Computational Linguistics and Chinese Language Processing
Vol. 15, No. 3-4, September/December 2010, pp. 193-218 193


# Using Linguistic Features to Predict Readability of Short Essays for Senior High School Students in Taiwan[1]

**Wei-Ti Kuo[*], Chao-Shainn Huang[*], and Chao-Lin Liu[*]**

**Abstract**

We investigated the problem of classifying short essays used in comprehension tests for senior high school students in Taiwan. The tests were for first and second year students, so the answers included only four categories, each for one semester of the first two years. A random-guess approach would achieve only 25% in accuracy for our problem. We analyzed three publicly available scores for readability, but did not find them directly applicable. By considering a wide array of features at the levels of word, sentence, and essay, we gradually improved the F measure achieved by our classifiers from 0.381 to 0.536.

**Keywords**: Computer-assisted Language Learning, Readability Analysis, Document Classification, Short Essays for Reading Comprehension.

## 1. Introduction

Reading is a key competence for language learners. For learners of English as a Second Language (ESL), reading provides a crucial channel for learners to integrate and exercise the knowledge of previously learned vocabulary and grammar. If we could provide appropriate material to ESL learners, they would receive individualized stimulus, maintain the motivation to learn, and benefit more from reading activities. Hence, researchers have been investigating the readability of articles and books for a long time (Flesch, 1948).

In recent decades, research about readability has not been confined to just classifying the readability of articles. In large-scale language tests that include a writing assessment, grading the writing of a large number of test takers is very time consuming. Moreover, maintaining a consistent grading standard over the group of graders is also a challenge. Hence, techniques

---

[1] This paper was converted from the technical report for an NSC-funded undergraduate practice project, NSC-98-2815-C-004-003-E, for Wei-Ti Kuo. This paper was revised from its Chinese version that was previously published in the ROCLING XXII conference (Huang *et al*., 2010).

[*] Department of Computer Science, National Chengchi University, Taipei, Taiwan
E-mail: {94703041, 97753007, chaolin}@nccu.edu.tw
The author for correspondence is Chao-Lin Liu.

for automated grading were studied and introduced in the Scholastic Aptitude Test (**SAT**[2]) in the USA (Burstein *et al*., 2003; Attali & Burstein, 2006; Chang *et al*., 2006).

In a broader sense, the problems of determining the readability of articles and judging the scores of essays are specialized instances of text classification. They are similar in that text materials are categorized based on some selected metrics, and they differ in the implications of the classification results.

Early work in readability analysis considered the frequency of words, number of sentences, and length of sentences (Flesch, 1948; Kincaid *et al*., 1975; Chall & Dale, 1995). These methods may seem deficient nowadays, but it was not easy to consider all conceivable factors when the training corpora and the computing power were not sufficient. Other factors clearly are relevant to readability (Bailin & Grafstein, 2001), and one may consider more lexical level information, such as the hypernyms and hyponyms of words in an article, to determine the readability (Lin *et al*., 2009). Higher levels of information, such as the structure of the articles, semantic information, and cognition-related connotation, may also be included in readability analysis (Crossley *et al*., 2008).

Depending on the purpose of classifying the textual material, a classifier should consider factors of various aspects. Linguistic features are obvious candidates, but psycholinguistic, educational, and cultural factors are important as well. Moreover, characteristics of the readers and writers of the essays should also be considered. Classifications of articles written by native speakers and non-native speakers might be quite different. Good reading materials for second graders of native and non-native speakers would vary in terms of their vocabulary and content.

In this study, we examine short essays that were designed for reading comprehension tests at the high school level in Taiwan. Essays were classified based on a comprehensive list of lexical and syntactic features that were extracted from the words, sentences, and paragraphs in a given essay. The essays used in the experiments were realistic; therefore, they were limited in regards to the available amount. We focused on 845 tests for the first four semesters in high school, so essays were classified into four categories that corresponded to the semester of the examinee. We explored the applications of several machine learning models for the classification task, and the best $F_1$ measure (Witten & Frank, 2005) that we achieved was only 0.536.

We understand that there is room to improve our work, in terms of both the scale of experiments and the achieved results of accuracy. The current experience, however, supports a popular viewpoint that lexical and syntactic information about the short essays are

[2] We highlighted acronyms of phrases and special terms with boldface and blue text to help readers find their meanings.

instrumental but are not sufficient for predicting readability (Bailin & Grafstein, 2001). Some deep analysis is required to achieve better results. For instance, the set of a reading comprehension test consists of a short essay and questions for the students to answer. The set of a reading comprehension test may be considered more difficult because of its questions, not just because of its essay. Analyzing the questions is a major step for us to complete the current study.

We introduce the data source and their preprocessing in Section 2, deal with the extracted lexical features in Section 3, discuss the syntactic features in Section 4, present and compare the effects of using different combinations of the features to predict the readability in Section 5, and make some concluding remarks in Section 6.

## 2. Background

To make the results of this study close to reality, we obtained essays for comprehension tests for students at senior high schools in Taiwan. The essays were retrieved from the item pool that was designed for the San-Min version (三民版) of English courses, and the item pool was published in the 96th school year. The 96th school year spanned August 2007 to July 2008.

The item pool was designed for preparing competence examinations that are similar to the SAT in the USA. Students apply for college during the fifth semester in high school in Taiwan. Hence, the contents of the item pool covered only English for the first two years in senior high school and we treated a semester as a level in our experiments.

The goal of our work was to determine the level of the short essay of a given comprehension test. Namely, we classified an essay into one of four possible levels.

Table 1 shows the number of essays that we gathered from the item pool. The original essays were classified according to their levels and "tracks". The test items were designed for three tracks of English courses. The first track was designed by Ling-Hsia Chen (陳凌霞) of National Taiwan University, and we denote this track as NTUC in Table 1. The other two tracks were designed by Kwock-Ping John Tse (謝國平) of Providence University. One of these two tracks was more recent than the other. We denote the relatively more recent one as PUTN and the older as PUTO.

The words used in the comprehension tests were chosen based on the expected competence of the students. In Taiwan, the Ministry of Education (MOE) has issued a ruling about what words middle school graduates are expected to be acquainted with (MOE, 2008). Partially because of this constraint, essays for the comprehension tests contained Chinese translations for selected words. The numbers of the essays that did not contain Chinese translations were counted, and the totals are placed under the column "No Hints". The total number of Chinese words that appeared in the essays was placed under "Chinese Hints" in

Table 1. Chinese translations were provided in the essays for special nouns, such as names, places, and medical terms, in order to avoid the disturbance of these challenging words against comprehension.

***Table 1. Data source***

| | NTUC | PUTN | PUTO | Row Total | No Hints | Chinese Hints (words) |
|---|---|---|---|---|---|---|
| Level 1 | 47 | 117 | 36 | 200 | 124 | 142 |
| Level 2 | 64 | 127 | 36 | 227 | 199 | 45 |
| Level 3 | 48 | 127 | 36 | 211 | 148 | 151 |
| Level 4 | 45 | 126 | 36 | 207 | 198 | 14 |
| Total | 204 | 497 | 144 | 845 | 669 | 352 |

The appearance of Chinese translations could be considered as a noise in the original data, but it could also be considered as a feature. We took the latter position in some of our experiments and ignored the Chinese translations in some experiments. The statistics in Table 1 suggested that the appearance of Chinese translations was related to the levels. On average, there were fewer Chinese translations for the second semester of each school year.

Figure 1 shows the major steps we used to convert an essay into a feature vector. We first removed and recorded the Chinese translations from the original essay, as we discussed in the previous paragraphs. The remaining English texts were then processed by the Stanford Part-of-Speech (POS) tagger[3] and the Stanford parser[4] to extract the lexical and syntactic features. Except for the Stanford NLP tools, we relied on word lists that were selected by experts (*cf.* Section 3.1), the CMU Pronouncing dictionary[5] (*cf.* Section 3.2), and Dr.eye[6] dictionary (*cf.* Section 3.3) to broaden the types of lexical level information that we could extract.

Some linguistic features intuitively are related to the difficulty of essays, *e.g.*, the number of sentences, the number of words, the popularity (frequency) of words, the number of senses a word can carry, and the number of complex sentences. We applied tools and dictionaries for analyzing the linguistic features to create feature vectors (*cf.* Section 4).

