# On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition

**Sowmya Vajjala**
Seminar für Sprachwissenschaft
Universität Tübingen
sowmya@sfs.uni-tuebingen.de

**Detmar Meurers**
Seminar für Sprachwissenschaft
Universität Tübingen
dm@sfs.uni-tuebingen.de

## Abstract

We investigate the problem of readability assessment using a range of lexical and syntactic features and study their impact on predicting the grade level of texts. As empirical basis, we combined two web-based text sources, Weekly Reader and BBC Bitesize, targeting different age groups, to cover a broad range of school grades. On the conceptual side, we explore the use of lexical and syntactic measures originally designed to measure language development in the production of second language learners. We show that the developmental measures from Second Language Acquisition (SLA) research when combined with traditional readability features such as word length and sentence length provide a good indication of text readability across different grades. The resulting classifiers significantly outperform the previous approaches on readability classification, reaching a classification accuracy of 93.3%.

## 1 Introduction

Reading plays an important role in the development of first and second language skills, and it is one of the most important means of obtaining information about any subject, in and outside of school. However, teachers often find it difficult to obtain texts appropriate to the reading level of their students, on a given topic. In many cases, they end up modifying or creating texts, which takes significant time and effort. In addition to such a traditional school setting, finding texts at the appropriate reading level is also important in a wide range of real-life contexts involving people with intellectual disabilities, dyslexics, immigrant populations, and second or foreign language learners.

Readability-based text classification, when used as a ranking parameter in a search engine, can help in retrieving texts that suit a particular target reading level for a given query topic. In the context of language learning, a language aware search engine (Ott and Meurers, 2010) that includes readability classification can facilitate the selection of texts from the web that are appropriate for the students in terms of form and content. This is one of the main motivations underlying our research.

Readability assessment has a long history (DuBay, 2006). Traditionally, only a limited set of surface features such as word length and sentence length were considered to derive a formula for readability. More recently, advances in computational linguistics made it possible to automatically extract a wider range of language features from text. This facilitated building machine learning models that estimate the reading level of a text. On the other hand, there has also been an on-going stream of research on reading and text complexity in other areas such as Second Language Acquisition (SLA) research and psycholinguistics.

In SLA research, a range of measures have been proposed to study the development of complexity in the language produced by learners. These measures are used to evaluate the oral or written production abilities of language learners. The aim of readability classification, on the other hand, is to retrieve texts to be comprehended by readers at a par-



*The 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 163–173,
Montréal, Canada, June 3-8, 2012. 

ticular level. Since we want to classify and retrieve texts for learners of different age groups, we hypothesized that these SLA-based complexity measures of learner production, when used as features for readability classification will improve the performance of the classifiers. In this paper, we show that this approach indeed results in a significant performance improvement compared to previous research.

We used the WeeklyReader website[1] as one of the text used in previous research. We combined it with texts crawled from the BBC-Bitesize website[2], which provides texts for a different age group. The combined corpus, *WeeBit*, covers a comparatively larger range of ages than covered before.

To summarize, the contributions of this paper are:

- We adapt measures from second language acquisition research to readability classification and show that the overall classification accuracies of an approach including these features significantly outperforms previous approaches.

- We extend the most widely used WeeklyReader corpus by combining it with another corpus that is graded for a different age-group, thereby creating a larger and more diverse corpus as basis for future research.

The paper is organized as follows: Section 2 describes related work on reading level classification to put our work in context. Section 3 introduces the corpora we used. Section 4 describes the features we considered in detail. Section 5 presents the approach and discusses the results. Section 6 provides a summary and points to future work.

## 2 Related Work

The traditional readability formulae made use of a limited number of surface features, such as the average sentence length and the average word length in characters or syllables (Kincaid et al., 1975; Coleman and Liau, 1975). Some works also made use of lists of "difficult" words, typically based on frequency counts, to estimate readability of texts (Dale and Chall, 1948; Chall and Dale, 1995; Stenner, 1996). Dubay (2006) provides a broad survey of traditional approaches to readability assessment. Although the features considered appear *shallow* in terms of linguistic modeling, they have been popular for many years and are widely used.

