Cognitively Motivated Features for Readability Assessment

Lijun Feng
The City University of New York,
Graduate Center
New York, NY, USA
lijun7.feng@gmail.com

Noémie Elhadad
Columbia University
New York, NY, USA
noemie@dbmi.columbia.edu

Matt Huenerfauth
The City University of New York,
Queens College & Graduate Center
New York, NY, USA
matt@cs.qc.cuny.edu

Abstract

We investigate linguistic features that correlate with the readability of texts for adults with intellectual disabilities (ID). Based on a corpus of texts (including some experimentally measured for comprehension by adults with ID), we analyze the significance of novel discourse-level features related to the cognitive factors underlying our users' literacy challenges. We develop and evaluate a tool for automatically rating the readability of texts for these users. Our experiments show that our discourse-level, cognitively-motivated features improve automatic readability assessment.

1 Introduction

Assessing the degree of readability of a text has been a field of research as early as the 1920's. Dale and Chall define readability as "the sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at optimal speed, and find it interesting" (Dale and Chall, 1949). It has long been acknowledged that readability is a function of text characteristics, but also of the readers themselves. The literacy skills of the readers, their motivations, background knowledge, and other internal characteristics play an important role in determining whether a text is readable for a particular group of people. In our work, we investigate how to assess the readability of a text for people with intellectual disabilities (ID).

Previous work in automatic readability assessment has focused on generic features of a text at the lexical and syntactic level. While such features are essential, we argue that audience-specific features that model the cognitive characteristics of a user group can improve the accuracy of a readability assessment tool. The contributions of this paper are: (1) we present a corpus of texts with readability judgments from adults with ID; (2) we propose a set of cognitively-motivated features which operate at the discourse level; (3) we evaluate the utility of these features in predicting readability for adults with ID.

Our framework is to create tools that benefit people with intellectual disabilities (ID), specifically those classified in the "mild level" of mental retardation, IQ scores 55-70. About 3% of the U.S. population has intelligence test scores of 70 or lower (U.S. Census Bureau, 2000). People with ID face challenges in reading literacy. They are better at decoding words (sounding them out) than at comprehending their meaning (Drew & Hardman, 2004), and most read below their mental age-level (Katims, 2000). Our research addresses two literacy impairments that distinguish people with ID from other low-literacy adults: limitations in (1) working memory and (2) discourse representation. People with ID have problems remembering and inferring information from text (Fowler, 1998). They have a slower speed of semantic encoding and thus units are lost from the working memory before they are processed (Perfetti & Lesgold, 1977; Hickson-Bilsky, 1985). People with ID also have trouble building cohesive representations of discourse (Hickson-Bilsky, 1985). As less information is integrated into the mental representation of the current discourse, less is comprehended.

Adults with ID are limited in their choice of reading material. Most texts that they can readily understand are targeted at the level of readability of children. However, the topics of these texts often fail to match their interests since they are meant for younger readers. Because of the mismatch between their literacy and their interests, users may not read for pleasure and therefore miss valuable reading-skills practice time. In a feasibility study we conducted with adults

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 229–237,
Athens, Greece, 30 March – 3 April 2009.

with ID, we asked participants what they enjoyed learning or reading about. The majority of our subjects mentioned enjoying watching the news, in particular local news. Many mentioned they were interested in information that would be relevant to their daily lives. While for some genres, human editors can prepare texts for these users, this is not practical for news sources that are frequently updated and specific to a limited geographic area (like local news). Our goal is to create an automatic metric to predict the readability of local news articles for adults with ID. Because of the low levels of written literacy among our target users, we intend to focus on comprehension of texts displayed on a computer screen and read aloud by text-to-speech software; although some users may depend on the text-to-speech software, we use the term *readability*.

This paper is organized as follows. Section 2 presents related work on readability assessment. Section 3 states our research hypotheses and describes our methodology. Section 4 focuses on the data sets used in our experiments, while section 5 describes the feature set we used for readability assessment along with a corpus-based analysis of each feature. Section 6 describes a readability assessment tool and reports on evaluation. Section 7 discusses the implications of the work and proposes direction for future work.

