Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications

Sowmya Vajjala and Detmar Meurers
LEAD Graduate School and Seminar für Sprachwissenschaft, Universität Tübingen

ABSTRACT

Readability assessment can play a role in the evaluation of a simplification algorithm as well as in the identification of what to simplify. While some previous research used traditional readability formulas to evaluate text simplification, there is little research into the utility of readability assessment for identifying and analyzing sentence level targets for text simplification. We explore this aspect in our paper by first constructing a readability model that is generalizable across corpora and across genres and later adapting this model to make sentence-level readability judgments.

We start with experiments establishing that the readability model integrating a broad range of linguistic features works well at a document level, performing on par with the best systems on a standard test corpus. The model then is confirmed to be portable to different text genres. Moving from documents to sentences, we investigate the model's ability to correctly identify the difference in reading level between a sentence and its human simplified version. We conclude that readability models can be useful for identifying simplification targets for human writers and for evaluating machine generated simplifications.

Keywords: generalizability of readability models, readability assessment, sentence readability, simplification evaluation, text simplification.

To appear in: *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*, edited by Thomas François and Delphine Bernhard, 2014.

INTRODUCTION

Automatic text simplification is the process of simplifying the form of a text while preserving its meaning. The goal is to obtain a text that is less challenging to read or process for the target users or systems. Early research into text simplification started with Chandrasekar et al. (1996)'s approach for splitting long sentences into multiple short sentences to improve parsing efficiency. Application scenarios for human users, such as text simplification for aphasics, dyslexics and language learners, have also gained some attention in the past few years (e.g., Canning et al., 1999). Complementing such direct applications of text simplification, it is also seen as an important component of other applications such as summarization, information extraction, question answering and information retrieval (e.g., Klebanov et al., 2004). While most of the machine-oriented approaches focused on deriving short sentences to enable more efficient processing for the machine, approaches targeting human users emphasized the simplification of specific lexical items and syntactic constructs that are known to be difficult for the intended target population. A recent strand of work on text simplification views the task from a machine translation perspective (e.g., Zhu et al., 2010; Wubben et al., 2012). An alternative strand pursues text simplification by identifying the transformations needed to simplify a sentence and specifying the articulation points for performing these transformations (e.g., Medero & Ostendorf, 2011). While most approaches make use of any opportunity to simplify a targeted form, the downside of such opportunity-driven simplification is that any change to a text may also negatively impact its naturalness and meaningfulness. In our research we therefore explore the question which sentences constitute the best targets for simplification. In this paper we explore automatic readability assessment for determining readability at a sentence level, with the goal of identifying target sentences for text simplification.

Automatic readability assessment, i.e., the task of assessing the reading difficulty of a text for a target population can be useful for text simplification in two ways: for evaluation and for target identification. It was sometimes used as a measure to evaluate the performance of a text simplification system, in the form of traditional readability formulae (e.g., Siddharthan, 2004). But the use of a robust readability model to compare readability at sentence level to the best of our knowledge has not been the target of previous research. As far as we see, such use of readability models can play an important role in identifying which sentences need to be simplified and which sentential transformations would simplify the text.

We start with the construction of a robust readability model that employs a wide range of features. We first establish the validity of the model by performing cross-corpus and genre-wise evaluation and compare our performance against the results reported in the literature. We then move from the document level to the sentence level and explore the utility of readability assessment for identifying the reading level of sentences. We evaluate the sentence level performance in terms of the ability of the model to distinguish between the reading levels of original and simplified versions in the right order. To our knowledge, this is the first work which considers readability assessment as a part of text simplification and performs an evaluation in context on several independent test sets at both document and sentence level.

To summarize, the specific goals of this paper are:

a) to construct a robust readability model and establish its cross-corpus and cross-genre portability.

b) to apply this model to sentences instead of documents and to study the utility of readability models for identifying sentential simplifications provided by human authors.

The paper is organized as follows: We first survey the related work on readability and text simplification and address how this paper links these research areas. Next, we present our experimental setup, corpora and features. This is followed by a discussion of our readability model and its performance across different corpora and domains. In the next section, we discuss our experiments on sentence level readability assessment and the results. We conclude the paper with a discussion of the results and pointers to directions for future work.

RELATED WORK

Readability Assessment

Automatic assessment of the reading level of a text or a human reader has a long history that spans diverse fields of research owing to a wide spectrum of possible application scenarios. Some of them include: assessing the textbooks of language learners, providing reading material that students can understand, addressing the language needs of second language adult and child learners and providing accessible text for people with various forms of cognitive disabilities.

In its eight decade long history, researchers explored various features that contribute to readability from different points of view. Traditional readability models were based on formulae created using easily computable textual measures such as word length, sentence length and word lists (e.g., Kincaid et al., 1975; Chall & Dale, 1995; Dubay, 2006). Although traditional readability formulae have been very popular and are still being used frequently (e.g., MS Word has Flesch-Kincaid reading ease as a measure of readability), recent research (e.g., Perfetti et al., 2012) showed that systems that relied on a broad set of features instead of surface features performed the best on real life datasets like Common Core Standards texts (http://www.corestandards.org/).

Computational linguistic approaches to readability assessment generally use natural language processing tools to extract a variety of features and build machine learning models for classifying texts into different reading levels. The classes of features which were used include language models (e.g., Collins-Thompson & Callan, 2005), syntactic features (e.g., Heilman et al., 2007), coherence and cohesion features (e.g., Graesser et al., 2012), cognitively motivated features (e.g., Feng et al., 2009), features derived from second language acquisition research (Vajjala & Meurers, 2012), language specific morphological features (e.g., François & Watrin, 2011; Hancke et al., 2012), language specific semantic features (Von der Brück et al., 2008), and aspects of genre (e.g., Futagi et al., 2007). The analysis and modeling of human sentence processing difficulty in psycholinguistics (e.g., Boston et al., 2008), and the features such as surprisal and construal investigated there, essentially constitutes a further, related field of inquiry.

Compared to the considerable research into constructing readability models, there is much less research on studying the applicability of these models to real-world applications. Some of the past research focused on issues such as detecting the reading levels of searchers through their queries (Liu et al., 2004), detecting the reading level of texts suiting a target audience (e.g., Pera & Ng, 2012; Ma et al., 2012), studying the distribution of reading levels across web texts (e.g., Martin & Gottron, 2012; Vajjala & Meurers, 2013), and combining reading level assessment with topic classification (e.g., Heilman et al., 2008a; Kim et al., 2012). Traditional readability formulae were used as a measure for evaluating text simplification in the past (e.g., Siddharthan, 2004; Jonnalagadda et al., 2009). Using measures of text readability as

one of the means to pick possible syntactic transformations for text simplification started to receive attention only recently (e.g., Medero & Ostendorf, 2011; Štajner et al., 2013). Our work can be seen as continuing this line of research, shifting readability assessment's focus from documents to sentences.

Text Simplification

Research in text simplification started with the idea of splitting sentences into shorter pieces to improve the parser performance (Chandrasekar et al., 1996; Chandrasekar & Srinivas, 1997). Other early work focused on developing rule-based systems to simplify newspaper text for aphasic users (Carroll et al., 1998; Canning et al., 1999). Siddharthan (2002, 2003, 2004) developed a theory and approach to text simplification that also considered discourse structure and preservation of text cohesion in to account. More recent research in this direction primarily considered simplification as monolingual translation, using a parallel sentence-aligned corpus such as Wikipedia–Simple Wikipedia (e.g., Specia, 2010; Zhu et al., 2010; Bach et al., 2011; Coster & Kauchak, 2011; Woodsend & Lapata, 2011).

