulation of information. A large body of literature that focuses on language comprehension differences between individuals with and without ID have found that the cognitive processes of persons with ID during reading comprehension are similar to those of people without ID (Bilsky, 1985; Merrill et al., 2003; Stanovich, 1985). However, their comprehension performance is qualitatively and quantitatively poorer compared with their peers without ID. Among several specific reading difficulties that have been identified as characteristic to individuals with ID, many are assumed to be closely related to working memory capacity.

Several studies have suggested that persons with ID appear to have specific difficulty with phonological decoding (reading by sounding out) (Cawley and Parmar, 1995; Cohen, 1982; Jenkinson, 1992; Mason, 1976, 1977, 1978), which may fundamentally impair their reading (Stanovich, 1985), because literature on persons without ID indicates a strong link between phonological processing ability and reading acquisition (Barron, 1980, 1981; Hogaboam and Perfetti, 1978). Deficit on these skills may play an important limiting factor for reading development later, especially when focus of reading shifts from word identification to sentence processing and more complex discourse comprehension. Fowler (1998) pointed out that phonological difficulties may be a crucial factor in limiting syntactic

development, and what appear to be semantic production problems may ultimately depend on well-specified phonological representations.

The ability to actively and strategically apply one's semantic knowledge to facilitate comprehension activities is considered crucial in understanding differences in individual comprehension performance. In many empirical studies, individuals with ID were observed to show deficits in various aspects of semantic processing. On many semantic information tasks that involved active retrieval and evaluation of conceptual information, individuals with ID appeared to have difficulties with spontaneous activation of appropriate semantic knowledge during the comprehension process: they did not appear to access related background knowledge to facilitate discourse comprehension as readily, regularly or effectively as persons without ID (Bos and Tierney, 1980; Davies et al., 1981; Glidden and Mar, 1978; Merrill and Bilsky, 1990). This kind of deficiency in semantic processing skills would impair comprehension performance of persons with ID directly, because constructing coherent memory representation requires making frequent inferences by applying background knowledge. Less strategic search and retrieval of semantic related information would result in obvious limitations in the accuracy and efficiency of inferential processing (Fowler, 1998). Researchers suggest that specific reading difficulties exhibited in semantic processing by individuals with ID could be related to deficiencies in mechanisms of working memory that facilitate and promote discourse coherence by enhancing relevant information and suppressing less relevant or inappropriate information. It could also be attributed to the extreme limitations of working memory capacity that are typically associated with individuals with ID. If working memory capacity is insufficient and more of the

available resources are used for on-line language processing activities, the retrieval process is slower and less strategic and language comprehension is likely to suffer.

Because many language processing difficulties that have been observed to be specific to individuals with ID are generally assumed to be related to working memory, it has been suggested that working memory may be the single most reliable predictor of reading ability among individuals with ID (Conners, 2003; Merrill et al., 2003).

2.3 Motivations

It is difficult to find reading materials for individuals with MID that are (1) of interest to them and (2) at the right reading level. Reading materials at lower reading levels are typically written for children, and texts written for adults without disabilities often require a high level of linguistic skills and sufficient real world knowledge, which these individuals often lack. The lack of appropriate reading materials may also discourage adults with ID from practicing reading, thus diminishing their already low literacy skills.

The need to identify or reformulate texts suitable for lower reading levels is not unique to people with ID. Children, second language learners, and adults with low literacy skills can also benefit from such texts. However, manually adapting written texts is both time and labor intensive. In the past decade, Natural Language Processing (NLP) techniques have been used to develop automatic text simplification systems to assist not only other NLP tasks like parsing, machine translation, information retrieval, and text summarization (Chandrasekar et al., 1996; Chandrasekar and Srinivas,

1997; Klebanov et al., 2004; Siddharthan, 2004), but also to assist human readers with low literacy or various language impairments, such as aphasia and deafness (Carroll et al., 1998, 1999; Devlin, 1999; Devlin and Unthank, 2006; Inui et al., 2003; Williams and Reiter, 2005). Research has focused mainly on lexical and syntactic simplification. Lexical simplification often uses word frequency or predefined word lists to identify difficult words and replaces them with less formidable synonyms. Syntactic simplification often uses dependency-tree structures and pattern recognition techniques to identify allegedly difficult syntactic constructs, which are assumed to include relative clauses and passive voice, and which may even include conjoined sentences. Transformation rules are then applied that change these constructs into shorter or plainer sentences and as a result they are thought to be easier to understand.

People with MID would certainly benefit from texts simplified in this fashion. However, synonym-replacement and syntax-tree simplification alone cannot fully cover the needs of this group of users, because, in addition to challenges that come from lexical and syntactic factors, they have other difficulties with processing written information. As discussed above, many of these difficulties arise from discourse processing. Moreover, most earlier text simplification systems process input text one sentence at a time, which inevitably results in increased length of the simplified document, because long and complex sentences are often split into multiple shorter sentences. The resulting increased length of the whole document can pose another challenge to the already limited working capacity of readers with MID because it requires processing and storing more information.

Our research on readability is partly motivated by an envisioned long-

term project on designing and implementing an automatic text simplification system that modifies a text at the discourse level to meet the special needs of the underrepresented group of individuals with MID. In addition to lexical and syntactic simplification, the envisioned discourse simplification entails high-level semantic simplification, whereby the most relevant information is retained and less relevant information simplified or completely left out (Feng, 2008). In designing such a system, we face several open, foundational questions, which are both self-contained and crucial for further research, and thus form a stand-alone dissertation project. There are two major research questions that are at the center of the design and implementation of such a text simplification system (Inui et al., 2003): (1) How do we identify which portions of a text will pose difficulty for our users? (2) When there are several possible simplification choices, how do we decide which is the optimal one to choose for our users? Ideally, a reliable automatic readability assessment tool would help solve both questions. Readability assessment is an important issue in designing and evaluating an automatic text simplification system. A reliable readability assessment tool can aid automatic text simplification in many ways. Depending on the needs of the application, such a tool can be used before the start of the simplification process to select among texts on similar topics the easiest one to begin with. This would be the case for our envisioned text simplification system for adults with ID. During the simplification process, such a tool can be used to evaluate and identify text portions that are particularly difficult for the target users; when there is more than one simplification choice, readability assessment can be made on each resultant text, so that an optimal one is chosen for the given user. It can also provide objective evaluation for the

system's performance by measuring the change of reading difficulty of a text before and after simplification process.

2.4 Applications

We envision that our research on developing an automatic readability assessment tool can be useful in a variety of applications. For instance, in educational settings, school children, second language learners, adults with low literacy can use our tool to select reading material that is of their interest and tailored to their varying reading proficiency. Similarly, language instructors can use this tool to select teaching material effectively that is at appropriate level of reading difficulty for target readers.

An automatic readability assessment tool can also be useful for many automated NLP systems. It can be used to rank documents by reading difficulty for automated systems such as text simplification, text summarization, machine translation and other text generation systems. For example, as a reprocessing step, such a tool can be used to select documents that are at appropriate level of reading difficulty among those on similar topic for the target system to begin with. More importantly, such a tool can be used to provide efficient evaluation measure for systems' performance.

Take text simplification as example, the need of automatic and reliable readability measure is not unique to our envisioned discourse level text simplification system. The lack of automatic and objective evaluation measure is a common problem faced by many existing text simplification systems. Many of them rely on subjective human judgment or traditional readability formulas such as Flesch-Kincaid scores (see section 3.1) to eval-

uate the system's performance. Human readability judgment is not only time consuming, it is also a tricky issue, several studies reported that human readability judgment may be correlated with reading time (Lapata, 2006; Miller and Kintsch, 1980), but did not show significance with actual comprehension performance (Miller and Kintsch, 1980). Traditional readability formulas such as Flesch-Kincaid scores have been proven to be highly unreliable by several recent studies (Collins-Thompson and Callan, 2004; Feng, 2008; Schwarm and Ostendorf, 2005; Si and Callan, 2001). A reliable tool that can accurately assess the change of reduction in reading difficulty before and after simplification process is clearly in need.

Similarly, we can use our tool to check the quality of text generated by systems such as text summarization, machine translation and text ordering system. One of many important aspects to look at when evaluating the quality of text generated by automated systems is coherence. It is commonly agreed that coherent texts are easier to read. One of many ways to check the coherence of resultant texts is compare their reading difficulty before and after change. Our automatic readability assessment tool can be well suited for this task.

Chapter 3

Relevant Literature & Previous Work

Extensive research has been conducted in the past 80 years to understand what affects the readability of a text and how to assess its reading difficulty. To make it easier for people to judge the reading difficulty of a text, grade levels or number of years of education required to completely understand a text are commonly used as index for reading difficulty. Although over one hundred readability formulas have been developed over the years, little progress has been made to quantify important factors that affect text readability until recently. In the following sections, we first summarize the characteristics and limitations of traditional readability metrics and recent statistical development in the field. We then discuss how our current work differs from previous research both in goals and methodology.

3.1 Traditional Readability Metrics

Many traditional readability metrics use simple linear functions with two or three shallow language features to model the readability of a given text. The features studied commonly focus on two factors: lexical and syntactic. Lexical features, intended to measure the difficulty of words, often look at three factors: the number of syllables a word contains, the number of characters a word contains, and word frequency. Words with more syllables or characters are considered to be harder. Frequently used words are supposed to be easier than those that are less frequently encountered. Syntactic features are even more limited: the complexity of sentences is solely judge by their average length in words.

These characteristics can be observed in many popular traditional metrics. For example, the widely used Flesch Reading Ease and the Flesch-Kincaid grade level formulas (Flesch, 1979) use average sentence length and average syllables per word to calculate the grade level of a text. Similarly, the Gunning FOG (Gunning, 1952) and the SMOG (McLaughlin, 1969) index use average sentence length and the percentage of words with at least three syllable as parameters. Syllable counting is not an easy task; to automate the formula, the Automated Readability Index (Senter and Smith, 1967) counts the number of characters per word instead to determine word difficulty. Different from the syllabic approach, the Dale-Chall formula (Dale and Chall, 1949) made an advance in measuring lexical difficulty by introducing a list of common words familiar for 4th-grade students. It uses the percentage of difficult words (words that do not appear in the list) and average sentence length to predict the grade level of a text. Since 1995, the "new Dale-Chall

formula" (Chall, 1995) has expanded the common word list from 763 to 3000 words.

To make the point clearer, we list some widely used traditional readability formulas below:

- Flesch Reading Ease:

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

- Flesch-Kincaid Grade Level:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59$$

- Gunning FOG Formula:

$$0.4\left[\left(\frac{\text{words}}{\text{sentence}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

  where "complex words" are defined by words with three or more syllables.

- SMOG Formula:

$$\text{grade} = 1.0430\sqrt{30\left(\frac{\text{number of polysyllables}}{\text{number of sentences}}\right)} + 3.1291$$

  where "polysyllables" are defined by words of three or more syllables.

- Automated Readability Index:

$$4.71\left(\frac{\text{total number of characters}}{\text{total number of words}}\right)+0.5\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right)-21.43$$

- The Dale-Chall Formula:

$$\text{Raw Score} = 0.1579\ (\text{percent of difficult words}) + 0.0496\ (\text{average sentence length}) + 3.6365$$

Despite the importance of the topic and the amount of research it has attracted over the past 60 years, little progress has been made over the traditional readability metrics to quantify our understanding of text readability besides the two limited shallow factors discussed above. Even the recently developed and widely used Lexile Scale relies only on word frequency and sentence length to predict text comprehensibility (Stenner, 1996). The popularity of these two semantic factors in traditional readability metrics is by no means a coincidence. Stenner et al. (1983) has analyzed more than 50 lexical variables and did extensive correlation tests to find out that word frequency and sentence length have the most predictive power in ranking the reading difficulty of texts contained in their experiment data. These traditional metrics are widely used, especially in educational settings, partly also because they are simple and easy to calculate. However, the limitations of these metrics are obvious. They overweighted the impact of word frequency and sentence length on text comprehensibility and systematically ignored many other important factors that are crucial to reading

comprehension, such as syntactic constituents, the structure of the text, local and global discourse coherence across the text, familiarity of the discourse topic to the reader, readers' prior knowledge and motivation to read, etc. Moreover, the number of syllables per word, which acts as a reliable proxy for word frequency, and sentence length do not always capture the reading complexity of a text accurately. Davison and Kantor (1982) has illustrated that reduced sentence length can result in increasing reading difficulty or vice versa, thus opposed strongly against manipulation of sentence length to conform a text to certain level of readability defined by formulas. Several recent studies in the field (Collins-Thompson and Callan, 2004; Feng et al., 2009; Petersen and Ostendorf, 2009; Si and Callan, 2001) have also confirmed that traditional metrics are not reliable. Si and Callan (2001), for example, reported that traditional metrics cannot capture content information and often misjudge the reading difficulty of scientific web documents.

3.2 Recent Statistical Approaches

Recent work on readability research benefits largely from NLP technology. Much progress has been made over traditional readability metrics by deploying sophisticated natural language processing techniques, such as parsing and statistical language modeling, to capture more complex linguistic features that have great impact on text comprehensibility.

Statistical language models estimate the probability of sequences and are widely used in many NLP applications to capture the regularity and patterns of natural language. Si and Callan (2001) used unigram language models to capture content information from scientific web pages. A linear model was

built combining language models with sentence length. Their experiment results show that the combined model is much accurate in predicting K-8 science Web pages than the Flesch-Kincaid readability metric. Collins-Thompson and Callan (2004) adopted similar language modeling approach in predicting reading difficulty of short passages and web documents. They used a Smoothed Unigram model to capture vocabulary variation across all grade levels contained in the corpus, which consists of 550 manually collected English documents ranging from 1 to 12 grade level. Different from Si and Callan (2001)'s model, their Smoothed Unigram model is purely vocabulary-based and does not contain any syntactic features. The classifier built with this model outperformed several traditional predictors, such as the percentage of difficult words using the revised Dale-Chale's common word list and the Flesch-Kincaid score, in predicting grade levels for web documents.

Although vocabulary-based unigram language models help capture important content information and variation of word usage, they do not capture syntactic information. Most recently, detailed analysis of syntactic complexity based on parse trees has been combined with language models and traditional measures in readability research (Heilman et al., 2007; Pitler and Nenkova, 2008; Schwarm and Ostendorf, 2005). Besides three traditional measures (average sentence length, average number of syllables per word and Flesch-Kincaid score), Schwarm and Ostendorf (2005) used Charniak's parser (Charniak, 2000) and higher order n-gram ($n = 3$) models over a combination of word and part-of-speech (POS) sequences to capture syntactic and semantic features. The four parse features include *average parse tree height, average number of noun phrases, average number of verb*

phrases, and average number of "SBAR"s (relative clauses). For the *n*-gram models, they used information gain to select words that are able to discriminate between classes, and replaced the rest of the words by their POS tags. By combining language model perplexity scores with syntactic and other features as predictive variables for a Support Vector Machine (SVM) classifier, they achieved better results than two widely used traditional metrics (Flesch-Kincaid and Lexile).

Subsequent work by Heilman et al. (2007) on readability measurement was motivated by pedagogical differences in first language (L1) and second language (L2) learning. They argue that grammatical features play a more important role in L2 texts than in L1 texts because, unlike L1 learners who learn grammar through natural interaction, L2 learners learn grammatical patterns explicitly from L2 textbooks. They built a unigram language model to predict reading difficulty for L1 texts. For L2 texts, they developed two sets of grammatical features: the first one consists of various patterns of relationships between nodes in the parse trees of a text; the second one consists of POS-tagged grammatical functions of words in a sentence, such as verb tenses. They reported that while the language modeling approach was more effective for measuring both L1 and L2 texts than grammar-based predictions, combining both produced more accurate results.

So far, all the work discussed above was limited to the study of lexical and syntactic features with regard to text comprehensibility. As illustrated in section 2.1, cognitive science reveals that the most important process during reading comprehension lie in discourse comprehension, which entails making appropriate inferences from concepts and propositions, connecting and/or integrating related information to construct a coherent

memory representation. The only work that attempted to tackle readability at discourse level, to our knowledge, is by Barzilay and Lapata (2008) and Pitler and Nenkova (2008).

Barzilay and Lapata (2008) designed and implemented an entity-grid model to capture the distribution of entity transition patterns at sentence to sentence level. Using the this model, each sentence of a text is abstracted by four possible grammatical functions of salient entities contained within the text: whether a salient entity serves as a subject ("S"), an object ("Os") in the sentence, or none of both ("X"), or not present ("-"). The transition patterns of these grammatical functions is then computed at sentence to sentence level for each entity (for in-depth detail of entity-grid model, see section 6.1.4). It is believed that these distribution patterns of salient entities capture certain characteristics of local discourse coherence. Their work was not motivated by text readability, but rather by other NLP tasks related to text generation, such as text ordering and summary coherence rating. In addition to the experiments on these tasks, they subsequently tested the usefulness of the local discourse coherence features generated by the grid model in a style classification task: differentiating the original texts from their simplified versions contained in the Britannica corpus (see chapter 5 for more details of this corpus). They drew comparisons with Schwarm and Ostendorf (2005) by replicating most of their features and enriching them with the local entity coherence features. They reported that adding local coherence features helps improve classification accuracy.

Pitler and Nenkova (2008) for the first time looked at readability factors at all three linguistic levels: lexical, syntactic and discourse. Their analysis of discourse factors largely benefited from the newly released Penn Discourse

Treebank (Prasad et al., 2008). They selected 30 articles that were used both in the Penn Treebank and the Penn Discourse Treebank (PDTB). In the PDTB, all discourse connectives and the relations between two adjacent sentences of a text were manually annotated. Each of the 30 articles was then read and rated by at least three college students ("On a scale of 1 to 5, 1 being the worst, 5 being the best, how well written is this text?"). The average scores were collected as gold standard and their task was defined as predicting this average rating – not grade level – for each article. They analyzed 6 classes of features: traditional readability factors such as *average number of characters per word, average sentence length, maximum number of words per sentence, document length*, vocabulary-based unigram features, four parsed syntax features as described in Schwarm and Ostendorf (2005), local entity coherence as described in Barzilay and Lapata (2008), elements of lexical cohesion, and discourse relations. The last two classes of features had never been used in previous research. The elements of lexical cohesion include five features: *average number of pronouns per sentence, average number of definite articles per sentence, average cosine similarity, word overlap and word overlap over just nouns and pronouns*. They reported that none of these five features correlate significantly with human readability ratings. To analyze discourse relations, they treated each text as a bag of relations governed by a multinomial model. In addition to this multinomial probability, there are three more concerning the number of discourse relations contained in each article: *total number of discourse relations, total number of explicit relations, total number of implicit relations*. They reported that, among all individual factors analyzed at all three linguistic levels, the likelihood of discourse relations with text length taken into account shows the strongest correlation

with human readability ratings ($r = .4835$). When experimenting with the combined features on the prediction of readability ratings, they reported that models containing the log likelihood of discourse relations, document length and vocabulary-based unigram feature gave the best results.

Pitler and Nenkova (2008) made a significant advancement in readability research by analyzing possible impact of discourse coherence related features on text readability. In this sense, their work is novel and inspiring, because it touched the core of text comprehension and showed a new direction in readability study that has been long overdue. However, their work is not without limitations. First of all, although their approach to the analysis of discourse relations – which counts for their most significant contribution in the field – is very desirable, it is not portable and cannot be adopted for any corpus other than the PDTB, because, to our knowledge, as they pointed out as well, there exists no robust systems yet that can automatically annotate discourse relations. This greatly limits the practical use of their findings. Secondly, the subjective human ratings they collected to conduct the study were more about text style than text readability, because the ratings were obtained solely from test participants' response to the question "how well is the text written". One needs to be aware that there does not exist any necessary implications between how well a text is written and how difficult or easy a text is to read. A well written text can be fairly complex or simple, and about a less well written text the same can, vice versa, be said. The nature of their gold standard determines that the predictive power of many factors identified in their study is more relevant for text style than for text readability, which fundamentally weakens the significance of their work with regard to readability research. Thirdly, and

consequently, because they rely only on limited subjective human ratings, their study lacks any objective measure. Moreover, 30 articles is too small of size to build and train effective and reliable prediction models.

Our research model draws from the strength of previous research; we make use of significant findings by past work and verify their applicability in our study. However, our approach differs from previous work both in goals and methodology (see chapter 4). We approach readability from a text comprehension point of view; in particular, we pay special attention to discourse processes that are crucial for constructing and maintaining local and global memory coherence of a text, which is key to successful text comprehension. We hypothesize that high level discourse features that reflect these important discourse processes during the reader's comprehension task can be useful in predicting the complexity of a text when combined with well studied lexical and syntactic features. Following this hypothesis, we propose to apply advanced NLP techniques to implement three classes of novel discourse features that have not been studied by any of the previous research. Our preliminary study has confirmed the positive contribution of these novel discourse features in readability research. Unlike previous research, which focuses only on intrinsic text properties, we view text comprehensibility as the result of the interaction between the text and the reader's prose processing ability. We integrate the characteristics of a given reader into our readability study by addressing constraints of working memory capacity placed on the reader's comprehension effort. Moreover, our study does not rely on a single measure of readability; rather, we will combine various proxies, such as paired original/simplified corpora, grade levels, subjective ratings by experts and users, and objective

observations in our user studies, to get at those underlying text properties that are associated with reading difficulties.

Chapter 4

Methods & Research Hypothesis

4.1 Methods

The primary goal of this thesis is to advance our understanding of, and quantify, what makes a text easy or difficult to read, in particular for readers with mild intellectual disabilities (MID). We combine novel NLP and machine learning techniques together with empirical studies to build and evaluate an automatic readability assessment tool with high performance.

The development of our automatic readability assessment tool consists of four major parts: data collection, feature extraction and implementation, building and evaluating the tool on labeled corpora, and test and evaluating the tool on unlabeled texts from different domain.

The main corpus for our study consists of texts with reading difficulty annotated by elementary grades level ranging from Grade 2 to 5. We use this corpus primarily to build and evaluate our automatic text readability assessment tool. Chapter 5 provides more details on this corpus and other corpora selected for the study.

We use NLP techniques to exploit a variety of text features at several linguistic levels, in particular **discourse features** based on density of entities, lexical chains, coreferential inference and those derived from entity grids, features based on **language modeling**, **part-of-speech**, **syntactic parse trees**, as well as **shallow features** that are used in traditional readability metrics. Chapter 6 presents techniques for feature extraction and describes the design and implementation of our features in detail.

We frame the assessment of text reading difficulty as a classification task. We use two machine learning packages known for efficient high-quality multi-class classification: LIBSVM (Chang and Lin, 2001) and the Weka machine learning toolkit (Hall et al., 2009), from which we choose Logistic Regression, SVM, J48 and OneR. as classifiers. We train various prediction models with the features implemented for this study and evaluate them using classification accuracy obtained from repeated 10-fold cross-validation. Classification accuracy is defined as the percentage of texts predicted with correct grade levels. We repeat each experiment 10 times and report the mean accuracy and its standard deviation. However, reading difficulties annotated by grade levels imply the ranking of grades assigned. A misclassification by more than one grade levels is more severe than a misclassification by only one grade level. To adjust our generic classification approach, we use multiple evaluation measures in addition to accuracy, including mean squared error, mean absolute error, number of misclassifications by more than one grade levels and number of misclassifications by one grade level.

We use two ways to assess how well our readability assessment tool generalizes to texts from different domain. We bear in mind that our text readability assessment tool is intended as a subcomponent for an envisioned

text simplification system designed for adult readers with mild intellectual disabilities. First, we manually created two corpora consisting of original and simplified texts adapted specifically for adults with mild intellectual disabilities. Our assumption on paired original/simplified texts is that simplified texts should be easier to read than the original one. We use our automatic readability assessment tool build on corpora annotated with grade levels to predict the reading difficulty of original and simplified texts contained in these two corpora. We evaluate how well our tool can differentiate between original and simplified texts. Second, we have experts rated the reading difficulty of paired texts, we develop statistical models to estimate latent difficulty derived from reading experiments involving adult participants with mild intellectual disabilities. We compare the correlations between grade level predictions by our tool, expert ratings, and inferred text difficulty for adult participants with mild intellectual disabilities.

Hence our general methodology relies on the following five proxies:

- **Grade levels** Grade levels indicate the number of years of education generally required to understand the text. It is generally understood that reading difficulty increases with grade level. They are a commonly accepted index for reading difficulty of a text, especially in educational settings, because the scale of grade levels make it easier for teachers, parents, librarians, and others to judge the readability level of various books and texts. Another reason to look at grade levels is that they have been widely used in previous research. Using the same measurement index would make it easier for us to draw comparisons between our metric and existing approaches.

- **Paired original/simplified texts** A common assumption is that simplified texts should be easier to read. Paired texts provide valuable clues on how texts with identical subject matter differ. During earlier stage of our modeling, we use paired texts to analyze and select features that distinguish the simplified texts most from the original ones.

- **Subjective ratings by experts** We ask experts who have linguistic expertise or specialize in working with adults with ID to rate text difficulty. The motivation behind this is as follows: (a) We rely on their expertise to help us identify factors that may play an important role in affecting reading difficulty for our users. (b) Subjective expert ratings are much more reliable and easier to obtain than from target users. We evaluate subjective ratings by checking inter-rater agreement, as well as correlation with grade levels and subject ratings and observations.

- **Objective observations in user studies** We will present our target users with texts at a variety of difficulty levels and record their reading times. Subjects will answer simple comprehension questions afterwards, and we will analyze the accuracy of their answers. This will give us the most direct clues about the difficulties faced by our target user group, even though we will need to account for per-subject and other effects. Details of these user studies are described in chapter 5 (also see Feng et al. (2009) and Huenerfauth et al. (2009)).

- **Subjective (introspective) ratings by users** This will probably be especially problematic in our study, as the users' subjective judgment may not be fully reliable because of their cognitive impairments. Many research questions remain open as how to design and conduct studies

with adults with ID to get effective and valid feedback. In Huenerfauth et al. (2009), we discuss how we address some of these issues together with subsequent experiment results. We are aware that subjective user feedback is not completely reliable, especially in our case, that is the reason why we have several proxies to perform multi-fold evaluation on our models. Despite all this, we believe direct user feedback is valuable in our user-specific study.

For developing our readability metric, we want to combine all the above observations to get at those underlying text properties that are associated with reading difficulties.

4.2 Research Hypothesis

Our research is guided by widely accepted theoretical framework established in cognitive science in understanding of cognitive processes underlying reading comprehension. This framework defines the goal of reading comprehension as actively constructing a coherent memory representation of a text by the reader. According to this theory, reading comprehension encompasses more than word identification and sentence processing. Much more important processes occur in discourse comprehension, which entails frequent activities such as resolving entities, inferring meaning from words and phrases, assessing and evaluating semantic relations among concepts and propositions and making connections among them, using background knowledge to generate appropriate inferences to fill in gaps, and integrating new information into existing semantic structure to achieve and maintain coherent memory representation of a text. Reading difficulties arise more

often from discourse comprehension rather than lexical or and syntactic processing. It is generally assumed that demands made by such discourse processing are related to readers' working memory capacity (Daneman and Carpenter, 1980). It is working memory capacity that underlies individual differences in language comprehension (Merrill et al., 2003). Moreover, Miller and Kintsch (1980) and Britton and Gulgoz (1991) have shown that the comprehensibility of a text can be well predicted by an analysis of the demands it makes of readers' working memory (Britton and Gulgoz, 1991; Lorch and van den Brock, 1997; Miller and Kintsch, 1980).

We base our research hypothesis on the theoretical framework illustrated as above and in section 2.3. We hypothesize that the amount of working memory burden inflicted by various discourse processes are crucial for constructing coherent memory representation of a text can be useful in predicting text comprehensibility. We believe that such working memory burden imposed by a text can be objectively measured by carefully selected linguistic factors contained within the text. Our study focuses on designing and extracting such features and analyzing their impact on text readability, both for general readers and readers with limited working memory capacity. We propose to design and implement four classes of novel discourse features that we think best reflect working memory burden posed on the reader's attempt to understand a text: ***density of entities***, ***lexical chains***, ***coreferential inference features*** and ***local entity coherence features***. We illustrate our hypothesis for the relation between the first three classes of discourse features – which have not be studied by previous research – and working memory capacity in the follow sections. ***Local entity coherence features*** have been explored by Barzilay and Lapata (2008) and Pitler and Nenkova

(2008) in readability related study. We have introduced them in chapter 3, chapter 6 will describe them in more detail.