2 Related Work on Readability Metrics

Many readability metrics have been established as a function of shallow features of texts, such as the number of syllables per word and number of words per sentence (Flesch, 1948; McLaughlin, 1969; Kincaid et al., 1975). These so-called traditional readability metrics are still used today in many settings and domains, in part because they are very easy to compute. Their results, however, are not always representative of the complexity of a text (Davison and Kantor, 1982). They can easily misrepresent the complexity of technical texts, or reveal themselves un-adapted to a set of readers with particular reading difficulties. Other formulas rely on lexical information; e.g., the New Dale-Chall readability formula consults a static, manually-built list of "easy" words to determine whether a text contains unfamiliar words (Chall and Dale, 1995).

Researchers in computational linguistics have investigated the use of statistical language models (unigram in particular) to capture the range of vocabulary from one grade level to another (Si and Callan, 2001; Collins-Thompson and Callan, 2004). These metrics predicted readability better than traditional formulas when tested against a corpus of web pages. The use of syntactic features was also investigated (Schwarm and Ostendorf, 2005; Heilman et al., 2007; Petersen and Ostendorf, 2009) in the assessment of text readability for English as a Second Language readers. While lexical features alone outperform syntactic features in classifying texts according to their reading levels, combining the lexical and syntactic features yields the best results.

Several elegant metrics that focus solely on the syntax of a text have also been developed. The Yngve (1960) measure, for instance, focuses on the depth of embedding of nodes in the parse tree; others use the ratio of terminal to non-terminal nodes in the parse tree of a sentence (Miller and Chomsky, 1963; Frazier, 1985). These metrics have been used to analyze the writing of potential Alzheimer's patients to detect mild cognitive impairments (Roark, Mitchell, and Hollingshead, 2007), thereby indicating that cognitively motivated features of text are valuable when creating tools for specific populations.

Barzilay and Lapata (2008) presented early work in investigating the use of discourse to distinguish abridged from original encyclopedia articles. Their focus, however, is on style detection rather than readability assessment *per se*. Coh-Metrix is a tool for automatically calculating text coherence based on features such as repetition of lexical items across sentences and latent semantic analysis (McNamara et al., 2006). The tool is based on comprehension data collected from children and college students.

Our research differs from related work in that we seek to produce an automatic readability metric that is tailored to the literacy skills of adults with ID. Because of the specific cognitive characteristics of these users, it is an open question whether existing readability metrics and features are useful for assessing readability for adults with ID. Many of these earlier metrics have focused on the task of assigning texts to particular elementary school grade levels. Traditional grade levels may not be the ideal way to score texts to indicate how readable they are for adults with ID. Other related work has used models of vocabulary (Collins-Thompson and Callan, 2004). Since we would like to use our tool to give adults with ID access to local news stories, we choose to keep our metric topic-independent.

Another difference between our approach and previous approaches is that we have designed the features used by our readability metric based on

the cognitive aspects of our target users. For example, these users are better at decoding words than at comprehending text meaning (Drew & Hardman, 2004); so, shallow features like "syllable count per word" or unigram models of word frequency (based on texts designed for children) may be less important indicators of reading difficulty. A critical challenge for our users is to create a cohesive representation of discourse. Due to their impairments in semantic encoding speed, our users may have particular difficulty with texts that place a significant burden on working memory (items fall out of memory before they can be semantically encoded).

While we focus on readability of texts, other projects have automatically generated texts for people with aphasia (Carroll et al., 1999) or low reading skills (Williams and Reiter, 2005).

3 Research Hypothesis and Methods

We hypothesize that the complexity of a text for adults with ID is related to the number of entities referred to in the text overall. If a paragraph or a text refers to too many entities at once, the reader has to work harder at mapping each entity to a semantic representation and deciding how each entity is related to others. On the other hand, when a text refers to few entities, less work is required both for semantic encoding and for integrating the entities into a cohesive mental representation. Section 5.2 discusses some novel discourse-level features (based on the "entity density" of a text) that we believe will correlate to comprehension by adults with ID.