More recently, the developments in computational linguistics made it possible to consider various lexical and syntactic features to automatically model readability. In some of the early works on statistical readability assessment, Si and Callan (2001) and Collins-Thompson and Callan (2004) reported the impact of using unigram language models to estimate the grade level of a given text. The models were built on a United States text book corpus.

Heilman et al. (2007; 2008b; 2008a) extended this approach and worked towards retrieving relevant reading materials for language learners in the REAP[3] project. They extended the above mentioned approach to include a set of manually and later automatically extracted grammatical features.

Schwarm and Ostendorf (2005) and Petersen and Ostendorf (2009) report on classification experiments with WeeklyReader data, considering statistical language models, traditional formulae, as well as certain basic parse tree features in building an SVM-based statistical model. Feng et al. (2010) and Feng (2010) went beyond lexical and syntactic features and studied the impact of several discourse-based features, comparing their performance on the WeeklyReader corpus.

While the vast majority of approaches have targeted English texts, some work on other languages such as German, Portuguese, French and Italian (vor der Brück et al., 2008; Aluisio et al., 2010; Francois and Watrin, 2011; Dell'Orletta et al., 2011) is starting to emerge. Parse-tree-based features have also been used to measure the complexity of spoken Swedish (Roll et al., 2007).

The process of text comprehension and the effect of factors such as the coherence of texts have also been intensively studied (e.g., Crossley et al., 2007a; 2007b; Graesser et al., 2004) and measures to analyze the text under this perspective have been implemented in the CohMetrix project.[4]

The DARPA Machine Reading program created

[1] http://www.weeklyreader.com
[2] http://www.bbc.co.uk/bitesize
[3] http://reap.cs.cmu.edu
[4] http://cohmetrix.memphis.edu

a corpus of general text readability containing various forms of human and machine generated texts (Strassel et al., 2010).[5] The aim of this program is to transform natural language texts into a format suitable for automatic processing by machines and to filter out poorly written documents based on the text *quality*. Kate et al. (2010) used this data set to build a coarse grained model of text readability.

While in this paper we focus on comparing computational linguistic approaches to readability assessment and improving the state of the art on a traditional and available data set, Nelson et al. (2012) compared several research and commercially available text difficulty assessment systems in support of the Common Core Standards' goal of providing students with texts at the appropriate level of difficulty throughout their schooling.[6]

Independent of the research on readability, the complexity of the texts *produced* by language learners has been extensively investigated in Second Language Acquisition (SLA) research (Housen and Kuiken, 2009). Recent approaches have automated and compared a number of such complexity measures for learner language, specifically in English as Second Language learner narratives (Lu, 2010; Lu, 2011b). So far, there is hardly any work on using such insights in computational linguistics, though, with the notable exception of Chen and Zechner (2011) using SLA features to evaluate spontaneous non-native speech. Given that graded corpora are also intended to be used by incremental age groups, we started to investigate whether the insights from SLA research can fruitfully be applied to readability classification.

## 3 Corpora

We used a combined corpus of WeeklyReader and BBC-Bitesize to develop a statistical model that classifies texts into five grade levels, based on the age groups.

**WeeklyReader**[7] is an educational newspaper, with articles targeted at four grade levels (Level 2, Level 3, Level 4, and Senior), corresponding to children between ages 7–8, 8–9, 9–10, and 9–12 years. The articles cover a wide range of non-fiction topics, from science to current affairs, written according to the grade level of the readers. The exact criterion of graded writing is not published by the magazine. We obtained permission to use the graded magazine articles and downloaded the archives in 11/2011.[8]

Though we used the same WeeklyReader text base as the previous works, the corpus is not identical since we downloaded our version more recently. Thus the archive contained more articles per level and some preprocessing may differ. The WeeklyReader magazine issues in addition to the actual articles include teacher guides, student quizzes, images and brain teaser games, which we did not include in the corpus. The distribution of articles after this preprocessing is shown in Table 1.

| *Grade Level* | *Age in Years* | *Number of Articles* | *Avg. Number of Sentences/Article* |
|---|---|---|---|
| Level 2 | 7–8 | 629 | 23.41 |
| Level 3 | 8–9 | 801 | 23.28 |
| Level 4 | 9–10 | 814 | 28.12 |
| Senior | 10-12 | 1325 | 31.21 |

Table 1: The *Weekly Reader* corpus

**BBC-Bitesize**[9] is a website with articles classified into four grade levels (KS1, KS2, KS3 and GCSE), corresponding to children between ages 5–7, 8–11, 11–14 and 14–16 years. The Bitesize corpus is freely available on the web, and we crawled it in 2009. Most of the articles at KS1 consisted of images and flash files and other audio-visual material, with little text. Hence, we did not include KS1 in our corpus. We also excluded pages that contained only images, audio, or video files without text.