4.2.1 Density of Entities

Conceptual information is often introduced in a text by entities, which consist of general nouns and named entities, such as people's names, locations, organizations, etc. Serving as major information carrier, entities are foundationally important during discourse processing. Established entities form basic components of concepts and propositions, on which higher level of semantic relations can be organized and arranged in order to facilitate and promote the construction of coherent mental representation. The number of entities introduced in a text represents the amount of information the reader needs to process and keep track of in order to understand the text. We hypothesize that the more entities are introduced into a text, the more demands they make of the reader's working memory capacity; for individuals with ID who suffer from impoverished working memory, the increasing demands of entity processing would become especially overwhelming.

4.2.2 Lexical Chains

According to semantic network structure theory, processed conceptual information and propositions stored in memory systems are organized and structured by various semantic relations that connect them. The ability to assess and evaluate semantic relations among concepts and propositions and make connections among them is crucial for building the semantic structure of the text during the on-line process of reading comprehension.

Using existing NLP technology, various semantic relations among entities – such as synonym, hypernym, hyponym, coordinate terms (siblings), etc. (Galley and McKeown, 2003) – can be automatically annotated. Based on these annotations, entities that are connected by certain semantic relations can be chained up through the text and form a lexical chain. These lexical chains often represent related concepts and propositions that are being introduced and reëmphasized or elaborated across the text. Some of the chains can even capture major local or global discourse topics that either interleave or overlap concurrently with each other. We hypothesize that the process of forming lexical chains, to some degree, mimics readers' necessary language comprehension activities during reading. They would need to go through similar processes to resolve discourse relations among entities and keep related discourse topics in mind in order to achieve a coherent semantic representation of the text by the end of reading. From an information processing point of view, these comprehension activities would make heavy demand on readers' working memory for resources to process, manipulate, organize and store text information. The more lexical chains there are and the longer the chains become, the more working memory capacity is required to accommodate them.

4.2.3 Coreferential Inferences

The importance of inferential processes during reading comprehension follows from the same theoretical framework as lexical chains, in which text comprehension is defined as the process of constructing a coherent memory representation. In natural language narratives, it is common for relations

among concepts and propositions to not be stated explicitly; sometimes it is even a stylistic necessity. In order to fill in semantic gaps so that connections among propositions or events can be reëstablished to maintain local or global coherence, readers are required to actively apply acquired prior background knowledge to disambiguate and make appropriate inferences. The inference processes involve searching and retrieving relevant information from various long- and short-term memory systems. It is generally believed that central propositions of a text are kept active in working memory (Conners, 2003; Kintsch and van Dijk, 1978). Because of the constraints that are placed on readers' working memory capacity, how efficiently and accurately readers can perform the search and retrieval task in memory will have a direct impact on their comprehension performance.

Extensive research has been conducted to investigate the role and nature of inferential processes during reading comprehension. Two types of inferences that have received considerable attention in cognitive science concern causal inferences and referential inferences (Lorch and van den Brock, 1997). It has been demonstrated that causally related events represent an important source of both local and global coherence in narrative (Trabasso et al., 1984; Trabasso and Suh, 1993). Moreover, researchers have found that whether an event is in the causal chain of the story and the number of causal connections it has are important determinants of the probability of recall of the event (Goldman and Varnhagen, 1986; Trabasso and van den Broek, 1985). However, we emphasize automatic computation of linguistic features for our intended readability metrics. To our best knowledge, we are not aware of any existing NLP systems that automatically annotate causal inferences contained in a text. There do exists several coreference resolution systems

for immediate use. For this reason, we focus on analysis of (co)referential inferences instead.

Referential relations are often established through anaphoric devices, such as pronominal references. The ability to resolve references is key to discourse comprehension, because a sentence cannot be understood without all anaphoric references being appropriately resolved to their referents. It has been well established that readers are conscientious in resolving references as soon as they are introduced in a text (Lorch and van den Brock, 1997). Empirical findings have indicated that the time it takes readers to process anaphoric expressions depends on factors affecting the ease with which the referent of the anaphora can be unambiguously identified (Ehrlich and Rayner, 1983; Sanford et al., 1977). We propose to use coreference resolution software to extract entities together with their pronominal references that are connected by various anaphoric devices across text to refer to the same object or person. The connected entities and pronominal references extracted will be formed into a coreferential chain in the order they appear in a text. We use the number of coreferential chains, chain length and the distance between each pronominal reference and its referent as indicators for working memory burden inflicted by the complexity of the inferential task. We hypothesize that the longer the referential distance is, the further back the reader needs to search in memory space for relevant information; the more references are made in a single chain, the more resources are required from working memory capacity to disambiguate. The same applies when the number of total coreferential chains in a text increases. We also hypothesize that inferential tasks would be especially hard for readers with intellectual disabilities. A large body of literature on comprehension

differences between individuals with and without ID has revealed that people with ID demonstrate deficiency in accessing semantic memory when confronted with inferential tasks (Fowler, 1998; Merrill and Jackson, 1992). They often do not appear to apply background knowledge as readily as persons without ID to actively resolve inferences (Conners, 2003; Merrill et al., 2003; Stanovich, 1985).

Chapter 5

Corpora

We have collected six corpora for our readability study. Table 5.1 shows statistics of each corpus. The following sections describe each of the corpus and explain how it is chosen/created and used in our study.

5.1 Labeled Corpus: WeeklyReader

In order to build our readability metric, we need data that is labeled with some reliable measurement of reading difficulty to train prediction models. Ideally, this data should contain texts that are of interest to our target users and do not require high literacy skills. High quality data satisfying all these needs is hard to obtain electronically. We contacted the Weekly Reader corporation[1], an on-line publisher producing magazines for elementary and high school students, and were granted access (October 2008) to their archived articles. Among the articles retrieved, only those intended for elementary students are labeled with grade levels ranging from grade 2 to 5.

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

Table 5.1: Corpora statistics.

Corpus		Nb Docs	Avg Nb Words/Doc	Avg Nb Sents/Doc	Avg Nb Words/Sents
WeeklyReader	G 2	174	128.27 ± 106.03	13.19 ± 10.77	9.54 ± 2.32
	G 3	289	171.96 ± 106.05	15.25 ± 9.59	11.39 ± 2.424
	G 4	428	278.03 ± 187.58	20.45 ± 14.22	13.67 ± 2.65
	G 5	542	335.56 ± 230.25	22.03 ± 15.13	15.28 ± 3.21
LocalNews2007	ori.	10	238.40 ± 110.78	11.70 ± 6.25	20.38 ± 5.18
	sim.	10	132.50 ± 52.59	10.40 ± 4.35	12.74 ± 3.06
LocalNews2008	ori.	11	389.82 ± 194.51	21.64 ± 12.58	18.76 ± 3.98
	sim.	11	198.27 ± 83.98	18.64 ± 9.48	11.34 ± 2.41
NewYorkTimes100	G 7	100	782.49 ± 291.26	31.74 ± 12.18	25.39 ± 5.18
Britannica	ori.	114	1924.18 ± 2871.06	87.46 ± 119.09	19.19 ± 3.31
	sim.	114	506.08 ± 233.85	36.46 ± 15.56	13.79 ± 1.34
LiteracyNet	ori.	115	410.02 ± 132.20	23.86 ± 9.304	17.76 ± 3.08
	sim.	115	281.77 ± 77.69	22.60 ± 6.22	12.75 ± 2.70

We selected only this portion of articles (1629 in total) to form our training corpus. These articles are intended to build children's general knowledge and help them practice reading skills, they cover a variety topics such as science, history, health, current events and so on. While pre-processing the texts, we found that many articles, especially those of low grade levels, consist of only basic word quizzes and math and other sort of simple comprehension questions, which is often in the format of a question followed by a few answer choices. Because many of our features are extracted based on complex syntactic and discourse processing and computation, texts consists of multiple choices do not provide much meaningful parsing or discourse information for us. Therefore we discarded some texts that are merely quizzes and puzzles and kept only 1433 full articles in the end. Column 2, 3, 4 of Table 5.2 show the details of the selection process. We will we use this corpus to build and train various readability predicting models with

Table 5.2: Statistics for the number of collected and kept documents of the WeeklyReader data.

	Our Data			Schwarm and
	retrieved	left out	kept	Ostendorf's data
Grade 2	285	111	174	351
Grade 3	316	27	289	589
Grade 4	454	26	428	766
Grade 5	574	32	542	691
Total	1629	196	1433	2397

subsets of features proposed and implemented in this thesis.

Data from the WeeklyReader has also been previously studied by Schwarm and Ostendorf (2005). We want to point out here that although we both used the Weekly Reader as the same source for our data, the two corpora are not identical in size or content. In Table 5.2, we listed the details of the data collected by Schwarm and Ostendorf (2005) for comparison. This will partly explain why their experiments results are not directly comparable to the performance of our metric. More about comparison with Schwarm and Ostendorf (2005) is discussed in Chapter 7.3.

5.2 LocalNews2007 and LocalNews2008

5.2.1 LocalNews2007

As discussed in section 2.3, the proposed thesis work on readability metrics was partly motivated by an envisioned long-term project on a discourse-level text simplification system designed for individuals with ID. To investigate as preliminary research whether users with ID would benefit from such a system, a collaborating research group between the City University of New

York (CUNY) and the Columbia University conducted a pilot study with a group of adults with ID in Fall 2007. The LocalNews2007 corpus resulted from that study.

During this preliminary study, the researchers interviewed experts on adults with ID in deciding domain genre and topics that might be of interest to the target users. Local news was suggested as favorable. Ten articles from various local news websites were thus collected for a feasibility study involving adults with ID.

The study was Wizard-of-Oz in nature. A human editor with expertise in automatic text summarization and knowledgeable with people with ID performed the text simplification for each of the 10 original new articles, with the goal of making the text more readable for adults with mild ID. The editor made the following types of changes to the original news stories: 1) breaking apart complex sentences; 2) unembedding information in complex prepositional phrases and reintegrating it as separate sentences; 3) replacing infrequent vocabulary items with more common/colloquial equivalents; 4) omitting sentences and phrases from the story that mention entities and phrases extraneous to the main theme of the article. For instance, the original sentence "They're installing an induction loop system in cabs that would allow passengers with hearing aids to tune in specifically to the driver's voice." was transformed into "They're installing a system in cabs. It would allow passengers with hearing aids to listen to the driver's voice."

We first conducted a pilot study with 14 adults with mild intellectual disabilities. From the 20 paired original and simplified texts, 10 were randomly selected and displayed on the screen to each of the test participant. The texts were selected in a way that no participant saw both the original

and its simplified version. A test participant read each article on the screen, they can scroll the screen up and down as needed. A text-to-speech software was also installed in the computers, in hope to ease the reading challenge for participants with ID. While a participant was reading, each word was read aloud to him by the software and highlighted on the screen. After they finished reading an article, participants were asked multiple-choice comprehension questions. The questions and multiple-choices were presented to the participants in paper version. In order to help them understand the questions better, each question and choices were read aloud to the participants while they read them.

This pilot study laid the groundwork for later experiments. Most importantly, it provided us an opportunity to interact with adults reader with ID directly, through which we had an estimation on varying reading ability of adults with ID and an empirical understanding of their comprehension process. It also helped us identify what worked and what still needs to be improved in the future for similar study. Our second experiment followed in 2009 benefited hugely from this study.

This study produced 10 pairs of original/simplified news articles, which we refer to as LocalNews2007. We use this corpus as part of unseen data to evaluate models constructed in Chapter 7. The details are discussed in Chapter 8.

5.2.2 LocalNews2008

We build our automatic text readability assessment tool with a corpus labeled with grade levels. It is a challenging task to evaluate our prediction

models on unseen data, because the text difficulty of unseen data is undetermined. One way to approach this problem is to have unseen data annotated with grade levels as well, so valid comparison can be made with model predictions. However, the criteria used by Weekly Reader are not published (Petersen and Ostendorf, 2009), we lack the guidelines for grade level annotation. Moreover, we would like to investigate how our machine-learning-based assessment tool models reading difficulty of texts, not just for general audience, but for adult readers with ID as well. For this purpose, we conducted a reading experiment following the pilot study with adults with ID, which produced a user specific corpus, called LocalNews2008. We describe the details in creating this corpus below.

The LocalNews2008 corpus was created in similar fashion as LocalNews2007, with improved methods added in the effort to obtain better quality of user responses to text comprehension questions. The pilot study conducted in 2007 revealed several questions important to our research which were not fully anticipated and prepared to address at the time. Through the interactions with the test participants during the pilot study, we observed that some of them either randomly picked a choice as an answer or selected one specific number of choice and gave it as an answer to all questions. This kind of behavior indicates that they either did not comprehend the texts well, or they were having difficulties understanding the questions and the multiple choices provided, or both. Moreover, we observed that test participants in general did not respond well to questions that have "*yes*", "*no*", "*it didn't say*" as multiple choices. This observation brings the awareness to us that question formats may affect test participants' responses as well. We cannot control factors arising from the text side that influence comprehen-

sion difficulty for our participants with ID, but we believe improvement can be made on the questions side to help test participants better understand what is being asked and what possible choices there are to choose from. This motivated our second round of user study in Spring 2008.

Major steps in preparation for the study, such as test participant recruitment, IRB protocol design, collecting and simplifying local news articles of interest, experiment design and so on are similar to that of the 2007 pilot study. We collected 11 local news articles, which were manually simplified by humans mimicking operations that would be restricted to the state-of-the-art text simplification system. Long and complex sentences were split into shorter simpler ones. Important information contained in complex prepositional phrases were unembedded and reintegrated in separate sentences. Infrequent words were replaced with common ones. Sentences and phrases that are not closely related to the central topic were omitted.

What makes it different from the 2007 pilot study is the new question formats and the subsequent new experiment design. This study carries two major goals with itself. The primary one is to gather comprehension responses directly from adult readers with ID, so text difficulty can be inferred for adults with ID based on their individual comprehension ability. Another goal is to investigate whether certain question formats solicit more qualitative responses from the test participants than others, given that the questions asked are either identical or comparable. In 2007 pilot study, we used several multiple-choice comprehension question types: questions with "yes" or "no" as answers, questions with answers in complete sentences, and questions with answers in short noun or verb phrases. We observed that many test participants did not respond well to "yes/no" type questions

or questions with long sentences as answers. In this study, we improved the formats of questions as follows. For each paired original/simplified articles, we selected six facts to ask questions, where these facts can be found both in the original text and its simplified version. In other words, each pair of original and simplified articles share the same set of six questions. For each question, we prepared three different question formats:

1. multiple-choice question with single words or short phrases as answer choices, the question was phrased short and simple, no complete sentences were used as answer choices;

2. multiple-choice question (identical to 1) with clip art images or photographs as answer choices, with English text captions placed below each image, the text captions are identical to the answer choices in 1;

3. multiple choice question with "*yes/no/it didn't say*" as answer choices, the question is often rephrased slightly different from 1 and 2, but the content of the question is still comparable with 1 and 2.

Figure 5.1 shows an example of three different formats designed for a single question. There are 11 pairs of original/simplified articles, for each pair we selected six facts to form questions, and for each question we designed three types of questions as described above, there are $11 \times 6 \times 3 = 198$ unique questions in total.

The experiment was conducted as follows. 20 adults with ID were recruited to participate the experiment. Each test participant was assigned 11 articles to read, the articles were arranged in a way that no test participant saw both the original and simplified article on the same topic, the number

MultipleChoice Version:
What product does the Kryptonite company make?
a. cars
b. bikes
c. locks and chains

ClipArt Version:
What product does the Kryptonite company make?

a. cars b. bikes c. locks and chains

TrueFalse Version:
Does the Kryptonite company make bikes?
a. yes
b. no
c. it didn't say

From Huenerfauth et al. (2009)

Figure 5.1: Example of three question types.

of original and simplified article assigned to each participant were balanced across all test participants, the order the articles were arranged in alternating complex simple order. Each article has six questions asking about six different facts of the article, two in multiple choice text formats, two in multiple clip art formats and two in "*yes/no*" formats. We made sure that no participant was asked about the same fact in more than one question type. Across the entire study, the number of questions in each of the three types (*multiple text choices, multiple clip art choices and yes/no type*) was ensured to be balanced across each test participant, each fact, and each simplified and original version of article. To account for the possibility that some participants may select a fixed choice as answer to all questions, we randomized the order of answer choices for each unique type of questions as well.

The rest of the experiment protocol was the same as the 2007 pilot study. Articles were presented on the computer screen to the participants. Text-to-speech software was installed to highlight each word while it was read aloud to the participant. After finishing reading the article, the participant answered six comprehension questions, questions and relevant clip-art images were all presented to the participant in paper version.

In our recent paper (Huenerfauth et al., 2009), we published statistical results on investigating which question format is more likely to solicit valid feedback from the test participants. The method used for this investigation is to test whether user responses to each of the three question types were able to distinguish simplified texts from the original ones. We reported that none of the individual question types showed the ability to significantly distinguish simplified texts from the original ones, but when selectively combining the results of two question types – using multiple text choices for questions on facts such numbers or non-referring word and the rest with choices and multiple clip art choices – produced statistical significance in participants responses to simplified and original articles.

In Huenerfauth et al. (2009), we used a simplistic approach to measuring reading difficulty of texts: for each article, we gathered all responses from test participants who had read this article and scored it with the percentage of correct responses. The reading difficulty of each article is interpreted as a negative correlation with the score it was assigned: the higher the score, the easier it is to read, and vice versa.

The weakness of this simplistic approach lies in that it treated each response equally and independently and did not take individual's varying reading ability into account. In this thesis, we improve our previous ap-

proach by developing a hierarchical latent trait model that is appropriate to capture key aspects of the experimental design. We use this model to infer reading difficulty of each text in LocalNews2008 for adults with ID. This model not only takes individual reading abilities of participants and the difficulties of question items into account, it also captures two important aspects of the experiment. First, items are no longer independent, but are grouped by article and condition. Second, our model will reflect the fact that the set of comprehension questions for each article was identical for the complex and simplified versions. Section 8.4 presents this model in detail.

In addition to gathering comprehension responses from adult readers with ID for LocalNews2008, we also had all 22 articles rated by three experts using an independent number scale. In total, we have three different measures of reading difficulty for LocalNews2008: model predictions, expert ratings and inferred text difficulty for adult readers with ID. We use these three independent measures to evaluate our automatic readability assessment tool on unseen data and investigate the relations among them. Chapter 8 presents research on this topic.

5.3 NewYorkTimes100

The WeeklyReader corpus contains only texts labeled with limited range of grade levels (Grade 2 to 5). This inevitably limits the model's prediction ability when encountering texts with reading difficulty higher that Grade 5, the highest level of reading difficulty the model's can estimate. In order to test whether the features, with which the prediction models are built, are robust enough to generalize to unseen texts with reading difficulty

much higher than Grade 5, we created a corpus called NewYorkTimes100, which contains 100 original news articles manually selected from The New York Times. These 100 articles were carefully selected so that their reading difficulty is distinguishably much higher than texts labeled with grade 5 in the Weekly Reader corpus. We intend to mix this corpus with the WeeklyReader corpus as training data to build more robust models, so we assign Grade 7 to each of the 100 articles as class label. It is to note that the assigned grade level 7 is intended more as an artificial marker to differentiate from grade 5 texts in Weekly Reader rather than a true and accurate grade level: we assume that the reading difficulty of the easiest articles in NewYorkTimes100 corpus is at least comparable to grade 7, many articles within the corpus may have reading difficulty higher than grade 7.

5.4 Ulabeled Paired Corpora: Britannica and LiteracyNet

We approach readability from a text simplification point of view. The goal of text simplification is to rewrite a text – whether constructed by an expert, or automatically – in a way such that the level of reading difficulty of the resultant text is reduced compared with the original text. Therefore we assume that simplified texts should be easier to read compared with original ones.

Paired original/simplified texts are helpful in the early stage of our research, because they provide valuable resources for us to analyze characteristic changes of text properties before and after the simplification processes.

Some of these unique changes could be related to or result in change of reading difficulty, therefore they can be used as good predictors for readability assessment. In the early stage of our research, before the user specific corpora are available, we could use paired original/simplified texts to test feature effectiveness, that is to say, features that can distinguish simplified text the most from the original ones should be good readability predictors. For this purpose, we collected two comparable corpora from two on-line sources: Encyclopedia Britannica and literacynet.org. Another important reason to collect these two corpora is to make comparisons with previous study by Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009). In their work, they used these two corpora to develop language-modeling-based features. In Section 7.3 we will discuss in detail how we use these two corpora to replicate major features used in their study.

Both corpora consist of original texts written for adults, and manually simplified and abridged versions for children, second language learners and people with low literacy skills. The paired Britannica corpus was originally collected by Barzilay and Elhadad (2003). The corpus consists of 114 original encyclopedia articles about cities around the world, and their corresponding 114 simplified versions. The original articles are aimed at educated adult readers. The simplified versions are adapted for children with some details omitted and background information inserted.

The LiteracyNet corpus consists of 115 original local news articles and their corresponding 115 abridged versions manually adapted for children and adult second language learners. This corpus was made available through the Western/Pacific Literacy Network[2]. The source of the arti-

[2]`http://literacynet.org/cnnsf`

cles comes from the current and past CNN San Francisco bureau and CBS 5 - KPIX (CBS Broadcasting) news stories. This corpus is of particular interest to us, not only because it provides comparable original and simplified version of texts, but also because of its local news domain genre. Before creating our own user specific corpus, we conducted a survey in a 2007 pilot study. Because many test participants responded favorably to local news articles, we have decided to collect local news article that are of interest to our users to build our user specific corpus. Before the user specific corpus is available, the LiteracyNet corpus can provide us valuable stylistic information that is characteristic to news articles.

Although these two corpora are created for general users with no disabilities, we believe the paired corpora do demonstrate many common issues related to readability and text simplification. We use these two corpora to test feature significance for our automatic readability assessment tool in addition to making valid comparisons with previous work.

Chapter 6

Feature Extraction

In this chapter, we present features to be used for our automatic text readability assessment tool and discuss the techniques deployed to extract and implement them.

We propose the following 5 feature subsets, many of which result from refinement and improvement of previously studied features. We categorize them using the linguistic levels at which they are extracted as boundaries.

- **Discourse Features** (45)
- **Language-Modeling-based Perplexity Features** ($80 + 48 = 128$)
- **Parsed Syntactic Features** (21)
- **Part-Of-Speech-based (POS) Features** (64)
- **Shallow Features** (9)

Discourse features are one of the major contributions this thesis makes toward the advancement of current readability research. As discussed in Chapter 3, previous study on readability has been mostly limited at lexical

Table 6.1: Discourse features.

Feature group	num. of features
Entity-Density Features	16
Lexical Chain Features	6
Coreference Inference Features	7
Entity-Grid Features	16
total	45

and syntactic level. Only recently, Pitler and Nenkova (2008) attempted to analyze discourse relations when addressing a readability related problem, which is to determine how well is a text is written. However, their approach can not be used in automatic readability assessment, because, as they pointed out as well, there exists no robust systems yet that can automatically annotate discourse relations. Their analysis relied solely on manual annotations. In this thesis, we deploy sophisticated NLP techniques to extract four subsets of features automatically from various linguistic levels and study their effectiveness for readability prediction task. Table 6.1 lists these features to be extracted and implemented in Section 6.1.

In addition to cognitively motivated novel discourse features, we improve and refine previously studied features at syntactic and lexical levels and continue explore language-modeling-based features. For comparisons purposes later in Chapter 7, we also replicate and expand 48 LM-based features and 6 out-of-vocabulary (OOV) features from Schwarm and Ostendorf's work (2005). In total, we implement 273 features. The following sections describe the design and implementation of these features in detail.

6.1 Discourse Features

We implement four subsets of discourse features: **entity-density features, lexical-chain features, coreference inference features** and **entity grid features**. The first three subsets of features are novel and have not been studied by other researchers before. In our early work (Feng et al., 2009), we have published results on entity-density features and lexical-chain features for readers with intellectual disabilities (Feng et al., 2009). Entity-grid features have been studied by Barzilay and Lapata (2008) in a stylistic classification task. Pitler and Nenkova (2008) used the same features to evaluate how well a text is written. We replicate this set of features for grade level prediction task. The following sections describe the design and implementation details of these features.

6.1.1 Entity-Density Features

We define our entities as a union of named entities and the rest of general nouns (nouns and proper nouns) contained in a text. We used open source OpenNLP's[1] name-finding tool to extract named entities, such as names of persons, locations and organizations. We extract nouns by examining the leaf nodes from the output of the Charniak's Parser, where each leaf node consists of a pair of a word and its part-of-speech tag. For each document, we first extract general nouns based on their POS-tags. We then extract the named entities from the output of openNLP's name finder. We remove those general nouns that appear in the named entities. The remaining nouns are

[1] `http://opennlp.sourceforge.net/`

Table 6.2: Entity density features.

	Feature description
1	total number of entities per document
2	total number of unique entities per document
3	percentage of entities entities per document
4	percentage of unique entities per document
5	average number of entities per sentence
6	average number of unique entities per sentence
7	percentage of named entities per document
8	average number of named entities per sentences
9	percentage of named entities in total entities
10	percentage of general nouns in total entities
11	percentage of general nouns per document
12	average number of general nouns per sentence
13	percentage of remaining nouns per document
14	average number of remaining nouns per sentence
15	percentage of overlapping nouns per document
16	average number of of overlapping nouns per sentence

then joined with the named entities to form the complete set of our version of entities.

Based on the collected set of entities, we implemented 16 features as described in Table 6.2. We refer to them as entity-density features henceforth. The difference between "entities" and "unique entities" is that "entities" include duplicate mentions of the same nouns or named entities in a document, while "unique entities" treat duplicate mentions of each noun or named entities only once. In Table 6.2, "overlapping nouns" refer to general nouns that appear in named entities; "remaining nouns" refer to the set of general nouns with overlapping nouns removed. In our early work (Feng et al., 2009), we implemented only four entity-density features (see 1 to 4 in Table 6.2), based on which we conducted more refined analysis later and implemented 12 new features (see 5 to 16 in Table 6.2). Our experimental results show that these 12 newly implemented features significantly improves the predictive power of the previous four features (Feng et al., 2010).

Table 6.3: Lexical chain features.

	Feature description
1	total number of lexical chains in a document
2	average lexical chain length measured by the number of words captured in a chain
3	average lexical chain span measured by the smallest and highest index of words captured in a chain
4	number of chains with span equal to or greater than half of the document
5	number of active chains per word
6	number of active chains per entity

6.1.2 Lexical Chain Features

Research indicates that people with ID do not appear to access and assimilate semantically related knowledge to facilitate language comprehension as readily as people without ID (Conners, 2003; Davies et al., 1981; Glidden and Mar, 1978; Merrill et al., 2003; Stanovich, 1985). Our entity-density features do not require the ability to assess and evaluate the semantic relations among the nouns and named entities. To better measure the working memory burden of a text for people with ID from the perspective of semantic association of words, particularly nouns, during reading comprehension, we used the output of a lexical chaining tool "LexChainer" (Galley and McKeown, 2003) to build a set of lexical chain features.

LexChainer produces chains of words connected by six semantic relations: synonymy, hypernym, hypony, meronym, holonym and coordinate terms (siblings) (Galley and McKeown, 2003). Our hypothesis is that important conceptual and topical information recurring throughout a text is likely to be captured by these lexical chains. In order to construct a coherent semantic representation of a text, it is necessary that a reader keeps semantic related discourse units in his/her working memory throughout the whole reading comprehension process.

34 transplant
35 operation
46 transplant
92 medication
98 operation
217 therapy

Figure 6.1: An example of a lexical chain.

Table 6.3 shows the six lexical chain features that were implemented to reflect possible working memory burden inflicted by the task of retrieving and assessing semantic network associations during reading comprehension process.

Figure 6.1 gives an example of a lexical chain we extracted from a sample text based on the output of LexChainer. There are six semantically related words captured in this chain, the numbers on the left indicate the token index of the corresponding word in the document. Based on this example, the length of this chain is 6, the span of the chain is $217 - 34 + 1 = 184$.

We believe these features may indicate the number of entities/concepts that a reader must keep in mind during a document and the subset of very important entities/concepts that are the main topic of the document. The average length and average span of the lexical chains in a document (aLCL and aLCS) may also indicate how many of the chains in the document are short-lived, which may mean that they are ancillary entities/concepts, not the main topics.

The final two features use the concept of an "active" chain. At a particular location in a text, we define a lexical chain to be "active" if the span (between the first and last noun in the lexical chain) includes the current location. We expect these features may indicate the total number of concepts that

the reader needs to keep in mind during a specific moment in time when reading a text. Measuring the average number of concepts that the reader of a text must keep in mind may suggest the working memory burden of the text over time. We were unsure if individual words or individual noun phrases in the document should be used as the basic unit of "time" for the purpose of averaging the number of active lexical chains; so, we included both features.