Some basic features could be extracted easily. We calculated the number of sentences ($N$) in an essay and collected the following features: the number of tokens ($f_1$), the number of punctuations ($f_2$), the number of tokens and punctuations ($f_3=f_1+f_2$), the average number of

[3] http://nlp.stanford.edu/software/tagger.shtml

[4] http://nlp.stanford.edu/software/lex-parser.shtml

[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[6] http://www.dreye.com/

tokens per sentence ($f_4$=$f_1/N$), the average number of punctuations per sentence ($f_5$= $f_2/N$), and the average number of tokens and punctuations per sentence ($f_6$=$f_3/N$).

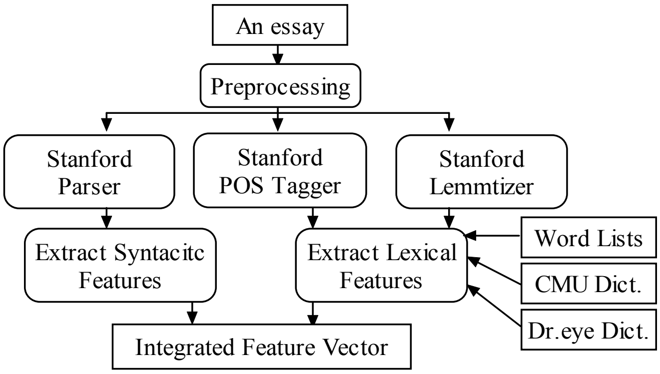


***Figure 1. Converting an essay into an instance***

## 3. Lexical Level Features

Words are the basic building blocks of essays. For ESL learners, learning basic vocabulary is an important first step into the world of English. According to the MOE's standards of course design for elementary education (MOE, 2008), graduates from middle schools should have learned and should be able to apply 1200 basic English words in daily conversations. In this section, we explain various types of lexical level features that we extracted from words in an essay.

### 3.1 Word Lists

Due to the crucial role of individual words in learning English, experts compiled different word lists for different purposes. We employed three lists in our work. Table 2 shows the detailed statistics of the NTNU, GETP, and CEEC word lists.

Professors at National Taiwan Normal University compiled a list of words for a competition related to English words, and we refer to this list as the **NTNU** list[7]. The NTNU list classifies words into three major groups – elementary, middle, and senior high schools – that are further divided for the targeted grades. For instance, "E34" is for the third and the fourth grades in elementary schools; and M3 is for the third year in middle school.

The General English Proficiency Test[8] (**GEPT**) is a standardized test accepted by domestic and some international institutions. To provide references for test takers, the GEPT offers word lists for different levels of test takers. Three of the lists were relevant to our work:

[7] http://vq.ie.ntnu.edu.tw/
[8] https://www.gept.org.tw/

Elementary, Intermediate, and High-Intermediate. These three lists include words that people who have graduated from middle schools, high schools, and colleges (non-English majors), respectively, should have learned.

The College Entrance Examination Center[9] (**CEEC**) is an institution for managing the college entrance examinations in Taiwan. The word list is designed for graduates of high schools and includes nearly 9000 words. This CEEC list contains 6 grades.

***Table 2. Statistics about word lists***

| Word Lists | Level | # of Words | Total # of Words |
|---|---|---|---|
| NTNU | E34 | 498 | 6041 |
| | E5 | 250 | |
| | E6 | 250 | |
| | M1 | 350 | |
| | M2 | 350 | |
| | M3 | 407 | |
| | S1 | 936 | |
| | S2 | 1500 | |
| | S3 | 1500 | |
| GEPT | Elementary | 2184 | 7853 |
| | Intermediate | 2560 | |
| | High-Intermediate | 3109 | |
| CEEC | G1 | 1775 | 8976 |
| | G2 | 1490 | |
| | G3 | 1472 | |
| | G4 | 1350 | |
| | G5 | 1543 | |
| | G6 | 1346 | |

We employed the Stanford NLP tools to tokenize the strings in an essay, as we illustrated in Figure 1. We lemmatized the tokens and identified their POS tags. After this step, we looked up the word lists to see which level the tokens belonged to and updated the frequencies of the levels. Similar to how Dale-Chall dealt with their word list (Dale & Chall, 1995), a word not belonging to any level was considered to belong to the "difficult" level, which is an additional level not listed in Table 2.

9 http://www.ceec.edu.tw/

We created feature vectors based on the NTNU, GEPT, and CEEC lists separately. With the above procedure, we created 10 features for an essay when we considered the NTNU list – 9 levels in Table 2 and one "difficult" level. Analogously, we had 4 and 7 features for the GEPT and CEEC lists, respectively.

We expected these features to be useful for the essay classification under the premise that, if an essay contains more words in the more advanced levels, the essay should be more difficult.

## 3.2 Pronunciation

For ESL learners in Taiwan, an English word with relatively more syllables is generally more difficult to remember and pronounce. This is partially due to the fact that Chinese is a tonal language and students may not be used to words with several syllables yet.

Based on this observation, we thought it might be worthwhile to explore the influence of the number of syllables on the readability of essays. Although not all long words are difficult and not all short words are easy, it was interesting to explore the intuitive impression.

After obtaining the lemmatized tokens in an essay, we looked at the CMU Pronouncing dictionary (**CMUPD**) to find the number of syllables in the tokens. The CMUPD contains more than 125000 words. The pronunciation of an English word is represented with English letters and numbers. The pronunciation of "university" is shown below.

**Y UW2 N AH0 V ER1 S AH0 T IY0**

Vowels and consonants are separated in CMUPD, and only vowels are followed by digits. The digits indicate stresses: 0 for no stress, 1 for primary stress, and 2 for secondary stress.

Given the CMUPD phoneme notation, we could compute the number of syllables in an English word and the total number of vowels and consonants in a word. Take "university" as an example. This token has 5 syllables and 10 vowels and consonants. In our corpus, a token may have at most 7 syllables and at most 16 vowels and consonants. If a token was not covered by CMUPD, we would record that this token had no syllables, no vowels, and no consonants.[10]

[10] We employed distributions of some random variables as features in this paper, and we generally used larger numbers to denote relatively more difficult cases. For instance, when creating features for word lists, larger indices indicated higher grade and more challenging words. Here, we converted the number of syllables of a word into a sequence of features. The first feature denoted the number of words with one syllable, the second feature denoted the number of words with two syllables, *etc*. We used the zero-th feature to denote the number of words not covered by CMUPD. This would not confuse the classifiers that we tried in Section 5 because the semantics of the order of the features was not explicit to the classifiers.

For a given essay, we would record the frequencies of tokens that have *i* syllables and *j* vowels and consonants, where *i* is in the range [0, 7] and *j* is in the range [0, 16]. Therefore, we had 25 features related to pronunciation of tokens in an essay.

## 3.3 Lexical Ambiguity

Ambiguity may not be just a problem for natural language processing of computers; it could be a problem for ESL learners as well. Many English words carry multiple possible meanings. If an essay contains many words with multiple possible meanings, its contents may become relatively difficult to understand. Based on this intuition, we considered the distribution of the numbers of translated senses of words in an essay as features.

Finding the number of translated senses of an English word took a little bit of work. Using the Stanford POS tagger, we could find the POS of a token. The POS tag followed the Penn TreeBank convention[11]. Also, we used Dr.eye to find the Chinese translations of English words. Dr.eye only has a very rough POS system: noun, transitive verb, intransitive verb, adjective, adverb, preposition, pronoun, conjunction, and determiner. Therefore, we had to convert a POS tag in the Penn TreeBank system into a category in Dr.eye. We employed the classification in a CEEC publication[12], and considered only 8 different POS tags. The conversion of POS tags was conducted based on the mapping listed in Table 3.