To cover a broad range of non-overlapping age groups, we used Level 2, Level 3 and Level 4 from WeeklyReader and KS3 and GCSE from Bitesize data respectively and built a combined corpus covering learners aged 7 to 16 years. Note that while KS2 covers the age group of 8–11 years, Levels 2, 3, and

[5] The corpus is apparently intended to be available for public use, but does not yet seem to be so; we so far were unsuccessful in obtaining more information from the authors.

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

[7] http://www.weeklyreader.com

[8] A license to use the texts on the website for research can be obtained for a small fee from *support@weeklyreader.com*. To support comparable research, we will share the exact corpus we used with other researchers who have obtained a license to use the WeeklyReader materials.

[9] http://www.bbc.co.uk/bitesize

4 together cover ages 7–10 years. Similarly, the Senior Level overlaps with Level 4 and KS3. Hence, we excluded KS2 and Senior from the combined corpus. We will refer to the combined five-level corpus we created in this way as **WeeBit**. The distribution of articles in the combined *WeeBit* corpus after preprocessing and removing the overlapping grade levels, is shown in Table 2.

| *Grade Level* | *Age in Years* | *Number of Articles* | *Avg. Number of Sentences/Article* |
|---|---|---|---|
| Level 2 | 7–8 | 629 | 23.41 |
| Level 3 | 8–9 | 801 | 23.28 |
| Level 4 | 9–10 | 814 | 28.12 |
| KS3 | 11–14 | 644 | 22.71 |
| GCSE | 14–16 | 3500 | 27.85 |

Table 2: The *WeeBit* corpus

To avoid a classification bias towards a class with more training examples during, for each level in the *WeeBit* corpus, 500 documents were taken as training set and 125 documents were taken as test set. In total, we trained on a set of 2500 documents and used a test set of 625 documents, spanning across five grade levels.

## 4 Features

To build our classification models, we combined features used in previous research with other parse tree features as well as lexical richness and syntactic complexity features from SLA research. We group the features into three broad categories: lexical, syntactic and traditional features.

### 4.1 Lexical Features

Word n-grams have been frequently used as lexical features in the previous research (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005).[10] POS n-grams as well as POS-tag ratio features have also been used in some of the later works (Feng et al., 2010; Petersen and Ostendorf, 2009).

In the SLA context, independent of the readability research, Lu (2011a) studied the relationship of **lexical richness** to the quality of English as Second Language (ESL) learners' oral narratives and analyzed the distribution of three dimensions of lexical richness (lexical density, sophistication and variation) in them using various metrics proposed in the language acquisition literature. Those measures were used to analyze a large scale corpus of Chinese learners of English. We adapted some of the metrics from this research as our lexical features:

*Type-Token Ratio (TTR)* is the ratio of number of word types (T) to total number word tokens in a text (N). It has been widely used as a measure of lexical diversity or lexical variation in language acquisition studies. However, since it is dependent on the text size, various alternative transformations of TTR came into existence. We considered Root TTR (T/$\sqrt{N}$), Corrected TTR (T/$\sqrt{2N}$), Bilogarithmic TTR (Log T/Log N) and Uber Index ($Log^2T/Log(N/T)$).

Another recent TTR variant we considered, which is not a part of Lu (2011a), is the Measure of Textual Lexical Diversity (MTLD; McCarthy and Jarvis, 2010). It is a TTR-based approach that is not affected by text length. It is evaluated sequentially, as the mean length of string sequences that maintain a default Type-Token Ratio value. That is, the TTR is calculated at each word. When the default TTR value is reached, the MTLD count increases by one and TTR evaluations are again reset. McCarthy and Jarvis (2010) considered the default TTR as 0.72 and we continued with the same default.