Lexical simplification was explored as an independent task both in terms of identifying as well as ranking lexical substitutes (Yatskar et al., 2010; Biran et al., 2011). The SemEval-2012 task on English lexical simplification (Specia et al., 2012) studied various features for ranking lexical substitutes by their simplicity. Although most of the work focused on English, research on other languages is starting to emerge (e.g., Spanish: Bott & Saggion, 2011; Danish: Klerke & Søgaard, 2012; Basque: Aranzabe et al., 2012; Italian: Barlacchi & Tonelli, 2013). Corpus studies to identify the nature of simplification (e.g., Petersen & Ostendorf, 2007; Bott & Saggion, 2011) and characteristic features of simplifications (e.g.,; Allen, 2009; Gasperin et al., 2009; Medero & Ostendorf, 2011; Štajner et al., 2013) have also been performed in this connection.

Despite a considerable amount of interest in text simplification in the recent past, the question what needs to be simplified is far from well-explored. The primary focus has either been on splitting the sentences or on simplifying everything as much as possible. However, in certain application scenarios, such as those providing reading assistance to learners or aiding the preparation of simplified texts for teachers or providing writing assistance, more precise suggestions on the reading level of sentences and ways to make them simpler could be very useful. Another under explored area of research is the evaluation of text simplification. It is typically done either using traditional readability formulae or using BLEU scores (in machine translation approaches). But the utility of modern readability models to evaluate simplification was not studied.

While application of readability model at the sentence level was also not explored in detail in previous research, there is some related work. Bormuth (1966) studied readability formulae and linguistic characteristics of smaller language units such as words, independent clauses and sentences by conducting cloze tests using twenty passages. He observed a good correlation between the cloze test results as indicators of comprehension difficulty and some of their linguistic variables even at a sentence level. More recently, Napoles & Dredze (2010) applied a binary document classification model trained on full documents drawn from Wikipedia and Simple Wikipedia directly to individual sentences, assuming that all sentences in Wikipedia are hard and all in Simple Wikipedia are simple (a simplifying assumption which we found to be false).

OUR APPROACH

The purpose of the research presented in the following is two-fold: to build a robust readability model that performs well in cross-corpus and cross-genre validation, and to study its utility for the task of text simplification by zooming into sentences. First, we explain the methods we used to construct our models i.e., the corpora, features and our experimental setup.

Corpora

For training and testing our readability models, we used six corpora:

WeeBit corpus: The WeeBit corpus we originally compiled for Vajjala & Meurers (2012) consists of texts at five reading levels, with 625 documents per level, covering language learners of age groups 7–16 yrs. It is a compilation consisting of two sub-corpora: WeeklyReader and BBC BiteSize. The articles primarily consist of informational news texts rewritten to suit the children belonging to different grade levels. In Vajjala & Meurers (2013), we found that readability models built on the WeeBit corpus generalized well to various web-corpora.

Common Core Standards corpus[1]: This is a corpus consisting of 168 English texts belonging to four genres that serve as exemplars for the Common Core Standards reading initiative of the U.S. education system. This corpus was introduced as an evaluation corpus for readability models in the recent past (Sheehan et al., 2010; Perfetti et al., 2012; Landauer & Way, 2012; Flor et al., 2013). We use this corpus to test our readability model and to evaluate its performance across genres. We also use this corpus to compare our model with other existing readability systems.

TASA corpus: This is a corpus consisting of about 37.000 texts annotated with their reading level in terms of DRP (Degrees of Reading Power)[2] scale assigned by Touchstone Applied Science Associates Inc. (TASA). The score typically ranges from 30–80. The corpus was created in 1995 from 6.333 textbooks, fiction and non-fiction works used in schools and colleges throughout the United States, with the aim of estimating the frequency of words at different grade levels. It consists of texts with a mean length of 250–300 words covering nine content areas: business, health, home economics, industrial arts, language arts, miscellaneous, science, social studies, and uncategorized. The corpus is widely used in Latent Semantic Analysis[3] and was used as an evaluation corpus in some of the Coh-Metrix[4] readability analyses studies (e.g., Graesser et al., 2012). We use this corpus for evaluation and to test the adaptability of our features to different topic categories.

Math Readability corpus[5]: The corpus was created by Zhao & Kan (2010) and consists of 120 Math web pages annotated with the reading level. It was created using a crowd-sourcing setup, where people were asked to evaluate the conceptual difficulty of math web pages on a scale of one to seven, with one indicating primary school level and seven advanced university level. While this conceptual difficulty labeling is not specific to linguistic complexity, we in-

[1] The texts were extracted from Appendix-B of the Common Core Standards description, excluding the items categorized as poetry (http://www.corestandards.org/assets/Appendix_B.pdf). The texts are classified into grade bands that were designated by expert evaluators.

[2] Degrees of Reading Power (DRP) program: http://textcomplexity.questarai.com/getdrp

[3] http://lsa.colorado.edu/spaces.html

[4] Coh-Metrix: http://cohmetrix.memphis.edu

[5] Math Readability Corpus: http://wing.comp.nus.edu.sg/downloads/mwc

clude it here to test the limits of a generic readability approach. Thus, we use this corpus to perform cross-corpus evaluation as well as to train a model to verify the generalizability of our feature set across different genres.

Wikipedia–Simple Wikipedia corpus: We use the sentence-aligned Wiki-SimpleWiki corpus created by Zhu et al. (2010) to verify the robustness of our readability model when we move from documents to sentences. It consists of approximately 100,000 unsimplified–simplified sentence pairs.

One Stop English corpus: This is a corpus consisting of thirty articles from onestopenglish.com that are parallel versions of ten articles at three reading levels (beginner, intermediate and advanced). The articles are originally from the Guardian newspaper and are manually rewritten by experts for teaching at three levels. We compiled this corpus by crawling some of the freely accessible articles and later aligned them manually, first by document and later by sentence.[6] We used this corpus for evaluating our readability model's performance in identifying target sentences for simplification and comparing it with the choices made by the human experts who wrote the simplified versions.

Features

We explored a wide range of features for developing our readability model. They can be broadly classified into four categories: lexical richness and POS features, syntactic complexity features, word characteristics features and surface features.

Lexical Richness and POS Features: We adapted a range of measures of lexical richness from Second Language Acquisition research, including the type-token ratio, corrected type-token ratio and measures of lexical variation (noun, verb, adjective, adverb and modifier variation). In addition, this feature set also includes the density of different parts of speech (POS) in the texts to study the relation of POS density with the overall score. The POS information was extracted using the Stanford Tagger (Toutanova & Klein, 2003). The features from this subset are adapted from Vajjala & Meurers (2012).

Syntactic Complexity Features: Syntactic complexity measures from SLA research along with other parse tree based features proved be useful for readability classification (cf. Vajjala & Meurers, 2012; 2013). We adapted the following measures for this study: mean lengths of various production units, measures of co-ordination and sub-ordination, the presence of particular syntactic structures, number of phrases of various categories, average lengths of phrases, parse tree height and number of constituents per subtree. We used the BerkeleyParser (Petrov & Klein, 2007) for generating parse trees and the Tregex (Levy & Andrew, 2006) pattern matcher to count the occurrence of various syntactic patterns.

Word Characteristics Features: While the previous two feature sets are primarily based on SLA research, we additionally explored word characteristic features hypothesizing that information about the morpho-syntactic properties of words and their psychological characteristics along with information on their age-of-acquisition may be useful for readability assessment. So, we constructed a set of features based on the information provided by two widely used psycholinguistic databases and a new database with age-of-acquisition norms for English words.