***Table 3. Converting a POS in Penn TreeBank system to Dr.eye's category***

| POS tags | Stanford POS Tagger | Dr.eye |
|---|---|---|
| Noun | NN, NNS, NNP, NNPS | n. |
| Verb | MD, VB, VBD, VBG, VBN, VBP, VBZ | vt., vi. |
| Adjective | CD, JJ, JJR, JJRS | a. |
| Adverb | EX, RB, RBR, RBS, RP, WRB | ad. |
| Preposition | IN, TO | prep. |
| Pronoun | DT, PRP, PRP$, WDT, WP, WP$, WRB | pron. |
| Conjunction | CC, IN | conj. |
| Determiner | DT | art. |

Note that the conversion was imperfect. The POS "IN" could be mapped to conjunction and preposition. When we encountered a token with "IN," we checked Dr.eye to see if the token could be used as a conjunction. If yes, that token was considered a conjunction. Otherwise, the token was considered a preposition.

---

[11] ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz

[12] http://www.ceec.edu.tw/Research/paper_doc/ce37/6.pdf

In Dr.eye, an English word can have at most 43 translated senses. We considered the number of translated senses as a feature. A token that could not be found in Dr.eye would be considered to have no translated senses. Hence, the distribution of the number of translated senses of tokens in an essay consisted of 44 numbers.

Figure 2 shows the entry for “divide” in Dr.eye. Assuming that we have a “divide/VBD” in an essay; we would know that this “divide” was a verb and would consider that this word had 8 possible translated senses.

**“divide”**

**vt. (及物動詞 transitive verb)**

1. 分,劃分[(+into/from)]
2. 分發;分享 [(+between/among/with)]
3. 分配[(+between)]
4. 【數】除[(+by/into)]
5. 使對立,分裂
6. 使分開,使隔開[(+from)]

**vi. (不及物動詞 intransitive verb)**

1. 分開
2. 分裂;意見分歧

**n. (名詞 noun)**

1. 分歧,不和[S][(+between)]
2. 分水嶺[C]

***Figure 2. The entry for “divide” in Dr.eye***

## 4. Syntactic Level Features

We collected information not just about the words in an essay, but we also attempted to find useful syntactic information as features for the classification task. This is necessary because simple words in complex sentences may not be easy to understand.

A sentence may be complex for different reasons. We considered the **depth**s of parse trees as an indication. Figure 3 shows a parse tree for the sentence, “I liked playing basketball when I was young.” Let the root, *i.e.*, `ROOT`, of the tree be Level 0, and its child node, *i.e.*, `S`, be Level 1. The deepest node in this tree is Level 9. We refer to the level of the deepest node in a tree as its depth.

We parsed sentences in our corpus with the Stanford parser (using the PCFG grammar file EnglishPCFG.ser.gz) and asked for only the parse trees with the highest score. In our corpus, the depth of the deepest tree was 31. We used features to represent the distribution of the depths of parse trees in an essay: ($d_0$, $d_1$, $d_2$, …, $d_8$). We increased $d_k$ by 1 if the depth of a parse tree was $k$ when $k$<8; and increased $d_8$ by 1 if the depth of a parse tree was larger than

8[13].

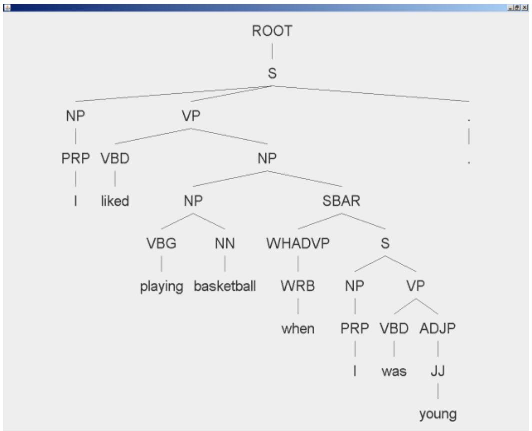


***Figure 3. A sample parse tree***

Given an essay, we could analyze every sentence to obtain its depth, and we recorded the average depth and the distribution of the depths of sentences in this essay.

Other than its depth, a sentence may be complex because it employs some rarely used grammatical relationships. The parse tree in Figure 3 includes several grammatical relationships: "S → NP VP .," "VP →VBD NP," "SBAR → WHADVP S," VP → VBD ADJP," *etc*. If one or more of these relationships are rare, the sentence may be difficult to read, rendering the essay not easy to understand.

We employed a corpus-based approach to determine whether or not a grammatical relationship was rare. We collected more than 7000 sentences from web sites that provide educational resources. They included "Shi Yuan You Grammar"[14], "1200 Fundamental English sentences"[15], "Learning Resources for Middle Schoolers"[16], and "I-Lan County Language Resources for Middle Schoolers"[17]. We parsed the collected sentences and recorded the frequencies of the grammatical relationships in these sentences.

We observed 985 grammatical relationships in these 7000+ sentences. Only 8 relationships occurred more than 1000 times, and 62 relationships took place more than 100 times.

As the span of the frequencies was wide and the distribution of the frequencies was

---

[13] Although one might expect that $d_0$ and $d_1$ should not appear in regular essays, we left these possibilities to avoid weird strings that might appear in our corpus.

[14] http://tw.myblog.yahoo.com/jw!GFGhGimWHxN4wRWXG1UDIL_XSA--/

[15] http://hk.geocities.com/cnlyhhp/eng.htm

[16] http://siro.moe.edu.tw/fip/index.php

[17] http://140.111.66.37/english/ (last visited 2010/8/14, but not functioning at the time of writing)

irregular, we quantized the ranges of the frequencies into 6 segments by the frequency binning method (Witten and Frank, 2005). The 985 relationships that we observed appeared at 127 different frequencies. We ordered them from frequent to infrequent ones and treated relationships that appeared the same number of times as the same relationship. Each segment contained 21 different frequencies (except the last segment, which covered 22 frequencies). We could consider these 6 segments of rules as “very frequent,” “frequent,” “slightly frequent,” “slightly infrequent,” “infrequent,” and “very infrequent”. (The choice of 6 was arbitrary. We did not try other selections.)

Given the above procedure, we could generate a vector of 7 components that considered the “rareness” of grammatical relationships in a sentence: {“very frequent,” “frequent,” “slightly frequent,” “slightly infrequent,” “infrequent,” “very infrequent,” “unseen”}. In a sentence containing 8 grammatical relationships, 2 very frequent, 4 frequent, 1 infrequent, 1 very infrequent, and 1 unseen, in our training corpus, we would convert it to {2, 4, 0, 0, 1, 1, 1}.

For an essay with many sentences, we could generate a 7-item vector for each sentence, and we took the average of every item to create the 7-item vector for the essay. An essay that includes a relatively larger number of rare grammatical relationships may be more difficult to read.

## 5. Experimental Evaluation

We classified the short essays reported in Section 2 with Weka (Witten and Frank, 2005), using different combinations of features reported in Sections 3 and 4.

Before we evaluated our methods, we acquired the SMOG scores of our essays via a Web-based service[18]. Equation (1) shows the score function for SMOG, where *m* represents the number of polysyllables and *n* is the number of sentences. A word is considered a polysyllable if it contains three or more syllables. Essays with higher SMOG scores are relatively harder to read.

$$1.043 \times \sqrt{m \times \frac{30}{n}} + 3.1291 \tag{1}$$

Table 4 shows basic statistics about the SMOG scores of our essays. The smallest, largest, and average SMOG scores increased with the levels of the essays quite impressively. This is probably a good reason for the popularity of this simple formula.

[18] Simple Measure of Goobledygook (SMOG). http://www.harrymclaughlin.com/SMOG.htm

***Table 4. Basic statistics of SMOG scores***

| | Level 1 | Level 2 | Level 3 | Level 4 |
|---|---|---|---|---|
| Smallest SMOG score | 6.59 | 7.22 | 6.75 | 7.3 |
| Largest SMOG score | 17.11 | 19.88 | 22.75 | 22.09 |
| Average SMOG score | 10.889 | 11.822 | 12.554 | 12.757 |

Nevertheless, if we looked into the details of the scores for individual essays, we would realize that assessing the readability of an individual essay is not easy. Figure 4 shows distributions of the SMOG scores of our essays of different levels. We quantized the SMOG with 0.5 as an interval, and accumulated the essays within an interval to draw the chart. The vertical axis shows the proportion of essays of a level for a given SMOG score interval (on the horizontal axis). Although the chart is quite complex to read, the curves clearly show that essays of easier levels may have higher SMOG scores than essays of harder levels.