Considering nouns, adjectives, non-modal and non-auxiliary verbs and adverbs as lexical items, Lu (2011a) studied various syntactic category based word ratio measures. *Lexical variation* is defined as the ratio of the number of lexical types to lexical tokens. Other variants of lexical variation studied in Lu (2011a) included noun, adjective, modifier, adverb and verb variations, which represent the proportion of the words of the respective categories compared to all lexical words in the document. Alternative measures of verb variation, namely Verb Variation-1 ($T_{verb}/N_{verb}$), Squared Verb Variation-1 ($T^2_{verb}/N_{verb}$) and Corrected Verb Variation-1 ($T_{verb}/\sqrt{2N_{verb}}$) are also studied in the literature. We considered all these measures of lexical variation as a part of our lexical features. We have also included *Lexical Density*, which is the ratio of the number of lexical items in relation to the total number of words in a text.

[10] In the readability literature, n-grams are traditionally discussed as lexical features. N-grams beyond unigrams naturally also encode aspects of syntax.

In addition to these measures from the SLA literature, in our lexical features we included the average number of syllables per word (NumSyll) and the average number of characters per word (NumChar), which are used as word-level indicators of text complexity in various traditional formulae (Kincaid et al., 1975; Coleman and Liau, 1975).

Finally, we included the proportion of words in the text which are found on the *Academic Word List* as another lexical feature. It refers to the word list created by Coxhead (2000), which contains a list of most frequent words found in the academic texts.[11] The list does not include the most frequent words in the English language as such. The words in this list are specific to academic contexts. It was intended to be used both by teachers and students as a measure of vocabulary acquisition. We use it as an additional lexical feature in our work – and it turned out to be one of the most predictive features.

All the lexical features we considered in this work are listed in Table 3. The SLA based lexical features are referred to as SLALEX in the table. Of these, six features *CTTR, RTTR, SVV1, CVV1, AdvV, ModV* were shown by Lu (2011b) to correlate best with the learner data. We will refer to them as BESTLEX-SLA, highlighted in italics in the table.

**Lexical Features from SLA research (SLALEX)**
– Lexical Density (LD)
– Type-Token Ratio (TTR)
– *Corrected TTR (CTTR)*
– *Root TTR (RTTR)*
– Bilogarithmic TTR (LogTTR)
– Uber Index (Uber)
– Lexical Word Variation (LV)
– Verb Variation-1 (VV1)
– *Squared VV1 (SVV1)*
– *Corrected VV1 (CVV1)*
– Verb Variation 2 (VV2)
– Noun Variation (NV)
– Adjective Variation (AdjV)
– *Adverb Variation (AdvV)*
– *Modifier Variation (ModV)*
– Mean Textual Lexical Density (MTLD)

**Other Lexical Features**
– Proportion of words in AWL (AWL)
– Avg. Num. Characters per word (NumChar)
– Avg. Num. Syllables per word (NumSyll)

Table 3: Lexical Features (LEXFEATURES)

[11] http://en.wikipedia.org/wiki/Academic_Word_List

## 4.2 Syntactic Features

Schwarm and Ostendorf (2005) implemented four parse tree features (average parse tree height, average number of SBARs, NPs per sentence and VPs per sentence) in their work. Feng (2010) considered more syntactic features, adding the average lengths of phrases (NP, VP and PP) per sentence in words and characters, and the total number of respective phrases in the document. In our work, we started with reconsidering the above mentioned syntactic features.

In addition, we included measures of syntactic complexity from the SLA literature. Lu (2010) selected 14 measures from a large set of measures used to monitor the syntactic development in language learners. He then used these measures in the analysis of syntactic complexity in second language writing and showed that some of them correlate well with the syntactic development of adult Chinese learners of English. They are grouped into five broad categories:

The first set consists of three measures of syntactic complexity based on the length of a unit at the sentential, clausal and T-unit level respectively. The definitions for sentence, clause and T-unit were adapted from the SLA literature. While a *sentence* is considered to be a group of words delimited with punctuation mark, a *clause* is any structure with a subject and a finite verb. Finally, a *T-unit* is characterized as one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it.

The second type of measure targets sentence complexity. Clauses per sentence is considered as a sentence complexity measure.