[6] http://www.onestopenglish.com is a teacher's resource website administered by MacMillan Education Group, one of the leading publishers of English teaching materials.

The Celex Lexical Database (Baayen et al., 1995) for English consists of information on the orthography, phonology, morphology, syntax and frequency for more than 50,000 English lemmas. The morphological properties of words in Celex include information about the derivational, inflectional and compositional features of the words along with information about their morphological origins and complexity. Syntactic properties of the words in Celex describe the various attributes of a word depending on its parts of speech. We used the proportion of occurrences per text of various morphological and syntactic properties of words as features (e.g., the ratio of transitive verbs, complex morphological words, and vocative nouns to the number of words that had Celex entries). Words in the document that are not included in the Celex database were ignored from this calculation. For the texts we analyzed, 40-50% of the lemmas were found in the Celex database. In all, we used the 35 morphological and 49 syntactic properties that were expressed using character or numeric codes in the Celex database as features for our task and excluded word frequency statistics and properties which consisted of word strings. More details about the morphological and syntactic properties of the lemmas can be found in the Celex user manual[7].

The MRC Psycholinguistic Database (Wilson, 1988) is a machine-readable dictionary with around 1.5 million words along with their 26 linguistic and psychological attributes. It is a freely available online resource (http://www.psych.rl.ac.uk). We used the measures of word familiarity, concreteness, imageability, meaningfulness and age of acquisition from this database as our features.

Kuperman et al. (2012) compiled a database of age-of-acquisition ratings for over 50000 English words (freely available at: http://crr.ugent.be/archives/806) through crowd sourcing. They compared the ratings with several other age-of-acquisition norms that are also accessible through the database. We included all the Age of Acquisition (AoA) ratings as features.

Finally, this feature group also includes an encoding of the number of senses per word, calculated using the MIT Java Wordnet Interface[8]. We excluded auxiliary verbs for this calculation, as they tend to have multiple senses that do not necessarily contribute to reading difficulty.

While the features based on hand-crafted lexical resources are limited by the size of the respective databases, they capture a different type of information compared to the other feature categories we study in this paper. As we will see in the later sections, some of these features indeed received high weights in the regression model, confirming that a potential lack of coverage is not a problem invalidating these features in practice. In terms of the reading level most impacted by the lack of coverage, one would assume it to mostly impact the higher reading levels given that those include less common words, which therefore are also covered less in the lexical resources. However, the features adapted from the SLA complexity literature should provide a good coverage of the properties distinguishing the more complex reading levels. Experiments we conducted on TV subtitles (Vajjala and Meurers, 2014b) confirm that the feature set is better at distinguishing the more complex levels.

Surface features: This final group consists of two traditional features, the average sentence length in words and number of sentences per document.

---

[7] The CELEX lexical database is available from LDC at: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC96L14 and Celex user manual can be consulted at: http://catalog.ldc.upenn.edu/docs/LDC96L14

[8] MIT Java-Wordnet Interface: http://projects.csail.mit.edu/jwi

General Experimental Setup

Automatic readability assessment is typically treated as a classification problem. Heilman et al. (2008) and Aluisio et al. (2010) experimented with different kinds of statistical models for readability assessment, including regression, using large feature sets. Since regression helps us in identifying reading levels on a numeric scale in a way that allows us to also identify the documents falling between levels, we also consider readability assessment as a regression problem. We considered only linear models since they are most readily interpretable.

We built regression models using two approaches: linear regression and a support vector regression algorithm, SMOReg (Sequential Minimal Optimization regression) with the default PolyKernel. The default exponent for PolyKernel in WEKA is 1, which makes it a linear kernel. Since the SMO regression (SMOReg) performed slightly better in terms of prediction error, we used it for the rest of our experiments. We relied on the WEKA machine learning toolkit (Witten & Frank, 2005) for training and testing our models.

For training and testing on the same corpus, we used Pearson correlation and Root Mean Square Error (RMSE) as performance evaluation measures. However, while performing cross-corpus evaluations, since the scales used in the various datasets are different, we used Spearman's rank correlation coefficient along with Pearson correlation.

For the final set of experiments related to identifying sentential simplifications, we report the percentage of cases where i) our system rightly identified the order of reading levels between unsimplified and simplified versions correctly, ii) where the order was reversed and iii) where no change in readability was identified although the sentence was altered/simplified. We also compared the sentence pairs using Wilcoxon's signed rank test (Wilcoxon, 1945) to verify if there really is a difference between the means of the reading levels between unsimplified and simplified versions as assigned by our readability model.

EXPERIMENT 1: DOCUMENT LEVEL READABILITY MODEL

We used the WeeBit corpus introduced above to train our primary document level readability model. Since we model readability as regression, we mapped the five reading levels in the WeeBit corpus to a scale of 1–5. We trained two regression models using the Weka toolkit: one with linear regression and one with SMO regression, both using normalized feature values. We used the entire feature set introduced above, which consists of a total of 151 features. In a 10 fold cross validation experiment, linear regression had a correlation 0.92 and an RMSE of 0.57, whereas SMOReg had a correlation of 0.92 and an RMSE of 0.53.

As a baseline comparison, we trained a model with only the traditional surface features (average sentence length and number of sentences per document). The model achieved a correlation of 0.6 and an RMSE of 1.13 for Linear Regression and a correlation of 0.6 and RMSE of 1.16 for SMOReg. Clearly, the model with the full linguistic feature set performs much better than a model using only surface features.

To facilitate a concise discussion, for the following experiments we will only report the SMOReg results. Unlike Linear Regression, SMOReg does not involve a feature selection step. However, it assigns very low to zero weights to features that do not contribute to the model. Table 1 shows the five features with the highest positive and negative weights as assigned by the SMOReg model. Table 2 illustrates some of the features that were assigned very low weights by the model.

Table 1: Top 10 Features with high weight in the WeeBit corpus trained SMOReg model

Feature description	Weight	Feature class (source)
Word Familiarity	+0.8215	Word Characteristics (MRC Psycho-linguistic DB)
Age-of-acquisition	+0.73	Word Characteristics (Kuperman et al., 2012)
Modifier variation	-0.61	Lexical Richness and POS
Co-ordinate phrases per t-unit[9]	+0.5979	Syntactic Complexity
Proportion of words whose morphology is irrelevant (e.g., words like "nowadays")	+0.56	Word Characteristics (Celex)
Dependent clauses per t-unit	+0.5451	Syntactic Complexity
Proportion of verbs	-0.53	Lexical Richness and POS
Proportion of pronouns	+0.5097	Lexical Richness and POS
Proportion of countable Nouns	+0.4694	Word Characteristics (Celex)
Noun variation	+0.4672	Lexical Richness and POS

Table 2: Some features with very low weight in the WeeBit corpus trained SMOReg model

Feature description	Weight	Feature class (source)
Word concreteness	+0.0001	Word Characteristics (MRC Psycholinguistic DB)
Proportion of words that can be predicative adjectives	+0.0002	Word Characteristics (Celex)
Pavio meaningfulness of a word	+0.0017	Word Characteristics (MRC Psycholinguistic DB)
VPs/Sentence	+0.0152	Syntactic Complexity
Proportion of interjections	-0.0024	Lexical Richness and POS
Proportion of proper nouns	-0.0038	Lexical Richness and POS
Mean length of a clause	-0.0216	Syntactic Complexity
Co-ordinate phrases per clause	-0.0509	Syntactic Complexity
Proportion of expressive adverbs	0	Word Characteristics (Celex)
Proportion of demonstrative pronouns	0	Word Characteristics (Celex)

Among the top features we find lexical, syntactic, and word characteristic features, with age of acquisition and word familiarity being at the top, followed by the variability of the modifier use and several syntactic complexification aspects, such as the use of coordinate phrases and dependent clauses per t-unit. The uninformative features also include word specific features such as concreteness and meaningfulness of a word and syntactic complexification aspects such as the mean length of a clause. The heterogeneous nature of the features that are found to be useful for readability classification supports our strategy to explore a rich linguistic feature basis on which to build readability models.