6.1.3 Coreferential Inference Features

Relations among concepts and propositions are often not stated explicitly in a text. The constructive nature of building a coherent semantic representation of a text requires a reader to actively retrieve and assess previously processed information to generate appropriate inferences when conceptual information is not stated explicitly. Automatically resolving implicit discourse relations is a hard problem. Therefore, we focus on one particular type, referential relations, which are often established through anaphoric devices, e.g. pronominal references. The ability to resolve referential relations is important for text comprehension.

We use OpenNLP[2] to resolve coreferences. Entities and pronominal references that occur across the text and refer to the same person or object are extracted and formed into a coreference chain. Based on the chains extracted, we implement seven features as listed in Table 6.4. The chain length, chain span and active chains are defined in a similar way to the lexical chain features. Inference distance is the difference between the index

[2] `http://opennlp.sourceforge.net/`

Table 6.4: Coreference chain features.

	Feature description
1	total number of coreference chains per document
2	avg. num. of coreferences per chain
3	avg. chain span
4	num. of coref. chains with span $\geq$ half doc. length
5	avg. inference distance per chain
6	num. of active coreference chains per word
7	num. of active coreference chains per entity

of the referent and that of its pronominal reference. If the same referent occurs more than once in a chain, the index of the closest occurrence is used when computing the inference distance.

Figure 6.2 gives an example of a coreference chain extracted from a sample text from the LocalNews2008 corpus. The two numbers on the left of each phrase are the token indices indicating the start and the end of that phrase. The phrases are sorted by their starting indices in ascending order. The first phrase is the referent, which the rest of the phrases in the chain refer back to.

Based on the coreference chains constructed from the output of the coreference resolution tool, we are implementing the following seven reference chain related features:

The first six features are similar to that of the lexical chains. In the above example, four references are made to the first phrase by “Robles”, “her”,

99	111	Athena Robles , who opened the store with fellow artist Anna Stein
372	373	Robles
433	434	her
436	437	she
441	442	her

Figure 6.2: An example of a coreference chain.

"she" and "her" in the chain. The span of the chain is $442 - 99 = 343$ (words). For each word in the text, we examine whether the span of each reference chain crosses the index of this word. If the chain span passes through this word, we consider this chain active for this word. The number of active chains per entity is implemented in a similar way. For each entity (here entity is defined as named entities plus the rest of general nouns) in the text, we examine if the span of each reference chain passes through this entity. If it does, this chain is active for this entity.

Inference distance is measured by number of words (or tokens). For each reference after the referent (the first phrase) in the sorted reference chain, we compare if this reference and the referent are exact string match. If they don't match, we assume a inference needs to be made. The distance of inference is obtained by subtracting the start index of the referent from the end index of this phrase. For example, each of the four inference distance in the above example is obtained this way: $d_1 = 373 - 99 = 274$; $d_2 = 434 - 99 = 335$; $d_1 = 437 - 99 = 338$; $d_1 = 442 - 99 = 343$. And the average inference distance of this chain is $(274 + 335 + 338 + 343)/4 = 322.5(words)$. If the reference and the referent (the first phrase appear in the reference chain) match exactly, we update the location of the referent, the current reference become the referent. We compute the inference distance of the subsequent references in the same way, with the base of the referent updated to the location of the current reference.

6.1.4 Entity Grid Features

Coherent texts are easier to read. Several computational models have been developed to represent and measure discourse coherence (Barzilay and Lapata, 2008; Elsner et al., 2007; Lapata and Barzilay, 2005; Soricut and Marcu, 2006) for NLP tasks such as text ordering and text generation. Although these models are not intended directly for readability research, Barzilay and Lapata (2008) have reported that distributional properties of local entities generated by their grid models are useful in detecting original texts from their simplified versions when combined with well studied lexical and syntactic features. This approach was subsequently pursued by Pitler and Nenkova (2008) in their readability study. We implement these entity grid features and study their effectiveness in automatic readability assessment.

Barzilay and Lapata's entity-grid model is based on the assumption that the distribution of entities in locally coherent texts exhibits certain regularities. Discourse representation is a challenging task, it often requires manually specified rules and extensive semantic knowledge engineering. To overcome these limitations, the grid model approach focuses on three simple linguistic properties that are tightly linked to local discourse coherence and can be easily extracted from or analyzed on a parsed text: syntax, coreference resolution and salience. According to entity-based theories, discourse coherence is achieved by the way discourse entities are introduced and discussed subsequently (Grosz et al., 1995). Across discourse utterances, some entities are more salient than others and display different patterns in their grammatical roles. For instance, salient entities are more likely to appear in prominent syntactic positions such as participant and object,

1. *Free trade is flourishing once more in the Financial District – the hippie commune variety, that is.*

2. *No money is exchanged at the Free Store, which recently opened at 99 Nassau St., and all the merchandise – which ranges from jewelry and vintage clothing to knickknacks – is literally priceless.*

3. *To New Yorkers hit hard by the recession, the price and the timing couldn't be more right.*

4. *"It's amazing when we tell customers, 'Yes, you can take anything and it's free,' " said Athena Robles, who opened the store with fellow artist Anna Stein. "It's a good time to do a project like this, especially near Wall Street. No one has any money now."*

Figure 6.3: A fragment of text for grid computation.

Table 6.5: Entity grid representation for text document shown in Figure 6.3.

	1	2	3	4
TRADE	S	-	-	-
DISTRICT	X	-	-	-
VARIETY	X	-	-	-
MONEY	-	S	-	-
STORE	-	X	-	O
ST.	-	X	-	-
MERCHANDISE	-	S	-	-
CLOTHING	-	X	-	-
KNICKKNACKS	-	X	-	-
YORKERS	-	-	X	-
RECESSION	-	-	X	-
PRICE	-	-	S	-
TIMING	-	-	S	-
WE	-	-	-	S
CUSTOMERS	-	-	-	O
ANYTHING	-	-	-	O
ROBLES	-	-	-	X
STEIN	-	-	-	X
TIME	-	-	-	-
PROJECT	-	-	-	-
THIS	-	-	-	-
STREET	-	-	-	-
ONE	-	-	-	-

Table 6.6: Distribution of entity grid transition patterns.

SS	SO	SX	S-	OS	OO	OX	O-	XS	XO	XX	X-	-S	-O	-X	--
0	0	0	0.07	0	0	0	0	0	0	0	0.12	0.07	0.34	0.12	0.58

and to be introduced in a main clause (Barzilay and Lapata, 2008). The grid model is constructed to distinguish salient entities from the rest and capture the distribution of their specific syntactic transitions at sentence level across a text.

Using their model, each text is abstracted and represented by an entity grid that captures the distribution of entity patterns at the level of sentence-to-sentence transitions. The entity grid is a two-dimensional array, with one dimension corresponding to the salient entities extracted from the text, and the other corresponding to each sentence of the text. Each grid cell corresponds to the grammatical role of the specified entity in the specified sentence: whether it is a participant (S), object (O), neither of the two (X), or absent from the sentence (-).

We use the Brown Coherence Toolkit (v0.2) (Elsner et al., 2007), which was built based on the work of Lapata and Barzilay (2005), to generate entity grid representation for syntactically parsed texts. Table 6.5 shows the entity grid representation for a text fragment in Figure 6.3. The distribution of entity transition patterns between two adjacent sentences is shown in Table 6.6. Based on the entity grid in Table 6.5, there are 69 entity transitions in total, and there are 5 "**S-**" transitions, so the distribution for the "**S-**" pattern is 0.07. Our local entity coherence features consist of the distribution probabilities of all 16 entity transition patterns.

6.2 Language-Modeling-Based Perplexity Features

Language modeling (LM) has been used in many recent statistical approaches to readability research (Collins-Thompson and Callan, 2004; Heil-

man et al., 2007; Pitler and Nenkova, 2008; Schwarm and Ostendorf, 2005; Si and Callan, 2001). Our LM-based perplexity features are inspired by Schwarm and Ostendorf's work (Petersen and Ostendorf, 2009; Schwarm and Ostendorf, 2005), a study that is closely related to ours. They used data from the same source – the Weekly Reader – for their study.

In their approach, they first used information gain (IC) (Yang and Pedersen, 1997) as a feature selection scheme. Words with high information gain are kept and the remaining words are replaced with their parts of speech. They then used n-gram language models with smoothing to characterize the resulting mixed word/POS sequence. Two paired complex/simplified corpora – Britannica and LiteracyNet as described in Section 5.4 – were chosen to train language models. They divided these two paired corpora into four smaller subsets, each of them contains only either the original texts of Britannica or LiteracyNet, or the simplified corresponding texts of the Britannica or LiteracyNet and trained three language models (unigram, bigram and trigram) on each of the smaller corpus, resulting in 12 language models. These 12 language models were then used to score each text in the Weekly Reader corpus by perplexity resulting in 12 perplexity features. They reported that this approach was more successful than training LMs on text sequences of word labels alone, though without providing supporting statistics.

It is worth pointing out that their LMs were not trained on domain-specific data the Weekly Reader, but rather on two unrelated paired corpora (Britannica and LiteracyNet). This seems counter-intuitive, because training LMs directly on the Weekly Reader data would provide more class-specific information for the reading level prediction task. They justified this choice

by stating that splitting limited Weekly Reader data for training and testing purposes resulted in unsuccessful performance.

In this thesis, we describe how we overcome this problem by using a hold-one-out approach to train domain-specific LMs directly on our Weekly Reader corpus. We use grade levels to divide our Weekly Reader corpus into four smaller subsets, that each one contains only Grade 2, Grade 3, Grade 4 or Grade 5 texts. For each text t in a specific subset, we dynamically train n-gram ($n = 5$) language models on data formed by remaining texts in this subset together with texts in the rest three subsets.

Petersen and Ostendorf (2009) reported from their recent work that, in terms of feature selection, the use of POS tags was much more effective than using a single generic word label, and mixed word/POS sequence resulting from information gain approach led to better performance than word-based models alone. However, they did not provide statistical details to show how significant the improvement was. Since we have made advancement over their work by training domain-specific LMs, we are interested in finding out whether their claim on choices of feature selection still holds for our LMs, because comparing feature effectiveness and feature selection choices is one of major research topics of this thesis. For comparison's sake, we construct four types of text sequences for texts in our WeeklyReader corpus:

- IG: mixed word/POS sequence resulting from information gain
- textOnly: sequence of generic word labels alone
- posOnly: sequence of POS tags alone
- tagged: sequence of word tagged by it POS

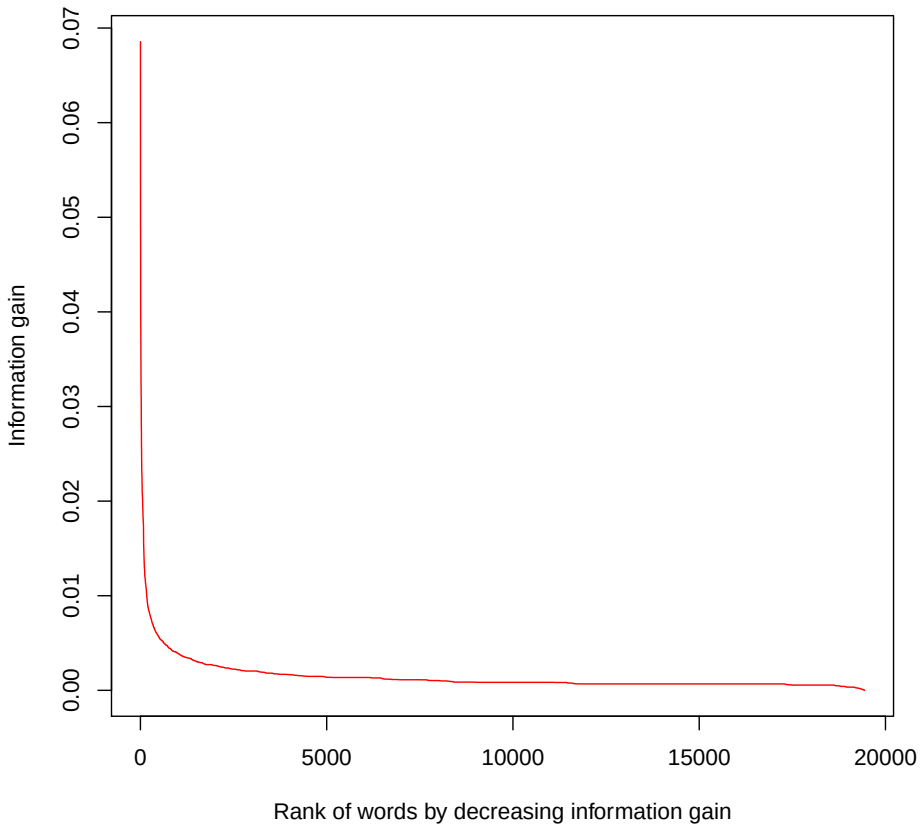


Figure 6.4: Ranked information gain of words for feature selection.

For IG text sequence, we adopt Petersen and Ostendorf's information gain approach for feature selection. Figure 6.4 shows the information gain of all the words contained in the Weekly Reader corpus ranked by decreasing order. After multiple experiments based on manual inspection above the "knee" of the information gain curve, the threshold of 0.0077 yields the best results. We use this threshold to select features for IG text sequence. Words with information gain above 0.0077 are kept, and others with information gain lower than 0.0077 are replaced by their POS tags. The selected features consist of 276 words and 46 POS tokens.

We use the SRI Language Modeling Toolkit[3] (with Good-Turing dis-

[3] `http://www.speech.sri.com/projects/srilm/`

counting and Katz backoff for smoothing) to train LMs. Using each of these four types of text sequences, we train 5 language models (1- to 5-gram) by held-one-out approach, resulting in $4 \times 5 \times 4 = 80$ perplexity features for each text tested in our WeeklyReader corpus.

In order to compare performance differences of perplexity features resulting from LMs trained on domain-specific data, i.e. Weekly Reader, and those obtained from LMs trained on cross-domain data, i.e. Britannica and LiteracyNet, as did by Petersen and Ostendorf (2009), we construct the same four text sequence as described above for Britannica and LiteracyNet. We then break Britannica and LiteracyNet into four subsets of smaller corpora as described at the beginning of this section and use the same SRI toolkit to train three LMs (uni-, bi- and trigram) using each of the four text sequences, resulting in $3 \times 4 \times 4 = 48$ LMs. For each text in WeeklyReader, we use these 48 LMs to compute its perplexity, resulting in 48 perplexity features. We refer to these 80 features collectively as 5gramWR henceforth.

In addition to implementing Schwarm and Ostendorf's information-gain approach, we also build LMs based on three other types of text sequences for comparison purposes. These included: word-token-only sequence (i.e., the original text), POS-only sequence, and paired word-POS sequence. For each grade level, we use the SRI Language Modeling Toolkit (again with Good-Turing discounting and Katz backoff for smoothing) to train 5 language models (1- to 5-gram) using each of the four text sequences, resulting in $4 \times 5 \times 4 = 80$ perplexity features for each text tested. We refer to these 48 features collectively as 3gramBL henceforth, which include the 12 perplexity features replicated from the work by Petersen and Ostendorf (2009).

6.3 Parsed Syntactic Features

Recent approaches to readability have utilized natural language processing techniques such as probabilistic parsers to analyze syntactic features of texts and reported their positive contributions. Schwarm and Ostendorf (2005) studied four parse tree features (average parse tree height, average number of SBARs, noun phrases, and verb phrases per sentences). We implemented these and additional features, using the Charniak parser (Charniak, 2000). Our parsed syntactic features focus on clauses (SBAR), noun phrases (NP), verb phrases (VP) and prepositional phrases (PP). For each phrase, we implement four features: total number of the phrases per document, average number of phrases per sentence, and average phrase length measured by number of words and characters respectively. In addition to average tree height, we implement two non-terminal-node-based features: average number of non-terminal nodes per parse tree, and average number of non-terminal nodes per word (terminal node).

6.4 POS Features

Part-of-speech-based grammatical features were shown to be useful in readability prediction (Heilman et al., 2007; Leroy et al., 2008). To extend prior work, we systematically studied a number of common categories of words and investigated to what extent they are related to a text's complexity. We focus primarily on five classes of words (nouns, verbs, adjectives, adverbs, and prepositions) and two broad categories (content words, function words). Nouns include general nouns and proper nouns. Verbs include past tenses,

present participles, past participles and modals in addition to infinitives, present 3rd person singular forms and all forms of auxiliary verbs. Content words include nouns, verbs, numerals, adjectives, and adverbs; the remaining types are function words. An additional feature is the percentage of function words, which is broadly defined as any word that is not a noun, verb, adjective or adverb. This feature is inspired by the work of Leroy et al. (2008). In examining the reading difficulty of medical texts, they noted a strong negative correlation between the user ratings and the percentage of function words. They attributed this phenomenon to the hypothesis that function words often do not carry content information, a higher percentage of function words would space out content words and make sentences easier to read (Leroy et al., 2008).

The part of speech of each word is obtained from examining the leaf node based on the output of Charniak's parser, where each leaf node consists of a word and its part of speech. We group words based on their POS labels. For each class of words (12 subgroups), we implement five features, resulting in 60 features. For example, for the adjective class, we implemented the following five features: percentage of adjectives (tokens) per document, percentage of unique adjectives (types) per document, ratio of unique adjectives per total unique words in a document, average number of adjectives per sentence and average number of unique adjectives per sentence. For infinitives and function words, we implement two features each: average number of the target words per sentence and percentage of the target words per document. In total we have 64 POS-based features.

6.5 Shallow Features

Shallow features refer to those used by traditional readability metrics, such as Flesch-Kincaid Grade Level (Flesch, 1979), SMOG (McLaughlin, 1969), Gunning FOG (Gunning, 1952), etc. Although recent readability studies have strived to take advantage of NLP techniques, little has been revealed about the predictive power of shallow features. Shallow features, which are limited to superficial text properties, are computationally much less expensive than syntactic or discourse features. To enable a comparison against more advanced features, we implement 9 shallow features as listed in Table 6.7, most of which have been frequently used by traditional readability metrics.

Table 6.7: Shallow Features.

1	average number of syllables per word
2	percentage of poly-syll. words per doc.
3	average number of poly-syll. words per sent.
4	average number of characters per word
5	Chall-Dale difficult words rate per doc.
6	average number of words per sentence
7	average number of characters per sentence
8	Flesch-Kincaid score
9	total number of words per document

6.6 Other Features

For comparison, we replicated 6 out-of-vocabulary features described in Schwarm and Ostendorf (2005). For each text in the WeeklyReader corpus, these 6 features are computed using the most common 100, 200 and 500 word tokens and types calculated from Grade 2 texts in WeeklyReader. We refer to them as OOV features henceforth.

Chapter 7

Automatic Readability Assessment

7.1 Introduction

In this chapter, we present research on building and evaluating an automatic readability assessment tool on WeeklyReader, a corpus annotated with grade levels ranging from Grade 2 to 5.

One important aspect of recent work on automatic readability assessment centers on statistical learning techniques and evaluation measures. Prediction of reading difficulty has been taken by researchers in the field as either a classification task (Barzilay and Lapata, 2008; Collins-Thompson and Callan, 2004; Feng et al., 2010; Heilman et al., 2007; Schwarm and Ostendorf, 2005; Si and Callan, 2001) or regression task (Feng et al., 2009; Pitler and Nenkova, 2008), or for comparison's sake, both (Aluisio et al., 2010; Petersen and Ostendorf, 2009). After many studies with different methodology, there is no consensus on which statistical learning model is more appropriate for the task of readability prediction. The choice of learning technique often depends on multiple factors, such as the nature of annotated reading

difficulty for training data, the audience and the specific applications.

Evaluation measures of readability prediction depend on which statistical model is chosen. Accuracy, precision, recall and F-measure are often used for the classification task. Mean square error and mean absolute error are often used for the regression task, these measures can be used for classification task as adjustment as well. As Collins-Thompson and Callan (2004) rightly pointed out, readability prediction lies in an interesting region between classification and regression, with close connections to ordinal regression (MacCullagh, 1980) and discriminative ranking models (Crammer and Singer, 2001). Classification accuracy does not fully reflect the fact that a misclassification of more than 1 grade level is more severe than an error of a single level. Mean square error can be used as a justification in addition to classification accuracy rate. Correlation coefficients are also frequently used by both classification and regression task to analyze the relations between prediction results and golden standard.

In order to decide which learning techniques are more appropriate and accurate to model reading difficulty of texts annotated with elementary grade levels, we have experimented with various statistical models on our WeeklyReader corpus, including linear regression (R); standard classification (LIBSVM and Logistic Regression and SVM from Weka), which assumes no relation between grade levels; and ordinal regression/classification (provided by Weka, with Logistic Regression and SMO as base function), which assumes that the grade levels are ordered. Our experiments show that, measured by mean squared error and classification accuracy, linear regression models perform considerably poorer than classification models. Measured by accuracy and F-measure, ordinal classifiers perform comparable or worse

than standard classifiers. Based on this observation, we decide to treat automatic readability prediction as a classification task. We use two machine learning packages known for efficient high-quality multi-class classification: LIBSVM (Chang and Lin, 2001) and the Weka machine learning toolkit (Hall et al., 2009), from which we choose several classifiers based on various functions, such as Logistic Regression, SMO (a support vector machine based on sequentially minimized optimization), J48 (decision-tree-based classifier), and OneR. We train and evaluate various prediction models using the features described in Chapter 6.

All experiments presented in this chapter follow the same design: we evaluate classification accuracy of each model using repeated 10-fold cross-validation on the Weekly Reader corpus. Classification accuracy is defined as the percentage of texts predicted with correct grade levels. We repeat each experiment 10 times, each time with a full run of 10-fold cross-validation. We report the mean accuracy and its standard deviation. We also use mean square error in addition to classification accuracy as a justification to penalize misclassification by more than 1 grade level.

The main focus of this chapter is to study the effectiveness of features in terms of their impact on predicting reading difficulty indexed by grade levels. As discussed in Chapter 6, we group the features by the linguistic levels from which they are extracted. The comparison of feature effectiveness is systematically conducted both within a selected linguistic level and across all levels. In Section 7.2, we evaluate and compare the predictive power of individual features within specific feature subsets. In Section 7.2.6, we evaluate and compare how combined feature subsets extracted from various linguistic levels impact grade level predictions differently. In Section

7.3, we present techniques for feature selection to achieve optimal model performance and compare our results with previous studies.

7.2 A Comparison of Features

7.2.1 Discourse Features

In section 6.1, we described the implementation of four novel discourse features, which include entity-density features, lexical-chain features, co-reference-inference features and entity-grid features. We refer to them henceforth as "entity", "lex", "coref" and "egrid" as subsets of features respectively.

In this section, we examine the usefulness of these four subsets of discourse feature in modeling reading difficulty of texts in terms of grade levels. We use LIBSVM and Logistic regression from the Weka toolkit to train classifiers with each feature subset on the WeeklyReader corpus. Each classifier is evaluated by repeated 10-fold cross-validation. The mean and standard deviation of classification accuracy and F-measure generated by each classifier are presented in Table 7.1.

Table 7.1: Classification accuracy generated by subsets of discourse features on WeeklyReader.

	LIBSVM		Logistic Regression	
Feature Set	Accuracy (%)	F-Measure	Accuracy (%)	F-Measure
entity	**59.63** $\pm$ 0.632	0.595 $\pm$ 0.006	57.59 $\pm$ 0.375	0.571 $\pm$ 0.004
coref	40.93 $\pm$ 0.839	0.386 $\pm$ 0.008	42.19 $\pm$ 0.238	0.390 $\pm$ 0.003
lex	45.86 $\pm$ 0.815	0.454 $\pm$ 0.009	42.58 $\pm$ 0.241	0.386 $\pm$ 0.002
egrid	45.92 $\pm$ 1.155	0.422 $\pm$ 0.011	42.14 $\pm$ 0.457	0.367 $\pm$ 0.006
all	60.50 $\pm$ 0.990	0.602 $\pm$ 0.010	58.79 $\pm$ 0.703	0.584 $\pm$ 0.007

We see that, with some fluctuation, the classification accuracy generated by LIBSVM and Logistic Regression is in general consistent across all feature subsets. Among the four subsets of discourse features, entity-density features perform significantly better than the other three feature sets and generate the highest classification accuracy (LIBSVM: 59.63%, Logistic Regression: 57.59%). While Logistic Regression results show that there is not much performance difference among lexical chain, coreference inference, and entity grid features, classification accuracy of LIBSVM models indicates that lexical chain features and entity grid features are better in predicting text readability than coreference inference features. We find that combining all four sets of discourse features improves the overall performance, but the improvement is not very significant compared with models trained with entity-density features alone. Using LIBSVM, the mean accuracy is improved by 0.87% from 59.63% to 60.50%.

To investigate whether certain combinations of these four feature subsets yield better performance, we train classifiers with all combinations of them and present the results in Table 7.2. Both LIBSVM and Logistic Regression models indicate that, when combined with either coreference-inference or lexical-chain features, or both, the predictive power (measured by classification accuracy) of entity-density features decreases. However, when entity-density features are combined with entity-grid features alone, the models trained with this combination of features achieve the best performance (61.26% accuracy by LIBSVM and 59.37% accuracy by Logistic Regression). These results are even better than the performance of models trained with all four feature subsets combined (60.50% by LIBSVM and 58.79% by Logistic Regression).

Table 7.2: Accuracy generated by combinations of discourse feature subsets on WeeklyReader.

Feature Set	LIBSVM Accuracy (%)	Logistic Reg. Accuracy (%)
entity+coref	58.40 ± 0.521	56.55 ± 0.454
entity+lex	59.28 ± 0.899	56.99 ± 0.616
entity+egrid	$\mathbf{61.26} \pm 1.209$	$\mathbf{59.37} \pm 0.659$
coref+lex	46.80 ± 1.303	44.40 ± 0.349
coref+egrid	49.58 ± 1.014	46.72 ± 0.563
lex+egrid	52.36 ± 0.519	46.55 ± 0.511
entity+coref+lex	58.27 ± 1.086	56.50 ± 0.461
ent+cor+egrid	59.87 ± 0.801	58.95 ± 0.436
ent+lex+egrid	59.31 ± 1.328	58.97 ± 0.394
cor+lex+egrid	53.07 ± 0.980	47.64 ± 0.406
all	60.50 ± 0.990	58.79 ± 0.703

To further analyze how these four subsets of features model reading difficulty of texts differently at individual grade level, we conduct detailed analyses based on predictions generated by LIBSVM classifiers and present results in Table 7.3. Figure 7.1 shows the corresponding histogram view of the table. We find that, while at each of the Grade 2, 3 and 4 levels, the classification accuracy generated by entity-density features are significantly better than the rest of the three features sets, however, at Grade 5 level, the entity-grid features generate the highest accuracy (80.96%), outperforming entity-density features with statistically significant improvement. This may well explain what we observed earlier that the only combination of entity grid features and entity-density features leads to better performance than that generated by entity-density features alone. When combining all four subsets of features together, we see that the classifier generates optimal accuracy for Grade 3 (57.09%), which is significantly better than all the rest of four classifiers constructed with each of the individual feature subsets. At Grade 2 and 4 level, combining all features leads to worse performance

compared with entity-density features. At Grade 5 level, combining all features generates accuracy 74.24%, which is better than entity-density features (71.77%), but significantly lower than entity grid features (80.96%).

To summarize, within the four subsets of discourse features, we have the following key observations:

- Among all four subsets of features, entity-density features exhibit the most significant discriminative power in modeling text reading difficulty.
- Combining all discourse features together leads to overall improvement. However, the best performance is achieved by combining entity-density features and entity grid features together.
- Analysis at grade level reveals that entity-density features generate the highest accuracy for Grade 2 (57.41%) and 4 (50.09%); combining all features produces the best performance for Grade 3 (57.09%); and entity grid features generate the highest accuracy for Grade 5 (80.96%).

Table 7.3: Grade level prediction accuracy by LIBSVM classifiers trained with discourse feature subsets.