The third set of measures reflect the amount of subordination in the sentence. They include clauses per T-unit, complex T-units per T-unit, dependent clauses per clause and dependent clauses per T-unit. A *complex T-unit* is considered as any T-unit that contains a dependent clause.

The fourth type of measures measured the amount of co-ordination in a sentence. They consist of co-

ordinate phrases per clause and co-ordinate phases per T-unit. Any adjective, verb, adverb or noun phrase that dominates a co-ordinating conjunction is considered a co-ordinate phrase.

The fifth type of measures represented the relationship between specific syntactic structures and larger production units. They include complex nominals per clause, complex nominals per T-unit and verb phrases per T-unit. Complex nominals are comprised of a) nouns plus adjective, possessive, prepositional phrase, relative clause, participle or appositive, b) nominal clauses, c) gerunds and infinitives in subject positions.

We implemented these 14 syntactic measures as features in building our classification models, in addition to existing features. Eight of these features *(MLC, MLT, CP/C, CP/T, CN/C, CN/T, MLS, VP/T)* were argued to correlate best with language development. We refer to this subset of eight as BESTSYNSLA, shown in italics in Table 4. We will see in section 5 that a set including those features also holds good predictive power for classifying graded texts.

We also included the number of dependent clauses, complex T-units, and co-ordinate phrases per sentence as additional syntactic features. Table 4 summarizes the syntactic features used in this paper.

### 4.3 "Traditional" Features

The average number of characters per word (NumChar), the average number of syllables per word (NumSyll), and the average sentence length in words (MLS) have been used to derive formulae for readability in the past. We refer to them as *Traditional Features* below. We included MLS in the syntactic features and NumChar, and NumSyll in the Lexical features. We also included two popular readability formulae, Flesch-Kincaid score (Kincaid et al., 1975) and Coleman-Liau readability formula (Coleman and Liau, 1975), as additional features. The latter will be referred as *Coleman* below, and both formulas together as *Traditional Formulae*.

**Syntactic features from SLA research (SLASYN)**
– *Mean length of clause (MLC)*
– *Mean length of a sentence (MLS)*
– *Mean length of T-unit (MLT)*
– Num. of Clauses per Sentence (C/S)
– Num. of T-Units per sentence (T/S)
– Num. of Clauses per T-unit (C/T)
– Num. of Complex-T-Units per T-unit (CT/T)
– Dependent Clause to Clause Ratio (DC/C)
– Dependent Clause to T-unit Ratio (DC/T)
– *Co-ordinate Phrases per Clause (CP/C)*
– *Co-ordinate Phrases per T-unit (CP/T)*
– *Complex Nominals per Clause (CN/C)*
– *Complex Nominals per T-unit (CN/T)*
– *Verb phrases per T-unit (VP/T)*

**Other Syntactic features**
– Num. NPs per sentence (NumNP)
– Num. VPs per sentence (NumVP)
– Num. PPs per sentence (NumPP))
– Avg. length of a NP (NPSize)
– Avg. length of a VP (VPSize)
– Avg. length of a PP (PPSize)
– Num. Dependent Clauses per sentence (NumDC)
– Num. Complex-T units per sentence (NumCT)
– Num. Co-ordinate Phrases per sentence (CoOrd)
– Num. SBARs per sentence (NumSBAR)
– Avg. Parse Tree Height (TreeHeight)

Table 4: Syntactic features (SYNFEATURES)

## 5 Experiments and Evaluation

We used the Berkeley Parser (Petrov and Klein, 2007) with the standard model they provide for building syntactic parse trees and defined the patterns for extracting various syntactic features from the trees using the Tregex pattern matcher (Levy and Andrew, 2006). More details about the patterns from the SLA literature and their definitions can be found in Lu (2010). We used the OpenNLP[12] tagger to get POS tag information and calculate Lexical Richness features. We used the WEKA (Hall et al., 2009) toolkit for our classification experiments. We explored different classification algorithms such as Decision Trees, Support Vector Machines, and Logistic Regression. The Multi-Layer Perceptron (MLP)-classifier performed best with various combinations of features, so we focus on reporting the results for that algorithm.