[9] The definition for t-unit follows the conventional definition given in Hunt (1970), which is used in L2 writing studies. It is defined as: "one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it.''

A model with so many features can be prone to overfitting. Although performing a 10-fold cross-validation addresses this issue to some extent, establishing that the model performs well on cross-corpus evaluations would strengthen the claim that the model does not overfit. We therefore tested our SMOReg model on three other external corpora explained in the Corpora section: the Common Core corpus, TASA corpus and Math readability corpus. Table 3(a) reports the performance of our model with these three corpora in terms of Pearson correlation (r) and Spearman's rank correlation (ϱ). Table 3(b) reports the performance of another model that takes only surface features into account on these test sets.

Table 3(a): WeeBit model with all features, tested on other standard readability corpora

Dataset	Size	"Grade" scale	Pearson corr.	Spearman's rank corr. ϱ
CommonCore	168 documents	2–12	0.61	0.69
TASA Corpus	~37k documents	10–100	0.83	0.86
Math Readability	120 documents	1–7	0.19	0.29

Table 3(b): WeeBit model with surface features, tested on other standard readability corpora

Dataset	Pearson corr.	Spearman's rank corr. ϱ
CommonCore	0.40	0.50
TASA Corpus	0.56	0.72
Math Readability	-0.09	-0.13

The model with all features generalized well to Common Core (r=0.6 and ϱ=0.69) and TASA (r=0.83 and ϱ=0.86), but was not so effective on the Math Readability Corpus (r=0.19 and ϱ=0.29). This may be due to the nature of the ratings in the Math Readability corpus, for which the raters were encoding the conceptual difficulty of the problem, not the linguistic complexity of its formulation. Since our model only considers features related to linguistic complexity, the fact that the model does not perform well on the Math Readability corpus probably only indicates that such conceptual mathematical complexity ratings and linguistic complexity are distinct.

Looking at the results on the Common Core Standards texts and the TASA corpus, to put the results of our readability model into context, we need a frame of reference. Perfetti et al. (2012) compared the performances of six proprietary text difficulty metrics on five test sets. Since the Common Core standards dataset is a part of this study, it gives us a way to compare our system performance against seven proprietary systems. The systems compared in this study are:

- Lexile (Metametrics, http://www.lexile.com)
- ATOS (Renaissance Learning, http://www.renlearn.com/atos)
- DRP analyzer (Questar Assessment Inc., http://www.questarai.com/Products/DRPProgram)

- REAP (Carnegie Mellon University, http://reap.cs.cmu.edu)
- SourceRater (Educational Testing Service, cf. Text Evaluator: https://texteval-pilot.ets.org/TextEvaluator)
- Pearson Reading Maturity Metric (Pearson Knowledge Technologies, http://www.readingmaturity.com)
- Coh-Metrix (University of Memphis, http://cohmetrix.memphis.edu)

More details on the individual systems can be found in Perfetti et al. (2012). Complementing this study, Flor et al. (2013) also used the grade level annotations of the Common Core standards test set to compare the *Lexical Tightness* measure they introduce, the Flesch-Kincaid Grade level formula, and the text length as a surface baseline. While Perfetti et al. (2012) report their comparison in terms of Spearman's rank correlation ϱ, Flor et al. (2013) provide the Pearson correlation. To enable comparison with all of them, we report both of the measures for our models. Table 4 lists the performance of various systems on Common Core data as reported in the two papers and contrasts them with the results for our models. Since the Coh-Metrix performance was only reported graphically in Perfetti et al. (2012), the correlation values listed in Table 5 are approximate.

Table 4: A comparison of various readability models on Common Core Standards texts

System	**Pearson corr.**	**Spearman's rank corr. ϱ**	**Source of information**
REAP	-	0.543	Perfetti et al. (2012)
ATOS	-	0.592	
DRP	-	0.527	
Lexile	-	0.502	
Reading Maturity	-	0.690	
SourceRater	-	0.756	
Lexical Tightness	-0.441**	-	Flor et al. (2013)
Text Length	0.360**	-	
FKGrade Level	0.487*	-	
Our Model	**0.61**	**0.69**	
Using only surface features	**0.40****	**0.54**	

** = p<0.01, * = p<0.05

Table 5: Coh-Metrix performance with Common Core Standards texts (Perfetti et al. 2012)

Coh-metrix Dimension	Spearman's corr. ϱ with CommonCore data
Narrativity	~ -0.08
Referential Cohesion	~ -0.2
Syntactic Simplicity	~ -0.45
Word Concreteness	~ -0.4
Deep Cohesion	~ +0.08

As can be seen from the table, our readability model with all features performs on par with the best performing systems in the study and is outperformed only by the SourceRater system developed by Educational Testing Service (ETS), which uses a combination of a cognitively oriented feature set and psychometric methods. In terms of Pearson correlation, our model performs better than other systems that reported the measure. Also note that our approach correlated well both with a formula-based corpus (TASA) linked to relatively shallow features as well as with a corpus that was created manually by experts (Common Core Standards texts).

Generalizability of the feature set

It is clear from the above results that the readability *model* trained on the WeeBit corpus generalizes well across several standard datasets. Another aspect we wanted to investigate is the generalizability of the *feature set* used. In other words, building a model on WeeBit and testing it on other datasets establishes that the model (consisting of the features and their weights) is generalizable to a certain extent. However, how informative are the observations captured by the feature set in general? To answer this question, we trained and tested models with the same feature set for several corpora using 10-fold cross-validation. Table 6 presents the performance of our feature set on the different corpora.

Table 6: 10-fold cross validation results for models with all our features on the different data sets

Corpus	Description	Pearson corr.	RMSE
WeeBit	625 documents per level, with 5 levels	0.92*	0.53
CommonCore	168 documents, scale of 2–12	0.59*	2.69
Math Readability Corpus	120 documents, scale of 1–7	0.51*	1.73
TASA Corpus	~37k documents, scale of 28–110	0.97*	1.77

* = $p<0.0001$

The correlations and RMSEs for models trained on CommonCore and Math Readability datasets, though statistically significant ($p<0.001$), are low in comparison with the WeeBit dataset based model. This could be because of the fact that we are dealing with much smaller datasets that in addition make use of a larger scale range, i.e., a sparse data problem arising from few instances for a large set of possible values. With the TASA corpus, which was huge and had a wide score range, the model again trained well, despite the fact that we do not con-

sider any features encoding the measures used in their formula (except sentence length)[10]. Apart from the above experiments, this feature set was also useful in building efficient models for classifying television programs into suitable age-groups, based on their subtitles (Vajjala & Meurers, 2014b). In sum, given enough training data, our readability feature set can be used to build a good model for a wide range of datasets annotated with reading-level judgments.

EXPERIMENT 2: GENRE EFFECTS IN READABILITY ASSESSMENT

Sheehan et al. (2009, 2010) studied the effect of text genre on readability assessment and established that the genre of a text influences readability assessment. Flor et al. (2013) also showed there was a difference in the performance of their readability model across different genres. Hence, to determine the genre dependence of our model, we studied the genre-wise performance of the WeeBit model on the Common Core Standards data. We chose this dataset as it includes genre annotations, while at the same time ensuring comparability with previous research (Flor et al., 2013). Table 7 presents the Pearson and Spearman correlations for the different genres of text in this dataset.