To test our hypothesis, we used the following methodology. We collected four corpora (as described in Section 4). Three of them (Britannica, LiteracyNet and WeeklyReader) have been examined in previous work on readability. The fourth (LocalNews) is novel and results from a user study we conducted with adults with ID. We then analyzed how significant each feature is on our Britannica and LiteracyNet corpora. Finally, we combined the significant features into a linear regression model and experimented with several feature combinations. We evaluated our model on the WeeklyReader and LocalNews corpora.

4 Corpora and Readability Judgments

To study how certain linguistic features indicate the readability of a text, we collected a corpus of English text at different levels of readability. An ideal corpus for our research would contain texts that have been written specifically for our audience of adults with intellectual disabilities – in particular if such texts were paired with alternate versions of each text written for a general audience. We are not aware of such texts available electronically, and so we have instead mostly collected texts written for an audience of children. The texts come from online and commercial sources, and some have been analyzed previously by text simplification researchers (Petersen and Ostendorf, 2009). Our corpus also contains some novel texts produced as part of an experimental study involving adults with ID.

4.1 Paired and Graded Generic Corpora: Britannica, LiteracyNet, and Weekly Reader

The first section of our corpus (which we refer to as Britannica) has 228 articles from the Encyclopedia Britannica, originally collected by (Barzilay and Elhadad, 2003). This consists of 114 articles in two forms: original articles written for adults and corresponding articles rewritten for an audience of children. While the texts are paired, the content of the texts is not identical: some details are omitted from the child version, and additional background is sometimes inserted. The resulting corpus is comparable in content.

Because we are particularly interested in making local news articles accessible to adults with ID, we collected a second paired corpus, which we refer to as LiteracyNet, consisting of 115 news articles made available through (Western/Pacific Literacy Network / LiteracyNet, 2008). The collection of local CNN stories is available in an original and simplified/abridged form (230 total news articles) designed for use in literacy education.

The third corpus we collected (Weekly Reader) was obtained from the Weekly Reader corporation (Weekly Reader, 2008). It contains articles for students in elementary school. Each text is labeled with its target grade level (grade 2: 174 articles, grade 3: 289 articles, grade 4: 428 articles, grade 5: 542 articles). Overall, the corpus has 1433 articles. (U.S. elementary school grades 2 to 5 generally are for children ages 7 to 10.)

The corpora discussed above are similar to those used by Petersen and Ostendorf (2009). While the focus of our research is adults with ID, most of the texts discussed in this section have been simplified or written by human authors to be readable for children. Despite the texts being intended for a different audience than the focus of our research, we still believe these texts to be

of value. It is rare to encounter electronically available corpora in which an original and a simplified version of a text is paired (as in the Britannica and LiteracyNet corpora) or texts labeled as being at specific levels of readability (as in the Weekly Reader corpus).

4.2 Readability-Specific Corpus: LocalNews

The final section of our corpus contains local news articles that are labeled with comprehension scores. These texts were produced for a feasibility study involving adults with ID. Each text was read by adults with ID, who then answered comprehension questions to measure their understanding of the texts. Unlike the previous corpora, LocalNews is novel and was not investigated by previous research in readability.

After obtaining university approval for our experimental protocol and informed consent process, we conducted a study with 14 adults with mild intellectual disabilities who participate in daytime educational programs in the New York area. Participants were presented with ten articles collected from various local New York based news websites. Some subjects saw the original form of an article and others saw a simplified form (edited by a human author); no subject saw both versions. The texts were presented in random order using software that displayed the text on the screen, read it aloud using text-to-speech software, and highlighted each word as it was read. Afterward, subjects were asked aloud multiple-choice comprehension questions. We defined the readability score of a story as the percentage of correct answers averaged across the subjects who read that particular story.