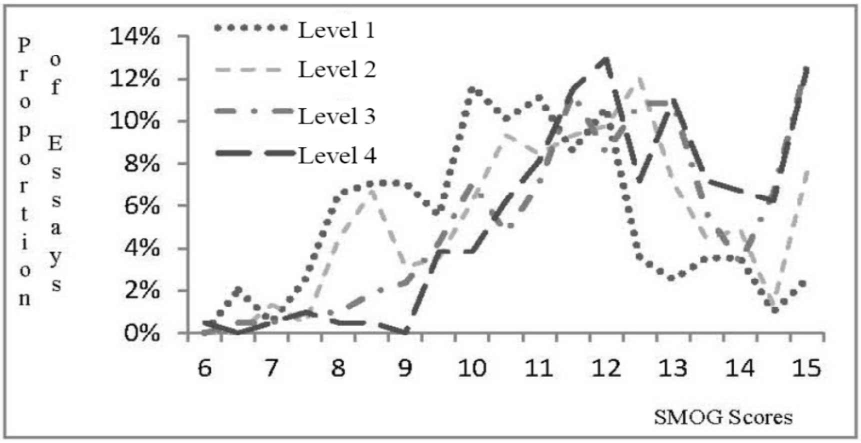


***Figure 4. Distributions of SMOG scores for different levels of essays***

## 5.1 Basic Features and Measures of the Prediction Quality

Since we had several different types of features, we grouped them to streamline our experiments. **Group A** consisted of features discussed in Sections 3.2 and 3.3: 8 features of the distribution of the number of syllables, 17 features of the distribution of the number of vowels and consonants, and 44 features of the distribution of lexical ambiguities.

**Group B** consisted of $f_4$, $f_5$, and $f_6$ in Section 2; the average depth; and the distribution of the depths of parse trees in an essay in Section 4. In total, we have 36 (=3+1+32) features in this group.

**Group C** consisted of the word lists in Section 3.1. We use **Ca**, **Cb**, and **Cc** to represent features generated based on the NTNU, GEPT, and CEEC word lists, respectively.

Whenever necessary, we normalized the statistics with the number of words and the number of sentences in a given essay. This is an important step to reduce the impact of different lengths of essays. In Group A, features about pronunciation were normalized by the number of words; the feature about the distribution of the number of lexical ambiguities would be normalized by the number of words. In Group B, the distribution of the depths of parse trees would have been normalized by the number of sentences in the essay. In Group C, the word counts of different levels would be normalized by the total number of words in the essay.

We ran 10-fold cross-validation with features in Group A, B, Ca, Cb, and Cc separately, using the J48 decision tree model, LMT decision tree model, Artificial Neural Networks (ANNs), and Ridor rules leaner. We did not do a random restart when we ran ANNs, and we set the number of epochs to 500 and learning rate to 0.3.

We measured the classification quality with the $F_1$ measure. $F_1$ measure is the harmonic average of recall rate and precision rate for a classification task (*cf.* Witten & Frank, 2005), and it is usually referred as the **F measure**. The **recall rate** achieved in a classification task is the proportion of instances that belong to the targeted classes captured by the classifier. The **precision rate** achieved in a classification task is the proportion of correct decisions of the classifier when it classifies instances as the targeted class.

## 5.2 Performance Achieved by the Basic Features

Table 5 shows the F measures achieved by individual groups of features. The best F measure was achieved when we used Cb with LMT, and the worst F measure occurred when we used Cc with J48. Table 5 also shows the column and row averages. The column averages indicate the effectiveness of a feature group, and the row averages show the effectiveness of a classifier.

***Table 5. F measures achieved by individual groups of features***

| | A | B | Ca | Cb | Cc | Average |
|---|---|---|---|---|---|---|
| J48 | 0.297 | 0.270 | 0.297 | 0.335 | **0.248** | 0.289 |
| LMT | 0.334 | 0.318 | 0.300 | **0.353** | 0.264 | 0.314 |
| ANN | 0.278 | 0.291 | 0.340 | 0.323 | 0.268 | 0.300 |
| Ridor | 0.293 | 0.291 | 0.307 | 0.304 | 0.261 | 0.291 |
| Average | 0.301 | 0.293 | 0.311 | 0.329 | 0.260 | 0.299 |

When the feature groups were applied separately, Cc might offer inferior effects because it contained words specifically for senior-high school levels and could not provide sufficient information about relatively easier words. The column averages indicate that using Ca or Cb word lists achieved better classification quality than not using word lists, *i.e.*, A and B.

Comparing the averages, we found that Cb and LMT are, respectively, the best individual feature group and the best classifier in Table 5.

Recall that we classified essays into one of four possible levels. Hence, a purely random guess is expected to achieve only 25% in accuracy. Although there are no good ways to compare F measures and accuracy directly, the F measures listed in Table 5 were not very encouraging.

Table 6 shows the F measures that were achieved when we combined the basic features to predict the readability. Again, the results were achieved in 10-fold cross-validations. The best F measure was 0.381 when we combined Groups B, Ca, and Cb in the predication task. The worst F measure was 0.261 when we combined Groups A, B, and Ca in the task. Again, the row averages indicate that the best classifier is LMT.

***Table 6. F measures achieved by combining basic features***

| | A+B | A+Ca | A+Cb | A+Cc | B+Ca | B+Cb | B+Cc | Ca+Cb | Ca+Cc | Cb+Cc | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| J48 | 0.281 | 0.266 | 0.275 | 0.293 | 0.293 | 0.299 | 0.274 | 0.3 | 0.306 | 0.312 | 0.290 |
| LMT | 0.335 | 0.345 | 0.337 | 0.346 | 0.344 | 0.341 | 0.3 | 0.348 | 0.338 | 0.364 | 0.340 |
| ANN | 0.283 | 0.348 | 0.33 | 0.318 | 0.315 | 0.319 | 0.303 | 0.346 | 0.347 | 0.324 | 0.323 |
| Ridor | 0.288 | 0.291 | 0.323 | 0.312 | 0.322 | 0.356 | 0.253 | 0.319 | 0.341 | 0.346 | 0.315 |
| Average | 0.297 | 0.313 | 0.316 | 0.317 | 0.319 | 0.329 | 0.283 | 0.328 | 0.333 | 0.337 | **0.317** |
| | A+B+Ca | A+B+Cb | A+B+Cc | A+Ca+Cb | A+Ca+Cc | A+Cb+Cc | B+Ca+Cb | B+Ca+Cc | B+Cb+Cc | Ca+Cb+Cc | Average |
| J48 | **0.261** | 0.303 | 0.301 | 0.291 | 0.307 | 0.325 | 0.303 | 0.295 | 0.304 | 0.286 | 0.298 |
| LMT | 0.331 | 0.327 | 0.35 | 0.341 | 0.321 | 0.35 | **0.381** | 0.349 | 0.366 | 0.358 | 0.347 |
| ANN | 0.359 | 0.309 | 0.328 | 0.313 | 0.33 | 0.323 | 0.319 | 0.352 | 0.314 | 0.357 | 0.330 |
| Ridor | 0.321 | 0.329 | 0.305 | 0.33 | 0.31 | 0.323 | 0.307 | 0.299 | 0.302 | 0.326 | 0.315 |
| Average | 0.318 | 0.317 | 0.321 | 0.319 | 0.317 | 0.330 | 0.328 | 0.324 | 0.322 | 0.332 | **0.323** |

| | A+B+Ca+Cb | A+B+Ca+Cc | A+B+Cb+Cc | A+Ca+Cb+Cc | B+Ca+Cb+Cc | A+B+Ca+Cb+Cc | Average |
|---|---|---|---|---|---|---|---|
| J48 | 0.305 | 0.305 | 0.313 | 0.302 | 0.317 | 0.33 | 0.312 |
| LMT | 0.329 | 0.353 | 0.369 | 0.341 | 0.373 | 0.358 | 0.354 |
| ANN | 0.335 | 0.335 | 0.362 | 0.338 | 0.333 | 0.324 | 0.338 |
| Ridor | 0.349 | 0.314 | 0.329 | 0.334 | 0.354 | 0.362 | 0.340 |
| Average | 0.330 | 0.327 | 0.343 | 0.329 | 0.344 | 0.344 | **0.336** |

Using more features allowed us to achieve better results. The best possible F measure increased from 0.353 in Table 5 to 0.381. The overall average of Table 5 is 0.299, indicating the average performance of our classifiers when we used only one feature group. The overall average of the first (upper) part of Table 6 is 0.317, the overall average of the second part is

0.323, and the overall average of the third part is 0.336. These averages show the average performance when using two groups, three groups, and more than three groups of features. Hence, we observed that using more features groups led to steady improvement in the average prediction quality.