Table 7: Performance of the WeeBit model on the genre-specific subsections of the Common Core data

Genre	# Docs	Pearson corr.	Spearman's corr. ϱ
Speech	13	0.41	0.35
Mix	44	0.61	0.69
Literature	56	0.44	0.51
Informative	55	**0.71**	**0.76**

Since the WeeBit dataset primarily consists of non-fiction articles on news events, as expected the model performs best for informational texts (ϱ=0.76). The performance is at the level of the average performance of the best commercial system SourceRater on the overall Common Core Standards data set (Table 4).

Considering that we next turn to sentence level experiments on informational texts (Wiki-Simple Wiki and OneStopEnglish sentences), we can also observe that our readability model can be safely applied to such data without worrying about a genre effect.

To verify whether the features themselves facilitate building good models across genres, we trained multiple regression models using the TASA corpus, which includes a genre annotation for most of its texts. Table 8 summarizes the results of the 10-fold cross-validation experiments for the different genres in terms of Pearson correlation and RMSE.

[10] To be sure, we also trained and tested a model not using the sentence length feature and indeed found that it resulted in the exact same performance on the TASA corpus.

Table 8: Genre-specific readability models with TASA corpus, all features

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

All the models resulted in a high correlation and low RMSE. Hence, as the results indicate, the feature set is sufficiently general and informative as information source when building genre-specific readability models.

Our experiments so far have established that our readability model and the feature set generalize well across corpora and across genres. Now, we will move on to answer the question what happens when we move from documents to sentences to be able to pursue the overarching question: How good is the readability model in identifying target sentences for text simplification?

EXPERIMENT 3: MOVING FROM DOCUMENTS TO SENTENCES

We applied the WeeBit readability model to two datasets annotated with sentential readability levels to explore whether the model is capable of distinguishing individual sentences in terms of their readability. To the best of our knowledge, the performance of document-level readability models when applied to the sentence level has not yet been investigated, with the exception of Bormuth (1966). While Bormuth (1966) performed experiments using cloze tests on a small dataset of 20 test passages, we perform experiments using a feature-rich model on the large Wikipedia–Simple Wikipedia dataset with sentence level readability annotation and on the OneStopEnglish corpus consisting of sentences simplified across three reading levels.

Reading level distribution for Wikipedia–Simple Wikipedia sentences:

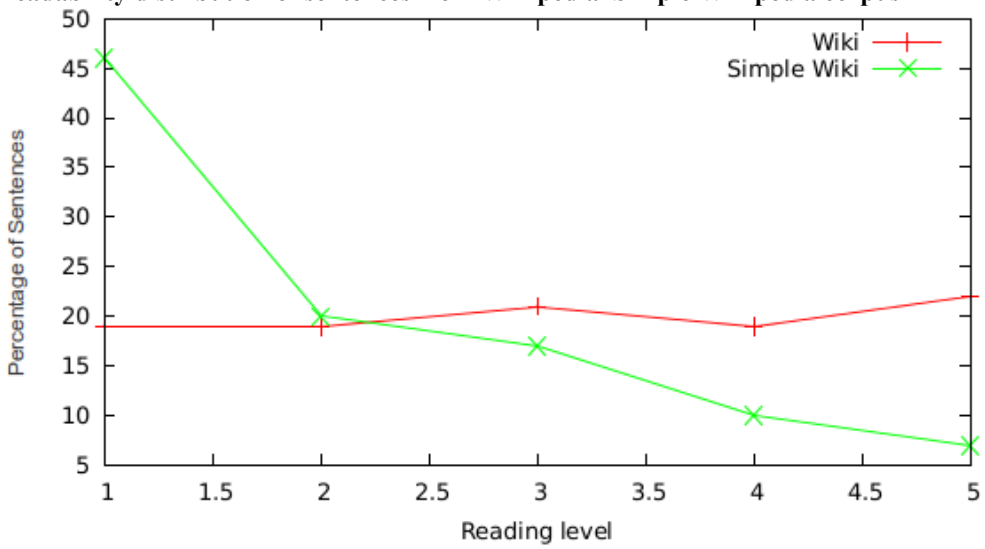
Our first test dataset consists of 100k sentence pairs of the sentence-aligned Wikipedia–Simple Wikipedia corpus. This dataset was produced by Zhu et al. (2010) based on parallel Wikipedia–Simple Wikipedia article pairs written by human users that were aligned at the sentence level. We excluded the sentences that remained unchanged in both versions. As readability classes, we used 1 for Simple Wikipedia and 2 for Wikipedia.

We first explored the possibility of constructing a binary classification model for “simple” vs. “not simple” sentences. For this, we sampled the training set so that it contains equal number of instances per category (i.e., the random baseline for binary classification will be 50%). We experimented with different classification algorithms (SMO, Logistic Regression, Random Forest), with different feature subsets, and with varying training set sizes. However, the highest classification accuracy we achieved for this binary classification task was only 66% (with SMO). With sentence length alone, we achieved a classification accuracy of 60% (with SMO). To determine the cause for this poor performance of the feature set that performed so

well at the document level, we investigated the distribution of reading levels of the Wikipedia–Simple Wikipedia sentences as determined by our WeeBit readability model. In other words, we ran our WeeBit readability model on each of the Wikipedia–Simple Wikipedia sentences as a test set.

Figure 1 shows the distribution of Wikipedia and Simple Wikipedia sentences according to the reading levels assigned by the WeeBit model.

Figure 1: Readability distribution of sentences from Wikipedia–Simple Wikipedia corpus



The figure confirms the expectation that Simple Wikipedia includes more sentences at lower reading levels, though a smaller percentage of sentences is also determined to be at higher reading levels. On the other hand, the distribution of the reading level in the regular Wikipedia is relatively uniform, with sentences belonging to all reading levels.

The overall distribution of sentences raises the question whether a binary, absolute classification of sentences into simple as label of all sentences from Simple Wikipedia and difficult for all the sentences in the regular Wikipedia is meaningful for this data set. While we, of course, need to take the picture in Figure 1 with a grain of salt, given that they only reflect labeling performed by the WeeBit model, it seems clear that some of the sentences in the Wikipedia set are simpler than some of the sentences in the Simple Wikipedia set. In light of this, the disappointing results of the classification experiment mentioned above at least partly results from the fact that binary classification does not do justice to the nature of the Wikipedia–Simple Wikipedia data set. In place of an absolute, binary classification of sentences, it would be more meaningful for this data set to determine a relative labeling, which for each aligned sentence pair determines which of the sentences is simpler and which is harder.

The fact that the Wikipedia and Simple Wikipedia from which the aligned sentences were collected are collaboratively prepared resources may also be reflected in the nature and consistency of the simplification performed. Hence, we next tested our readability model with another corpus that consists of articles that were simplified by human experts to obtain versions at three reading levels for each article.

An evaluation based on the OneStopEnglish corpus

The OneStopEnglish corpus consists of 30 articles consisting of 10 parallel articles at three reading levels each: beginner, intermediate and advanced. Human experts simplified the original articles to obtain the beginner and the intermediate versions of each. Table 9 presents some characteristic properties of the corpus.