A human editor performed the text simplification 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: breaking apart complex sentences, unembedding information in complex prepositional phrases and reintegrating it as separate sentences, replacing infrequent vocabulary items with more common/colloquial equivalents, 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.*"

This corpus of local news articles that have been human edited and scored for comprehension by adults with ID is small in size (20 news articles), but we consider it a valuable resource. Unlike the texts that have been simplified for children (the rest of our corpus), these texts have been rated for readability by actual adults with ID. Furthermore, comprehension scores are derived from actual reader comprehension tests, rather than self-perceived comprehension. Because of the small size of this part of our corpus, however, we primarily use it for evaluation purposes (not for training the readability models).

5 Linguistic Features and Readability

We now describe the set of features we investigated for assessing readability automatically. Table 1 contains a list of the features – including a short code name for each feature which may be used throughout this paper. We have begun by implementing the simple features used by the Flesh-Kincaid and FOG metrics: average number of words per sentence, average number of syllables per word, and percentage of words in the document with 3+ syllables.

5.1 Basic Features Used in Earlier Work

We have also implemented features inspired by earlier research on readability. Petersen and Ostendorf (2009) included features calculated from parsing the sentences in their corpus using the Charniak parser (Charniak, 2000): average parse tree height, average number of noun phrases per sentence, average number of verb phrases per sentence, and average number of SBARs per sentence. We have implemented versions of most of these parse-tree-related features for our project. We also parse the sentences in our corpus using Charniak's parser and calculate the following features listed in Table 1: aNP, aN, aVP, aAdj, aSBr, aPP, nNP, nN, nVP, nAdj, nSBr, and nPP.

5.2 Novel Cognitively-Motivated Features

Because of the special reading characteristics of our target users, we have designed a set of cognitively motivated features to predict readability of texts for adults with ID. We have discussed how working memory limits the semantic encoding of new information by these users; so, our features indicate the number of entities in a text that the reader must keep in mind while reading each sentence and throughout the entire document. It is our hypothesis that this "entity density" of a

Code	Feature
aWPS	average number of words per sentence
aSPW	average number of syllables per word
%3+S	% of words in document with 3+ syllables
aNP	avg. num. NPs per sentence
aN	avg. num. common+proper nouns per sentence
aVP	avg. num. VPs per sentence
aAdj	avg. num. Adjectives per sentence
aSBr	avg. num. SBARs per sentence
aPP	avg. num. prepositional phrases per sentence
nNP	total number of NPs per sentence
nN	total num. of common+proper nouns in document
nVP	total number of VPs in the document
nAdj	total number of Adjectives in the document
nSBr	total number of SBARs in the document
nPP	total num. of prepositional phrases in document
nEM	number of entity mentions in document
nUE	number of unique entities in document
aEM	avg. num. entity mentions per sentence
aUE	avg. num. unique entities per sentence
nLC	number of lexical chains in document
nLC2	num. lex. chains, span > half document length
aLCL	average lexical chain length
aLCS	average lexical chain span
aLCw	avg. num. lexical chains active at each word
aLCn	avg. num. lexical chains active at each NP

Table 1: Implemented Features

text plays an important role in the difficulty of that text for readers with intellectual disabilities.

The first set of features incorporates the LingPipe named entity detection software (Alias-i, 2008), which detects three types of entities: person, location, and organization. We also use the part-of-speech tagger in LingPipe to identify the common nouns in the document, and we find the union of the common nouns and the named entity noun phrases in the text. The union of these two sets is our definition of "entity" for this set of features. We count both the total number of "entity mentions" in a text (each token appearance of an entity) and the total number of unique entities (exact-string-match duplicates only counted once). Table 1 lists these features: nEM, nUE, aEM, and aUE. We count the totals per document to capture how many entities the reader must keep track of while reading the document. We also expect sentences with more entities to be more difficult for our users to semantically encode due to working memory limitations; so, we also count the averages per sentence to capture how many entities the reader must keep in mind to understand each sentence.