[12] http://opennlp.apache.org

| Feature set | # Features | Classifier Performance | |
|---|---|---|---|
| | | Accuracy | RMSE |
| Traditional Formulae | 2 | 38.8% | 0.36 |
| Traditional Features | 3 | 70.3% | 0.25 |
| Trad. Features + Trad. formulae | 5 | 72.3% | 0.32 |
| SLALEX | 16 | 68.1% | 0.29 |
| SLASYN | 14 | 71.2% | 0.28 |
| SLALEX + SLASYN | 30 | **82.3%** | 0.23 |
| BEST10SYN | 10 | 69.9% | 0.28 |
| All Syntactic Features | 25 | 75.3% | 0.27 |
| BEST10LEX | 10 | 82.4% | 0.22 |
| All Lexical Features | 19 | 86.7% | 0.20 |
| BEST10ALL | 10 | 89.7% | 0.18 |
| All features | 46 | **93.3%** | 0.15 |

Table 5: Classification results for WeeBit Corpus

### 5.1 Evaluation Metrics

We report our results in terms of classification accuracy and root mean square error.

*Classification accuracy* refers to the percentage of instances in the test set that are classified correctly. The correct classifications include both true positives and true negatives. However, accuracy does not reflect how close the prediction is to the actual value. A difference between expected and predicted values of one grade level is treated the same way as the difference of, e.g., four grade levels.

*Root mean square error (RMSE)* is a measure which gives a better picture of this difference. RMSE is the square root of empirical mean of the squared prediction errors. It is frequently used as a measure to estimate the deviation of an observed value from the expected value. In readability assessment, it can be understood as the average difference between the predicted grade level and the expected grade level.

### 5.2 Feature Combinations

Complementing our experiments comparing the different lexical and syntactic features and their combination, we also used WEKA's information-gain-based feature selection algorithm, and selected the Top-10 best features using the ranker method.

When all features were considered, the top 10 most predictive features were found to be: (*NumChar, NumSyll, MLS, AWL, ModVar, CoOrd, Coleman, DC/C, CN/C,and AdvVar*), which are referred to as BEST10ALL in the table.

Considering the 25 syntactic features alone, the 10 most predictive features were: (*MLS, CoOrd, DC/C, CN/C, CP/C, NumPP, VPSize, C/T, CN/T and NumVP*), referred to as BEST10SYN in the table.

The 10 most predictive features amongst all the lexical features were: (*NumChar, NumSyll, AWL, ModV, AdvV, AdjV, LV, VV1, NV and SVV1*). They are referred to as BEST10LEX in the table.

Although the traditionally used features (*NumChar, NumSyll, MLS*) seem to be the most predictive, it can be seen from the other top ranked features, that there is significant overlap between the best features identified by WEKA and the features which Lu (2010; 2011b) identified as correlating best with language development (shown in italics in Table 3 and Table 4), which supports our hypothesis that the SLA-based measures are useful features for readability classification of non-learner text too.

### 5.3 Results

Table 5 shows the results of our classification experiments using WEKA's Multi-Layer Perceptron algorithm with different combinations of features. Combining all features results in the best accuracy of 93.3%, which is a large improvement over the current state of the art in readability classification reported on the WeeklyReader corpus (74.01% by Feng et al., 2010). It should, however, be kept

| | # Features | Highest reported accuracy |
|---|---|---|
| **Previous work (on WeeklyReader)** | | |
| (Feng et al., 2010) | 122 | 74.01% |
| (Petersen and Ostendorf, 2009) | 25 | 63.18% |
| Syntactic features only (Petersen and Ostendorf, 2009) | 4 | 50.91% |
| **Our Results (on WeeklyReader alone)** | | |
| Syntactic features from (Petersen and Ostendorf, 2009) | 4 | 50.68% |
| All our Syntactic Features | 25 | 64.3% |
| All our Lexical Features | 19 | 84.1% |
| All our Features | 46 | 91.3% |
| **Our Results (on WeeBit)** | | |
| All our Syntactic Features | 25 | 75.3% |
| All our Lexical Features | 19 | 86.7% |
| All our Features | 46 | 93.3% |

Table 6: Overall Results and Comparison with Previous Work

in mind that the improvement is achieved on the WeeBit corpus which is an extension of the WeeklyReader corpus previously used. Interestingly, the result of 89.7% for BEST10ALL, the top 10 features chosen by the WEKA ranker, are quite close to our best result, with a very small number of features.