Figure 5 shows the trends of improving performance for our classifiers when we employed more feature groups. The legends show the number of feature groups used with the classifiers, where “>3” indicates four or five groups.

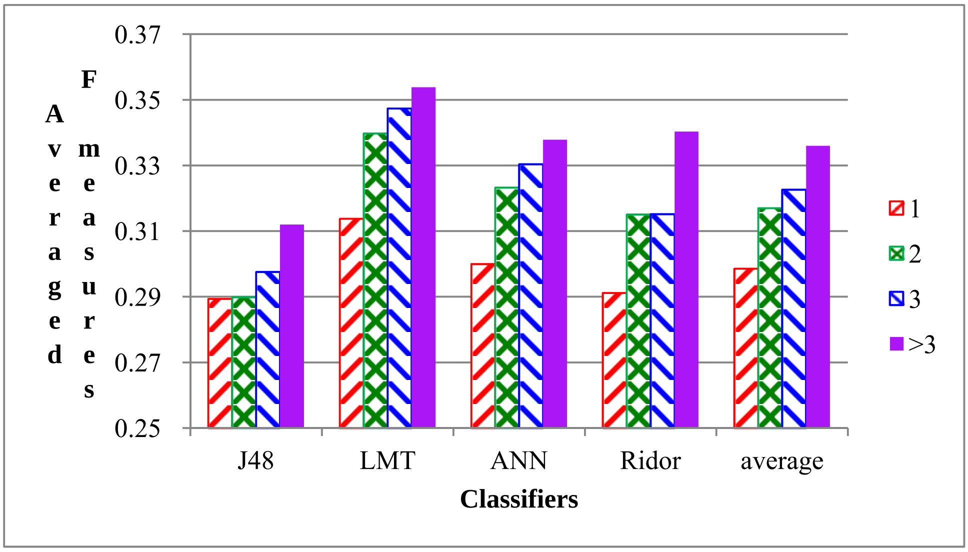


***Figure 5. Using more feature groups improve the prediction quality on average***

## 5.3 Frequencies of Grammatical Relationships

We refer to the frequency distribution of the grammatical relationships (Section 4) as **Group D**. Assume that there are two sentences in an essay, and the frequency distributions of their grammatical relationships are {0,0,2,0,0,1,3} and {1,1,0,5,5,0,4}. There are 22 grammatical relationships in this example. We add these distributions and divide each item by 22 to acquire a normalized distribution {0.045, 0.045, 0.091, 0.227, 0.227, 0.045, 0.318}.

We repeated the six experiments in Table 6. We considered the three experiments that had the best F measures (B+Ca+Cb, B+Ca+Cb+Cc, and Cb+Cc) and three word lists adding Groups A and B. The upper part of Table 7 is copied from the data in Table 6, and the lower part of Table 7 shows the F measures of the new experiments.

***Table 7. Effects of including Group D***

| Before | A+B+Ca | A+B+Cb | A+B+Cc | B+Ca+Cb | B+Ca+Cb+Cc | Cb+Cc | Average |
|---|---|---|---|---|---|---|---|
| J48 | **0.261** | 0.303 | 0.301 | 0.303 | 0.317 | 0.312 | 0.300 |
| LMT | 0.331 | 0.327 | 0.350 | **0.381** | 0.373 | 0.364 | 0.354 |
| ANN | 0.359 | 0.309 | 0.328 | 0.319 | 0.333 | 0.324 | 0.329 |
| Ridor | 0.321 | 0.329 | 0.305 | 0.307 | 0.354 | 0.346 | 0.327 |
| Average | 0.318 | 0.317 | 0.321 | 0.328 | 0.344 | 0.337 | 0.327 |
| **After** | **A+B+Ca** | **A+B+Cb** | **A+B+Cc** | **B+Ca+Cb** | **B+Ca+Cb+Cc** | **Cb+Cc** | **Average** |
| J48 | **0.251** | 0.294 | 0.294 | 0.309 | 0.318 | 0.309 | 0.296 |
| LMT | 0.346 | 0.342 | 0.325 | 0.343 | **0.357** | 0.345 | 0.343 |
| ANN | 0.327 | 0.339 | 0.306 | 0.308 | 0.351 | 0.327 | 0.326 |
| Ridor | 0.320 | 0.302 | 0.302 | 0.346 | 0.346 | 0.326 | 0.324 |
| Average | 0.311 | 0.319 | 0.307 | 0.327 | 0.343 | 0.327 | 0.322 |

Evidently, adding Group D in these experiments did not change the F measures significantly. Possible reasons for the observed irrelevancy include the fact that we determined the distributions based on another corpus of ours (Section 4), whose contents were designed for middle school students. The distribution of grammatical relationships in a corpus for middle schools may not be closely relevant to the readability of essays for senior high schools. Another possible reason is that Group D is in fact not relevant to readability.

## 5.4 Essay-Level Features

Although we normalized many features by the total number of sentences and the total number of words in an essay, we wondered about the potential contributions of the essay-level features. They include the total number of sentences, the total depth of parse trees, the number of tokens ($f_1$, Section 2), the number of punctuations ($f_2$, Section 2), the number of tokens and punctuations ($f_3$, Section 2), and the number of Chinese hints (Table 1 in Section 2); we refer to them as **Group E**.

We repeated the same set of experiments that we conducted for Table 7. This time, both Group D and Group E were used. Table 8 shows the F measures that we observed. The statistics suggest that using Group D and Group E helped us improve the prediction quality. As we have discussed in Section 2 about Table 1, the appearance of Chinese hints is noticeably related to the levels of the short essays. Hence, the improvement introduced by Group E was not very surprising.

***Table 8. Effects of including Groups D and E***

| | A+B+Ca | A+B+Cb | A+B+Cc | B+Ca+Cb | B+Ca+Cb+Cc | Cb+Cc | Average |
|---|---|---|---|---|---|---|---|
| J48 | 0.338 | 0.314 | 0.349 | 0.333 | 0.331 | 0.347 | 0.335 |
| LMT | 0.412 | 0.405 | 0.374 | 0.423 | **0.425** | 0.412 | 0.409 |
| ANN | 0.370 | 0.353 | 0.345 | 0.352 | 0.402 | 0.363 | 0.364 |
| Ridor | 0.337 | 0.36 | **0.312** | 0.353 | 0.377 | 0.341 | 0.347 |
| Average | 0.364 | 0.358 | 0.345 | 0.365 | 0.384 | 0.366 | 0.364 |

## 5.5 Distribution of Parts of Speech

It was suggested that we explore the influence of the distribution of the POS tags of the words in an essay. We considered the eight categories of POS tags in Section 3.3 to create features. We added these new features and repeated the experiment B+Ca+Cb+Cc+D+E in Table 8. Table 9 shows a comparison of the achieved F measures before and after adding the distribution.

***Table 9. Influences of distribution of POSes***

| | B+Ca+Cb+Cc+D+E | After adding dist. of POSes |
|---|---|---|
| J48 | 0.331 | 0.349 |
| LMT | 0.425 | 0.425 |
| ANN | 0.402 | 0.346 |
| Ridor | 0.377 | 0.343 |

With this limited scale of experiment, we could not reach a decisive conclusion about the effectiveness of the distribution of POS tags. The observed insignificance may result from the distribution of POS tags possibly remaining steady if we study the distribution in a large corpus (Shih, 2000) or might result from the distribution not being relevant to readability.