To measure the working memory burden of a text, we'd like to capture the number of discourse entities that a reader must keep in mind. However, the "unique entities" identified by the named entity recognition tool may not be a perfect representation of this – several unique entities may actually refer to the same real-world entity under discussion. To better model how multiple noun phrases in a text refer to the same entity or concept, we have also built features using lexical chains (Galley and McKeown, 2003). Lexical chains link nouns in a document connected by relations like synonymy or hyponomy; chains can indicate concepts that recur throughout a text. A lexical chain has both a length (number of noun phrases it includes) and a span (number of words in the document between the first noun phrase at the beginning of the chain and the last noun phrase that is part of the chain). We calculate the number of lexical chains in the document (nLC) and those with a span greater than half the document length (nLC2). 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 in Table 1 (aLCw and aLCe) 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.

5.3 Testing the Significance of Features

To select which features to include in our automatic readability assessment tool (in Section 6),

we analyzed the documents in our paired corpora (Britannica and LiteracyNet). Because they contain a complex and a simplified version of each article, we can examine differences in readability while holding the topic and genre constant. We calculated the value of each feature for each document, and we used a paired t-test to determine if the difference between the complex and simple documents was significant for that corpus.

Table 2 contains the results of this feature selection process; the columns in the table indicate the values for the following corpora: Britannica complex, Britannica simple, LiteracyNet complex, and LiteracyNet simple. An asterisk appears in the "Sig" column if the difference between the feature values for the complex vs. simple documents is statistically significant for that corpus (significance level: $p<0.00001$).

The only two features which did not show a significant difference ($p>0.01$) between the complex and simple versions of the articles were: average lexical chain length (aLCL) and number of lexical chains with span greater than half the document length (nLC2). The lack of significance for aLCL may be explained by the vast majority of lexical chains containing few members; complex articles contained more of these chains – but their chains did not contain more members. In the case of nLC2, over 80% of the articles in each category contained no lexical chains whose span was greater than half the document length. The rarity of a lexical chain spanning the majority of a document may have led to there being no significant difference between complex/simple.

Feature	Brit. Com.	Brit. Simp.	Sig	LitN. Com.	LitN. Simp.	Sig
aWPS	20.13	14.37	*	17.97	12.95	*
aSPW	1.708	1.655	*	1.501	1.455	*
%3+S	0.196	0.177	*	0.12	0.101	*
aNP	8.363	6.018	*	6.519	4.691	*
aN	7.024	5.215	*	5.319	3.929	*
aVP	2.334	1.868	*	3.806	2.964	*
aAdj	1.95	1.281	*	1.214	0.876	*
aSBr	0.266	0.205	*	0.793	0.523	*
aPP	2.858	1.936	*	1.791	1.22	*
nNP	798	219.2	*	150.2	102.9	*
nN	668.4	190.4	*	121.4	85.75	*
nVP	242.8	69.19	*	88.2	65.52	*
nAdj	205	47.32	*	28.11	19.04	*
nSBr	31.33	7.623	*	18.16	11.43	*
nPP	284.7	70.75	*	41.06	26.79	*
nEM	624.2	172.7	*	115.2	82.83	*
nUE	355	117	*	81.56	54.94	*
aEM	6.441	4.745	*	5.035	3.789	*
aUE	4.579	3.305	*	3.581	2.55	*
nLC	59.21	17.57	*	12.43	8.617	*
nLC2	0.175	0.211		0.191	0.226	
aLCL	3.009	3.022		2.817	2.847	
aLCS	357	246.1	*	271.9	202.9	*
aLCw	1.803	1.358	*	1.407	1.091	*
aLCn	1.852	1.42	*	1.53	1.201	*

Table 2: Feature Values of Paired Corpora

6 A Readability Assessment Tool

After testing the significance of features using paired corpora, we used linear regression and our graded corpus (Weekly Reader) to build a readability assessment tool. To evaluate the tool's usefulness for adults with ID, we test the correlation of its scores with the LocalNews corpus.