Lexical features seem to perform better than syntactic features when considered separately. However, this better performance of lexical features was mainly due to the addition of the traditionally used features *NumChar* and *NumSyll*. So it is no wonder that these shallow features have been used in the traditional readability formulae for such a long time; but the predictive power of the traditional formulae as features by themselves is poor (38.8%), in line with the conclusions drawn in previous research (Schwarm and Ostendorf, 2005; Feng et al., 2010) about the Flesch-Kincaid and Dale-Chall formulae. Interestingly, *Coleman*, which was not considered in those previous approaches, was ranked among the Top-10 most predictive features by the WEKA ranker. So it holds a good predictive power when used as one of the features for the classifier.

We also studied the impact of SLA based features alone on readability classification. The performance of the SLA based lexical features (SLALEX) and syntactic features (SLASYN) when considered separately are still in a comparable range with the previously reported results on readability classification (68.1% and 71.2% respectively). However, combining both of them resulted in an accuracy of 82.3%, which is a considerable improvement over previously reported results. It again adds weight to the initial hypothesis that SLA based features can be useful for readability classification.

### 5.4 Comparison with previous work

Table 6 provides an overall comparison of the accuracies obtained for the key features sets in our work with the best results reported in the literature for the WeeklyReader corpus. However, since our classification experiments were carried out with a newly compiled corpus extending the WeeklyReader data, such a direct comparison is not particularly meaningful by itself. To address this issue, we explored two avenues.

Firstly, we ran additional experiments, training and testing on the WeeklyReader data only, including the four levels used in previous work on that corpus. A summary of the results can be seen in Table 6. Our approach with 46 features results in 91.3% accuracy on the WeeklyReader corpus, compared to 74.01% as the best previous WeeklyReader result, reported by Feng et al. (2010) for their much larger feature set (122 features).

In order to verify the impact of our choice of features, we also did a replication of the parsed syntactic feature measures reported by (Schwarm and Ostendorf, 2005) on the WeeklyReader corpus and obtained essentially the same accuracy as the one pub-

lished (50.7% vs. 50.91%), supporting the comparability of the WeeklyReader data used. The significant performance increase we reported thus seems to be due to the new features we integrated from the SLA literature.

Secondly, we were interested in the impact of the training size on the results. We therefore investigated how good our best approach (using all features) is on a training corpus that is comparable to the WeeklyReader corpus used in previous work in terms of the number of documents per class. When we took 1400 WeeklyReader documents distributed into four classes as described in Feng et al. (2010), we obtained an accuracy of 84.2%, compared to the 74.01% they reported as best result. Using 2500 documents distributed into four classes as in Petersen and Ostendorf (2009) we obtained 88.4%, compared to their best result of 63.18%. Given that the original corpora used are not available, these WeeklyReader corpora with the same source, number of documents, and size of classes are as close as we can get to a direct comparison. In the future, the availability of the WeeBit corpus will support a more direct comparison of approaches.

In sum, the above experiments seem to indicate that the set of features and classifier used in our approach play an important role in the resulting significant increase in accuracy.

## 6 Conclusion and Discussion

We created a new corpus, *WeeBit*, by combining texts from two graded web sources WeeklyReader and BBC Bitesize. The resulting text corpus is larger and covers more grade levels, spanning the age group between 7 and 16 years. We hope that the availability of this graded corpus will be useful as an empirical basis for future studies in automatic readability assessment.[13]

We studied the impact of various lexical and syntactic features and explored their performance in combination with features encoding syntactic complexity and lexical richness that were inspired by Second Language Acquisition research. Our experiments show that not only the full set of features, but also specific manually or automatically selected subsets of features provide results significantly improving on the previously published state of the art in automatic readability assessment. There also seems to be a clear correlation between the good predictors according to SLA research on language learning and those that performed well in text classification.