## 5.6 Articles with Chinese Hints

In Section 5.4, we investigated the contribution of using the number of Chinese hints as a feature for the classification task. Now, we explore the implications of whether an essay had Chinese hints or not on the predictability of its readability. We separated the essays into two sub-groups: those having Chinese hints and those having no Chinese hints; we then repeated the experiments for Table 5 and Table 6. Note that we removed the Chinese hints when we classified the essays that originally contained Chinese hints.

Table 10 and Table 11 show the F measures observed when we repeated the experiments with essays that originally contained Chinese hints. It was quite surprising to find that all F measures in Table 10 and Table 11 are better than their counterparts in Table 5 and Table 6, without any exceptions. The best F measure is now 0.536.

***Table 10. Using individual groups for essays with Chinese hints***

| | A | B | Ca | Cb | Cc |
|---|---|---|---|---|---|
| J48 | 0.423 | 0.364 | 0.472 | 0.428 | 0.382 |
| LMT | 0.435 | 0.404 | 0.494 | 0.466 | 0.363 |
| ANN | 0.429 | 0.396 | 0.467 | **0.52** | 0.396 |
| Ridor | **0.353** | 0.365 | 0.364 | 0.424 | 0.385 |

***Table 11. Using mixed groups for essays with Chinese hints***

| | A+B | A+Ca | A+Cb | A+Cc | B+Ca | B+Cb | B+Cc | Ca+Cb | Ca+Cc | Cb+Cc |
|---|---|---|---|---|---|---|---|---|---|---|
| J48 | 0.342 | 0.335 | 0.406 | 0.364 | 0.427 | 0.349 | 0.343 | 0.437 | 0.443 | 0.406 |
| LMT | 0.4 | 0.479 | 0.432 | 0.458 | 0.487 | 0.507 | 0.402 | 0.49 | 0.493 | 0.493 |
| ANN | 0.404 | 0.406 | 0.424 | 0.471 | 0.457 | 0.389 | 0.422 | 0.475 | 0.406 | 0.439 |
| Ridor | 0.395 | 0.375 | 0.413 | 0.364 | 0.381 | 0.424 | 0.366 | 0.462 | 0.401 | 0.456 |
| | A+B+Ca | A+B+Cb | A+B+Cc | A+Ca+Cb | A+Ca+Cc | A+Cb+Cc | B+Ca+Cb | B+Ca+Cc | B+Cb+Cc | Ca+Cb+Cc |
| J48 | 0.345 | 0.342 | 0.342 | 0.412 | 0.348 | 0.394 | 0.353 | 0.441 | 0.391 | 0.429 |
| LMT | 0.46 | 0.477 | 0.391 | 0.449 | 0.489 | 0.457 | 0.485 | 0.438 | 0.507 | **0.536** |
| ANN | 0.42 | 0.4 | 0.364 | 0.444 | 0.444 | 0.44 | 0.421 | 0.457 | 0.436 | 0.402 |
| Ridor | 0.397 | 0.435 | 0.355 | 0.374 | 0.377 | 0.45 | 0.431 | 0.438 | 0.369 | 0.457 |

| | A+B+Ca+Cb | A+B+Ca+Cc | A+B+Cb+Cc | A+Ca+Cb+Cc | B+Ca+Cb+Cc | A+B+Ca+Cb+Cc |
|---|---|---|---|---|---|---|
| J48 | **0.33** | 0.369 | 0.353 | 0.419 | 0.348 | 0.371 |
| LMT | 0.471 | 0.47 | 0.49 | 0.46 | 0.465 | 0.458 |
| ANN | 0.448 | 0.422 | 0.442 | 0.48 | 0.382 | 0.453 |
| Ridor | 0.412 | 0.387 | 0.473 | 0.35 | 0.424 | 0.416 |

Table 12 and Table 13 show the F measures observed when we repeated the experiments with essays that did not contain Chinese hints originally. Most of the F measures in Table 12 and Table 13 are better than their counterparts in Table 5 and Table 6, but some of them became worse. The best F measure in Table 13 is better than the best one in Table 6, but it is just 0.414.

***Table 12. Using individual groups for essays without Chinese hints***

| | A | B | Ca | Cb | Cc |
|---|---|---|---|---|---|
| J48 | 0.297 | **0.254** | 0.344 | 0.315 | 0.297 |
| LMT | 0.356 | 0.295 | 0.378 | 0.349 | 0.308 |
| ANN | 0.358 | 0.286 | **0.417** | 0.372 | 0.28 |
| Ridor | 0.324 | 0.3 | 0.351 | 0.378 | 0.276 |

***Table 13. Using mixed groups for essays without Chinese hints***

| | A+B | A+Ca | A+Cb | A+Cc | B+Ca | B+Cb | B+Cc | Ca+Cb | Ca+Cc | Cb+Cc |
|---|---|---|---|---|---|---|---|---|---|---|
| J48 | 0.265 | 0.320 | 0.320 | 0.280 | 0.350 | 0.341 | **0.256** | 0.386 | 0.345 | 0.382 |
| LMT | 0.345 | 0.381 | 0.401 | 0.375 | 0.387 | 0.366 | 0.324 | 0.378 | 0.400 | 0.392 |
| ANN | 0.327 | 0.413 | 0.368 | 0.339 | 0.323 | 0.337 | 0.267 | 0.403 | 0.383 | 0.366 |
| Ridor | 0.334 | 0.374 | 0.353 | 0.323 | 0.356 | 0.341 | 0.317 | 0.384 | 0.377 | 0.381 |
| | A+B+Ca | A+B+Cb | A+B+Cc | A+Ca+Cb | A+Ca+Cc | A+Cb+Cc | B+Ca+Cb | B+Ca+Cc | B+Cb+Cc | Ca+Cb+Cc |
| J48 | 0.322 | 0.327 | 0.307 | 0.330 | 0.316 | 0.372 | 0.334 | 0.356 | 0.347 | 0.359 |
| LMT | 0.384 | 0.356 | 0.335 | 0.391 | 0.393 | 0.393 | 0.399 | 0.382 | **0.414** | 0.385 |
| ANN | 0.365 | 0.352 | 0.362 | 0.393 | 0.382 | 0.343 | 0.347 | 0.349 | 0.35 | 0.404 |
| Ridor | 0.349 | 0.36 | 0.382 | 0.340 | 0.335 | 0.368 | 0.391 | 0.333 | 0.33 | 0.367 |

| | A+B+Ca+Cb | A+B+Ca+Cc | A+B+Cb+Cc | A+Ca+Cb+Cc | B+Ca+Cb+Cc | A+B+Ca+Cb+Cc |
|---|---|---|---|---|---|---|
| J48 | 0.352 | 0.294 | 0.348 | 0.380 | 0.310 | 0.345 |
| LMT | 0.386 | 0.388 | 0.369 | 0.381 | 0.396 | 0.400 |
| ANN | 0.380 | 0.379 | 0.349 | 0.390 | 0.353 | 0.389 |
| Ridor | 0.352 | 0.346 | 0.367 | 0.344 | 0.375 | 0.334 |

The F measures reported in Tables 5, 6, 10, 11, 12, and 13 suggested that the natures of essays with and without Chinese hints are different. The chart in Figure 6 shows the average performance of our classifiers when we used 1, 2, 3, and more than 3 feature groups to classify the essays that originally contained Chinese hints. The chart in Figure 7 shows the trends for predicting the levels of the essays that did not contain Chinese hints originally. The charts in Figures 5, 6, and 7 indicate that we achieved the worst performance when we mixed the essays in the corpus. If we separated those essays with and without Chinese hints, we achieved better results for both sub-groups on average. This is quite an interesting discovery, but we do not have a good explanation for this phenomenon.