6.1 Versions of Our Model

We began our evaluation by implementing three versions of our automatic readability assessment tool. The first version uses only those features studied by previous researchers (aWPS, aSPW, %3+S, aNP, aN, aVP, aAdj, aSBr, aPP, nNP, nN, nVP, nAdj, nSBr, nPP). The second version uses only our novel cognitively motivated features (section 5.2). The third version uses the union of both sets of features. By building three versions of the tool, we can compare the relative impact of our novel cognitively-motivated features. For all versions, we have only included those features that showed a significant difference between the complex and simple articles in our paired corpora (as discussed in section 5.3).

6.2 Learning Technique and Training Data

Early work on automatic readability analysis framed the problem as a classification task: creating multiple classifiers for labeling a text as being one of several elementary school grade levels (Collins-Thompson and Callan, 2004). Because we are focusing on a unique user group with special reading challenges, we do not know *a priori* what level of text difficulty is ideal for our users. We would not know where to draw category boundaries for classification. We also prefer that our assessment tool assign numerical difficulty scores to texts. Thus, after creating this tool, we can conduct further reading comprehension experiments with adults with ID to determine what threshold (for readability scores assigned by our tool) is appropriate for our users.

To select features for our model, we used our paired corpora (Britannica and LiteracyNet) to measure the significance of each feature. Now that we are training a model, we make use of our *graded* corpus (articles from Weekly Reader). This corpus contains articles that have each been labeled with an elementary school grade level for which it was written. We divide this corpus – using 80% of articles as training data and 20% as testing data. We model the grade level of the articles using linear regression; our model is implemented using R (R Development Core Team, 2008).

6.3 Evaluation of Our Readability Tool

We conducted two rounds of training and evaluation of our three regression models. We also compare our models to a baseline readability assessment tool: the popular Flesh-Kincaid Grade Level index (Kincaid et al., 1975).

In the first round of evaluation, we trained and tested our regression models on the Weekly Reader corpus. This round of evaluation helped to determine whether our feature-set and regression technique were successfully modeling those aspects of the texts that were relevant to their grade level. Our results from this round of evaluation are presented in the form of average error scores. (For each article in the Weekly Reader testing data, we calculate the difference between the output score of the model and the correct grade-level for that article.) Table 3 presents the average error results for the baseline system and our three regression models. We can see that the model trained on the shallow and parse-related features out-performs the model trained only on our novel features; however, the best model overall is the one is trained on all of the features. This model predicts the grade level of Weekly Reader articles to within roughly 0.565 grade levels on average.

Readability Model (or baseline)	Average Error
Baseline: Flesh-Kincaid Index	2.569
Basic Features Only	0.6032
Cognitively Motivated Features Only	0.6110
Basic + Cognitively-Motiv. Features	0.5650

Table 3: Predicting Grade Level of Weekly Reader

In our second round of evaluation, we trained the regression model on the Weekly Reader corpus, but we tested it against the LocalNews corpus. We measured the correlation between our regression models' output and the comprehension scores of adults with ID on each text. For this reason, we do not calculate the "average error"; instead, we simply measure the correlation between the models' output and the comprehension scores. (We expect negative correlations because comprehension scores should increase as the predicted grade level of the text goes down.)

Table 4 presents the correlations for our three models and the baseline system in the form of Pearson's R-values. We see a surprising result: the model trained only on the cognitively-motivated features is more tightly correlated with the comprehension scores of the adults with ID. While the model trained on all features was better at assigning grade levels to Weekly Reader articles, when we tested it on the local news articles from our user-study, it was not the top-performing model. This result suggests that the shallow and parse-related features of texts designed for children (the Weekly Reader articles, our training data) are not the best predictors of text readability for adults with ID.