Table 9: OneStopEnglish corpus description

Reading level	Avg. number of sentences per document	Avg. sentence length (# words)
Beginner	38.6	14.7
Intermediate	41.6	16.8
Advanced	45.7	18.8

We first applied our readability model to this corpus at a document level in order to be able to compare the document level performance to the sentence level analysis later. A comparison between two documents (or sentences) in terms of the reading level assigned by our model can result in three cases:

1. The model identifies the relationship between them correctly; that is, the human-simplified version is assigned a lower reading level than the version from which it was simplified.
2. The model puts both the simplified and unsimplified versions at the same reading level.
3. The model identifies a reverse relation between them; that is, the human simplified version is identified as being at a higher reading level than its unsimplified version.

Only the third case clearly represents an error. How to interpret the second case depends on the granularity of the analysis that one wants to be able to perform. We thus do not count it as an error and proceed to compare the first two cases to the third case, first at the document-level and then at the sentence-level.

In the document level analysis, in 62% of the cases, our WeeBit trained model with 10-fold CV correctly identified a drop in the reading level (*advanced → intermediate*, *intermediate → elementary*, or *advanced → elementary*). In 36% of the cases, the model did not identify a difference in the reading levels, and in 2% of the cases, the model identified the simplified version to be at a higher reading level than the text from which it was simplified. That is, the model's overall accuracy (in terms of the above-explained error avoidance) was 98% (62+36). The model performed even better in terms of identifying *advanced → elementary* simplification, correctly determining 90% of the simplified documents to be at a lower level than the advanced versions and 10% to be at the same reading level. No documents were rated in the reverse order for this case (100% accuracy).

To perform a similar analysis at a sentence level, we need a sentence-level alignment across the three levels. Thus we manually aligned the three versions of each text at the sentence level. Some sentences remained unchanged in a simpler version, and some sentences were removed completely. Table 10 shows some statistics on the percentage of sentences that remained unchanged or were removed completely when simplification was performed across different reading levels.

Table 10: OneStopEnglish corpus: Unchanged and removed sentences across simplifications performed

	Adv. → Inter.	*Inter. → Elementary*	*Adv. → Elementary*
sentences unchanged	39.0%	30.8%	21.3%
sentences removed	14.6%	11.9%	17.5%

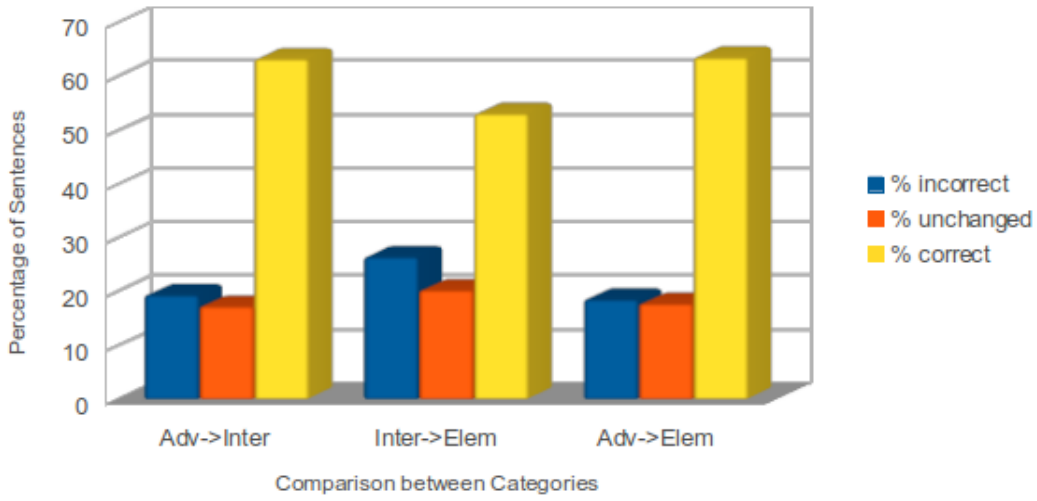
On average, around 30% of the sentences were removed by the human editors when a document was rewritten to a lower reading level, and about 15% of the sentences remained unchanged. There also were instances where simplification resulted in sentences being split into two. Overall sentences were mostly simplified by replacing, removing, or rewriting parts of the sentence rather than splitting or removing the sentence completely. Table 11 shows an example sentence across the three reading levels together with the reading levels assigned by our model to these sentences.

Table 11: An example sentence from three reading levels, in the OneStopEnglish corpus

Sentence	**Actual reading level**	**Reading level assigned by our model**
In Beijing, mourners and admirers made their way to lay flowers and light candles at the Apple Store.	*Advanced*	3.6
In Beijing, mourners and admirers came to lay flowers and light candles at the Apple Store.	*Intermediate*	2.3
In Beijing, people went to the Apple Store with flowers and candles.	*Elementary*	1.0

We studied the ability of our model to distinguish between sentences across reading levels by considering only those sentences that were altered during the process of rewriting. In total, based on the 10 articles in three versions we analyzed approximately 1,200 sentences annotated with three reading levels. We manually aligned the corresponding sentences and for the three types of simplification pairs (*advanced → intermediate*, *intermediate → elementary*, or *advanced → elementary*) studied the percentage of correctly identified reading level orders, sentence pairs where the model did not identify a difference, and finally those pairs where the model misidentified the order. Figure 2 summarizes the results for the three simplification pairs in the form of a bar chart.

Figure 2: Performance of WeeBit model on OneStopEnglish Corpus Sentences



The figure shows that excluding the unchanged and removed sentences, on average the model identified the difference in reading levels after sentential simplification in the correct order for about 60% of the cases in all the three possible simplification pairs and identified no difference in 18% of the cases. That is, the accuracy of this model is 78% (60+18). Most of the 18% of cases where no difference was identified consist of lexical simplifications where only a word or two was replaced by another word. This is missed by our readability model given that it does not include features that would facilitate the identification of lexical simplification, apart from modeling the morphological and part-of-speech properties of the words and their age-of-acquisition. Adding lexical frequency information should help eliminate this weakness.

Compared to the document level results on this corpus (98–100% accuracy), the sentence level model achieved an accuracy of only 78%. However, there are no established benchmarks against which we could compare the performance of our model on sentence level data. As a baseline model for comparison, we trained sentence-level classification models using SMO algorithm, with sentence length as a feature on the OSE corpus that was split into sentences i.e., we trained a model in which the training instances are feature vectors of individual sentences belonging to three categories – beginner, intermediate and advanced. We used 10 fold CV as the evaluation method.

We also investigated the case where we ignored any difference in the reading levels of simplified and unsimplified versions that was <0.5. Table 12 summarizes the results.

Table 12: Sentences identified when the difference between reading levels of sentences across levels <0.5

	Adv. → Inter.	***Inter. → Elementary***	***Adv. → Elementary***
correctly identified sentences	29.7%	28.8%	19.7%
wrongly identified sentences	47.3%	44.0%	46.3%

Where the model failed to predict the expected reading order, more than 40% of the cases had a prediction difference of less than 0.5 between simplified and unsimplified versions. However, in the cases where the model predicted the order correctly, the difference was <0.5 in only about 26% of the cases. Clearly, the size of the difference in reading level determined by the model is a relevant aspect that deserves to be investigated further in future work.

To assess if there is a difference in the mean ranks of matched sentence pairs in terms of their reading level, we performed the Wilcoxon signed-rank test on the three simplification pairs: *advanced → intermediate*, *intermediate → elementary* and *advanced → elementary*. The Wilcoxon signed rank test shows that our model assigned higher reading levels to sentences in advanced level (median=2.95) than those at intermediate level (median = 2.15) ($z=7.82$, $p<0.0001$, $r=0.57$). For the *intermediate* (median=2.5) → *elementary* (median=1.9) sentence pairs, the test again rejected the null hypothesis that median difference between the pairs is zero ($z=5.34$, $p<0.0001$, $r=0.39$). The test also rejected the null hypothesis for the third sentence pair of *advanced* (median=3.1) → *beginner* (median=1.9) ($z=8.89$, $p<0.0001$, $r=0.62$).