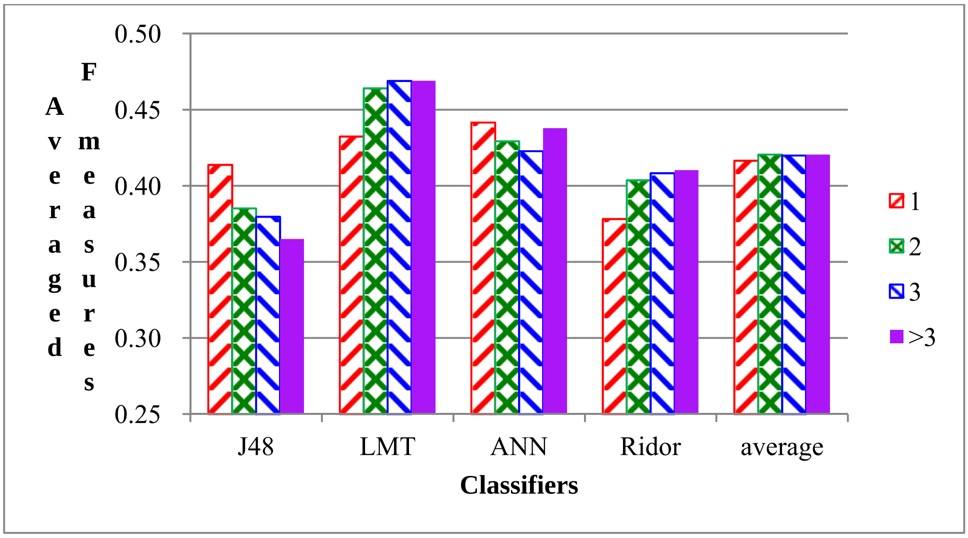


***Figure 6. Predicting readability of essays with Chinese hints was easier***

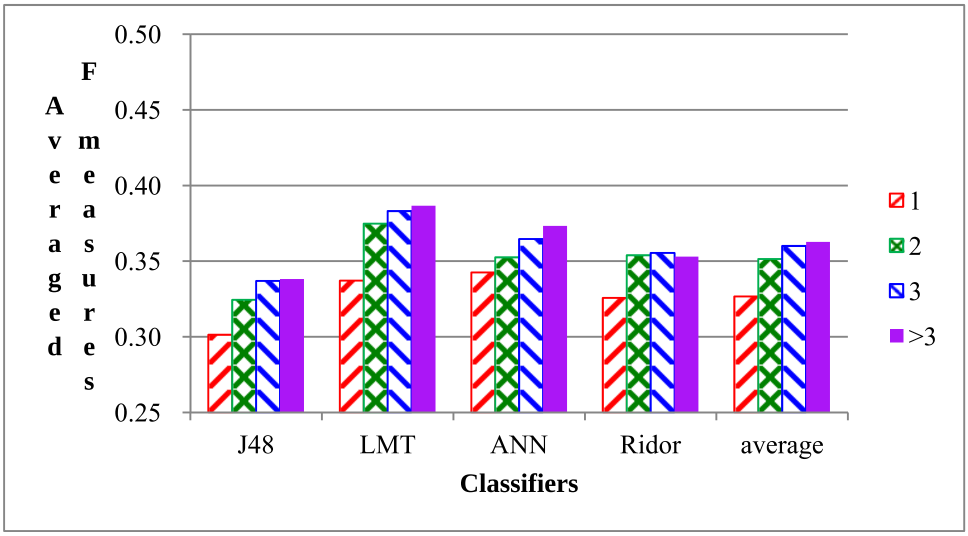


***Figure 7. Predicting readability of essays without Chinese hints was harder***

## 5.7 More Experiments with Syntactic Features

Finally, we explored some conjectural features at the syntax level, and we referred to them as **Group F**. We parsed our corpus with the Stanford parser to collect some statistics: (1) VBN appeared at most 4 times; (2) VP appeared at most 6 times; (3) MD appeared at most 3 times. Hence, we could use 16 features to describe the distributions of VBN, VP, and MD in an essay. In addition, we could use binary features to encode whether an essay contained ADJP, ADVP, and CONJP. This gave us 3 features. Adding the features for distribution of depth (9 features) and grammatical relationships (7 features), we had a total 35 features in Group F.

We repeated the experiments for Tables 11 and 13, after adding Group F to the combinations of features. Results reported in Table 14 are for essays that originally contained Chinese hints, and results reported in Table 15 are for essays that did not contain Chinese hints.

***Table 14. Results of adding syntactic features for essays with Chinese hints***

| | F | A+F | B+F | Ca+F | Cb+F | Cc+F | A+B+F | A+Ca+F | A+Cb+F | A+Cc+F |
|---|---|---|---|---|---|---|---|---|---|---|
| J48 | 0.340 | 0.374 | 0.388 | 0.383 | 0.417 | 0.33 | 0.354 | 0.393 | 0.399 | 0.386 |
| LMT | 0.438 | 0.478 | 0.389 | 0.472 | 0.476 | 0.447 | 0.426 | 0.501 | 0.484 | 0.467 |
| ANN | 0.352 | 0.422 | 0.375 | 0.363 | 0.368 | 0.412 | 0.410 | 0.440 | 0.454 | 0.414 |
| Ridor | 0.390 | 0.426 | 0.373 | 0.328 | 0.385 | 0.298 | 0.385 | 0.403 | 0.386 | 0.377 |

| | B+Ca+F | B+Cb+F | B+Cc+F | Ca+Cb+F | Ca+Cc+F | Cb+Cc+F | A+B+Ca+F |
|---|---|---|---|---|---|---|---|
| J48 | 0.321 | 0.348 | 0.345 | 0.386 | 0.390 | 0.428 | 0.321 |
| LMT | 0.450 | 0.529 | 0.385 | 0.461 | 0.457 | 0.484 | 0.459 |
| ANN | 0.376 | 0.400 | 0.363 | 0.380 | 0.418 | 0.435 | 0.408 |
| Ridor | 0.416 | 0.381 | 0.341 | 0.418 | 0.361 | 0.395 | 0.368 |

| | A+B+Cb+F | A+B+Cc+F | A+Ca+Cb+F | A+Ca+Cc+F | A+Cb+Cc+F | B+Ca+Cb+F |
|---|---|---|---|---|---|---|
| J48 | 0.360 | 0.341 | 0.488 | 0.408 | 0.412 | 0.378 |
| LMT | 0.482 | 0.461 | 0.482 | **0.530** | 0.494 | 0.465 |
| ANN | 0.448 | 0.378 | 0.417 | 0.430 | 0.430 | 0.380 |
| Ridor | 0.415 | 0.373 | 0.423 | 0.384 | 0.393 | 0.384 |

| | B+Ca+Cc+F | B+Cb+Cc+F | Ca+Cb+Cc+F | A+B+Ca+Cb+F | A+B+Ca+Cc+F |
|---|---|---|---|---|---|
| J48 | 0.346 | 0.383 | 0.388 | 0.385 | **0.306** |
| LMT | 0.452 | 0.522 | 0.496 | 0.443 | 0.482 |
| ANN | 0.417 | 0.427 | 0.417 | 0.427 | 0.415 |
| Ridor | 0.345 | 0.400 | 0.423 | 0.385 | 0.404 |

| | A+B+Cb+Cc+F | A+Ca+Cb+Cc+F | B+Ca+Cb+Cc+F | A+B+Ca+Cb+Cc+F |
|---|---|---|---|---|
| J48 | 0.381 | 0.447 | 0.403 | 0.379 |
| LMT | 0.474 | 0.480 | 0.502 | 0.472 |
| ANN | 0.408 | 0.460 | 0.392 | 0.407 |
| Ridor | 0.435 | 0.404 | 0.375 | 0.427 |