Readability Model (or baseline)	Pearson's R
Baseline: Flesh-Kincaid Index	-0.270
Basic Features Only	-0.283
Cognitively Motivated Features Only	-0.352
Basic + Cognitively-Motiv. Features	-0.342

Table 4: Correlation to User-Study Comprehension

7 Discussion

Based on the cognitive and literacy skills of adults with ID, we designed novel features that were useful in assessing the readability of texts for these users. The results of our study have supported our hypothesis that the complexity of a text for adults with ID is related to the number of entities referred to in the text. These "entity density" features enabled us to build models that were better at predicting text readability for adults with intellectual disabilities.

This study has also demonstrated the value of collecting readability judgments from target users when designing a readability assessment tool. The results in Table 4 suggest that models trained on corpora containing texts designed for children may not always lead to accurate models of the readability of texts for other groups of low-literacy users. Using features targeting specific aspects of literacy impairment have allowed us to make better use of children's texts when designing a model for adults with ID.

7.1 Future Work

In order to study more features and models of readability, we will require more testing data for tracking progress of our readability regression

models. Our current study has illustrated the usefulness of texts that have been evaluated by adults with ID, and we therefore plan to increase the size of this corpus in future work. In addition to using this corpus for evaluation, we may want to use it to *train* our regression models. For this study, we trained on Weekly Reader text labeled with elementary school grade levels, but this is not ideal. Texts designed for children may differ from those that are best for adults with ID, and "grade levels" may not be the best way to rank/rate text readability for these users. While our user-study comprehension-test corpus is currently too small for training, we intend to grow the size of this corpus in future work.

We also plan on refining our cognitively motivated features for measuring the difficulty of a text for our users. Currently, we use lexical chain software to link noun phrases in a document that may refer to similar entities/concepts. In future work, we plan to use co-reference resolution software to model how multiple "entity mentions" may refer to a single discourse entity.

For comparison purposes, we plan to implement other features that have been used in earlier readability assessment systems. For example, Petersen and Ostendorf (2009) created lists of the most common words from the Weekly Reader articles, and they used the percentage of words in a document not on this list as a feature.

The overall goal of our research is to develop a software system that can automatically simplify the reading level of local news articles and present them in an accessible way to adults with ID. Our automatic readability assessment tool will be a component in this future text simplification system. We have therefore preferred to include features in our tool that focus on aspects of the text that can be modified during a simplification process. In future work, we will study how to use our readability assessment tool to guide how a text revision system decides to modify a text to increase its readability for these users.

7.2 Summary of Contributions

We have contributed to research on automatic readability assessment by designing a new method for assessing the complexity of a text at the level of discourse. Our novel "entity density" features are based on named entity and lexical chain software, and they are inspired by the cognitive underpinnings of the literacy challenges of adults with ID – specifically, the role of slow semantic encoding and working memory limitations. We have demonstrated the usefulness of these novel features in modeling the grade level of elementary school texts and in correlating to readability judgments from adults with ID.

Another contribution of our work is the collection of an initial corpus of texts of local news stories that have been manually simplified by a human editor. Both the original and the simplified versions of these stories have been evaluated by adults with intellectual disabilities. We have used these comprehension scores in the evaluation phase of this study, and we have suggested how constructing a larger corpus of such articles could be useful for training readability tools.

More broadly, this project has demonstrated how focusing on a specific user population, analyzing their cognitive skills, and involving them in a user-study has led to new insights in modeling text readability. As Dale and Chall's definition (1949) originally argued, characteristics of the reader are central to the issue of readability. We believe our user-focused research paradigm may be used to drive further advances in readability assessment for other groups of users.

Acknowledgements

We thank the Weekly Reader Corporation for making its corpus available for our research. We are grateful to Martin Jansche for his assistance with the statistical data analysis and regression.

References

Alias-i. 2008. LingPipe 3.6.0. http://alias-i.com/lingpipe (accessed October 1, 2008)

Barzilay, R., Elhadad, N., 2003. Sentence alignment for monolingual comparable corpora. In *Proc EMNLP*, pp. 25-32.