Feature Set	Grade 2	Grade 3	Grade 4	Grade 5
entity	57.41 ± 4.601	52.32 ± 1.400	50.09 ± 1.522	71.77 ± 1.312
coref	17.18 ± 2.063	26.57 ± 1.913	26.47 ± 1.514	67.62 ± 1.814
lex	34.27 ± 3.813	38.46 ± 2.265	35.87 ± 1.705	61.40 ± 1.198
egrid	27.36 ± 3.021	23.94 ± 2.176	23.93 ± 1.919	$\mathbf{80.96} \pm 2.326$
all	52.70 ± 4.802	57.09 ± 2.952	48.55 ± 2.614	74.24 ± 1.579

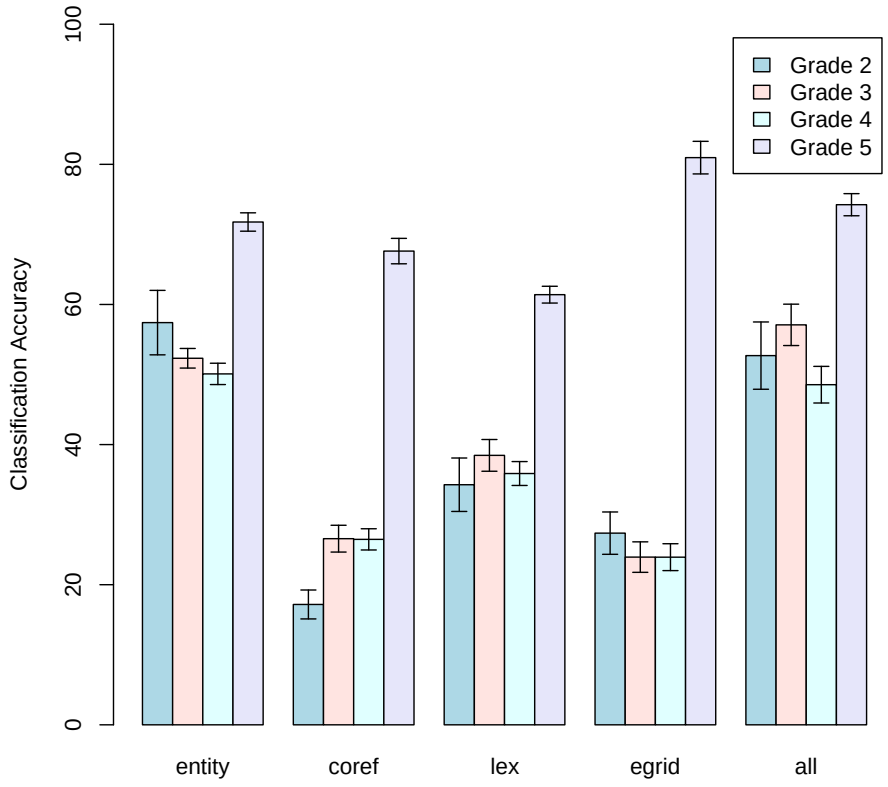


Figure 7.1: Grade Level predictions by LIBSVM classifiers trained with discourse feature subsets.

7.2.2 Perplexity: n-Gram Language Modeling Features

In Section 6.2, we discussed the language-modeling approach by Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009) in readability study. Inspired by their work, we proposed techniques to extract perplexity features using language models (LM) with domain knowledge (trained directly on the WeeklyReader corpus), as opposed to their LMs obtained from cross-domain corpora. To systematically investigate these language modeling related issues, we described four feature selection schemes to prepare text sequences for training corpora, they include:

- IG: sequence of mixed words and POS tags selected by their information gain;
- textOnly: sequence of generic word labels only;
- posOnly: sequence of POS tags only;
- tagged: sequence of paired words and their POS tags.

In this section, we conduct comprehensive experiments to examine how these two different LM approaches – cross-domain vs in-domain – and various feature selection schemes affect model performance differently in detecting and predicting reading difficulty in terms of grade levels.

Before discussing experiment results in details, we briefly lay out the characteristics of each group of features. We first implemented four subsets of perplexity features by adopting Schwarm and Ostendorf's (2005)cross-domain approach: we split Britannica and LiteracyNet into four smaller corpora, each containing either original or simplified texts only. For each

corpus, we used the above feature selection schemes to prepare four types of versions of text sequences. We then trained three LMs (unigram, bigram and trigram) on each version of the corpora, resulting in $4 \times 12 = 48$ LMs. We then used these 48 LMs to compute the perplexity for each text contained in the WeeklyReader corpus and obtained 48 perplexity features. We use "3gramBL" to refer to these features collectively, each subset of features is denoted by the specific feature selection scheme as described above.

However, measuring the information gain of words on the target test data (the Weekly Reader corpus) yet training language models on data that do not provide much class specific information (the Britannica and LiteracyNet corpora) seems to be counter intuitive. Petersen and Ostendorf justified this by reporting that splitting the limited Weekly Reader data that they obtained for LM and SVM training were unsuccessful due to the small size of the resulting data sets (Petersen and Ostendorf, 2006; Schwarm and Ostendorf, 2005). To overcome the problem of limited size of the Weekly Reader data, we use a hold-one-out approach instead of splitting a subset of data aside for LM training purposes. We partitioned the entire WeeklyReader corpus into four subsets, each consisting texts labeled with one specific grade level (2, 3, 4 and 5).Similarly, for each smaller corpus, we prepared four versions of text sequences using the feature selection schemes described above. The WeeklyReader corpus consists of 1433 texts in total. For each text tested, we train a n-gram (n=5) LM on the remaining 1432 texts and compute the perplexity of the selected text. This results in $5 \times 4 \times 4 = 80$ perplexity features, we refer to them collectively as "5gramWR" features. For comparison's sake later, we also extracted a subset of these features, which consists of perplexity features based on unigrams, bigrams and

Table 7.4: Accuracy generated by 3gramBL features on WeeklyReader.

Feature Set	LIBSVM Accuracy (%)	Logistic Reg. Accuracy (%)
IG	52.21 ± 0.832	51.89 ± 0.405
textOnly	45.57 ± 0.805	43.91 ± 0.411
posOnly	49.62 ± 0.510	46.74 ± 0.382
tagged	44.91 ± 1.173	44.09 ± 0.360
all	53.61 ± 0.847	52.97 ± 0.514

trigrams. We refer to this subset of "5gramWR" features as "3gramWR".

To examine the effectiveness of various feature subsets in detecting and classifying reading difficulty in terms of grade levels, we construct a set of classifiers with these features using LIBSVM and Logistic Regression from the Weka toolkit. We evaluate model performance by repeated 10-fold cross-validation and report the mean and standard deviation of classification accuracy.

Table 7.4 summarizes the model performance generated by 3gramBL features using cross-domain approach. In general, LIBSVM classifiers perform better than Logistic Regression. The classification accuracy generated by both types of classifiers consistently show that LMs trained with information gain approach (LIBSVM: 52.21%, Logistic Reg.: 51.89%) outperform LMs trained with the generic word labels (LIBSVM: 45.57%, Logistic Reg.: 43.91%), POS labels (LIBSVM: 49.62%, Logistic Reg.: 46.74%), and paired word/POS text sequence (LIBSVM: 44.91%, Logistic Reg.: 44.09%). This observation agrees with Schwarm and Ostendorf's claim that information gain is a better feature selection technique than POS labels. We also observe that combining all four subsets of perplexity features together achieves the best performance, resulting in 53.61% accuracy using LIBSVM.

Table 7.5: Accuracy generated by 5gramWR features on WeeklyReader.

Feature Set	LIBSVM Accuracy (%)	Logistic Reg. Accuracy (%)
IG	62.52 ± 1.202	62.14 ± 0.510
textOnly	60.17 ± 1.206	60.31 ± 0.559
posOnly	56.21 ± 2.354	57.64 ± 0.391
tagged	60.38 ± 0.820	59.00 ± 0.367
all	68.38 ± 0.929	66.82 ± 0.448

We run similar experiments with 5gramWR features obtained from LMs trained on the WeeklyReader corpus directly. The results are presented in Table 7.5. Not surprisingly, we find that LMs trained with domain-knowledge are much more effective than LMs trained on the Britannica and LiteracyNet corpora, as in Schwarm and Ostendorf's approach. Among the four feature selection techniques, we see that LMs trained with information gain approach (IG) generate the best performance, resulting in 62.52% accuracy using LIBSVM and 62.14% accuracy using Logistic Regression. Combining all features together leads to the highest accuracy (LIBSVM: 68.38%, Logistic Reg.: 66.82%). These two observations are consistent with what we have seen from Table 7.4. However, while LMs trained with POS tags on cross-domain corpora, e.g. Britannica and LiteracyNet, demonstrate stronger discriminative power than LMs trained with generic word sequence and paired word/POS sequence, this is not the case when using in-domain approach. Using WeeklyReader corpus, LMs trained on generic word sequence and paired word/POS tags both outperform LMs trained on the POS sequence alone with significant margin. Moreover, we also notice that training LMs on word labels alone or paired word/POS sequences achieved similar classification accuracy to the IG approach, while avoiding the complicated feature selection of the IG approach.

To make fair comparisons with LMs trained with 3gramBL features, we also extract a subset from 5gramWR features that consists of perplexity features obtained from unigrams, bigrams and trigrams. We refer to this subset of features as "3gramWR". We use the same experiment design to train and test classifiers using LIBSVM. The summative contrasting results are presented in Table 7.6. We see that, among all four feature subsets, LMs trained with 3gramWR features are still significantly better than LMs trained with 3gramBL features. The performance of LMs trained with 3gramWR is close to that of LMs trained with 5gramWR. we see that increasing n (n-gram) from 3 to 5 leads to slight performance gain for IG, textOnly and tagged approaches. However, for posOnly approach, 3gramBL features generates higher accuracy (57.17%) than 5gramWR features (56.21%). The histogram view of the comparisons between accuracy generated by 3gramBL and 5gramWR features is presented in Figure 7.2.

To summarize, we made the following key observations within language-modeling-based perplexity features:

- LMs trained on the WeeklyReader corpus with domain knowledge performed much more effective than LMs trained on cross-domain corpora, i.e. Britannic and LiteracyNet.
- LMs trained with information gain approach perform better than LMs trained with generic word labels, POS labels and word/POS pairs.
- However, when using the WeeklyReader corpus directly, LMs trained on word labels alone or paired word/POS sequences performed with similar classification accuracy to the IG approach, while avoiding the complicated feature selection of the IG approach.

Table 7.6: Comparison of accuracy generated by LIBSVM classifiers trained with 3gramBL, 3gramWR and 5gramWR on WeeklyReader.

Feature Set	3gramBL	3gramWR	5gramWR
IG	52.21 ± 0.832	61.98 ± 1.076	62.52 ± 1.202
textOnly	45.57 ± 0.805	60.08 ± 1.228	60.17 ± 1.206
posOnly	49.62 ± 0.510	57.17 ± 1.106	56.21 ± 2.354
tagged	44.91 ± 1.173	59.67 ± 0.829	60.38 ± 0.820
all	53.61 ± 0.847	68.02 ± 1.342	68.38 ± 0.929

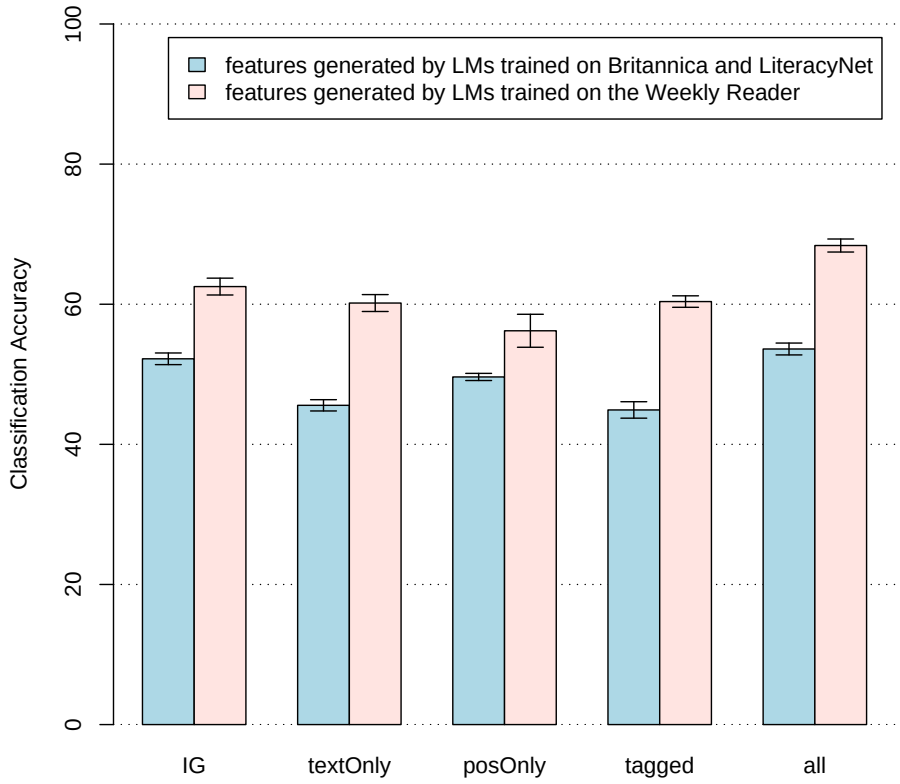


Figure 7.2: Comparison of accuracy generated by LIBSVM classifiers trained with 3gramBL and 5gramWR on WeeklyReader.

- While IG features performed the best among all four subsets of perplexity features, generating 62.52% accuracy using LIBSVM, combining all features together leads to much better improvement, resulting in 68.38% accuracy.

7.2.3 POS Features

Previous research by Heilman et al. (2007) reported that grammar-based features as indicated by part of speech tags can improve prediction accuracy of readability when combined with vocabulary-based features (unigram language modeling). Their grammar-based features focused mostly on various verb forms such as the present, progressive, past, perfect and continuous tenses. To avoid having varying sentence length confounded with these features, they used a per-word measure (the rate of particular POS occurrences per 100 words). In their study, they did not justify why they weighted various verb forms particularly over other part of speech labels, such as nouns and adjectives, which are often considered as important contributing factors to text complexity.

In this section, we systematically analyze the predictive power of a set of part-of-speech-based features, using various weighting schemes. We focus on the following five categories: nouns (consisting of general nouns and proper nouns), verb (a collapsed class consisting of verbs in various forms and modals), adjectives, adverbs and prepositions. In addition, we also include two broad collapsed categories: content words, which consists of nouns, verbs, cardinals, adjectives and adverbs, and function words, which consist of any words not included in the content words.

POS features studied by Heilman et al. (2007) were measured by counting average number of particular POS tags per 100 words, information on generic words was filtered out. In our approach, we implement POS features using both generic words and their POS tags. We first categorize words into groups by their POS tags, such as adjectives and nouns. Within a specific group, for instance nouns, we use information of generic words to count how many unique nouns there in total. The count of unique number of nouns within noun group adds more useful information related to readability to just the total count of nouns. For example, if there are two texts A and B with roughly the same document length and same amount of adjectives. In text A, many nouns occur repeatedly across the text, resulting in a small number of unique nouns, which may imply that the text is focused on certain entities and is easy to follow because of repeated information. Text B contains about the same amount of nouns as text A, however, most of the nouns occur only once or twice, resulting in much big number of unique nouns, which may make text B hard to read, because a lot of new information carried by large number of unique nouns requires more resources to process. Thus by introducing unique count of a particular group of words, we maintain certain characteristics of the text which would otherwise be blurred by simply counting collective POS tags.

Moreover, we use more weighting schemes as described in Section 6.4 for each class of POS selected by this study. Take the class of nouns as example, we implemented the following five features: percentage of nouns (tokens) per document, percentage of unique nouns (types) per document, ratio of unique nouns per total unique words in a document, average number of nouns per sentence and average number of unique nouns per sentence.

Table 7.7: Comparison of our features and Heilman et al.'s study (2007) based on noun class. Accuracy generated by by Logistic Regression classifiers.

Feature sets	Accuracy (%)
Average number of nouns per 100 words	38.66 ± 0.187
noun-based features using new counting and weighting	57.01 ± 0.256

We first examine to what extent our new approach has advantage over the simple count of specific POS tags per 100 words. We take noun-based features as example and train two Logistic Regression classifiers on the WeeklyReader. Table 7.7 shows that our new approach makes significant improvement over the simplistic approach by previous study, bring the classification accuracy from 38.66% to 57.01%.

To examine the predictive power of various word classes, we train and test a number of LIBSVM and Logistic Regression classifier with these features on WeeklyReader. The classification accuracy generated by these models are presented in Table 7.8. We find that, among the five word classes investigated, noun-based features generate the highest classification accuracy (58.15% by LIBSVM), which is consistent with what we have observed earlier about entity-density features. Another notable observation is that prepositions demonstrate higher discriminative power than adjectives and adverbs. Models trained with preposition-based features perform close to those trained with noun-based features. Among the two broader categories, content words (which include nouns) demonstrate higher predictive power than function words (which include prepositions).

Table 7.8: Accuracy generated by POS features on WeeklyReader.

	LIBSVM		Logistic Regression	
Feature Set	Accuracy (%)	F-Measure	Accuracy (%)	F-Measure
nouns	58.15 ± 0.862	0.573 ± 0.008	57.01 ± 0.256	0.563 ± 0.003
verbs	54.40 ± 1.029	0.533 ± 0.011	55.1 ± 0.291	0.540 ± 0.003
adjectives	53.87 ± 1.128	0.528 ± 0.011	52.75 ± 0.427	0.512 ± 0.005
adverbs	52.66 ± 0.970	0.516 ± 0.009	50.54 ± 0.327	0.474 ± 0.004
prepositions	56.77 ± 1.278	0.561 ± 0.013	54.13 ± 0.312	0.536 ± 0.003
content words	56.84 ± 1.072	0.563 ± 0.011	56.18 ± 0.213	0.552 ± 0.002
function words	52.19 ± 1.494	0.519 ± 0.015	50.95 ± 0.298	0.487 ± 0.003
all combined	59.82 ± 1.235	0.594 ± 0.012	57.86 ± 0.547	0.576 ± 0.006

From Table 7.8 we see that, although combining all 64 features together leads to best performance, resulting in 59.82% accuracy, the improvement is not significant compared with the accuracy generated by the model trained with 10 noun-based features alone (58.15%). We also find that model trained with content words, which include nouns, performs poorer compared with model trained with noun features alone. This indicates that combining certain other word class with nouns deteriorates the discriminative power of nouns. Considering that nouns demonstrate the most significant predictive power among all word classes, it is worth taking a closer look at how nouns interact with other word classes. We train a set of classifiers with various combinations of nouns and other major word classes using LIBSVM and and compare model performance with accuracy generated by nouns alone. The results are presented in Table 7.9. We find that combining nouns with verbs or prepositions deteriorates the discriminative power of nouns. However, combining adjectives or adverbs with nouns leads to performance gain. This finding can be useful for optimal feature selection, which we will discuss in depth in Section 7.3.4.

Table 7.9: Accuracy generated by LIBSVM classifiers trained with combinations of nouns and other word classes on WeeklyReader.

	LIBSVM	
Feature Set	Accuracy (%)	F-Measure
nouns	58.15 ± 0.862	0.573 ± 0.008
nouns + prepositions	57.05 ± 1.027	0.564 ± 0.011
nouns + adjectives	58.39 ± 0.845	0.579 ± 0.008
nouns + adverbs	58.42 ± 1.378	0.577 ± 0.014
nouns + verbs	57.58 ± 0.837	0.570 ± 0.008

To conclude this section, we have the following important findings:

- Our experiment results show that systematically designing and implementing POS-based features using our new counting and weighting schemes makes significant improvement over previous study.
- Among all word classes studied, nouns exhibit the most significant discriminative power. Combining nouns with verbs or prepositions deteriorates nouns' predictive power, combining nouns with adjectives or adverbs leads to performance gain.
- To our surprise, we find that prepositions demonstrate higher discriminative power than adjectives and adverbs. Models trained with preposition-based features perform close to those trained with noun-based features.
- Among the two broader categories, content words demonstrate higher predictive power than function words.
- Models trained with nouns and adverbs (15 features) perform close to those trained with all 64 features.

7.2.4 Parsed Syntactic Features

Previous study by Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009) has explored four syntactic features, they include include average parse tree height, average number of noun phrases per sentence, average number of verb phrases per sentence and average number of SBARs (a parsing marker for relative clauses) per sentence. In this thesis, we refine and enrich their work by introducing new measures for parsed phrases and two non-terminal-node-based features. As described in Section 6.3, for each phrase, we introduce phrasal length measured by number of words and characters to add additional syntactic information. We use average number of non-terminal nodes per sentence and average number of non-terminal nodes per word to capture the complexity of a parse tree.

We give a brief rationale why certain phrases are selected for the study. Among many parsed phrasal categories, noun phrases and verb phrases are the most common ones which form the basic syntactic constituents of a sentence. Because of this, it is reasonable to believe that these two phrases should be distributed stably across all four grade levels (2-5) and would not be too grade-specific. Relative clauses (indicated by parsing marker "SBARs") are subordinate clauses which are often used to modify noun phrases. They are often perceived to add more complexity to sentence precessing during reading. In section 7.2.3, we have observed that prepositions demonstrate significant discriminative power over other POS labels such as adjectives, adverbs and function words. It would be interesting to further study whether parsed prepositional phrases have competitive predictive power.

Table 7.10: Comparison of our augmented syntactic features with Schwarm & Ostendorf's study (2005). Accuracy generated by LIBSVM classifiers on WeeklyReader.

Feature Set	#. Feat.	Accuracy (%)
Schwarm et al.	4	50.68 ± 0.812
Our approach	21	57.79 ± 1.023

There are also several other common phrasal categories which could be useful in providing further syntactic information, such as wh-question clauses (parse marker "SBARQ"), wh-noun phrases ("WHNP"), wh-prepositional phrases ("WHPP"). However, a closer look at the parsed Weekly Reader data reveals that "WHPP"s and "SBARQ"s tend to be sparse, especially in texts of lower grades. And "WHNP" are often overshadowed by "SBAR"s already. We also observe that adjective phrases ("ADJP") and adverbial phrases ("ADVP") often consist of a single word. A study of which at the phrasal level would pretty much overlap with the analysis we have done in section 7.2.3 on the part-of-speech based features.

We first compare to what extent our augmented syntactic features improve over the four previously studied features. We use LIBSVM to train two classifiers with these two feature sets and evaluate their performance by repeated 10-fold cross-validation and present classification accuracy in 7.10. We see that the LIBSVM classifier trained with our expanded set of syntactic features scored 7 points higher than the one trained on only the original four features, improving from 50.68% to 57.79%.

Table 7.11 shows a detailed comparison of particular parsed syntactic features. We see that the two non-terminal-node-based features (average number of non-terminal nodes per tree and average number of non-terminal

Table 7.11: Accuracy generated by syntactic features on WeeklyReader.

	LIBSVM		Logistic Regression	
Feature Set	Accuracy (%)	F-Measure	Accuracy (%)	F-Measure
ratio Of Nodes	53.02 ± 0.571	0.517 ± 0.017	51.80 ± 0.171	0.495 ± 0.002
avg. tree height	44.26 ± 0.914	0.418 ± 0.008	43.45 ± 0.269	0.339 ± 0.003
SBARs	44.42 ± 1.074	0.418 ± 0.011	43.50 ± 0.386	0.386 ± 0.005
NPs	51.56 ± 1.054	0.504 ± 0.011	48.14 ± 0.408	0.463 ± 0.004
VPs	53.07 ± 0.597	0.531 ± 0.006	48.67 ± 0.484	0.469 ± 0.005
PPs	49.36 ± 1.277	0.482 ± 0.013	46.47 ± 0.374	0.443 ± 0.004
all	57.79 ± 1.023	0.578 ± 0.010	54.11 ± 0.473	0.530 ± 0.005

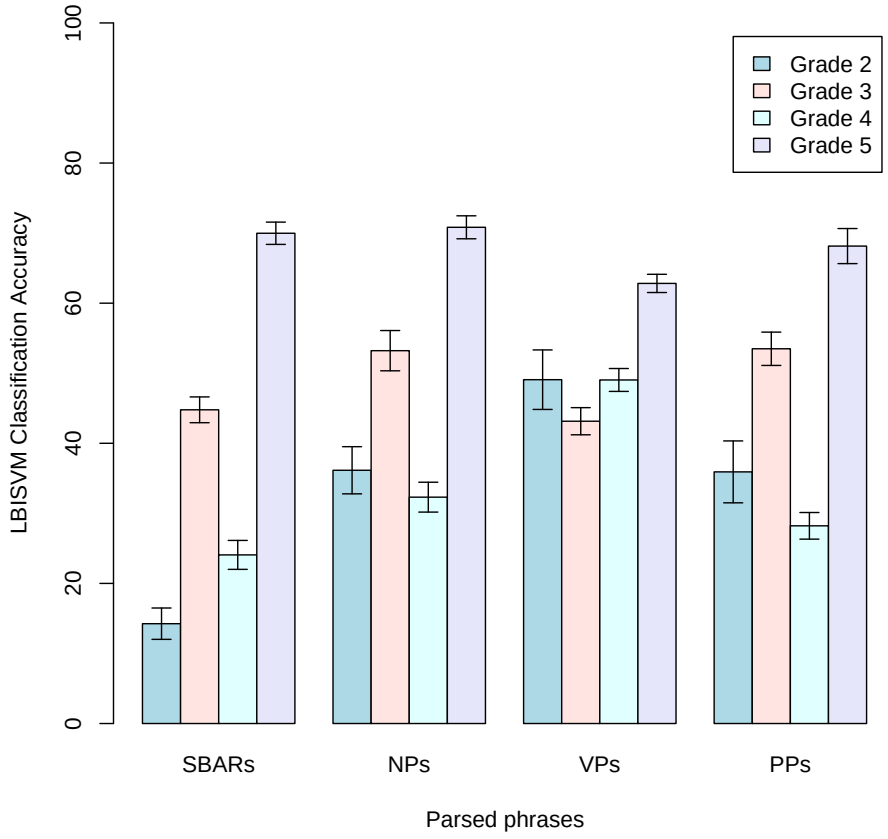


Figure 7.3: Grade level accuracy generated by LIBSVM classifiers trained with parsed syntactic features on WeeklyReader.

nodes per word) have significantly higher discriminative power (53.02% accuracy) than average tree height, which only generates 44.26% accuracy. Among SBARs, NPs, VPs and PPs, our experiment show that VPs have the strongest discriminative power, generating 53.07% accuracy. NPs and PPs perform slightly worse than VPs, generating 51.56% and 49.36% accuracy respectively. Compared with POS-based prepositional features examined in Section 7.2.3, which generate 56.77% accuracy using LIBSVM, the predictive power of prepositional phrases (PPs) are not so competitive. To our surprise, SBARs perform the poorest among all four phrases studied. A LIBSVM classifier trained with SBARs generates only 44.42% accuracy. Combining all 21 syntactic features together leads to best performance, resulting in 57.79% accuracy.

SBARs would be commonly perceived as good indicators for syntactic complexity, because, according to popular wisdom, relative clauses would make sentence processing more challenging, therefore SBARs could serve as effective predictors for text readability. However, based on our experiment results, this common belief is questionable. To find out what could be a reasonable explanation for what we have observed, we take a closer look at the predictive power of parsed phrases in modeling text difficulty at individual grade level. Figure 7.3 shows the bar plot of grade-level classification accuracy generated by LIBSVM classifiers trained with each of the four parsed phrasal features. Comparing the prediction accuracy of all classifiers at individual grade level, we find that there is not much statistically significant differences between model performances at Grade 3, 4 and 5; however, at Grade 2 level, SBARs generate accuracy considerably lower than NPs, VPs and PPs.

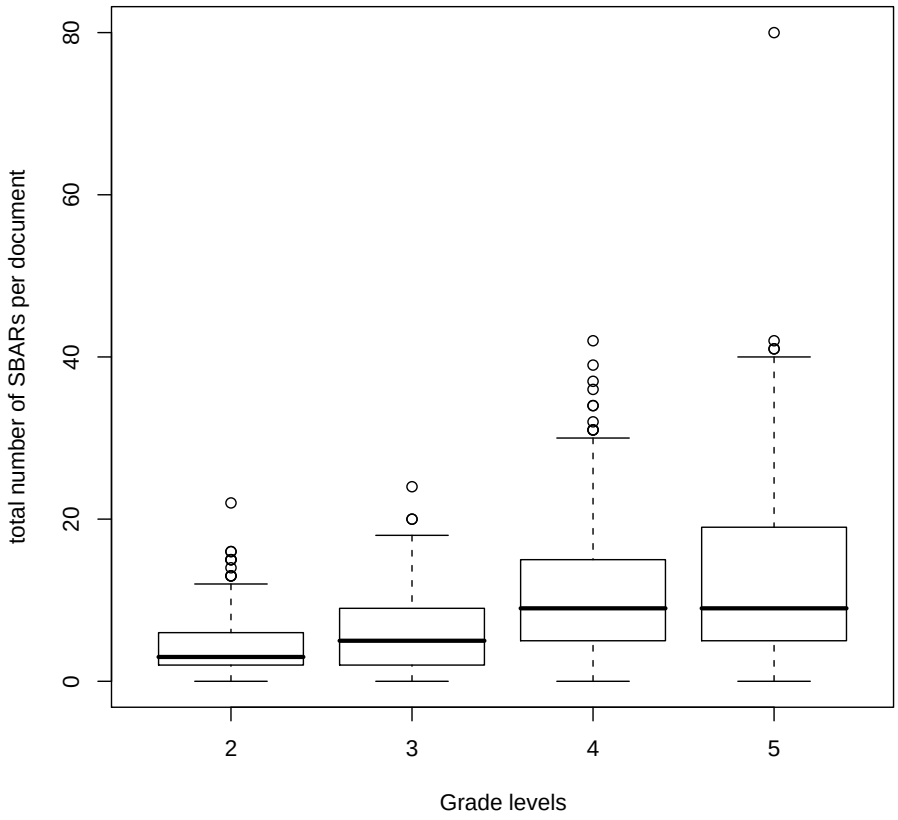


Figure 7.4: Grade-level distribution of total number of SBARs per document in the Weekly Reader Corpus.