Comparing in terms of accuracy again, *advanced → intermediate* (80.6%) and *advanced →* elementary (81.4%) pairs performed better than *intermediate → elementary* (73.5%) pair (Figure 2). Some of this may stem from the fact that the sentences at the higher reading levels are longer, thus providing more information for most of the features. Under this perspective, the drop in accuracy for the more elementary, shorter sentences then is rooted in the same data

sparsity issues causing the drop in accuracy for the sentence-level approach compared to the performance on texts.

Interestingly, some sentences that our model identified as a higher reading level remained unchanged throughout the three versions. While some of those complex sentences may have been overlooked by the human editors, others probably express ideas for which the language does not offer simpler ways to express them (without rewriting a broader passage). Although our model would still highlight such sentences, in an interactive system it would be possible to leave them unchanged if the simplifying author decides that no further simplification is possible in this context. Given that the readability model places texts on a scale, it also is possible to use this scale to determine for which sentences automatic simplification is attempted and which are highlighted as targets for manual simplification.

Concluding this analysis of sentence level simplifications, using two test sets we showed that the readability model can successfully identify distinctions between manually simplified versions of sentences with an accuracy of about 60% (70–80% considering the pairs where the reading level was identified to be the same) depending on the degree of simplification (e.g., *advanced → intermediate* vs. *advanced → elementary*). Although the accuracies could be higher, these initial experiments support the research perspective that readability models can be used to identify differences in reading levels at a sentence level.

CONCLUSIONS

In this paper, we explored the utility of readability assessment in the context of evaluating sentence level text simplification. We addressed the issue by considering two research questions: a) Does the document level readability model and its features generalize well across various corpora and genres? b) If they do, will the model also work as well at a sentence level and identify the distinctions between different levels of sentential simplification?

We constructed a readability model using a wide range of features that take the lexical and syntactic properties of the text into account. To verify the validity of the model, we tested its working on three corpora, created in different ways: The Common Core corpus was created as a part of the Common Core Standards Initiative by teachers and experts in educational research in the United States. The TASA corpus was created using the Degrees of Reading Power (DRP) reading scale formula that primarily considers only traditional metrics like word length, sentence length and difficult words in to account. The third corpus, containing Math readability annotations, was created by crowd sourcing.

Our readability model proved to be effective in a cross-corpus evaluation with the first two test corpora and displayed comparable performance with commercially available systems (Table 4 and 5). Although the trained model does not generalize well in cross-genre evaluation (Table 7), the features themselves generalized to various genres and succeeded in building effective genre specific readability models (Table 8). These experiments established that our readability approach is fairly generalizable and not limited to our primary corpus.

In a second step, we applied this model at a sentence level instead of at the document level. We first tested the model on the Wikipedia–Simple Wikipedia sentence aligned data and observed that the model identified a wide range of reading levels even within these categories. While the percentage of sentences belonging to a higher reading level was much less in Simple Wikipedia sentences, the Wikipedia sentences showed a relatively uniform distribution

across all reading levels (Figure 1).

We applied the same model to a corpus of documents from OneStopEnglish.com, which contains articles manually simplified by experts into three reading levels. We manually aligned the sentences across reading levels and studied the accuracy of the model to identify the differences in reading levels between simplified and unsimplified sentences. Our analysis showed that our model successfully identifies the differences between original and simplified versions more than 60% of the time, but identified no difference in 18% of the cases, and rated the simplified version to be of higher reading level than unsimplified version in 22% of the cases. This analysis showed that readability models such as the ones we developed can meaningfully serve as tools for identifying target sentences for simplification. The experiment also confirms that readability models may be relevant for evaluating simplification.

Future work

Our current focus is on improving the sentential reading level prediction. A data-driven approach to feature engineering also may be useful in improving the reading level prediction accuracy in comparison with the human judgments. In terms of the readability model itself, we intend to study the cross-corpus and cross-domain adaptability in terms of the most predictive features across different models and observe how feature selection affects these models.

The next step will target the identification of the nature of the simplifications (e.g., splitting, lexical replacement, paraphrasing) and the respective articulation points. In this direction, we also plan to explore the utility of building a model that enables us to identify the cases in which multiple transformations are needed.

Acknowledgements

We would like to thank the anonymous reviewers for their very detailed and helpful comments. We would also like to thank Michael Flor for making the cleaned version of the Common Core Standards resources available. Our research was funded by the LEAD Graduate School (GSC 1028, http: //purl.org/lead), a project of the Excellence Initiative of the German federal and state governments, and the European Commission's 7th Framework Program under grant agreement number 238405 (CLARA).

REFERENCES

Allen, D. (2009). Using a corpus of simplified news texts to investigate features of the intuitive approach to simplification. Proceedings of the Corpus Linguistics Conference. 585-599.

Aluisio, S., Specia, L., Gasperin, C., & Scarton, C. (2010). Readability Assessment for Text Simplification. Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics. 1-9.

Aranzabe, M.J., de Ilarraza, A.D., & Gonzalez-Dios, I. (2012). First Approach to Automatic Text Simplification in Basque. Proceedings of the First workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA). 1-8.

Baayen, R.H., Piepenbrock, R. & Gulikers, L. (1995). The CELEX lexical database (CD-ROM). http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html.

Bach, N., Gao, Q., Vogel, S., & Waibel, A. (2011). TriS: A Statistical Sentence Simplifier with Log-linear Models and Margin-based Discriminative Training. Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP). 474-482.

Barlacchi, G., & Tonelli, S. (2013). ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing). 476-487.

Biran, O., Brody, S., & Elhadad, N. (2011). Putting it Simply: a Context-Aware Approach to Lexical Simplification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT). 496-501.

Bormuth, J.R. (1966). Readability: A new approach. Reading Research Quarterly, 1(3), 79-132.

Boston, M.F., Hale, J.T., Patil, U., Kliegl, R., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. Journal of Eye Movement Research, 2, 1-12.

Bott, S., & Saggion, H. (2011). Spanish Text Simplification: An Exploratory Study. Proceedings of the 27th Conference of the Spanish Society for Natural Language Processing. 87-95.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology. 7-10.

Canning, Y., Tait, J., Archibald, J., & Crawley, R. (1999). Cohesive Generation of Syntactically Simplified Newspaper Text. Proceedings of the Third International Workshop on Text, Speech and Dialogue. 145-150.

Chall, J.S., & Dale E. (1995). Readability Revisited: The New Dale-Chall Readability Formula. Brookeline Books.

Chandrasekar, R., Doran C., & Srinivas, B. (1996). Motivations and Methods for Text Simplification. In Proceedings of the 16th Conference on Computational linguistics (COLING). 1041-1044.

Chandrasekar, R., & Srinivas, B. (1997). Automatic Induction of Rules for Text Simplification. Knowledge Based Systems 10(1997). 183-190.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. Journal of the American Society for Information Science and Technology, 56. 1448-1462.

Coster, W., & Kauchak, D. (2011). Learning to Simplify Sentences Using Wikipedia. Proceedings of the Workshop on Monolingual Text-To-Text Generation. 1-9.

Crossley, S.A., Dufty, D.F., McCarthy, P.M., & McNamara, D.S. (2007). Toward a new readability: A mixed model approach. Proceedings of the 29th annual conference of the Cognitive Science Society. 197-202.