Barzilay R., Lapata, M., 2008. Modeling Local Coherence: An Entity-based Approach. *Computational Linguistics*. 34(1):1-34.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., Tait, J. 1999. Simplifying text for language-impaired readers. In *Proc. EACL Poster*, p. 269.

Chall, J.S., Dale, E., 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA.

Charniak, E. 2000. A maximum-entropy-inspired parser. In *Proc. NAACL*, pp. 132-139.

Collins-Thompson, K., and Callan, J. 2004. A language modeling approach to predicting reading difficulty. In *Proc. NAACL,* pp. 193-200.

Dale, E. and J. S. Chall. 1949. The concept of readability. *Elementary English* 26(23).

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

Drew, C.J., and Hardman, M.L. 2004. *Mental retardation: A lifespan approach to people with intellectual disabilities (8th ed.)*. Columbus, OH: Merrill.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221-233.

Fowler, A.E. 1998. Language in mental retardation. In Burack, Hodapp, and Zigler (Eds.), *Handbook of Mental Retardation and Development*. Cambridge, UK: Cambridge Univ. Press, pp. 290-333.

Frazier, L. 1985. *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, chapter Syntactic complexity, pp. 129-189. Cambridge University Press.

Galley, M., McKeown, K. 2003. Improving Word Sense Disambiguation in Lexical Chaining. In *Proc. IJCAI,* pp. 1486-1488.

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

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proc. NAACL*, pp. 460-467.

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

Katims, D.S. 2000. Literacy instruction for people with mental retardation: Historical highlights and contemporary analysis. *Education and Training in Mental Retardation and Developmental Disabilities*, 35(1): 3-15.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. 1975. Derivation of new readability formulas for Navy enlisted personnel, Research Branch Report 8-75, Millington, TN.

Kincaid, J., Fishburne, R., Rodgers, R., and Chisson, B. 1975. Derivation of new readability formulas for navy enlisted personnel. Technical report, Research Branch Report 8-75, U.S. Naval Air Station.

McLaughlin, G.H. 1969. SMOG grading - a new readability formula. *Journal of Reading*, 12(8):639-646.

McNamara, D.S., Ozuru, Y., Graesser, A.C., & Louwerse, M. (2006) Validating Coh-Metrix., In *Proc. Conference of the Cognitive Science Society*, pp. 573.

Miller, G., and Chomsky, N. 1963. *Handbook of Mathematical Psychology*, chapter Finatary models of language users, pp. 419-491. Wiley.

Perfetti, C., and Lesgold, A. 1977. Cognitive Processes in Comprehension, chapter Discourse Comprehension and sources of individual differences. Erlbaum.

Petersen, S.E., Ostendorf, M. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23: 89-106.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Roark, B., Mitchell, M., and Hollingshead, K. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proc. ACL Workshop on Biological, Translational, and Clinical Language Processing (BioNLP'07)*, pp. 1-8.

Schwarm, S., and Ostendorf, M. 2005. Reading level assessment using support vector machines and statistical language models. In *Proc. ACL*, pp. 523-530.

Si, L., and Callan, J. 2001. A statistical model for scientific readability. In *Proc. CIKM*, pp. 574-576.

Stenner, A.J. 1996. Measuring reading comprehension with the Lexile framework. 4th North American Conference on Adolescent/Adult Literacy.

U.S. Census Bureau. 2000. *Projections of the total resident population by five-year age groups* and *sex, with special age categories: Middle series 2025-2045*. Washington: U.S. Census Bureau, Populations Projections Program, Population Division.

Weekly Reader, 2008. http://www.weeklyreader.com (Accessed Oct., 2008).

Western/Pacific Literacy Network / Literacyworks, 2008. CNN SF learning resources. http://literacynet.org/cnnsf/ (Accessed Oct., 2008).

Williams, S., Reiter, E. 2005. Generating readable texts for readers with low basic skills. In *Proc. European Workshop on Natural Language Generation*, pp. 140-147.

Yngve, V. 1960. A model and a hypothesis for language structure. *American Philosophical Society*, 104: 446-466.