Following this hint, we take a closer look at the distribution variation of SBARs in the WeeklyReader corpus. We use the boxplot in Figure 7.4 to show the grade-level distribution of total number of SBARs per document. Boxplot is a very informative data plot that essentially shows the five number summary: median (indicated by the darker line in the middle of the box), the first and third quartiles spread about the median (indicated by the lower and upper ends of the box respectively), max (indicated by the "whisker" above the box) and min (indicated by the "whisker" below the box) values of the data. We see from the boxplot that the distribution of SBARs at grade 2 level is much lower compared with those in texts at Grade 3, 4 and 5 level.

The sparse distribution of SBARs in Grade 2 texts is likely to be responsible for the poor performance of SBARs in general.

On a side note, based on what we observed from Table 7.11 and Figure 7.3, it is worth pointing out that, while the classification accuracy generated by SBARs, NPs and PPs fluctuates somewhat dramatically at all four grade levels, especially at Grade 2 and 4 levels, VPs perform relatively consistently across all four grades, generating the highest accuracy at Grade 2 and 4 levels. This observation indicates that VPs are not only the best performing but also the most robust ones among all four phrases investigated.

To conclude, we have the following key observations for parsed syntactic features:

- Our new approach to syntactic features by introducing phrasal length measures and non-terminal-node-based features makes significant improvement over previous study, generating 7% higher accuracy.
- We systematically studied the discriminative power of NPs, VPs, PPs and SBARs, among which we find that VPs are the best and stablest predictors. Next to VPs are NPs and PPs. SBARs appear to be least discriminative.
- Compared with significant predictive power observed from POS-based prepositions (56.77%), parsed prepositional phrases (PP) exhibit much lower discriminative power (49.36%).
- Sparse distribution of relative clauses in texts of lower grades is likely to be responsible for the poor performance of SBARs.

7.2.5 Shallow Features

Shallow features are limited to superficial text properties, such as average number of syllables per word and average sentence length measured by number of words. Due to lack of advanced natural language processing tools, these simplistic features were mostly explored by many traditional readability metrics, which often use a simple linear function with two or three variables to measure text difficulty. Although there has not been much corpus-based evidence to validate the reliability of traditional metrics, many of them are still popular even today, because the variables they use are simple and easy to calculate.

Benefiting from the advancement of NLP and statistical machine learning techniques, notable improvement has been made recently in developing robust readability assessment tools that are proven to be far better than traditional metrics. Take the Flesch-Kincaid Grade Level as an example, several recent studies have shown that the predictions by this metric are highly unreliable compared with existing state of the art (Collins-Thompson and Callan, 2004; Petersen and Ostendorf, 2006; Si and Callan, 2001).

The unreliability of traditional metrics could lie in two possible factors: either the oversimplified linear functions are not capable of assessing text complexity accurately; or the shallow features selected as predictors lack sufficient discriminative power. Although recent readability studies have strived to take advantage of NLP techniques, little has been revealed about the predictive power of shallow features.

In this section, we use advanced machine learning techniques to assess discriminative power of shallow features and enable a comparison against

Table 7.12: Accuracy generated by shallow features on WeeklyReader.

	Feature Set	Logistic Regression Accuracy (%)	J48 Accuracy (%)	OneR Accuracy (%)
1	avg. num. syll. per word	42.51 ± 0.264	42.41 ± 0.510	36.34 ± 0.316
2	fraction of poly-syll. words per doc.	40.36 ± 0.166	41.22 ± 0.556	39.14 ± 0.759
3	avg. num. poly-syll. words per sent.	45.70 ± 0.306	45.25 ± 0.290	44.15 ± 0.742
4	avg. num. chars per word	39.58 ± 0.239	39.78 ± 0.487	38.90 ± 0.723
5	ChallDale	42.26 ± 0.311	42.46 ± 0.381	38.90 ± 0.723
6	avg. sent. length by chars	50.65 ± 0.235	51.47 ± 0.565	48.93 ± 0.532
7	avg. sent. length by words	52.17 ± 0.193	52.18 ± 0.364	47.11 ± 0.802
8	total num. words per doc.	37.68 ± 0.254	46.43 ± 0.403	44.19 ± 0.679
9	Flesch-Kincaid scores	50.83 ± 0.144	53.52 ± 0.000	53.48 ± 0.064
	all combined	52.34 ± 0.242	52.99 ± 0.672	48.92 ± 0.837

more advanced features. For each of 9 shallow features described in Section 6.5, we use Logistic Regression, J48 (decision tree) and OneR (classifier for single feature, because many feature subsets examined in this section consist of one single feature) from the Weka machine learning toolkit to build classifiers. We then evaluate them on the entire WeeklyReader corpus using repeated 10-fold cross-validation and present some notable findings in Table 7.12.

We see that, among three types of classifiers, Logistic Regression generates best results for feature 1, 2 and 3, and J48 (decision tree) generates the highest accuracy for the rest of features. Across all classifiers, we find that average sentence length has dominating predictive power over lexical features. Using J48, average sentence length measured by characters generates 51.47% accuracy, average sentence length measured by words generate 52.18% accuracy, both perform much better than features based on syllable counting. Among three syllabic-based features, average number of poly-syllabic (≥ 3) words per sentence has much stronger discriminative power than average number of syllables per word and percentage of poly-

syllabic words per document. The rate of difficult words measured against the Chall-Dale common word list is also among low-performing features, generating only 42.46% accuracy using J48. The predictive power of text length measured by number of words appears to be a debatable. While the Logistic Regression classifier trained with this feature generates 37.68% accuracy, the lowest among all features, the accuracy generated by both J48 and OneR classifiers trained with the same feature indicates that document length is more effective than all other lexical-based features.

We have an interesting finding on the feature based on Flesch-Kincaid scores. The Flesch-Kincaid Grade Level score uses a fixed linear combination of average words per sentence and average syllables per word. We have pointed out earlier in this section that Flesch-Kincaid scores predict text readability very poorly. In our experiment, when tested on the entire Weekly Reader corpus, the scores generated by the Flesch-Kincaid Grade Level formula only predict 20 out of 1433 texts with correct grade levels, resulting a poor accuracy of 1.4%. However, we see from the above table that, when used as a feature for advanced machine learning tool, the Flesch-Kincaid scores demonstrate the most significant predictive power. Using J48 and OneR, Flesch-Kincaid scores alone generate 53.52% and 53.48% accuracy respectively, which is even higher than that by all 9 features combined. This is not completely surprising, because average sentence length, which demonstrates to be highly discriminative, contributes as an important variable to the Flesch-Kincaid Grade Level formula.

To conclude this section, we have the following notable findings on shallow features:

- Average sentence length exhibits dominating discriminative power over lexical-based shallow feature.
- Flesch-Kincaid scores perform poorly when used directly to model reading difficulty of texts. When applied as a feature in advanced machine learning tool, they demonstrate significant discriminative power comparable to average sentence length.

7.2.6 Comparison of Features across Linguistic Levels

In this section, we present results in Table 7.13 to compare the effectiveness of features across all linguistic levels in detecting and predicting reading difficulty of texts in terms of elementary grade levels. 5gramWR and 3gramBL are language-modeling-based perplexity features. 5gramWR combines 80 features obtained from LMs trained on the WeeklyReader corpus using all four feature selection schemes: IG, textOnly, posOnly and tagged. 3gramBL combines 48 features obtained from LMs trained on Britannica and LiteracyNet using the same feature selection schemes. Discourse features combine entity-density features, lexical chain features, coreferential inference features and entity grid features, resulting in 45 features in total. For comparison's sake later in Section 7.3, we replicated 6 out-of-vocabulary features (OOV) based on Schwarm and Ostendorf's work and include them in our study as well. The rest of three feature subsets – POS, syntactic and shallow – are self explanatory.

Table 7.13: Comparison of features across linguistic levels.

Feature Set	# Feat.	LIBSVM Accuracy (%)	Logistic Reg. Accuracy (%)	SMO Accuracy (%)
5gramWR	80	68.38 ± 0.929	66.82 ± 0.448	62.40 ± 0.344
discourse	45	60.50 ± 0.990	58.79 ± 0.703	57.47 ± 0.351
POS	64	59.82 ± 1.235	57.86 ± 0.547	57.12 ± 0.349
syntactic	21	57.79 ± 1.023	54.11 ± 0.473	51.64 ± 0.299
shallow	9	56.04 ± 1.364	52.34 ± 0.242	51.56 ± 0.404
OOV	6	55.85 ± 1.106	54.03 ± 0.207	51.20 ± 0.182
3gramBL	48	53.61 ± 0.847	52.97 ± 0.514	43.12 ± 0.316
all combined	273	72.21 ± 0.821	63.71 ± 0.576	70.90 ± 0.270

To compare the relative discriminative power the feature subsets examined, we use LIBSVM, Logistic Regression and SMO (a sequential minimal optimized SVM using polynomial kennel) from Weka toolkit to train and test with each set of features on the entire WeeklyReader corpus using repeated 10-fold cross-validation. We present the accuracy results in Table 7.13.

We find that, among all three type of classifiers, those trained using LIBSVM generate the best performance in terms of accuracy. 5gramWR features obtained from LMs trained directly on the WeeklyReader corpus exhibit the strongest discriminative power. Models trained with this feature subset generate significantly higher classification accuracy than the remaining models trained with single feature subset. However, 3gramBL features, extracted using same feature selection and language modeling techniques, but obtained from LMs trained on unrelated corpora, appear to be least useful among all single feature subsets.

Combined discourse features and POS features demonstrate comparable predictive power next to 5gramWR features. From experiment results we discussed in Section 7.2.1 and 7.2.3, we know that nouns, the word class with most discriminative power, contribute significantly to the good overall

performance of POS features. The high discriminative power of nouns in turn explains the good performance of entity-density features, which are primarily based on nouns and contribute the most to the high performance of combined discourse features. We observed in Section 7.2.1 that the remaining three subsets of discourse features, i.e. lexical chain features, coreferential inference features and entity grid features, do not appear to be very useful in modeling text difficulty in elementary grade levels. Therefore we conclude that nouns are the most significant contributor to the overall good performance of POS features and combined discourse features.

Table 7.13 also shows that the relative discriminative power of syntactic features is below 5gramWR, combined discourse features and POS features and slightly above shallow features.

As we stated clearly at the beginning of this chapter, accuracy as an evaluation measure is appropriate to the classification task, however it may not fully reflect some other aspects of the task in assessing reading difficulty of texts in elementary grade levels. For instance, different from other standard classification task, reading difficulty in our study is ranked by the grade levels assigned. A miss by more than one grade levels should not be treated equally, but rather more severe than a misclassification of one grade level. With this consideration in mind, we apply more evaluation measures to the prediction results by the best performing LIBSVM classifiers as shown in Table 7.14. In addition to accuracy, these measures include mean squared error (MSE), mean absolute error (MAE), total number of misclassification by just one grade levels (missOneGL), total number of misclassification by more than one grade levels (missMoreGL). We see from the table that, evaluated by multiple measures, the rank of relative

Table 7.14: Comparison of features across linguistic levels: multi-level evaluation.

Feature Set	Accuracy (%)	MSE	MAE	missOneGL	missMoreGL
5gramWR	68.38 ± 0.929	0.52 ± 0.02	0.38 ± 0.01	376 ± 12	78 ± 5
discourse	60.50 ± 0.990	0.68 ± 0.03	0.48 ± 0.01	463 ± 14	103 ± 8
POS	59.82 ± 1.235	0.73 ± 0.04	0.50 ± 0.02	457 ± 15	119 ± 8
syntactic	57.79 ± 1.023	0.74 ± 0.02	0.52 ± 0.01	496 ± 15	108 ± 6
shallow	56.04 ± 1.364	0.78 ± 0.04	0.54 ± 0.02	504 ± 13	126 ± 10
OOV	55.85 ± 1.106	0.72 ± 0.04	0.53 ± 0.02	524 ± 15	109 ± 14
3gramBL	53.61 ± 0.847	0.91 ± 0.04	0.60 ± 0.02	507 ± 11	158 ± 11
all combined	72.21 ± 0.821	0.39 ± 0.02	0.30 ± 0.01	325 ± 10	47 ± 7

discriminative power of all feature subsets remains consistent with what we obtained from analyzing based on accuracy alone.

In addition to comparing overall performance of individual feature subsets, we present detailed classification accuracy generated by these features at grade level in Table 7.15 based on predictions by LIBSVM classifiers. The barplot shown in Figure 7.5 presents the graphic view of the table. We see that, at individual grade level, the relative discriminative power of each feature subsets varies from grade to grade. Among all feature subsets, 5gramWR features generate the highest accuracy for Grade 2, 4 and 5. At Grade 3 level, to our surprise, syntactic features generate the highest accuracy. Between discourse features and POS features, discourse features generate much lower accuracy for Grade 2 texts than POS features. But at Grade 3 and 4 level, discourse features perform better than POS features. Both subsets of features generate the same accuracy for Grade 5 texts. Between syntactic and shallow features, shallow features outperform syntactic features at Grade 2 and 5 level, and syntactic features generate better accuracy for Grade 3 and 4 texts.

Table 7.15: Grade level accuracy generated by LIBSVM classifiers trained with major feature subsets on WeeklyReader.

Feature Set	Grade 2 Accuracy (%)	Grade 3 Accuracy (%)	Grade 4 Accuracy (%)	Grade 5 Accuracy (%)
5gramWR	70.75 ± 2.750	61.59 ± 1.976	58.97 ± 1.489	78.67 ± 1.634
discourse	52.70 ± 4.802	57.09 ± 2.952	48.55 ± 2.614	74.24 ± 1.579
POS	57.41 ± 6.272	53.04 ± 2.388	47.10 ± 1.526	74.24 ± 1.825
syntactic	49.48 ± 3.089	62.49 ± 1.819	49.02 ± 2.020	64.89 ± 2.045
shallow	50.52 ± 4.967	56.44 ± 3.093	43.86 ± 1.421	67.21 ± 2.331

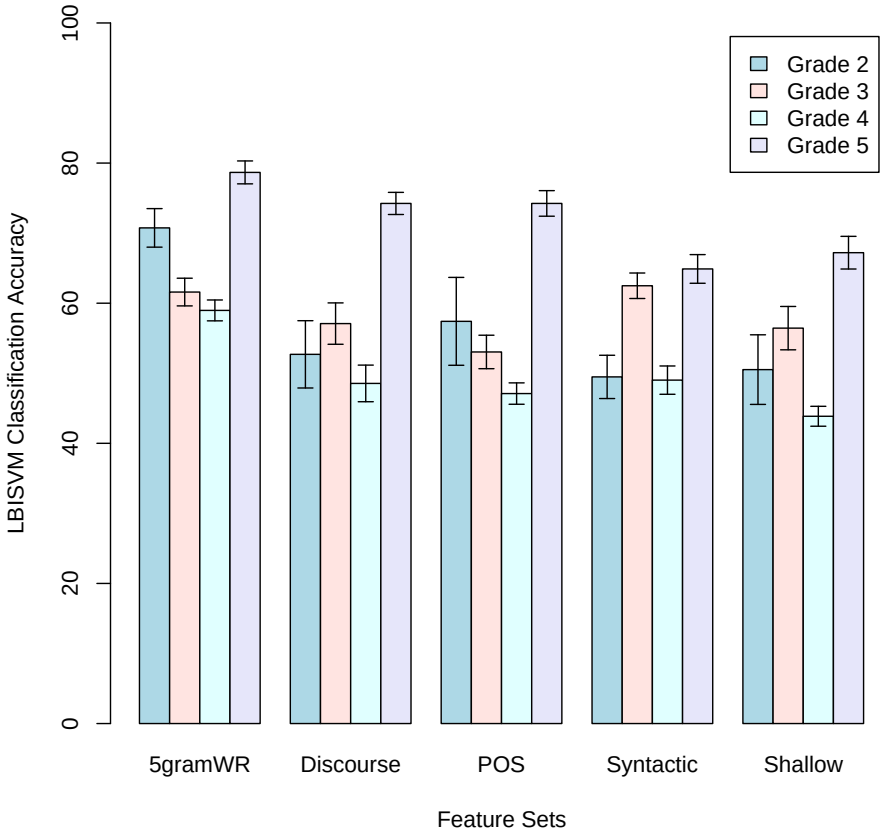


Figure 7.5: Grade level accuracy generated by LIBSVM classifiers trained with major feature subsets on WeeklyReader.

7.3 Comparison with Previous Studies

In this section, we compare our study with related work in the field. To set benchmark for comparisons, we first discuss our baseline performance in Section 7.3.1. We then compare the baseline with the widely used traditional metric Flesch-Kincaid Grade Level in Section 7.3.2.

As discussed earlier, the work by Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009) is closely related to our study. However, their corpus of study, though similar to ours, differs in size and content, plus our choices of machine learning tools are different from theirs, which makes their published results not directly comparable to ours. To make fair comparison possible, we replicate all features used in their study, train and test prediction models with these features using our corpus according to our experiment design. Section 7.3.3 describes the details of our replication of Schwarm and Ostendorf's features.

For comparison, we come up with three combinations of our features listed as follows:

- **All features**: a naive combination of all features
- **WekaFS**: a subset of features selected by Weka's feature selection tool
- **AddOneBest**: a subset of features resulted from group-wise add-one-best greedy approach

Section 7.3.4 describes the details of how the last two combinations are obtained.

We choose three classifiers – LIBSVM, Logistic Regression and SMO from weka toolkit – to train and test prediction models with Petersen

et al.'s features and the three combinations of our features on the entire Weekly Reader corpus. The performances of the models are evaluated by classification accuracy using 10-fold cross-validation. Each model is run 10 times. The experiment results are compared and discussed in Section 7.3.6.

7.3.1 Baseline

We use Weka's ZeroR classifier to obtain the baseline performance, which simply predicts the majority class. In our case, the majority class is the one that contains 542 texts labeled with Grade 5. The trivial baseline predicts 542 out of 1433 texts (or 37.8%) correctly.

7.3.2 Flesch-Kincaid Grade Level

With the trivial baseline in mind, we first compare our study with the widely-used Flesch-Kincaid Grade Level formula, which is a linear function of average words per sentence and average syllables per word that aims to predict the grade level of a text directly. Since this is a fixed formula with known coefficients, we evaluated it directly on our entire Weekly Reader corpus without cross-validation. We obtain the predicted grade level of a text by rounding the Flesch-Kincaid score to the nearest integer. For only 20 out of 1433 texts the predicted and labeled grade levels agree, resulting in a poor accuracy of 1.4%. By contrast, using the Flesch-Kincaid score as a feature of a simple logistic regression model achieves above 50% accuracy, as discussed in Section 7.2.5.

7.3.3 Replication of Schwarm and Ostendorf's Work (2005)

Previous work in the field that is closely related to ours are studies by Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009). However, their published results are not directly comparable to ours for two reasons: first, their corpus of study is different from ours; second, their experiment design and machine learning tools are different from ours.

As discussed in Section 5.1, although we both had the Weekly Reader as our data source, the size and the content of the two corpora differ from each other. In addition, we cleaned up our data by throwing out texts that consisted of puzzles and simple multiple-choice questions, which we think would not be meaningful for deeper syntactic parsing or discourse analysis.

In Schwarm and Ostendorf's work, they used binary classifiers from SVM^{light} (Joachims, 1999) to train a prediction model for each grade. The results they reported were based on one run of the model trained with 90% training data on 5% held-out testing data. The results were evaluated by precision, recall and F-measure at individual grade level.

Our experiment design differs from Schwarm and Ostendorf's as follows. To train, test and validate prediction models, we use various classifiers known for efficient multi-label classification, ranging from support vector machines to logistic regressors. The performance of each model is evaluated by 10-fold cross-validation using classification accuracy and F-measure, both across all grade levels and at individual grade level. Each model is run ten times, the results are reported in terms of mean and standard deviation.

To make fair comparison possible, we have replicated all features used in Schwarm and Ostendorf's study. The rationale is to build prediction

models with these replicated features on our Weekly Reader corpus and evaluate them according to our experiment design.

There are 25 features used in Schwarm and Ostendorf's study:

- 3 shallow features: average sentence length measured by number of words, average number of syllables per word and Flesch-Kincaid score
- 4 parse features: average parse tree height, average number of noun phrases, verb phrases and SBARs per sentence
- 6 out-of-vocabulary (OOV) rate scores

  The 6 OOV scores were computed using three common word lists: the most common 100, 200 and 500 words occurring in the lowest grade level – Grade 2 texts. Using each list, the percentage of word instances (tokens) and unique words (types) of each article that do not appear in this list is computed to form two OOV scores, resulting in six OOV features in total.
- 12 language-modeling-based perplexity scores

  Schwarm and Ostendorf used information gain theory to rank and select words from their version of Weekly Reader corpus. The corpora they used to train language models are two unrelated paired corpora: Britannica and LiteracyNet, which consist of paired original texts and their corresponding simplified versions adapted by human editors (see Section 5.4). Before training LMs, they used the words selected from the Weekly Reader corpus with high information gain to preprocess the training corpora. Words that appear in the selected word list were kept, and those that do not appear in the word list were replaced by

their POS tags. Using this scheme, they divided the preprocessed two paired corpora into four smaller subsets, each containing only the original texts or their abridged versions. They then trained three LMs (unigram, bigram and trigram) on each of the four smaller corpora, resulting in 12 LMs. Each article in the Weekly Reader corpus was then tested against these 12 LMs to obtain 12 perplexity scores.

We adopted Petersen and Ostendorf's information gain approach for feature selection and used Britannica and LiteracyNet to replicated all 12 perplexity features using our version of Weekly Reader corpus. The details of the replication procedures have been described in Section 7.2.2.

All 25 features replicated from Schwarm and Ostendorf's study are included as a subset in our features.

7.3.4 Model Optimization using Feature Selection

To compare with features replicated from Schwarm and Ostendorf's study, we come up with three combinations of our features. The first one is a naive combination of all 273 features, the other two are obtained from different feature selection techniques.

An unfiltered naive combination of all features has several drawbacks. These drawbacks become especially notable when the feature size gets bigger. The first problem is inefficiency. Increasing feature size slows down the learning process. It's time consuming to train learning models with a large number of features, especially with a big training corpus. Second, models trained with unfiltered large features do not provide much insight

as to which features are more important or less relevant and how they are related with each other. To build robust learning models with better interpretability, feature selection is often used in machine learning and statistics for improvement, where a subset of important features are kept, and most irrelevant or redundant features are removed.

From a theoretical point of view, in order to achieve an optimal subset of features that are most relevant to the model performance, an exhaustive search on all possible subsets are required. This is in general unrealistic when the feature size is big. In our case, we have 273 features in total, the computational cost for an exhaustive search would be 2^{273}, which is impractical to carry out.

To overcome this problem, ad hoc greedy algorithms are often used in machine learning and statistics to achieve a practically satisfactory feature subset. We describe below two popular greedy selection methods for our feature selection. The first one is a stepwise forward selection algorithm which takes a predefined subset of features as a group each time. We use this approach to investigate how subsets of features extracted from various linguistic levels interact with each other. The feature subset resulted from this algorithm is referred to as "Groupwise-Add-One-Best", abbreviated as "GAOB". The second one is also a greedy forward selection algorithm. Instead of taking a subset of features as a group each time, it evaluates each individual features at a time and decides whether to keep it or remove it from the satisfactory subset of features. We use Weka's automated attribute selection filter to carry out this algorithm, the resulted feature subset is referred to as "Weka-Feature-Selection", abbreviated as "WekaFS".

Groupwise-Add-One-Best

We use a groupwise greedy forward selection approach to investigate how features extracted from various linguistic levels interact with each other and which subsets of features, when combined together, achieve the best performance. The stepwise algorithm goes as follows: we treat each feature subset as a group and start with the best performing subset. In each round of the experiment, we add one feature subset at a time to the starting subset and build a model with the combined features. Each model is evaluated by classification accuracy using 10-fold cross-validation. At the end of each round of experiment, we compare which subset, when added, helps achieve the highest classification accuracy. If such a subset exists, it is selected to combine with the original starting feature subsets to the next round of experiments. When no satisfactory improvement is made in a round, the selection procedure is terminated. The selected feature combination prior the terminating round is returned as the optimal feature subset.

We have extracted and implemented 8 subsets of features at various linguistic levels, which include four subsets of discourse features **entity-density features**, **lexical-chain-based features**, **coreference inference features**, **entity-grid-based features**, **5gramWR** (perplexity features based on 5-gram LMs trained on the Weekly Reader corpus), **syntactic features**, **POS features**, and **shallow-features**. To provide a collective understanding of discourse features, we combined all four subsets of discourse features together as a separate subset of features and refer to it as "**discourse-Combined**".

In Section 7.2.3, we have experimented with the combinations of noun-based features with other POS categories. We found that combining features based on nouns and adverbs achieves classification accuracy as close as all 64 POS features combined (see Table 7.9). We consider this combination as a near optimal alternate subset for the POS features, which we refer to as "**NPNAdv**".

In addition, we replicated two subsets of features from Schwarm and Ostendorf's study (2005), which we refer to as **OOV features** (out-of-vocabulary rate) and **3gramBL** (perplexity features based on 3-gram LMs trained on the Britannica and LiteracyNet corpora).

In total, we have 12 predefined subsets of features to carry out the groupwise greedy forward feature selection algorithm. We use LIBSVM to build and test models of selected combination of feature subsets. We start with the best performing feature subset "**5gramWR**", which generates the highest classification accuracy among all individual feature subsets (see Table 7.13 in Section 7.2.6). After four rounds of experiments, the following combination of feature subsets are selected in that order as the satisfactory subset of features: **5gramWR + syntactic + OOV + NPNAdv**. We refer to this combination of feature subsets returned by the greedy selection algorithm as "Groupwise-Add-One-Best", abbreviated as "GAOB".

7.3.5 Weka-Feature-Selection

The "Groupwise-Add-One-Best" combination obtained from the experiments as described above provides a general understanding as to which feature subsets at what linguistic levels are most relevant to achieve satisfac-

tory model performance. However, since the algorithm treats each feature subset as a group, it is limited in providing useful information as to which specific features are most important.

The Weka machine learning toolkit provides a filter that automatically searches through all attributes (feature) at individual feature level to find the subset that works best for the prediction. The Weka's attribute (feature) selection process is straightforward: one only needs to specify which search method and evaluation measure to use. We choose cross-validation as evaluation measure and best-first forward search method to start the filter. Out of 273 features, the filter returned a subset of 28 features as the most relevant ones. We list these 28 features in Table 7.16 according to the linguistic levels at which they are extracted. We refer to this subset of features as "**Weka-Feature-Selection**", abbreviated as "**WekaFS**".

From the table we see that features selected by Weka's machine learning tool are at similar linguistic levels to those selected by the groupwise greedy algorithms, they include perplexity features obtained from LMs trained on the Weekly Reader corpus, parsed syntactic features, POS features and OOV features. In addition, the weka filter also selected four features from the subset of shallow features. Except the four shallow features and five OOV features, all of the remaining 19 features resulted from our newly implemented features, they have not been studied in any previous research. Most of the features selected are not a complete surprise, because, as described and discussed in Section 7.2, through comparisons of features at the level of each individual feature subsets, we already have a good understanding as to which are good-performing features.

Table 7.16: 28 features obtained from Weka feature selection.