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011). READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies. 73-83.

Dubay, W.H. (Ed.). (2006). The Classic Readability Studies. Costa Mesa: Impact Information.

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively Motivated Features for Readability Assessment. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009). 229-237.

François, T., & Watrin, P. (2011). On the Contribution of MWE-based Features to a Readability Formula for French as a Foreign Language. Proceedings of Recent Advances in Natural Language Processing (RANLP). 441-447.

Futagi, Y., Kostin, I.W., & Sheehan K.M. (2007). Reading Level Assessment for Literacy and Expository Texts. Proceedings of the 29th Annual Meeting of the Cognitive Science Society. 1853.

Gasperin, C., Specia, L., Pereira, T.F., & Aluisio, S.M. (2009). Learning When to Simplify Sentences for Natural Text Simplification. Encontro Nacional de Inteligência Artificial (ENIA-2009). 809-818.

Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2012). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. Educational Researcher, 40 (5), 223-234.

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability Classification for German using Lexical, Syntactic and Morphological Features. Proceedings of the 24th International Conference on Computational Linguistics (COLING). 1063-1080.

Heilman, M., Collins-Thompson K., Callan, J., & Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In Proceedings of the Human Language Technologies Conference (HLT). Association for Computational Linguistics. 460-467.

Heilman, M., Collins-Thompson K., & Eskenazi, M. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics. 71-79.

Heilman, M., Zhao, L., Pino, J., & Eskenazi, M. (2008). Retrieval of Reading Materials for Vocabulary and Reading Practice. Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA3). 80-88.

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? TESOL Quarterly, 4, 195–202.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C., & Gonzalez, G. (2009). Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (*NAACL* HLT). 177-180.

Kim, J.Y., Collins-Thompson, K., Bennett, P.N., & Dumais, S.T. (2012). Characterizing web content, user interests, and search behavior by reading level and topic. Proceedings of the fifth ACM International Conference on Web search and data mining (WSDM). 213-222.

Kincaid, J.P., Fishburne Jr., R.P., Rogers, R.L., and Chissom, B.S. (1975). Derivation of new Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Research Branch Report. Naval Technical Training Command. 8-75.

Klebanov, B.B., Knight, K. & Marcu, D. (2004). Text Simplification for Information-Seeking Applications. On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science, 735-747.

Klerke, S. & Søgaard A. (2012). Dsim, a Danish parallel corpus for text simplification. Proceedings of Language Resources and Evaluation Conference (LREC). 4015-4018.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, Marc. (2012). Age-of-acquisition ratings for 30,000 English words. Behavior Research Methods, 44, 978-990.

Landauer, T. & Way, D. (2012). Improving text complexity measurement through Reading Maturity metric. Annual Meeting of the National Council on Measurement in Education. http://www.pearsonassessments.com/hai/images/tmrs/Word_Maturity_and_Text_Complexity_NCME.pdf

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. 5th International Conference on Language Resources and Evaluation (LREC). 2231-2234.

Liu, X., Croft, W.B., Oh, P., & Hart, D. (2004). Automatic recognition of reading levels from user queries. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 548-549.

Ma., Y., Fosler-Lussier, E., & Lofthus, R. (2012). Ranking-based readability assessment for early primary children's literature. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 548-552.

Martin, L., & Gottron., T. (2012). Readability and the Web. Future Internet, 4, 238-252.

Medero, J., & Ostendorf, M. (2011). Identifying Targets for Syntactic Simplification. Proceedings of the International Workshop on Speech and Language Technology in Education (SLaTE 2011). http://ssli.ee.washington.edu/people/jmedero/publications/2011_slate.pdf

Napoles, C., & Dredze, M. (2010). Learning simple Wikipedia: a cogitation in ascertaining abecedarian language. Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids. 42-50.

Pera, M.S., & Ng, Y-K. (2012). BReK12: A Book Recommender for K-12 Users. The 35th International ACM SIGIR conference on research and development in Information Retrieval (SIGIR), 1037-1038.

Petersen, S.E., & Ostendorf, M. (2007). Text Simplification for Language Learners: A Corpus Analysis. Proceedings of Speech and Language Technology for Education (SLaTE). http://www.cs.cmu.edu/~max/mainpage_files/files/SLaTE07_Petersen_Text_Simplification.pdf

Petrov, S., & Klein, D. (2007). Improved Inference for Unlexicalized Parsing. Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. 404-411.

Sheehan, K.M., Kostin, I. & Futagi, Y. (2009). When Do Standard Approaches for Measuring Vocabulary Difficulty, Syntactic Complexity and Referential Cohesion Yield Biased Estimates of Text Difficulty? Proceedings of the 30th Annual Meeting of the Cognitive Science Society. 1978-1983.

Sheehan, K.M., Kostin, I., Futagi, Y., & Flor, M. (2010). Generating Automated Text Complexity Classifications That Are Aligned with Targeted Text Complexity Standards. ETS Research Report, RR-10-28.

Siddharthan, A. (2002). An Architecture for a Text Simplification System. Proceedings of the Language Engineering Conference (LEC). 64-71.
Siddharthan, A. (2003). Preserving Discourse Structure when Simplifying Text. Proceedings of the European Natural Language Generation Workshop (ENLG). 103-110.
Siddharthan, A. (2004). Syntactic simplification and text cohesion. PhD Thesis, University of Cambridge.
Specia, L. (2010). Translating from complex to simplified sentences. Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR'10). 30-39.
Specia, L., Jauhar, S.K., & Mihalcea, R. (2012). SemEval-2012 task 1: English Lexical Simplification. Proceedings of the 6th International Conference on Semantic Evaluation (SemEval). 347-355.
Štajner, S., Drndarevic, B., & Saggion, H. (2013). Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing). 17 (2). 251-262.
Toutanova, K., & Klein, D. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
Vajjala, S., & Meurers, D. (2012). On Improving the Accuracy of Readability Classification. In Proceedings of the seventh workshop on Innovative use of NLP for Building Educational Applications (BEA7). Association for Computational Linguistics. 163-173.
Vajjala, S., & Meurers, D. (2013). On The Applicability of Readability Models to Web Texts. In Proceedings of the second workshop on Predicting and Improving Text Readability for target reader populations (PITR). Association for Computational Linguistics. 59-68.
Vajjala, S., & Meurers, D. (2014a). Assessing the relative reading level of sentence pairs for text simplification. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Gothenburg, Sweden. Association for Computational Linguistics.
Vajjala, S., & Meurers, D. (2014b). Exploring Measures of "Readability" for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs. In Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations. Gothenburg, Sweden. Association for Computational Linguistics.
Vor der Brück, T., Hartrumpf, S., & Helbig, H. (2008). A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators. Informatica, 32, 429-435.
Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics Bulletin, 1(6), 80-83.
Wilson, M.D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. Behavioral Research Methods, Instruments and Computers, 20, 6-11.
Woodsend, K., & Lapata, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 409-420.
Wubben, S., van den Bosch, A. & Krahmer, E. Sentence Simplification by Monolingual Machine Translation, Proceedings of ACL 2012. 1015-1024.
Yatskar, M., Pang, B., Danescu-Niculescu-Mizil C., & Lee, L. (2010). For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. Proceedings of NAACL-HLT. 365-368.
Zhao, J., & Kan, M-Y. (2010). Domain-specific iterative readability computation. Proceedings of the 10th annual joint conference on Digital libraries. 205-214.
Zhu, Z., Bernhard, D., & Gurevych, I. (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. Proceedings of The 23rd International Conference on Computational Linguistics (COLING). 1353-1361.