Although the exact criteria based on which the individual corpora (WeeklyReader, BBC-Bitesize) were created is not known, it is possible that they were created with the well-known, traditional readability formulae in mind. It would be surprising if the two corpora, compiled in the US and Britain by different companies, were created with the same set of measures in mind, so the WeeBit corpus should be less affected. Still, it is possible that the reason the traditional features NumChar, NumSyll and MLS held such a strong predictive power is that these measures were considered when the texts were written. But removing these traditional features only strengthens the role of the other features and thereby the main point of the paper arguing for the usefuless of SLA developmental measures for readability classification.

As a part of our future work, we intend to revisit and study the impact of further classes of features employed in psycholinguistics and cognitive science research, such as those studied in Coh-Metrix (Graesser et al., 2004) or in the context of retrieving texts for specific groups of readers (Feng, 2010).

In terms of our overall application goal, we are currently studying the ability of the classification models we built to generalize to web data. We then plan to add the classification model to a language aware search engine (Ott and Meurers, 2010). Such a search engine may then also be able to integrate user feedback on the readability levels of webpages, to build a dynamic, online model of readability.

## 7 Acknowledgements

We thank the anonymous reviewers and workshop organizers for their feedback on the paper. The research leading to these results has received funding from the European Commission's 7th Framework Program under grant agreement number 238405 (CLARA).[14]

[13] As mentioned above, we will make the WeeBit corpus available to all researchers who have obtained the inexpensive research license from WeeklyReader.

[14] http://clara.uib.no

## References


Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California, June.

Jeanne S. Chall and Edgar Dale. 1995. *Readability Revisted: The New Dale-Chall Readability Formula*. Brookline Books.

Maio Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 722–731, Portland, Oregon, June.

Meri Coleman and T.L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*, Boston, USA.

Averil Coxhead. 2000. A new academic word list. *Teachers of English to Speakers of Other Languages*, 34(2):213–238.

Scott A. Crossley, David F. Dufty, Philip M. McCarthy, and Danielle S. McNamara. 2007a. Toward a new readability: A mixed model approach. In Danielle S. McNamara and Greg Trafton, editors, *Proceedings of the 29th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Scott A. Crossley, Max M. Louwerse, Philip M. McCarthy, and Danielle S. McNamara. 2007b. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational research bulletin; organ of the College of Education*, 27(1):11–28.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.*

Lijun Feng. 2010. *Automatic Readability Assessment*. Ph.D. thesis, City University of New York (CUNY).

Thomas Francois and Patrick Watrin. 2011. On the contribution of mwe-based features to a readability formula for french as a foreign language. In *Proceedings of Recent Advances in Natural Language Processing*, pages 441–447.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louweerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, 36:193–202.

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

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'07)*, pages 460–467, Rochester, New York.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008a. 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*, Columbus, Ohio.

Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. 2008b. Retrieval of reading materials for vocabulary and reading practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, pages 80–88, Columbus, Ohio.

Alex Housen and Folkert Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.

Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *23rd International Conference on Computational Linguistics (COLING 2010)*.

J. P. Kincaid, R. P. Jr. Fishburne, R. L. Rogers, and B. S Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2011a. A corpus-based evaluation of syntactic complexity measures as indices of college-level esl writers' language development. *TESOL Quarterly*, 45(1):36–62, March.

Xiaofei Lu. 2011b. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Languages Journal*. in press.

Philip McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

J. Nelson, C. Perfetti, D. Liben, and M. Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, The Council of Chief State School Officers.

Niels Ott and Detmar Meurers. 2010. Information retrieval for education: Making search engines language aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications*, 3(1–2):9–30.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:86–106.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April.

Mikael Roll, Johan Frid, and Merie Horne. 2007. Measuring syntactic complexity in spontaneous spoken swedish. *Language and Speech*, 50(2).

Sarah Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530, Ann Arbor, Michigan.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*, pages 574–576. ACM.

A. Jackson Stenner. 1996. Measuring reading comprehension with the lexile framework. In *Fourth North American Conference on Adolescent/Adult Literacy*.

Stephanie Strassel, Dan Adams, Henry Goldberg, Jonathan Herr, Ron Keesing, Daniel Oblinger, Heather Simpson, Robert Schrag, and Jonathan Wright. 2010. The darpa machine reading program - encouraging linguistic and reasoning research with a series of reading tasks. In *Language Resources and Evaluation (LREC)*, Malta.

Tim vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica*, 32(4):429—-435.