Short Code	Feature Description
shallow features:	
nonCommonWord	rate of diff. words against Dale-Chall's common wordlist
sentLenWord	avg. sentence length by words
sentLenChar	avg. sentence length by chars.
totalWords	doc. length by words
syntactic features:	
VPLenChar	avg. length of verb phrases by chars.
PPLenChar	avg. length of prepositional phrases by chars.
NP-VP-PPs	avg. number of NPs + VPs + PPs per doc.
POS features:	
uniqNounPerSent	avg. num. of unique nouns per sent.
uniqAdjPerSent	avg. num. of unique adjectives per sent.
uniqVerbsPerSent	avg. num. of unique verbs per sent.
PrepPerSent	avg. num. of prepositions per sent.
uniqContentPerSent	avg. num. of unique content words per sent.
FunctionPerSent	avg. num. of function words per sent.
OOV features:	
tok200	OOV by tokens against 200 common words of Grade 2 texts
tok600	OOV by tokens against 600 common words of Grade 2 texts
type100	OOV by types against 100 common words of Grade 2 texts
type200	OOV by types against 200 common words of Grade 2 texts
type600	OOV by types against 600 common words of Grade 2 texts
Perplexity features:	
ig2gWRGr2	IG bigram perplexity using LMs trained on Grade 2 texts
ig3gWRGr2	IG trigram perplexity using LMs trained on Grade 2 texts
ig4gWRGr2	IG 4-gram perplexity using LMs trained on Grade 2 texts
ig5gWRGr3	IG 5-gram perplexity using LMs trained on Grade 3 texts
ig1gWRGr4	IG unigram perplexity using LMs trained on Grade 4 texts
tok3gWRGr2	textOnly trigram perplexity using LMs trained on Grade 2 texts
tok2gWRGr4	textOnly bigram perplexity using LMs trained on Grade 4 texts
posOnly5gWRGr2	posOnly 5-gram perplexity using LMs trained on Grade 2 texts
posOnly5gWRGr3	posOnly 5-gram perplexity using LMs trained on Grade 3 texts
tagged2gWRGr5	tagged bigram perplexity using LMs trained on Grade 5 texts

Table 7.17: Comparison with previous work.

Baseline accuracy (majority class)		37.8		
Flesch-Kincaid Grade Level		1.4		
Feature Set	# Feat.	LIBSVM	Logistic Reg.	SMO
Schwarm	25	63.18 ± 1.664	60.50 ± 0.477	61.59 ± 0.286
All features	273	72.21 ± 0.821	63.71 ± 0.576	70.90 ± 0.270
WekaFS	28	70.06 ± 0.777	65.46 ± 0.336	64.49 ± 0.243
GAOB	122	74.01 ± 0.847	69.22 ± 0.411	69.33 ± 0.331

7.3.6 Results and Discussion

To compare the set of 25 features replicated from Schwarm and Ostendorf's study with the three combinations of our features (a naive combination of all features, Weka-Feature-Selection and Groupwise-Add-One-Best), we choose three classifiers known for their efficient multi-label classification to build and test prediction models on the entire Weekly Reader corpus: LIBSVM, Weka's Logistic Regression and SMO (SVM based on sequentially minimized optimization using linear kennel). The model performance is evaluated by classification accuracy using 10-fold cross-validation. Each model is run 10 times. The mean and standard deviation generated by each model are presented in Table 7.17 in comparison with baseline benchmark and Flesch-Kincaid Grade Level metric.

We see from Table 7.17 that all four feature sets perform much better than baseline accuracy (37.7%). Overall, across all three types of classifiers, models built with the three combination of our features perform significantly better than those trained with the features replicated from Schwarm and Ostendorf's study. Using LIBSVM, a naive combination of all features results in classification accuracy of 72%, which is 9% higher than accuracy

Table 7.18: Comparison with previous work: multi-level evaluation.

Feature Set	Accuracy	MSE	MAE	missOneGL	missMoreGL
Schwarm:	63.18 ± 1.664	0.55 ± 0.03	0.43 ± 0.02	454 ± 20	73 ± 6
All features	72.21 ± 0.821	0.43 ± 0.02	0.32 ± 0.01	343 ± 14	56 ± 5
WekaFS	70.06 ± 0.777	0.49 ± 0.02	0.36 ± 0.01	360 ± 9	69 ± 5
GAOB	74.01 ± 0.847	0.39 ± 0.02	0.30 ± 0.01	325 ± 10	47 ± 7

generated by the model trained with features replicated from Schwarm and Ostendorf's work (2005). This is not very surprising, since we are considering a greater variety of features than Schwarm and Ostendorf's previous study. The feature subset selected by Weka's automatic attribute (feature) selection filter that employs roughly the same number of features as Schwarm and Ostendorf's feature set still leads to significantly improved accuracy (70%). Our best results were obtained by group-wise add-one-best feature selection, resulting in 74% classification accuracy, with a near 11% improvement over the current state of the art.

In addition to accuracy, we use multi-level evaluation measures on predictions generated by the best performing LIBSVM classifier and present results in Table 7.18. These measures include mean squared error (MSE), mean absolute error (MAE), total number of misclassification by just one grade levels (missOneGL), total number of misclassification by more than one grade levels (missMoreGL). We see from the table that, evaluated by multiple measures, the discriminative power of all four combinations of features remains in the same order as we analyzed above using classification accuracy alone.

In addition to comparing the overall classification accuracy generated by the four feature sets as discussed above, we took the results generated by the best-performing LIBSVM models and performed detailed analysis

at individual grade level. Table 7.19 presents the statistics, the graphical view of which is illustrated by Figure 7.6. We find that, at Grade 2 level, while the three combinations of our features generate higher accuracy than the replication of Schwarm and Ostendorf's features, the performance differences among them are not statistically significant. However, we can see clearly from the barplot in Figure 7.6 that, at the level of Grade 3, 4 and 5, the three combinations of our features outperform the current state of the art with considerable margin, all improvements made are statistically significant.

At individual grade levels, the "Groupwise-Add-One-Best" combination (GAOB), which consists of "5gramWR", "syntactic", "OOV" and "NPNAdv" features and leaves out "shallow, "discourse" and "3gramBL" features, generates the highest classification accuracy at across all four grades. This implies that shallow features, 3gramBL features and discourse features do not seem to be very useful in building an accurate readability metric. 3gramBL features, as discussed in Section 7.2.2 and 7.3.3, were replicated from Schwarm and Ostendorf's previous study, where they obtained perplexity scores for Weekly Reader texts using LMs trained on unrelated data. Despite the complicated twist of information gain approach, the performance of this feature set is the poorest compared with other feature sets. It is not surprising that 3gramBL perplexity features are not useful in improving the overall model performance. Although shallow features do not seem to be relevant either in the GAOB combination where each individual feature subset was treated as a group during the feature selection process, its usefulness can not be completely ignored. When the importance or relevance of the features is evaluated at individual level, some of the

Table 7.19: Comparison with previous work: grade level accuracy based on predictions by LIBSVM classifiers on WeeklyReader.

Feature Set	Grade 2	Grade 3	Grade 4	Grade 5
Schwarm:	70.80 ± 3.888	59.65 ± 3.165	51.07 ± 2.564	72.18 ± 1.158
All features	74.02 ± 3.093	67.37 ± 1.444	65.35 ± 1.813	79.61 ± 0.862
WekaFS	74.43 ± 2.989	65.09 ± 1.936	64.07 ± 1.993	76.03 ± 1.225
GAOB	75.98 ± 4.206	69.27 ± 1.962	67.97 ± 1.634	80.66 ± 1.717

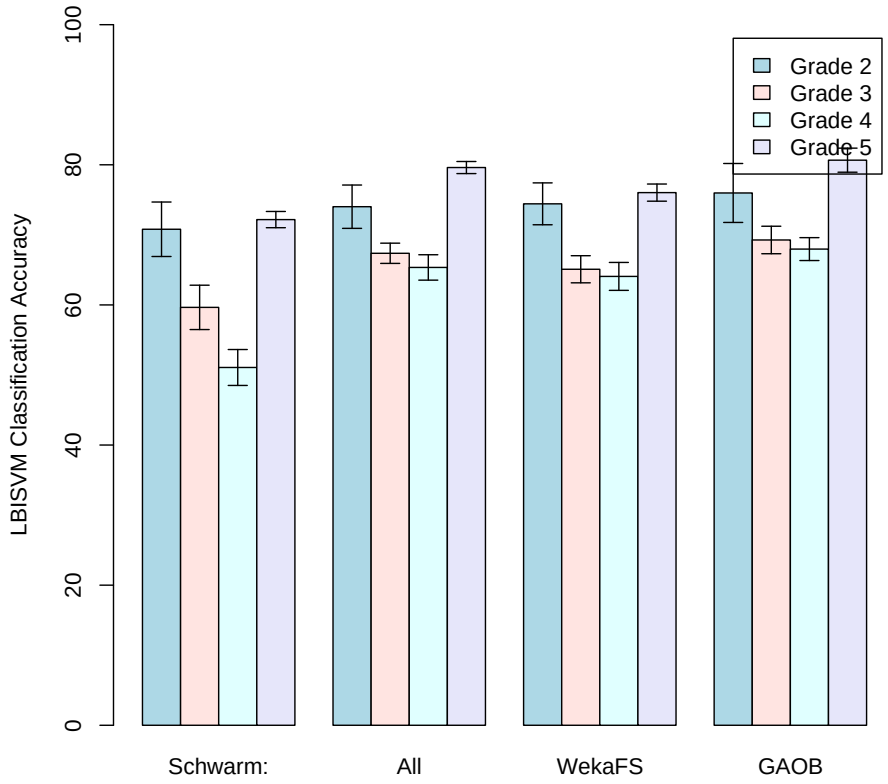


Figure 7.6: Comparison with previous work: grade level accuracy based on predictions by LIBSVM classifiers on WeeklyReader.

shallow features, such as average sentence length and Chall-Dale's rate of difficult words, were ranked highly enough to be picked by the Weka's attribute selection filter to form a minimized satisfactory feature subset.

We notice that none of the discourse features were selected by WekaFS or GAOB feature combinations. There could be several reasons. Through feature analysis in Section 7.2.1 we see that three out of four subsets of discourse features – lexical-chain-based features, coreferential inference features and entity-grid-based features – do not demonstrate strong predictive power. The reason could lie in the fact that the texts in the corpus we studied exhibit relatively low complexity, since they are aimed at primary-school students. Therefore the targeted features may appear sparsely in the texts. Entity density features, however, have generated significantly better classification accuracy (59.63%) comparable to POS-based noun features (58.15%). It is clear from our definition of entity density features – union of general nouns and named entities – that this feature set is highly correlated with POS-based noun features. Since both WekaFS and GAOB combinations have picked POS-based noun features to form the satisfactory feature subset, it would likely be redundant to include entity density features as well. In any case, it is still worth investigating whether these discourse features exhibit different discriminative power for texts at higher grade levels.

7.4 Conclusions

In this chapter, we examined the usefulness of features within and across various linguistic levels for predicting text readability in terms of assigning texts to elementary school grade levels. We implemented a set of discourse

features, enriched previous work by creating several new features, and systematically tested and analyzed the impact of these features.

We observed that POS features, in particular nouns and prepositions, have significant predictive power. The high discriminative power of nouns in turn explains the good performance of entity-density features, based primarily on nouns. In general, our selected POS features appear to be more correlated to text complexity than syntactic features, shallow features and most discourse features.

For parsed syntactic features, we found that features based on VPs appear to be more closely correlated with text complexity than other types of phrases. While SBARs are commonly perceived as good predictors for syntactic complexity, they did not prove very useful for predicting grade levels of texts in this study. We also find that the discriminative power of prepositional phrases decreases considerably compared with that of preposition-based POS features alone.

Among the 9 shallow features, which are used in various traditional readability formulas, we identified that average sentence length has dominating predictive power over all other lexical or syllable-based features. Flesch-Kincaid scores, though perform very poorly when used directly to measure reading difficulty of texts, demonstrates to have high discriminative power when used as a feature by advanced machine learning techniques to assess text readability.

Not surprisingly, among language modeling features, combined features obtained from LMs trained directly on the WeeklyReader corpus show high discriminative power, compared with features from LMs trained on unrelated corpora.

Discourse features, with the exception of entity density features, which are highly correlated with POS features, do not seem to be very useful in building an accurate readability metric. The reason could lie in the fact that the texts in the corpus we studied exhibit relatively low complexity, since they are aimed at primary-school students. We will investigate whether these discourse features exhibit different discriminative power for texts at higher grade levels in Section 8.2.4.

We compared our study with previous work, in particular Flesch-Kincaid Grade Level and recent work by Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009). A judicious combination of features examined here results in a significant improvement over the state of the art.

Chapter 8

Evaluation on Unseen Data

8.1 Introduction

In this chapter, we further evaluate our automatic text readability assessment tool on previously unseen data. We focus on investigating how well models built with grade levels generalize to unseen texts and what their limitations are. It is challenging to evaluate model performance on unseen data, because the reading difficulty of the texts contained within is unknown. We address this problem by two separate approaches: 1) we have experts annotate text difficulty of the same set of data; 2) we design a reading experiment based on the same set of data, recruit adult readers with ID to read assigned texts and answer simple comprehension questions. We then use a hierarchical latent trait model to infer text difficult based on test participants' reading ability. These two alternative measures for reading difficulty of texts allow us to evaluate model performances and investigate relations between grade level predictions, expert ratings and inferred text difficulty for adults with ID.

We use LocalNews2007 and LocalNews2008 as described in Section 5.2.1

and 5.2.2 to evaluate generalizing ability of our automatic text readability assessment tool constructed on the WeeklyReader corpus. We compare model performances and discuss their limitations in Section 8.2. We also provide technique to improve model performance by adding more texts with reading difficulty higher than Grade 5 into training data.

Because expert rating is expensive and time consuming, we only have 22 texts in LocalNews2008 rated by three experts. We discuss details of expert ratings in Section 8.3. In Section 8.4, we present a hierarchical latent trait model to infer reading difficulty of texts in LocalNews2008 for adults with ID. Section 8.5 presents a summative analysis on relations between grade level predictions, experts ratings and text difficulty inferred for adults with ID. We conclude in Section 8.6.

The two unseen corpora LocalNews2007 and LocalNews2008 consists of paired original and simplified text adapted specifically for adult readers with ID. The level of reading difficulty of simplified texts is considerably reduced compared with the original ones. When evaluating model predictions, expert ratings and inferred text difficulty for adults with ID separately, we focus on two aspects that are to the unique characteristics of relevant data: whether this particular measure of reading difficulty reflect that the reading level of simplified texts is lower than that of original ones? Whether this measure can distinguish between simplified texts and original ones?

8.2 Predict LocalNews2007 and LocalNews2008

8.2.1 Predictions by Models Trained on the Weekly Reader Alone

We have experimented with three standard classifiers (LIBSVM, Weka's Logistic Regression and SMO) and two meta ordinal classifiers (using Weka's Logistic Regression and SMO) in the previous chapter. Nearly all experiment results demonstrate consistently that, among all five types of classifiers, LIBSVM classifiers perform the best. Therefore we choose LIBSVM classification models trained with 11 feature subsets and 3 feature combinations to predict the LocalNews2008 data. In order to find out whether the models have the ability to differentiate between complex texts and their correspondingly simplified versions, we performed paired t-test on the model predictions of the original articles and their corresponding simplified texts. A p-value less than 0.05 indicates that the difference of predictions between complex articles and simplified ones is statistically significant. Table 8.1 presents the p-values obtained from the paired t-tests.

We see from the table that except models trained with OOV features, coreferential inference features and entity-grid-based features, all other models show strong ability to differentiate complex articles from the simplified ones ($p < 0.05$). It is especially notable that models generated by shallow features, syntactic features and combined discourse features (disc. comb.) demonstrate extraordinarily discriminative power ($p \leq 10^{-5}$). Among the four subsets of discourse features, models generated by entity density features and lexical-chain-based features showed strong ability to recognize

Table 8.1: p-values obtained from paired t-test on predictions of LocalNews2007 and LocalNews2008 by models trained on WeeklyReader alone.

Models	LocalNews2007	LocalNews2008
5gramWR	0.1039	0.0112
pos64	0.4344	0.0251
shallow	0.0187	0.00005
syntactic	0.0026	0.00005
disc. comb.	0.0026	0.00002
-entity	0.0249	0.0019
-coref	0.8321	0.0519
-lex	0.4679	0.0046
-egrid	0.7577	0.3705
all	0.0249	0.0041
GAOB	0.0249	0.0112

simplified texts from the original ones; by contrast, models generated by coreferential inference features and entity-grid-based features do not have such ability ($p > 0.05$). Models trained with a naive combination of all features as well as a feature subset obtained from groupwise-add-one-best approach also demonstrated significant discriminative power in differentiating between complex and simplified texts.

8.2.2 Limitation of Models

A closer look at the actual model predictions on the LocalNews2008 data reveals the limitation of the models trained solely on the Weekly Reader data. As the model predictions in Table 8.2 shows, the reading complexity of simplified articles, as indicated by the predicted grade levels, is considerably reduced compared with the original complex texts. But when we look at the predictions among complex articles, we see that, with only a few exceptions, nearly all models predicted the complexity of the 11 original articles with

Table 8.2: Predictions on NYLocalNews2008 by LIBSVM classifiers trained on WeeklyReader alone.

Code	5gWR	POS	shal	syn	enty	coref	lex	egrid	disc	all	GAOB	WkFS
ori.												
BT	5	5	5	5	5	5	5	5	5	5	5	5
BS	5	5	5	5	5	5	5	5	5	5	5	5
CO	5	5	4	4	5	5	4	5	5	5	5	5
DT	5	5	5	5	5	5	5	5	5	5	5	5
DR	5	3	5	5	5	5	5	5	5	5	4	5
HH	5	5	5	5	5	5	5	5	5	5	5	5
OS	5	5	5	5	5	5	5	5	5	5	5	5
PP	5	5	5	5	5	5	5	3	5	5	5	5
SC	5	5	4	5	5	5	5	5	5	5	5	5
ST	5	5	5	5	5	5	5	5	5	5	5	5
WH	5	5	5	5	5	5	5	5	5	5	5	5
sim.												
BT	5	5	4	4	5	5	5	5	5	5	5	5
BS	4	3	3	3	3	5	4	5	3	4	4	3
CO	4	3	3	3	3	5	4	5	3	4	4	4
DT	3	3	3	2	3	5	4	5	3	3	3	3
DR	3	5	5	3	5	5	5	2	4	5	4	4
HH	5	5	4	5	5	5	4	5	4	5	5	5
OS	3	3	3	3	3	3	5	5	3	3	3	3
PP	5	3	3	3	3	5	4	5	3	5	5	5
SC	4	3	3	3	3	3	3	3	3	3	3	3
ST	5	3	3	3	3	4	3	5	3	3	5	3
WH	5	5	4	4	5	4	3	4	4	4	4	4

grade 5. This observation points out that the complexity of these 11 articles clearly exceeds the highest grade level (grade 5) that these models can predict. As a matter of fact, these 11 original articles were hand-picked intentionally with varying levels of reading difficulty. Since the prediction models were trained on the Weekly Reader corpus, which covers only texts labeled with low level of grades ranging from 2 to 5 aiming at elementary students, they lack the ability to differentiate between texts with complexity higher than grade 5, resulting in flat predictions.

8.2.3 Model Improvement

To improve the models' ability to recognize texts with complexity higher than grade 5, we need to add extra training data into the Weekly Reader corpus with higher complexity. After carefully examining textual characteristics of texts labeled with grade 5 in the Weekly Reader corpus and those of the texts contained in the LocalNews2008 corpus, we manually selected 100 online articles from the New York Times that are distinguishably more complex than texts labeled with grade 5 in the Weekly Reader corpus. These 100 articles cover a variety of topics and are of similar or higher reading difficulty compared with the original articles in the LocalNews2008 corpus. Although the level of reading difficulty may vary within these 100 selected articles, it is safe to assume that their complexity is greater than grade 5 texts. To differentiate from grade 5 texts, we assign grade 7 to these 100 articles and refer to them as "NewYorkTimes100". It is to note that the assigned grade level 7 is intended more as an artificial marker to differentiate from grade 5 texts rather than a true and accurate grade level: we assume that the reading difficulty of the easiest articles in NewYorkTimes100 corpus is at least comparable to grade 7, many articles within the corpus may have reading difficulty higher than grade 7.

We mix "NewYorkTimes100" labeled with grade 7 and the Weekly Reader corpus together and retrain and test the models with the same feature sets using the same experiment design as we did earlier with the Weekly Reader corpus alone: we use LIBSVM to train and test prediction models with the selected feature sets, the model performance is evaluated by repeated (10 times) 10-fold cross-validation using mean and standard

Table 8.3: Comparison of accuracy generated by LIBSVM classifiers trained on WeeklyReader alone vs. on mixed WeeklyReader and NewYorkTimes100.

Feature Set	WR only	WR + NYT100	
	WR	WR	NYT100
5gramWR	68.38 ± 0.929	68.12 ± 1.160	91.30 ± 2.238
pos64	59.82 ± 1.235	59.10 ± 0.874	89.30 ± 2.410
syntactic	57.79 ± 1.023	56.99 ± 1.989	90.70 ± 2.830
shallow	56.04 ± 1.364	56.01 ± 0.986	80.30 ± 3.822
disc. comb.	60.50 ± 0.990	59.53 ± 1.152	89.60 ± 2.498
entity	59.63 ± 0.632	59.18 ± 0.663	85.40 ± 3.527
lex	45.86 ± 0.815	45.85 ± 0.612	65.80 ± 2.786
coref	40.93 ± 0.839	41.63 ± 1.000	28.60 ± 4.152
egrid	45.92 ± 1.155	43.36 ± 1.210	32.50 ± 4.653
All features	72.21 ± 0.821	71.86 ± 1.206	93.30 ± 1.552
GAOB	74.01 ± 0.847	73.84 ± 1.046	95.00 ± 2.280
WekaFS	70.06 ± 0.777	70.67 ± 0.879	91.70 ± 2.100

deviation of classification accuracy. Table 8.3 compares the performance differences of the models before and after adding NewYorkTimes100 data into the Weekly Reader corpus.

The evaluation of the new models trained with the mixed data (WeeklyReader plus NewYorkTimes100) focuses on the following two aspects: how does adding new data (NewYorkTimes100) affect the classification accuracy of texts in WeeklyReader corpus? And how do the models generalize to new data? We show related results in the last two columns of Table 8.3.

We see from the table that, in general, despite slight fluctuations, the overall classification accuracy generated by new models (trained with the mixed corpora) for texts ranging from grade 2 to 5 remains roughly the same compared with the models trained with the Weekly Reader corpus alone. This indicates that adding articles from the NewYorkTimes100 corpus with higher complexity to the Weekly Reader corpus does not have any significant

negative impact on the models' ability to classify grade 2 to 5 texts.

Moreover, we observe that models trained with the same selected feature sets on the mixed corpora are quite resilient in correctly recognizing new data with complexity higher than grade 5. At the individual level of feature subsets, models trained with 5gramWR generates the highest accuracy (91.3%). Models constructed with high level linguistic features, such as POS features, syntactic features and the combined discourse features, achieve close performance as well, generating accuracy around 90%. Compared with LM features and high level linguistic features, rudimentary features such as shallow features and out-of-vocabulary features (OOV) do not seem to be as powerful in predicting texts of high level complexity. Models trained with these two sets of features generate considerably lower accuracy.

Among the four subsets of discourse features, entity density features and lexical-chain-based features are much better at predicting texts of higher complexity than entity-grid-based features and coreferential inference features. While entity density features and lexical-chain-based features generate accuracy as high as 85.4% and 65.8%, entity-grid-based features and coreference features only generate 32.5% and 28.6% accuracy respectively.

We have previously discussed in Section 7.3.6 that, when tested only on the Weekly Reader corpus, among all four subsets of discourse features only entity density features have demonstrated strong discriminative power. The remaining three subsets of features do not seem to be very useful. We speculate that one of possible reasons for this is that the texts contained in the Weekly Reader corpus exhibit relatively low complexity, which results in sparse data captured by lexical-chain-based features, coreferential inference features and entity-grid-based features. Now after we have tested on the

mixed Weekly Reader and NewYorkTimes100 data – the latter contains texts with much higher complexity than the Weekly Reader – we observe considerable performance improvement made by lexical-chain-based features in classifying texts in the NewYorkTimes100 corpus with higher complexity; however, the accuracy generated by coreferential inference features and entity-grid-based features is decreased by more than 10% in classifying the NewYorkTimes100 texts compared with that of the Weekly Reader corpus. This observation can lead to the safe conclusion that entity density features and lexical-chain-based features are robust in classifying texts with varying level of reading difficulties. However, coreferential inference features and entity-grid-based features may not be very useful in predicting text complexity.

Aside from individual feature subsets, the two combinations of feature subsets generate even better accuracy for NewYorkTimes100 data. Models trained with a naive combination of all features generate 93.3% mean accuracy, and the combination obtained from groupwise-add-one-best approach achieves the best performance among all models, generating 95% accuracy.

In summary, we observe that adding NewYorkTimes100 texts into the training corpora does not affect classification accuracy for texts labeled with Grade 2 to 5 in the WeeklyReader corpus. Moreover, the improved models trained on the mixed corpora are robust in generalizing to texts in NewYorkTimes100, several models – those trained with 5gramWR, syntactic, all combined, GAOB and WekaFS feature – achieve accuracy above 90%. However, we are aware that news articles in NewYorkTimes100 are characteristically different from those labeled with elementary grades in WeeklyReader corpus in terms of genre, writing style, topics and complex-

ity. These factors may also contribute to the high recognition accuracy we observed in NewYorkTimes100 by our improved models.

8.2.4 Predictions by Improved Models

In this section, we use models trained with selected feature subsets on the mixed Weekly Reader and NewYorkTimes100 to predict the reading difficulty of texts in LocalNews2007 and LocalNews2008. Given same selected feature subsets, we compare the models trained on the mixed corpora and the models previously trained on the Weekly Reader corpus alone. We analyze whether the predictive ability of the newly trained models is improved in terms of differentiating between complex and simplified texts. We perform paired t-tests on predictions of the original and simplified articles by the new models. The obtained p-values indicate whether the prediction difference between the original articles and the simplified ones is statistically significant ($p \leq 0.05$). Table 8.4 presents the p-values based on the predictions by the new models and compares them with the p-values obtained from the predictions by the models trained with the Weekly Reader corpus alone.

From the table we see that, with only a few exceptions, adding texts with higher complexity into the Weekly Reader corpus significantly improves the models' ability in differentiating the original articles from the simplified ones. This observation applies both for LocalNews2007 and LocalNews2008. For LocalNews2007, by comparing the changes of p-values before and after adding the NewYorkTimes100 into the training data, we see that, for models that previously have demonstrated statistically signifi-

Table 8.4: P-values obtained from paired t-test on predictions of LocalNews2007 and LocalNews2008 by LIBSVM classifiers trained on WeeklyReader alone (WR) and mixed WeeklyReader and NewYorkTimes (WR-NYT).

	LocalNews2007		LocalNews2008	
Models	WR	WR_NYT	WR	WR_NYT
5gramWR	0.1039	0.0418	0.0112	0.0011
POS	0.4344	0.045	0.0251	0.0001
shallow	0.0187	0.0095	5.31e-05	3.591e-07
syntactic	0.0026	0.0012	5.31e-05	3.988e-06
disc. comb.	0.0026	8.498e-05	2.172e-05	8.137e-06
entity	0.0249	0.0086	0.0019	0.0002
coref	0.8321	0.0811	0.0519	0.0538
lex	0.4679	0.4679	0.0046	0.001
egrid	0.7577	0.7577	0.3705	0.3705
all	0.0249	NA	0.0041	0.0027
GAOB	0.0249	0.0063	0.0112	0.0026

cant ability in differentiating complex and simplified texts ($p < 0.05$) when trained on the Weekly Reader corpus alone, their differentiating ability is much further strengthened after adding NewYorkTimes100, which is indicated by significantly decreased p-values. We also observe that, before adding the NewYorkTimes100 corpus into the training data, models built with LM features (5gramWR), POS features and coreferential inference features do not exhibit the ability to differentiate between complex and simplified articles, because the corresponding p-values are greater than 0.05. After mixing the NewYorkTimes100 into the training data, the differentiating ability of models built with LM features and POS features becomes statistically significant ($p < 0.05$); although the differentiation ability of the model built with coreferential inference features is still not significant ($p = 0.0811$), the p-value obtained from the predictions by the new model is

considerably decreased. However, the discriminative power of models built with lexical-chain-based features and entity-grid-based features does not seem to be impacted by the change of training data, which in both cases remains statistically insignificant in differentiating between original and simplified texts.

Similar observations apply for the LocalNews2008. As discussed in Section 8.2.4, for LocalNews2008, most models – except for those trained with OOV features, coreference-based features and entity-grid-based features – have demonstrated significant differentiating ability when trained on the Weekly Reader corpus alone. After adding NewYorkTimes100 into training data, we observe that the p-values obtained from predictions by new models trained on the mixed corpora become much smaller, indicating that these models' discriminative power is further strengthened compared with the models trained with the same feature subsets on the Weekly Reader corpus alone.

When we take a closer look at the predictions on the LocalNews2008 by various models trained on the mixed corpora, as indicated by Table 8.5, we have a better understanding of the decreasing p-values that we observe after adding NewYorkTimes100 into the training data. Recall that the primary motivation for us to include more texts with complexity higher than grade 5 is not because the models trained on the Weekly Reader corpus alone can not differentiate between complex and simplified texts in LocalNews2008, as a matter of fact, most of the models have demonstrated significant differentiating ability already. The real reason lies in that old models trained on the Weekly Reader corpus alone are limited at predicting reading difficulty as high as grade 5 and can not go further up, resulting in flat predictions

Table 8.5: Predictions on LocalNews2008 by LIBSVM classifiers trained on mixed WeeklyReader and NewYorkTimes100.

Code	5gWR	POS	shal	syn	enty	coref	lex	egrid	disc	all	GAOB	WkFS
ori.												
BT	7	7	7	7	7	5	7	5	7	7	7	7
BS	7	5	5	5	5	5	5	5	5	7	7	5
CO	5	5	4	4	5	4	4	5	5	5	5	5
DT	5	5	5	5	5	7	5	5	5	7	5	5
DR	5	5	7	4	5	5	5	5	5	5	5	5
HH	5	7	7	7	7	5	7	5	7	5	5	7
OS	5	5	5	5	5	4	5	5	5	5	5	5
PP	5	5	5	5	5	5	5	3	5	5	5	5
SC	7	5	5	5	5	4	5	5	5	5	5	5
ST	7	7	5	5	7	5	5	5	7	5	5	5
WH	5	7	5	5	5	4	5	5	5	5	5	5
sim.												
BT	5	5	5	4	5	5	5	5	5	5	5	5
BS	4	3	3	3	3	5	4	5	3	4	3	3
CO	4	3	3	3	3	4	4	5	3	4	4	4
DT	3	3	3	2	3	4	4	5	3	3	3	3
DR	3	5	5	3	5	5	5	2	4	5	5	4
HH	5	5	4	5	5	4	4	5	4	5	5	5
OS	3	3	3	3	3	3	5	5	3	3	3	3
PP	5	3	3	3	3	3	4	5	3	5	5	5
SC	4	3	3	3	3	3	3	3	3	3	3	3
ST	5	2	3	3	3	4	3	5	3	3	3	3
WH	5	5	4	4	5	5	3	4	4	4	4	4

of grade 5 for almost all original articles contained in LocalNews2008. We see in Table 8.5 that, after adding the NewYorkTimes100 corpus into the training data, the newly trained models are able to dramatically change the previously flat grade 5 predictions for the 11 original articles in LocalNews2008, which are typically placed into at least two to three different levels of reading difficulty as indicated by grade 4, 5 and 7. Compared with the original articles, the reading difficulty of most simplified texts is considerably reduced as indicated by lower grades assigned.

In summary, when tested on unseen user-specific LocalNews2007 and LocalNews2008, we see expected and desired effects made by the new models retrained on the mixed Weekly Reader and NewYorkTimes100 corpora:

- The new models are able to classify unseen original articles with varying level of reading difficulty;
- The model predictions indicated by lower grades demonstrate that the level of reading difficulty of simplified texts is considerably reduced compared with their corresponding original articles;
- Compared with the old models trained on the Weekly Reader corpus alone, the new models have stronger ability in differentiating original articles from their manually simplified versions. Such differentiating ability observed in most of the new models is statistically significant.

We have observed major improvement of the new models trained on the mix corpora over the old models trained on the Weekly Reader corpus alone. For this reason, in the following section, we will use the predictions by these new models to analyze relations between model predictions, expert ratings and text complexity generated from users' comprehension performance.

Because expert rating is expensive in terms of time and reliable human resources consumed, we have only LocalNews2008 rated by three experts. In order to construct a complete set of analysis on correlations between machine predictions, expert ratings and user comprehension ability, our analysis and discussions below will center on LocalNews2008 alone.

8.3 A Correlation Study between Expert Ratings and Model Predictions

In this section, we analyze how model predictions are correlated with expert ratings. The testing corpus of the study is LocalNews2008, which contains 11 original local news articles and their corresponding abridged versions manually simplified by a human expert (see Section 5.2.2 for more details of the corpus). We use the new models trained with various feature subsets on the mixed Weekly Reader and NewYorkTimes100 corpora to predict the reading complexity of these 11 pair of articles in terms of grade levels. These grade levels range from grade 2 to 5, plus an artificial marker grade 7, which is used to indicate text complexity much higher than grade 5 and above. The predicted grade levels are used below to study how well machine predictions are correlated with human ratings.

We recruited three experts to rate the reading difficulty of the 11 pairs of articles in LocalNews2008. Expert A and expert B both have strong background in linguistics with PhD degrees. Expert A has work experience in language education. Expert B shares deep understanding in NLP development. Expert C is a graduate student in Psychology who had work experience with individuals with intellectual disabilities (ID). All three experts are well suited to annotate text complexity for LocalNews2008, a corpus tailored for the design and implementation of an NLP system intended for people with ID. We asked each expert to rate the level of reading difficulty for all texts in LocalNews2008 using a given number scale.

We gave all experts background knowledge on how LocalNews2008 is created. They were informed that there are 11 pairs of original and

simplified articles, each pair covers the same topic. The experts were not told the label of each article, whether it is simplified of original. To affect experts' judgment as little as possible, the annotation guidelines we gave them are rather open-ended: they should rate the reading difficulty of each article based on their own judgment. The only requirement is that they need to read all 22 articles to have general knowledge of them before rating.

We started with expert C, who was initially given a 10-point scale to rate, with 1 being the easiest and 10 being the most difficult to read. We later found that a number scale of 10 makes the rating challenging because it is difficulty to decide between too many numbers. To make the task easier, we later decided to reduce the 10-point scale to 5-point scale for expert A and B, with 1 being the easiest to read and 5 being the most difficult to read. Since we are interested in correlation analysis, although expert A and B used a rating scale different from expert C, this does not impact the correlation results. Table 8.6 presents the ratings by the three experts. The number scale used for expert rating is independent from grade levels used by machine predictions. Since we are interested in comparing how machine predictions are related to human ratings, various independent measures of reading difficulty are allowed in correlation study for broad comparisons.

Based on the experts' ratings, we first analyze whether their ratings can differentiate complex original articles from their corresponding simplified versions. We performed paired t-test on each set of the ratings and present the obtained p-values in Table 8.7. These p-values indicate that the ratings of all three experts are able to differentiate original articles from the simplified ones with high level confidence (see p-values in Table 8.7).

We bear in mind that human annotation is itself a complex task (Pe-

Table 8.6: Expert ratings using 5-point scale, with 5 being the most difficult to read, 1 being the easiest to read.

Code	Expert A	Expert B	Expert C
original			
BT	5	5	8
BS	4	4	8
CO	2	3	4
DT	5	5	8
DR	4	3	7
HH	5	5	7
OS	3	3	3
PP	3	3	7
SC	3	3	3
ST	4	3	5
WH	4	3	2
simplified			
BT	2	3	2
BS	2	3	6
CO	1	2	3
DT	2	4	5
DR	1	1	3
HH	2	3	4
OS	1	2	1
PP	1	2	2
SC	1	2	1
ST	1	2	3
WH	1	2	1

tersen and Ostendorf, 2009; Siddharthan, 2004). The criteria people use for annotation may differ from person to person, and those criteria can be influenced by various factors such as background, work experience, interests, specific knowledge relevant to the assigned task, etc. Before we compare the relations between human ratings and model predictions, we first analyze how the ratings among the three experts are correlated with each other. We compute the correlations and present Pearson's R in Table 8.8. We find that

Table 8.7: P-values obtained from paired t-test based on expert ratings. The p-values indicate that the expert ratings can differentiate between original and simplified texts with confidence ($p < 0.05$).

Expert A	3.313e-07
Expert B	3.988e-06
Expert C	0.0001643

Table 8.8: Correlations (Pearson's R) between expert ratings.

	Expert B	Expert C
Expert A	0.8490	0.7614
Expert B		0.7703

the ratings by two linguists (expert A and B) have the strongest correlation (0.85). The correlations between each of the linguists and the psychology graduate student who had work experience with people with ID are lower at the similar range (0.76 and 0.77).

We then compute and compare the correlations between machine predictions and expert ratings in Table 8.9. Among the three sets of human ratings, we find that the ratings by expert A show the strongest correlation with model predictions. The correlation level of expert B with model predictions is in general below expert A and above expert C, with a few exceptions: the correlations between expert C and predictions by models generated with shallow features and all features combined are stronger than that with expert B.

In our research, we pay special attention to discourse features that may be particularly challenging for adult readers with ID because of their unique characteristics. It would be interesting to see if this aspect is appropriately reflected in the ratings by expert C who had work experience with adults

Table 8.9: Correlations (Pearson's R) between machine predictions and expert ratings.

Models	Expert A	Expert B	Expert C
5gramWR	0.6032	0.4403	0.3913
pos64	0.7632	0.5074	0.3756
shallow	0.8126	0.5667	0.5796
syn	0.8197	0.6817	0.5231
disc. comb.	0.8369	0.6333	0.5356
entity	0.7495	0.5363	0.4694
coref	0.6077	0.5374	0.6823
lex	0.7644	0.6372	0.5576
egrid	0.3337	0.4279	0.1808
all	0.7327	0.5976	0.6435
addOneBest	0.6486	0.4706	0.5148

with ID. When we compare the correlations between the predictions by models generated with 5 sets of different discourse features and the ratings by three experts, we find that ratings by expert C produce a stronger correlation (0.68) with predictions by the model generated with coreferential inference features that expert A (0.61) and B (0.54). On all other four accounts, machine predictions are more closely correlated with the ratings by the two linguists than that of the psychology student. This observation seems to be supported by the statistics in Table 8.9 in general.

8.4 Reading Difficulty in Adults with ID: Analysis with a Hierarchical Latent Trait Model

This section is based, often verbatim, on joint work with Martin Jansche and Matt Huenerfauth (Jansche et al., 2010)

8.4.1 Introduction

In this thesis, our central focus is to design, implement and evaluate an automatic text readability assessment tool that predicts reading difficulty of a given text in terms of grade levels. We have developed and evaluated this tool in terms of a corpus of elementary texts annotated with grade levels (Feng et al., 2010). This tool is designed and developed not only just for applications targeted to general audience, we also intend to use it as a sub-component for an envisioned text simplification system that is designed specifically for adults readers with intellectual disabilities (ID). In order to adapt, evaluate, and refine our grade-level assessment tool for adult readers with ID, we have to deal with a key issue: how to determine the difficulty particular texts pose for adult readers with ID.

In Section 5.2, we described the creation of a user specific corpus LocalNews2008. This corpus is user-specific in a sense that (i) the simplified texts in LocalNews2008 were adapted specifically for adults readers with ID; (ii) and all texts in LocalNews2008 were read by participants with ID and evaluated by their actual comprehension performance. The experiment conducted to create LocalNews2008 carried multiple goals with itself, one of them was to investigate which of the three question formats – multiple

choices in text, clip art questions and "yes/no" questions in text – are more likely to solicit valid feedback from adult readers with ID. The analysis and results regarding this goal have been published in our recent work (Huenerfauth et al., 2009). Another important goal of the experiment is to gather participants' responses to comprehension questions, so that intrinsic difficulty of texts can be inferred for adult readers with ID. In Huenerfauth et al. (2009), we have not addressed explicitly as to how to achieve this goal in an appropriate way. Instead, we focused in that paper more on investigating which of the three question types are more likely to solicit readers' responses that can differentiate between original and simplified articles. Nevertheless, in order to solve this problem, a measure has to be chosen to indicate the level of reading difficulty of each article for test participants with ID. We dealt with this problem implicitly as follows: for each article, we gathered all responses from the test participants using a particular question type, we then scored it by the percentage of correct responses. It is assumed that the lower the percentage of correct responses is, the harder the article is for the participants to read, and vice versa.

The weakness of the above described approach lies in that it did not take the individual differences of the test participants' reading abilities into account. In this thesis, we improve our previous approach by applying a statistical model that is suitable for the task. More specifically, we apply a hierarchical latent trait model to a particular set of variables from the data that we gathered in the experiment as described in Huenerfauth et al. (2009). We use this model to infer participants' ability levels and the intrinsic difficulty of particular texts presented to them. We then use a correlation study to analyze relations between the inferred text difficulty for adult readers

with ID, grade levels predicted by our automatic readability assessment tool and expert ratings. The analysis of the study will be discussed in detail in Section 8.5

8.4.2 Experiment and Data

The reading material for the experiment conducted in Huenerfauth et al. (2009) consists of 11 original news-wire articles and 11 corresponding simplified versions. The original articles were selected from local news to ensure familiarity. The simplified versions were manually adapted by a human expert specifically for adult readers with ID. Participants were asked to read the articles and answer 6 basic factual comprehension questions for each article. For each simplified article, the questions were identical to the corresponding original version.

We recruited 20 adults with ID to participate in the experiment. Each participant was assigned 11 articles to read, some in their original and some in their simplified version. We made sure that no test participant saw both the original and simplified version of an article. The order of the articles and questions was randomized for each participant. The assignment of conditions – original vs. simplified version – was randomized with a margin constraint to ensure that all articles under both conditions would be presented to an equal number of participants.

Many more details of the experiment can be found in Huenerfauth et al. (2009). Here we concentrate on the participants' responses to the comprehension questions and what they tell us about the participants' abilities and the intrinsic difficulty of each article. Each observation recorded

in the experiment consists of the participant number $s \in \{1, \ldots, 20\}$, the article topic $a \in \{1, \ldots, 11\}$, the version of the article $v \in \{\text{com}, \text{sim}\}$ (complex/original vs. simplified), the question number $q \in \{1, \ldots, 6\}$, and an indicator $y^v_{s,a,q} \in \{0, 1\}$ of whether the participant's response to the comprehension question was correct (1) or incorrect (0). The total number of $1320 = 20 \times 11 \times 6$ observations is a consequence of the design where each of the 20 participants read 11 texts and answered 6 questions per text. Out of these 1320 observations, a small number of responses were not available because the experimenter ran out of time. We ended up with a total of 1296 usable observations.

8.4.3 Model and Computation

The above presentation of the experiment in terms of stimuli and question responses that are either correct or incorrect immediately suggests an analysis based on an item-response or latent trait model. A direct application of the Rasch model to our data assumes a univariate latent trait which expresses both the abilities α of participants and the difficulty θ of question items (see e.g. §14.3 of Gelman and Hill (2007)).

$$\Pr(y^v_{s,a,q} = 1) = \text{logit}^{-1}(\alpha_s - \theta_{a,v,q})$$
$$\alpha_s \sim \text{Normal}(\mu_\alpha, \sigma^2_\alpha) \qquad \theta_{a,v,q} \sim \text{Normal}(\mu_\theta, \sigma^2_\theta)$$

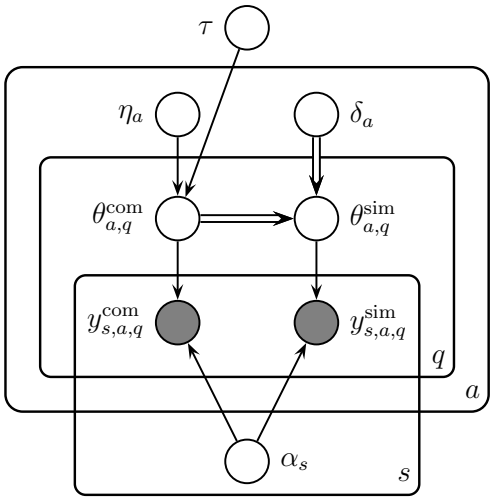
This model only captures some of the hierarchical structure inherent in the experimental design: each participant s is given a latent ability parameter α_s and each question item has a latent difficulty parameter. To the extent that a participant's ability exceeds an item's difficulty, the participant is

more likely to answer the item correctly. More precisely, the inverse of the logit link function transforms the difference between ability and difficulty to a probability, where a difference of zero means the participant has equal chance of answering the question correctly or incorrectly.

We now enrich this basic model with additional hierarchical structure to capture two additional aspects of the experimental design. First, items are no longer independent, but are grouped by article and condition. Second, our model will reflect the fact that the set of comprehension questions for each article was identical for the complex and simplified versions. We express this hierarchical structure in terms of additional latent variables in our model. Specifically, we assume:

- For each article a, a latent difficulty η_a. This can be thought of as the intrinsic difficulty of the original (complex) article.
- For each article a, a latent simplification amount δ_a. This expresses the reduction in difficulty when going from the original (complex) article to its simplified variant.
- For each article a and each associated question q, latent item difficulty $\theta^{\text{com}}_{a,q}$ and $\theta^{\text{sim}}_{a,q}$ for the complex and simplified versions, respectively, of the article.
- For each participant s, a latent ability α_s, as above in the Rasch model.

The full model then has the form shown in Figure 8.1. Here we follow the conventions of the BUGS language (Thomas, 2006) and assume that normal distributions are parametrized in terms of mean and precision. To save space, we write N for a normal distribution and g for the logit link function.



$$\tau \sim \text{Gamma}(0.001, 0.001)$$

$$\eta_a \sim \text{N}(0, 10^{-6}); \quad \delta_a \sim \text{N}(0, 10^{-6})$$

$$\theta^{\text{com}}_{a,q} \sim \text{N}(\eta_a, \tau); \quad \theta^{\text{sim}}_{a,q} = \theta^{\text{com}}_{a,q} - \delta_a$$

$$y^{v}_{s,a,q} \sim \text{Bernoulli}(g^{-1}(\alpha_s - \theta^{v}_{a,q}))$$

$$\alpha_s \sim \text{N}(0, 1)$$

Figure 8.1: Hierarchical latent trait model.

To ensure identifiability, we assume that the mean of the abilities α is known and fixed at zero.

The key property of our model lies in the way it imposes structure on item-level difficulty. We assume that each original/complex article has an inherent difficulty η_a. The item-level difficulty $\theta^{\text{com}}_{a,q}$ for the original version of the article are drawn from a normal distribution with mean η_a. For the simplified version of the article, we ask the exact same questions, hence we assume that the item-level difficulty $\theta^{\text{sim}}_{a,q}$ of each question is reduced by the same article-level amount δ_a, representing the reduction in difficulty due to the simplification of the article. The observed responses are assumed to be generated by a standard Rasch model that combines participant abilities and item difficulty.

The model was formally specified in the BUGS language. Computations were carried out by Gibbs sampling using the JAGS software package (Plummer, 2010), an open-source implementation very similar to classic

BUGS. In particular, we used 3 parallel Markov chains that ran for 10,000 iterations each, which took about three minutes on a Linux workstation with a 3.16 GHz Intel Core2 CPU. The first 5,000 iterations in each chain were discarded, after checking for approximate convergence. We monitored all unobserved variables and observed that the potential scale reduction factor $\hat{R}$ was less than 1.03 in all cases, indicating convergence (see e.g. §16 of Gelman and Hill (2007)). The last 5,000 iterations in each chain were recorded and analyzed using the CODA (Plummer et al., 2006) package for R. We checked the fit of the model by comparing the observed mean correct responses per article and per participant against the corresponding expected numbers under the model and found them to be uniformly close.

8.4.4 Results

For our work on text simplification, we were primarily interested in the quantities η_a, the difficulty of the 11 original articles, and $\eta_a - \delta_a$, which we take as the difficulty of the simplified articles. We computed marginal posterior means of each of these 22 quantities and used these as point estimates for subsequent computations. Table 8.10 presents the final results of the inferred intrinsic difficulty of the 11 original articles as well as the 11 corresponding simplified versions for test participants with ID.

Table 8.10: Intrinsic difficulty of texts inferred from test participants with ID. η_a represents the difficulty of the original articles, δ_a represents the amount of reduction in difficulty due to simplification process performed on the original articles, and $\eta_a - \delta_a$ represents the difficulty of simplified articles.

Code	η_a	δ_a	$\eta_a - \delta_a$
BT	0.009186	0.006919	0.002267
BS	0.422887	0.346392	0.076495
CO	0.046767	0.910710	−0.863943
DT	0.084888	0.591493	−0.506605
DR	−0.138620	−0.306111	0.167491
HH	−0.739079	−0.667760	−0.071319
OS	0.019599	0.510112	−0.490513
PP	−0.311692	−0.460704	0.149012
SC	−0.098520	0.735395	−0.833915
ST	0.747024	0.836578	−0.089554
WH	0.244301	0.067222	0.177079

8.5 Relations between Inferred Text Difficulty for Adults with ID, Expert Ratings and Model Predictions

Using the hierarchical latent trait model as described in the above section, we have obtained inferred text difficulty of LocalNews2008 for adult readers with ID. So far we have three different measures of reading difficulty from different perspectives: grade levels predicted by machine learning models, ratings by experts using a number scale, and statistical results inferred from actual comprehension responses from adults with ID, which take differences in individual reading ability into account. These three kinds of measures of reading difficulty are independent from each other. In Section 8.3, we

have analyzed the relations between expert ratings and machine predictions based on texts in LocalNews2008. In this section, with the addition of text difficulty inferred for adults with ID, we continue analyzing the relations among the three different measures and present summative results.

In addition to results presented in Section 8.3, we focus on investigating the following:

1. whether the inferred text difficulty for adults with ID can differentiate the original articles from the simplified ones;
2. how well is the inferred text difficulty correlated with expert ratings;
3. and how well is the inferred text difficulty correlated with various model predictions.

We have stated earlier that our assumption on paired original/simplified corpora, such as LocalNews2008, is that simplified texts should be easier to read. In other words, the level of reading difficulty of simplified texts should be lower than that of the original texts. In results presented in Section 8.2.4 (see Table 8.5) and 8.3 (see Table 8.6), both model predictions and expert ratings corroborate our assumption, moreover, both model predictions and expert rating can differentiate between original texts and simplified texts with high confidence. In order to find out whether the same holds for the inferred text difficulty for adults with ID, we performed a paired t-test on the inferred results and obtained a p-value of 0.1845, which indicates that, contrary to what we have observed from model predictions and expert ratings, the intrinsic text difficulties inferred from the comprehension responses by adult readers with ID can not differentiate between original texts and their simplified versions.

Table 8.11: Correlations (Pearson's R) between expert ratings and users' ability.

	Expert B	Expert C	Readers w/ID
Expert A	0.8490	0.7614	0.2594
Expert B		0.7703	0.0272
Expert C			0.1408

To answer the second question, we computed the correlations between the inferred text difficulty and each set of expert ratings and updated Table 8.8 with obtained Pearson's R. Table 8.11 presents summative results. We find that, while the Pearson's R indicates that there exist strong positive relations among the ratings by three experts, the inferred text difficulty for adults with ID does not seem to be closely related to any set of expert ratings, which can be seen from the comparatively much lower Pearson's R (0.26, 0.02, and 0.14 respectively). Even with expert C who has specific work experience with people with ID, the correlation is still very weak (0.14).

To answer the third question, we used models trained with various feature subsets on the mixed Weekly Reader and NewYorkTimes100 corpora Feng et al. (2010) to predict each of the 22 texts (original and simplified articles) in LocalNews2008 and computed the correlation (Pearon's R) between predicted grade levels and text difficulty inferred by the hierarchical latent trait model. Table 8.12 presents summative results in contrast to correlations between expert ratings and model prediction discussed in Section 8.3. We found that, with only one exception, the correlations between model predictions and inferred text difficulty are much weaker compared with correlations between expert ratings and model predictions. Compared with ratings by expert B, C and inferred text difficulty for adult readers with

Table 8.12: Correlations of machine predictions with expert ratings and user comprehension ability.

Models	Expert A	Expert B	Expert C	Readers w/ID
5gramWR	0.6032	0.4403	0.3913	0.5265
pos64	0.7632	0.5074	0.3756	0.4078
shallow	0.8126	0.5667	0.5796	0.1994
syn	0.8197	0.6817	0.5231	0.2010
disc. comb.	0.8369	0.6333	0.5356	0.3469
entity	0.7495	0.5363	0.4694	0.3978
coref	0.6077	0.5374	0.6823	0.3604
lex	0.7644	0.6372	0.5576	0.0543
egrid	0.3337	0.4279	0.1808	0.1248
all	0.7327	0.5976	0.6435	0.5040
GAOB	0.6486	0.4706	0.5148	0.4747

ID, ratings by expert A demonstrated the strongest positive correlations with 11 out of 12 sets of model predictions. In one exceptional case, we found a correlation of 0.53 between inferred text difficulty and predictions by language-modeling-based model (5gramWR), which is stronger than the correlations between model predictions and expert B (0.44) and expert C (0.39), but still lower than that with expert A (0.60).

8.6 Conclusions

We took models trained on labeled corpora and evaluated them on unseen data from LocalNews2007 and LocalNews2008, both consisting of paired original/simplified news articles. We observed that predictions on these data sets consistently show that the level of reading difficulty of simplified articles are lower than their corresponding original versions, which strengthens our hypothesis that simplified texts should be easier to read.

We further improved our models with additional training data from NewYorkTimes100 and tested them again on LocalNews2007 and LocalNews2008. The prediction results show that the models can generalize successfully (above 90% accuracy) when encountering unseen texts with reading difficulty higher than grade 5. This indicates that our models can be generalized with more grade levels predictions when suitable training data is available.

To compare our model predictions with other measures of reading difficulty, we have three experts rated all 22 texts in LocalNews2008. In addition, we conducted a reading experiment with adult readers with ID and have 20 test participants read articles in LocalNews2008 and answered comprehension questions. Based on the readers' actual responses to the comprehension questions, we used a hierarchical latent trait model to infer text difficulty for adults with ID. To investigate relations between model predictions, expert ratings and text difficulty inferred from the reading ability of adult readers with ID, we conducted a comprehensive correlation study and have the following observations:

- Model predictions and expert ratings are all able to differentiate between the original and simplified articles contained in LocalNews2008 with significant confidence. However, inferred text difficulty for adult readers with ID does not demonstrate such differentiating ability in recognizing original and simplified articles.

- Statistics from our study show that there exist strong positive correlations between model predictions and expert ratings. However, we

observed that the inferred text difficulty for adult readers with ID is not closely correlated with either expert ratings or model predictions.

The strong correlations observed between model predictions and expert ratings indicate two things: first, different measures of reading difficulty of texts can be translated within each other; second, model predictions comparable to human expert judgment demonstrate that our grade-level-based automatic text readability assessment tool can generalize reliably to unseen data. Although our current tool is limited by the fact that our training corpus is small and has only four grade levels at the lower range, we have successfully improved our models by mixing additional texts with higher reading difficulty into the training data. The fact that the improved models can recognize newly added texts with above 90% accuracy is promising and encouraging. This points to several ways in improving the tool in the future. For instance, the tool can be trained to predict a full range of grade levels when access to appropriate corpora becomes available. We are aware that access to corpora annotated with full range of grade levels for prediction task is limited. Alternatively, since we have observed strong correlations between grade level predictions and other measures of reading difficulty, such as number-scaled expert ratings, we could use texts labeled with coarsely defined reading difficulty, such as low, medium and high, to build and evaluate our readability assessment tool. Besides small text corpora with grade-level annotations that are available, much larger amounts of data from educational testing could potentially be harnessed for this purpose.

Our automatic text readability assessment tool is built with corpora labeled with reading difficulty annotated for general audience. Adapting,

evaluating, and refining our assessment tool for adult readers with ID requires an independent determination of the difficulty particular texts pose for this group of readers. This task is complicated by the fact that there are no large-scale text corpora annotated with difficulty levels for adult readers with ID. In our experiment, we used both expert ratings and direct comprehension feedback from adults with ID to estimate text difficulty. The results of analysis show that, in determining reading difficulty for adult readers with ID, expert ratings seem to be more reliable than text difficulty inferred from test participants' reading ability. As we have discussed above, this conclusion is supported by several forms of evidence we observed in our analysis. Expert ratings and model predictions all show that the level of reading difficulty of original texts is considerably reduced after simplification process, which corroborates our assumption that simplified texts should be easier to read. Moreover, expert ratings, similar to machine predictions, have strong differentiating ability to tell simplified texts apart from their original versions. Inferred text difficulty for adults with ID can not differentiate simplified texts from the original ones, the reason could lie in that the original texts were not simplified to adequate level to meet their low reading proficiency. In addition, results from correlation analysis show that there exists strong correlations between expert ratings and model predictions; even though the criteria the three experts used to annotate the reading difficulty of texts in LocalNews2008 may very likely differ from each other, their ratings are strongly correlated with each other as well. By contrast, the inferred text difficulty for adults with ID is neither strongly correlated with expert ratings nor model predictions. All these observations point out that, in future work, in order to create more suitable corpora for

readability related research targeted to adults with ID, it would be more practical and reliable to have experts annotate text difficulty.

However, direct feedback from target users is valuable to evaluate any applications that are intended for specific audience. Our reading experiment is a good example to validate whether adult readers with ID can benefit from simplified texts. Although our statistical analysis on gathered experiment data hints that simplified texts are not significantly beneficial for adult readers with ID, the reason for this could lie in several factors. The immediate one to begin with, as we have noticed during the reading experiment, is that the reading proficiency of some test participants is extremely low, such that even simplified texts are beyond their comprehension level. To design and develop a text simplification system that adult readers can really benefit from, further research needs to be done to assess target population's reading ability and determine the amount of simplification needed to meet target users' reading proficiency. The widely used computerized adaptive testing in educational and other settings is suggestive for future research. The hierarchical latent trait model we developed is generally useful – even without the added complexity of simplified texts – for inferring article-level difficulty from repeated observations based target readers' actual response to multiple comprehension questions per article. This deserves to become as ubiquitous in research on adults with ID as it already is in educational testing and other settings.

Chapter 9

Conclusions & Future Work

9.1 Summary and Conclusions

In this thesis, we present research on developing an automatic readability assessment tool with high performance in detecting and predicting reading difficulty of texts indexed by grade levels.

Our research is primarily motivated by the lack of efficient and accurate automatic evaluation tools for existing and envisioned text simplification systems. Many existing text simplification systems still rely on laborious human judgment or traditional readability metrics such as the Flesch-Kincaid Grade Level formula, which have been demonstrated to be highly unreliable by this study and several recent work in the field. We believe an automatic text readability assessment tool that accurately models reading difficulty of texts for target readers is essential to the development of a text simplification system that is envisioned to simplify texts from discourse levels, particularly for adult readers whose reading proficiency is limited due to various degrees of language impairments.

We approached readability from a text comprehension point of view, which emphasizes that the goal of reading is to actively construct a coherent mental representation of a text by the reader. According to established theories and frameworks on reading comprehension, reading difficulties often arise from discourse level comprehension rather than lexical tokenization and sentence processing. Our research paid special attention to text properties that play important roles for discourse level comprehension. We deployed various NLP techniques and implemented a set of discourse features, which include entity density features, lexical chain features, coreferential coherence features and entity grid features. In addition, we enriched previous work by creating new features at several linguistic levels.

We combined NLP and machine learning techniques to build an automatic readability assessment tool with high performance. We examined the usefulness of features within and across several linguistic levels for predicting text readability in terms of assigning texts to elementary school grade levels. We built various classifiers with subsets of these features and evaluated them using repeated 10-fold cross-validation on WeeklyReader, a corpus containing texts with grade levels ranging from Grade 2 to 5. Based on our experiment results, we made the following key observations and conclusions:

- Among all individual feature subsets examined, combined perplexity features obtained from LMs trained on WeeklyReader corpus demonstrate the most significant discriminative power, generating 68.38% accuracy.

In contrast with this, LMs trained on unrelated corpora appear to be least effective among all features examined.

Our experiment results support Schwarm and Ostendorf's (2005) claim that LMs trained using information gain (IG) approach outperform LMs trained on POS sequences. However, when trained on WeeklyReader directly, LMs trained with generic word labels or paired word/POS sequences achieved similar classification accuracy to the IG approach, while avoiding the complicated feature selection of the IG approach.

- POS features, in particular nouns, exhibit significant predictive power. The high discriminative power of nouns in turn explains the good performance of entity-density features, based primarily on nouns. Prepositions also demonstrate strong discriminative power.

  In general, our selected POS features appear to be more correlated to text complexity than syntactic features, shallow features and most discourse features.

- For parsed syntactic features, we found that verb phrases appear to be more closely correlated with text complexity than noun phrases and prepositional phrases. We also find that, compared with POS-based noun features and preposition features, the discriminative power of noun phrases and prepositional phrases decreases considerably.

  While SBARs are commonly perceived as good predictors for syntactic complexity, they did not prove very useful for predicting grade levels of texts in this study. The reason could lie in sparse distribution of SBARs in texts of lower grades.

- We re-assessed the usefulness of shallow features using advanced machine learning techniques. We identified that average sentence length has dominating predictive power over all other lexical or syllable-based features.

  We found that the effectiveness of shallow features is severely limited by the oversimplified linear functions that are frequently used by traditional readability metrics, such as the Flesch-Kincaid Grade Level formula. Our experiment results show that Flesch-Kincaid scores perform extremely poorly when used directly to measure text difficulty. However, when we used them as a feature to train classifiers using advanced machine learning techniques, they demonstrate significant discriminative power.

- Our hypotheses regarding the usefulness of discourse features were not uniformly supported by the data. While entity density features demonstrate significant discriminative power comparable to combined POS features, we found that lexical chain features, coreferential inference features and entity grid features do not appear to be very useful in detecting and predicting reading difficulty in elementary grade levels. When evaluated on texts with higher complexity, such as those contained in NewYorkTimes100 corpus, we found that the predictive power of lexical chain features increased considerably, coreferential inference features and entity grid features appear to be the least useful.

- Our research has led to an automatic readability assessment tool with high performance. Our best model trained with a judicious

combination of features generates 74% accuracy, outperforming the current state of art by nearly 11%.

Our automatic text readability assessment tool was built on corpora with reading difficulty of texts annotated for general audiences. As part of the primary motivation of this study, this tool is intended to be used as a subcomponent for an envisioned text simplification system for adult readers with mild intellectual disabilities (MID). Adapting, evaluating, and refining our assessment tool for this purpose requires an independent determination of the difficulty particular texts pose for adult readers with ID.

In order to evaluate how our automatic readability assessment tool models reading difficulty of texts for adult readers with ID, we conducted a pilot study followed by a reading experiment with adults with MID to create two user-specific corpora. For each text in the corpora, we asked test participants a set of comprehension questions and collected their responses to these questions. We then developed a hierarchical latent trait model that captures major aspects of the experimental design. Using this model, we inferred reading difficult of texts for test participants with MID based on their actual reading ability. We also had the same set of texts rated by three experts to establish another independent measure of reading difficulty. We trained classifiers on WeeklyReader and NewYorkTimes100 to predict reading difficulty of texts obtained from the user studies.

Based on these three types of independent measures of reading difficulty, we conducted a comprehensive correlation study to examine the relations between grade level predictions, expert ratings and readers' actual comprehension ability. We found that both our model predictions and expert

ratings are able to differentiate simplified texts from original ones with high confidence, which corroborates our assumption that simplified texts should be easier to read than the original ones. However, the intrinsic text difficulty inferred from test participants' actual reading ability was not able to recognize the differences between simplified and original texts, which indicates that comprehension feedback solicited from adults with MID, though valuable in providing insights to their actual reading proficiency, is not reliable in annotating reading difficulty of texts for adult readers with ID. Through correlation tests we observed that there exists strong positive relations between model predictions and expert ratings. However, inferred text difficulty for adults with ID does not demonstrate such strong relations with either model predictions or expert ratings.

These observations indicate that our automatic readability assessment tool built on corpora annotated with reading difficulty indexed in grade levels is highly generalizable to cross-domain data. They also point out that, in future readability studies targeted on adult readers with ID, expert ratings should be a more reliable source for annotating reading difficulty of text for this particular group of readers.

9.2 Contributions

To summarize, the research presented in this thesis has the following significant contributions:

- **Novel features and techniques**

  We designed, extracted and implemented three subsets of novel discourse features that have not been explored by previous work in readability research, they include entity-density features, lexical chain features, coreferential inference features. The novel discourse features are inspired by and in line with established frameworks and theories on text comprehension, in particular discourse comprehension. To implement these features, we deployed novel NLP techniques, such as named entity finder, semantic annotator (lexical chainer) and coreference resolution software to extract relevant information.

  Our experiment results show that entity density features and lexical chain features useful in detecting and predicting the reading difficulty of texts, especially those with higher level of complexity. We believe these features can be applicable to NLP tasks such as text cohesion and coherence, information extraction, text summarization, text generation, etc.

  In addition to feature extraction and implementation, we developed multiple techniques to measure text readability. First of all, we framed readability as a classification task. We used two machine learning packages known for efficient high-quality multi-class classification – LIBSVM and the Weka machine learning toolkit – to build classifiers

on WeeklyReader to detect and predict reading difficulty of texts in terms of grade levels. We used repeated 10-fold cross-validation to assess classifiers' performance. In addition to standard evaluation measures, such as classification accuracy, mean squared errors, we also computed the number of misclassification by one grade level and more than one grade levels to adjust our framing readability assessment as a generic classification task.

To evaluate the generalizability of our automatic text readability assessment tool on unseen data, we introduced two other independent measures of reading difficulty: expert ratings and intrinsic text difficulty inferred by a hierarchical latent model from observations on reader's actual comprehension ability. Based on these three independent measures, we conducted correlation studies to understand the relations among grade level predictions, expert ratings and text difficulty inferred on readers' reading ability.

- **Enrichment of previous work**

  In general, we enriched previous work by implementing a set of new features at several linguistic levels. Moreover, we conducted thorough experiments to examine and compare the usefulness of these features within and across linguistic levels in detecting and predicting reading difficulty of texts. Our detailed analyses on notable findings observed from our experiment results provide better and scientific understanding of what text properties are good proxies that model text difficulty more accurately. These analyses also provide a better understanding of feature effectiveness across linguistic boundaries.

More specifically, we advanced previously studied features related to readability in following ways:

- Above all, inspired by Schwarm and Ostendorf's work, we implemented a set of 80 new perplexity features obtained from LMs trained directly on WeeklyReader using various feature selection schemes. We systematically examined and compared the discriminative power of these language-modeling-based features using our hold-one-out approach with those obtained from previous research. We found that LMs trained on in-domain corpora appear to be much more effective than those trained on unrelated corpora.
- We systematically expanded 64 POS-based features and examined the predictive power of 7 major word classes (12 subgroups). Our combined POS features, in particular nouns, demonstrate to be more effective in modeling text difficulty than parsed syntactic features, shallow features and most of discourse features.
- We introduced new measures to capture syntactic complexity. We expanded the previously studied 4 syntactic features into 21 by introducing measurement of average phrasal length and ratios of terminal and non-terminal nodes per parse tree. Our experiment results show that the augmented features led to significant performance improvement.
- We used more advanced machine learning techniques to re-evaluate the usefulness of shallow features in detecting and predicting text difficulty. We identified that average sentence length

and Flesch-Kincaid scores have dominating predictive power over all other lexical or syllable-based features.

As a result, our enrichment of previous work together with our novel features led to an automatic readability assessment tool with state of the art performance.

- **Creation of two user-specific corpora for future study**

  In order to assist our research on adapting and refining our automatic readability assessment tool for adults with MID, we conducted a pilot study followed by a reading experiment with adult participants with MID. Through these studies, we created two small text corpora that consist of paired original news articles and their corresponding simplified versions adapted by experts specifically for adults with ID.

  The complexity of each text in these two corpora was evaluated by multiple test participants through comprehension questions. We collected test participants' responses to these questions. We then developed a hierarchical latent trait model to infer reading difficulty of each text based on the test participants' actual reading ability observed from their responses.

  Moreover, we had the same set of texts rated by three experts, one of them had work experience with adults with ID. We believe that these two unique corpora are useful for researchers in the community who are interested in text readability for adults with ID, because they bear valuable participative ratings of reading difficulty collected from

experts and adults with MID. We will make the corpora available to research community to facilitate further studies.

- **Hierarchical latent trait model**

  In order to infer text difficulty for adults with ID, we developed a hierarchical latent trait model that captures key aspects of the experimental design of our user studies. This model not only takes individual reading abilities of test participants and the difficulties of question items into account, it also captures two important aspects of our reading experiment design. First, question items are no longer independent, but are grouped by article and corresponding condition – simplified or complex. Second, this model reflects the fact that the set of comprehension questions for each article was identical for the complex and simplified versions.

  The hierarchical latent trait model from Section 8.4 is generally useful for inferring article-level difficulty from repeated observations based on target readers' actual responses to multiple comprehension questions per article. This deserves to become as ubiquitous in research on adults with ID as it already is in educational testing. This model may also be applied to other work involving reading comprehension that shares similar settings.

9.3 Future Work

The research on automatic readability assessment presented in this thesis can be improved and extended in several ways in future work.

A major obstacle we encountered in our study is limited access to appropriate text corpora. Because of this, we can not compare our results directly to Schwarm and Ostendorf's work. For the same reason – the WeeklyReader corpus we obtained can not be distributed – it is also hard for other researchers to replicate our work.

An ideal scenario for future work would be to have a large, validated, freely available corpora. At the moment, since we do not have access to annotation guidelines used for WeeklyReader corpus, we have no information to what extent grade level annotations in this corpus reflect reading difficulty as experienced by elementary school students. Therefore a key future direction for readability study is to create and validate large corpora of diverse texts annotated with reading difficulty for a well-defined group of readers. More effort should be directed to work on standardized annotation guideline and methods for validation with target group of readers. At the same time, we should also keep in mind that large data may already be available from educational testing, which can be used to readily for corpora creation.

Because of there being only limited appropriate text corpora, the current version of our tool can only detect and predict reading difficulty in elementary grade levels (Grade 2 to 5). We experimented with mixing texts with a higher level of complexity into our training corpora and observed that our tool is highly generalizable to new data. This observation is quite

encouraging. Another future direction is to adapt, refine and evaluate our models with larger corpora of more diverse texts and different target group of readers.

Another valuable finding of this research is that there exists strong correlation between reading difficulty predicted by our assessment tool in terms of grade levels and other independent measures, such as number scales used by expert ratings, and coarser scale of measures, such as complex and simplified. To differentiate between simple and complex texts among those on similar topics can be a useful application of our tool for other NLP tasks. We gathered four paired simplified/complex corpora in this study: Britannica, LiteracyNet, LocalNews2007 and LocalNews2008. An extension of our current work can be made to construct an automatic readability assessment tool that models text readability at a coarser level, such as simple and complex, instead of grade level predictions. The research presented in this thesis confirms that, among feature subsets extracted from various linguistic levels, language-modeling-based features exhibit the most discriminative power in detecting and classifying reading difficulty in terms of grade levels. In modeling text difficulty in terms of simple and complex, caution needs to be exercised in training language models, because there could be overlapping content in simple and complex texts. Including both versions of texts in training corpora or just one of them may have significant impact on the models performance. These aspects deserve to be studied further.

A major contribution of this thesis is to introduce and integrate novel discourse features into the study of readability. We found that many discourse features, such as entity density features and lexical chain features, are useful

in modeling text difficulties in terms of elementary grade levels. On the other hand, while expanding features based on existing literature, we found that our new feature design using different counting and weighting schemes made significant improvement over previous study. A clear direction for future work is to continue studying the effectiveness of current features on diverse texts and explore more features and feature design at various linguistic levels, and in particular at the discourse level.

Last but not least, we need to keep in mind that text readability is complex because it is not determined by intrinsic text properties alone, rather, it results from the interaction between the reader and the text. Factors arising from the reader's side are as important as the variety of text properties that we have studied and continue to explore. Future readability research should direct more attention on validating text readability experience by selected group of readers. For any readability assessment tool that is intended for a specific group of reader to be effective, it is important to study the characteristics of the target readers and incorporate these aspects into the development.

Bibliography

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *NAACL-HLT 2010: The 5th Workshop on Innovative Use of NLP for Building Educational Applications*.

J. R. Anderson and G. H. Bower. 1973. *Human Associative memory*. Winston, Washington.

R. Barron. 1980. Visual-orthographic and phonological strategies in reading and spelling. In U. Frith, editor, *Cognitive Processes in Spelling*, pages 195–213. Academic Press, New York.

R. Barron. 1981. Reading skill and reading strategies. In A. Lesgold and C. Perfetti, editors, *Interactive processes in reading*, pages 299–327. Erlbaum, Hillsdale, NJ.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Linda Hickson Bilsky. 1985. Comprehension and mental retardation. *International Review of Research in Mental Retardation*, 13:215–246.

J. Bormuth. 1966. Readability: A new approach. *Reading Research Quarterly*, 1:79–132.

C. S. Bos and R. J. Tierney. 1980. Inferential reading abilities of mildly mental retarded and nonretarded students. *American Journal of Mental Deficiency*, 89:75–82.

B. K. Britton and S. Gulgoz. 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.

B. K. Britton, L. VanDusen, S. M. Glynn, and D. Hemphill. 1990. The impact of inferences on instructional text. In A. C. Graesser and G. H. Bower, editors, *The Psychology of learning and motivation*, volume 25, pages 53–70. Academic Press, San Diego, CA.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of AAAI98 Workshop on Intergrating Artificial Intelligence and Assistive Technology*.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*.

J. F. Cawley and R. S. Parmar. 1995. Comparisons in reading and reading-related tasks among students with average intellectual ability and students with mild mental retardation. *Education and Training in Mental Retardation and Developmental Disabilities*, 30:118–129.

Jeanne Chall. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, pages 1041–1044.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge Based Systems*, 10:183–190.

Chih-Chung Chang and Chih-Jen Lin. 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the ACL*, pages 132–139.

R. Cohen. 1982. Individual differences in short-term memory. *International Review of Research in Mental Retardation*, 11:43–77.

A. M. Collins and E. F. Loftus. 1975. A spreading-activation theory of semantic procesing. *Psychological Review*, 82:407–428.

Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*.

Frances A. Conners. 2003. Reading skills and cognitive abilities of individuals with mental retardation. *International Review of Research in Mental Retardation*, 27:191–230.

K. Crammer and Y. Singer. 2001. Pranking with ranking. In *Neural Information Processing Systems (NIPS 2001)*, pages 641–647.

Edgar Dale and Jeanne Chall. 1949. The concept of readability. *Elementary English*, 26(23).

M. Daneman and P. A. Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19:450–466.

M. Daneman and P. M. Merickle. 1996. Working memory and language comprehension: a meta-analysis. *Psychonomic Bullitin and Review*, 3:422–433.

M. Daneman and T. Tardif. 1987. Working memory and reading skill re-examined. In Coltheart N, editor, *Attention and Performance XII*, pages 491–508. Erlbaum, Hillsdale, NJ.

D. Davies, R. Sperber, and C. McCauley. 1981. Intelligence-related differences in semantic processing speed. *Journal of Experimental Child Psychology*, 31:387–402.

Alice Davison and Robert N. Kantor. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17(2):187–209.

Siobhan Devlin. 1999. *Simplifying natural language for aphasic readers*. Ph.D. thesis, University of Sunderland, UK.

Siobhan Devlin and Gary Unthank. 2006. Posters and demos: Helping aphasic people process online information. In *Proceedings of the 8th*

international ACM SIGACCESS conference on Computers and accessibility Assets '06.

P. Dixon, J. LeFevre, and L. C. Twilley. 1988. Word knowledge and working memory as predictors of reading skill. *Journal of Educational Psychology*, 80:465–472.

Clifford J. Drew and Michael L. Hardman. 2004. *Mental retardation: A lifespan approach to people with intellectual disabilities*. Merrill, Columbus, OH.

L. M. Dunn. 1954. A comparison of the reading processes of mentally retarded and normal boys of the same mental age. *Monographs of the Society for Research in Child Development*, 19:7–99.

K. A. Ehrlich and K. Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22:75–87.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of the Conference on Human Language Technology and North American chapter of the Association for Computational Linguistics (HLT-NAACL 2007)*.

Lijun Feng. 2008. Automatic readability assessment for people with intellectual disabilities. In *10th ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2008)*. Halifax, Nova Scotia, Canada. Doctoral Consortium.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *The 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *The 23rd International Conference on Computational Linguistics*.

Rudolf Flesch. 1979. *How to write plain English*. Harper and Brothers, New York.

Anne E. Fowler. 1998. Language in mental retardation: Associations with and dissociations from general cognition. In *Handbook of mental retardation and development*, pages 290–333. Cambridge University Press.

C. H. Frederiksen. 1975. Acquisition of semantic information from discourse: Effects of repeated exposures. *Journal of Verbal Learning and Verbal Behavior*, 14:158–169.

C. Freeman. 1978. Readability and text structure: A view from linguistics. In P. Griffin and R. Shuy, editors, *Children's Functional Language and Education in the early Years*. Center for Applied Linguistics, Arlington, VA.

Michel Galley and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*.

Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge.

M. A. Gernsbacher. 1990. *Language Comprehension as Structure Building*. Erlbaum, Hillsdale, NJ.

M. A. Gernsbacher. 1997. Two decades of structure building. *Discourse Processes*, 23:265–304.

L Glidden and H. Mar. 1978. Availability and accessibility of information in the semantic memory of retarded and nonretarded adolescents. *Journal of Experimental Child Psychology*, 25:33–40.

S. R. Goldman and C. K. Varnhagen. 1986. Memory for embedded and sequential story structures. *Journal of Memory and Language*, 25:401–418.

J. W. Gourlay. 1978. This basal is easy to read – or is it? *The Reading Teacher*, 32:174–182.

W. S. Gray and B. Leary. 1935. *What makes a book readable*. Chicago University Press, Chicago.

Barbara Grosz, K. Joshi Aravind, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

C. A. Hale and J. G. Borkowski. 1991. Attention, memory and cognition. In R. W. Reese and L. Lipsett, editors, *Handbook of Mental Retardation*, pages 505–528. Pergamon, New York.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

M. W. Harm and M. S. Seidenberg. 1999. Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, 106:491–528.

S. E. Haviland and H. H. Clark. 1974. What's new? acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, 13:512–521.

Michael J. Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*.

Michael J. Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *ACL 2008: The 3rd Workshop on Innovative Use of NLP for Building Educational Applications*.

T. Hogaboam and C. Perfetti. 1978. Reading skill and the role of verbal experience in decoding. *Journal of Educational Psychology*, 70:717–729.

Matt Huenerfauth, Lijun Feng, and Noemie Elhadad. 2009. Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *11th ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2009)*. Pittsburgh, PA, USA.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, and Ryu Iida. 2003. Text simplification for reading assistance: A project note. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 9–16.

Martin Jansche, Lijun Feng, and Matt Huenerfauth. 2010. Reading difficulty in adults with intellectual disabilities: Analysis with a hierarchical latent trait model. In *12th ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2010)*.

J. C. Jenkinson. 1992. The use of letter position cues in the visual processing of words by children with an intellectual disability and nondisabled children. *International Journal of Disability, Development and Education*, 39:61–76.

Josephine C. Jenkinson. 1989. Word recognition and the nature of reading difficulty in children with an intellectual disability: A review. *International Journal of Disability, Development and Education*, 36(1):39–56.

T. Joachims. 1999. Making large-scale svm learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

F. W. Jones, K. Long, and W. M. L. Finlay. 2006. Assessing the reading comprehension of adults with learning disabilities. *Journal of Intellectual Disability Research*, 50:410–418.

M. A. Just and P. A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99:122–149.

David S. Katims. 2000. Literacy instruction for people with mental retardation: Historical highlights and contemporary analysis. *Education and Training in Mental Retardation and Developmental Disabilities*, 35:3–15.

J. King and M. A. Just. 1991. Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language*, 30:580–602.

W. Kintsch and T. A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85:363–394.

W. Kintsch and D. Vipond. 1979. Reading comprehension and readability in educational practice and psychological theory. In L. G. Nilsson, editor, *Perspectives on memory research*. Erlbaum, Hillsdale, NJ.

Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. 2004. Text simplification for information-seeking applications. In *On the Move to Meaningful Internet Systems*, volume 3290, pages 735–747. Springer-Verlag.

Mirella. Lapata. 2006. Automatic evaluation of information ordering: Kendalls tau. *Computational Linguistics*, 32(4):471–484.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'05)*, pages 1085–1090.

Gondy Leroy, Stephen Helmreich, James R. Cowie, Trudi Miller, and Wei Zheng. 2008. Evaluating online health information: Beyond readability formulas. In *AMIA 2008 Symposium Proceedings*.

R. F. Jr. Lorch and P. van den Brock. 1997. Understanding reading comprehension: current and future directions of cognitive science. *Contemporary Educational Psychology*, 22:213–246.

P. MacCullagh. 1980. Regressional models for ordinal data. *Journal of the Royal Statistical Society B*, 42:109–142.

J. M Mason. 1976. Testion pronunciation skill competence of both normal and retarded readers. *Mental Retardation*, 14:36–40.

J. M Mason. 1977. Questioning the notion of independent processing stages in reading. *Journal of Educational Psychology*, 69:288–297.

J. M Mason. 1978. Role of strategy in reading by mentally retarded persons. *American Journal of Mental Deficiency*, 82:467–473.

M. E. J. Masson and J. A. Miller. 1983. Working memory and individual differences in comprehension and memory of text. *Journal of Educational Psychology*, 75:314–318.

G. Harry McLaughlin. 1969. Smog grading — a new readability formula. *Journal of Reading*, 12(8):639–646.

Edward C. Merrill and Linda H. Bilsky. 1990. Individual differences in the representation of sentences in memory. *American Journal on Mental Retardation*, 95:68–76.

Edward C. Merrill and T. S. Jackson. 1992. Degree of associative relatedness and sentence processing by persons with and without mental retardation. *American Journal on Mental Retardation*, 97:173–185.

Edward C. Merrill, Regan Lookadoo, and Stacy Rilea. 2003. Memory, language comprehension and mental retardation. *International Review of Research in Mental Retardation*, 27:151–190.

M. A. Merrill. 1924. On the relation of intelligence to achievement in the case of mentally retarded children. *Comparative Psychology Monographs*, 11(10).

J. R. Miller and W. Kintsch. 1980. Readability and recall for short passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory*, 6:335–354.

Sarah E. Petersen and Mari Ostendorf. 2006. A machine learning approach to reading level assessment. Technical report, University of Washington CSE Technical Report.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:89–106.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

Martyn Plummer. 2010. JAGS (Just Another Gibbs Sampler), version 2.1.0.

Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. 2006. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn discourse treebank. In *The Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

M. B. Pulsifer. 1996. The neuropsychology of mental retardation. *Journal of the International Neuropsychological Society*, 2:159–176.

S. Jay Samuels. 2002. *Evidence-based Reading Instruction*, chapter The method of repeated readings, pages 85–90. International Reading Association.

A. J. Sanford, S. Farrod, and J. M. Boyle. 1977. An independence of mechanism in the origins of reading and classification-related semantic distance effects. *Memory and Cognition*, 5:214–220.

R. Schank. 1975. *Conceptual information processing*. Elsevier, Amsterdam.

Sarah E. Schwarm and Mari 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*.

R. J. Senter and E. A. Smith. 1967. Automated readability index. Technical report, Cincinnati University, Ohio.

D. G. Sheperd. 1967. Selected factors in the reading ability of educable mentally retarded boys. *American Journal of Mental Deficiency*, 71:563–570.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*.

Advaith Siddharthan. 2004. *Syntactic Simplification and Text Cohesion*. Ph.D. thesis, University of Cambridge.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Keith E. Stanovich. 1985. Cognitive determinants of reading in mentally retarded individuals. *International Review of Research in Mental Retardation*, 13:181–214.

A. J. Stenner. 1996. Measuring reading comprehension with the lexile framework. In *The fourth North America Conference on Adolescent/Adult Literacy*.

A. J. Stenner, M. Smith, and D. S. Burdick. 1983. Toward a theory of construct definition. *Journal of Educational Measurement*, 20:305–315.

Andrew Thomas. 2006. The BUGS language. *R News*, 6(1):17–21.

T. Trabasso, T. Secco, and P. van den Broek. 1984. Causal cohesion and story coherence. In H. Mandl, N. L. Stein, and T. Trabasso, editors, *Learning and Comprehension of Text*, pages 83–111. Erlbaum, Hillsdale, NJ.

T. Trabasso and S. Suh. 1993. Understanding text: Achieving explanatory coherence through on-line inferences and mental operations in working memory. *Discourse Processes*, 16:3–34.

T. Trabasso and P. van den Broek. 1985. Causal thinking and the representation of narrative events. *Journal of Memory and Language*, 24:612–630.

U.S. Census Bureau. 2006. American community survey/Puerto Rico community survey 2006 subject definitions. Available online at `http://www.census.gov/acs/www/Downloads/2006/usedata/Subject_Definitions.pdf`.

Sandra Williams and Ehud Reiter. 2005. Generating readable texts for readers with low basic skills. In *Proceeding of the 10th European Workshop on Natural Language Generation*. Aberdeen.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997)*, pages 421–429.