***Table 15. Results of adding syntactic features for essays without Chinese hints***

| | F | A+F | B+F | Ca+F | Cb+F | Cc+F | A+B+F | A+Ca+F | A+Cb+F | A+Cc+F |
|---|---|---|---|---|---|---|---|---|---|---|
| J48 | 0.279 | 0.330 | 0.262 | 0.314 | 0.323 | 0.289 | 0.298 | 0.314 | 0.344 | 0.286 |
| LMT | 0.277 | 0.348 | 0.308 | 0.374 | 0.361 | 0.307 | 0.341 | 0.343 | 0.384 | 0.362 |
| ANN | 0.265 | 0.371 | 0.279 | 0.336 | 0.315 | 0.301 | 0.339 | 0.370 | 0.388 | 0.365 |
| Ridor | 0.27 | 0.325 | 0.289 | 0.345 | 0.351 | 0.295 | 0.299 | 0.326 | 0.360 | 0.360 |

| | B+Ca+F | B+Cb+F | B+Cc+F | Ca+Cb+F | Ca+Cc+F | Cb+Cc+F | A+B+Ca+F |
|---|---|---|---|---|---|---|---|
| J48 | 0.307 | 0.328 | 0.290 | 0.356 | 0.326 | 0.329 | 0.262 |
| LMT | 0.366 | 0.371 | 0.326 | 0.395 | 0.378 | 0.362 | 0.353 |
| ANN | 0.348 | 0.327 | 0.299 | 0.346 | 0.348 | 0.343 | 0.375 |
| Ridor | 0.342 | 0.320 | **0.254** | 0.376 | 0.344 | 0.366 | 0.315 |

| | A+B+Cb+F | A+B+Cc+F | A+Ca+Cb+F | A+Ca+Cc+F | A+Cb+Cc+F | B+Ca+Cb+F |
|---|---|---|---|---|---|---|
| J48 | 0.343 | 0.302 | 0.337 | 0.325 | 0.343 | 0.288 |
| LMT | 0.388 | 0.340 | 0.386 | 0.340 | 0.374 | 0.396 |
| ANN | 0.370 | 0.365 | 0.378 | **0.402** | 0.384 | 0.350 |
| Ridor | 0.35 | 0.317 | 0.387 | 0.326 | 0.345 | 0.353 |

| | B+Ca+Cc+F | B+Cb+Cc+F | Ca+Cb+Cc+F | A+B+Ca+Cb+F | A+B+Ca+Cc+F |
|---|---|---|---|---|---|
| J48 | 0.275 | 0.329 | 0.350 | 0.341 | 0.292 |
| LMT | 0.348 | 0.370 | 0.377 | 0.378 | 0.350 |
| ANN | 0.374 | 0.307 | 0.383 | 0.374 | 0.386 |
| Ridor | 0.338 | 0.314 | 0.362 | 0.372 | 0.321 |

| | A+B+Cb+Cc+F | A+Ca+Cb+Cc+F | B+Ca+Cb+Cc+F | A+B+Ca+Cb+Cc+F |
|---|---|---|---|---|
| J48 | 0.351 | 0.346 | 0.326 | 0.344 |
| LMT | 0.380 | 0.370 | 0.389 | 0.393 |
| ANN | 0.376 | 0.406 | 0.338 | 0.378 |
| Ridor | 0.349 | 0.368 | 0.395 | 0.358 |

Although we wished to observe improved results when used these more complex features, the outcome was not encouraging. In general, the F measures in Table 14 were lower than their counterparts in Table 11. For instance, using B+Ca achieved 0.427 in Table 11, but using F+B+Ca achieved only 0.321 in Table 14. The same problem can be verified for corresponding numbers in Table 13 and Table 15. In fact, the drops from the numbers in Table 13 to the corresponding numbers in Table 15 were more severe.

Intuitively, considering syntactic features should have improved our results. Nevertheless, we probably did not choose the right features. Another possibility would be that the challenging levels of the short essays used in the comprehension tests in Taiwan simply did not relate to syntactic factors.

## 6. Concluding Remarks

A random classification of an essay into four categories would have achieved only 25% in accuracy on average. We considered features at the word, sentence, and essay levels in this classification task, and we found that it was possible to improve the F measure from 0.381 (Table 6) to 0.536 (Table 11). The best F measures were observed in 10-fold cross-validation tests for LMT in Weka. Not all classifiers achieved the same quality of classification. Among the four types of classifiers we used in this study, LMT performed the best on average.

The identified improvement was not small, but it was not significant enough either. The problem of determining levels of readability may not be as easy as the public scores suggested. We analyzed our corpus with the SMOG scores in Section 5.1, and found that the essays of supposedly more challenging levels may not have higher SMOG scores than the scores of the supposedly easer essays.

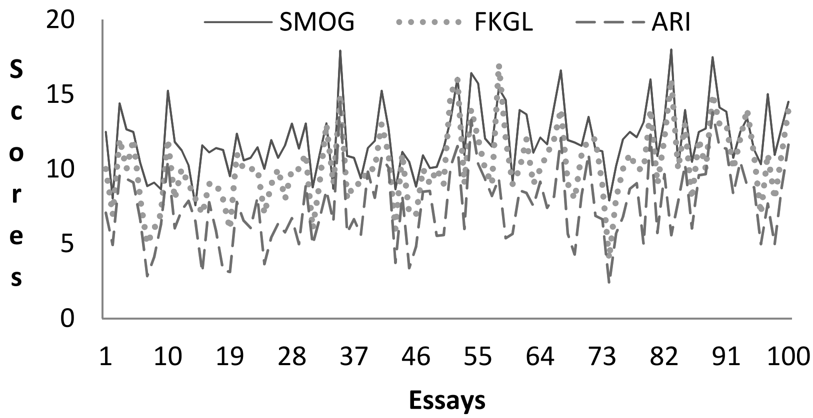


***Figure 8. Readability scores of more popular formulae***

We explored two additional scores for readability. In Figure 8, we show the SMOG, FKGL[19], and ARI[20] scores for 100 arbitrarily chosen essays from our corpus. The curves show rather strong similarity, which is not very surprising to us. These score functions rely mainly

[19] Flesch-Kincaid Grade Level. http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test

[20] Automated readability index. http://en.wikipedia.org/wiki/Automated_Readability_Index

on the word counts of different levels of words and the number of sentences in an essay. Hence, if using SMOG would not achieve good results for the classification task in our study (*cf*. Figure 4), then using the other two alternatives would not achieve much better results either.

One challenge to our work is whether we should consider only the short essays and classify the levels of the comprehension tests. A comprehension test contains the essay part and the question part. Obviously, we should take the questions into consideration in the classification task, which we have not begun yet. In addition, due to the "examination-centered" style of education in Taiwan, the same short essay may be reused in tests of students of higher classes. Such a reuse of short essays made our classification more difficult, because that made the "correct class" of an essay rather ambiguous.

Whether linguistic features were sufficient for the determination of readability of essays is also an issue. Understanding an essay may require domain-dependent knowledge that we have not attempted to encode with our features (Carrell, 1983). Culture-dependent issues may also play a role (Carrell, 1981). Hence, more features are needed to accomplish more improvement on the predication of readability, *e.g*. (Crossley *et al*., 2008; Zhang, 2008).

A review comment suggested that there might not be sufficient differences in the short essays used in the first and the second semesters of a school year, so trying to classify the short essays into three levels (each for a school year) may be more practical. Although we did not move our work in this direction, we think the suggestion is interesting.

A reviewer noticed an interesting crossing point in Figure 4. The SMOG score at 11.5 seems to be a major point for the curves in Figure 4 to intersect. A similar phenomenon appeared in Figure 8, where approximately half of the scores of the 100 essays were above 11.5. Whether 11.5 is the watershed of the easy and difficult essays is an interesting hypothesis to verify with a larger amount of essays.

## Acknowledgments

The work was supported in part by the funding from the National Science Council in Taiwan under the contracts NSC-97-2221-004-007, NSC-98-2815-C-004-003-E, and NSC-99-2221-004-007. The authors would like to thank Miss Min-Hua Lai for her technical support in this study and Professor Zhao-Ming Gao for his comments on an earlier report (Kuo *et al*., 2009) of this paper.

## References


Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater V.2, *Journal of Technology, Learning, and Assessment*, 4(3), 3-30.

Bailin, A. & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique, *Language and Communication*, 21(2), 285-301, 2001.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays, *IEEE Intelligent Systems*, 18(1), 32-39.

Carrell, P. L. (1981). Culture-specific schemata in L2 comprehension, *Selected Papers from the Ninth Illinois TESOL/BE Annual Convention, the First Midwest TESOL Conference*, 123-132.

Carrell, P. L. (1983). Some issues in studying the role of schemata or background knowledge in second language comprehension, *Reading in a Foreign Language*, 1(1), 81-92.

Chall, J. & Dale, E. (1995). *Readability Revisited: The new Dale-Chall Readability Formula*. Brookline Books.

Chang, T.-H., Lee, C.-H., & Chang, Y.-M. (2006). Enhancing automatic Chinese essay scoring system from figures-of-speech, *Proceedings of the Twentieth Pacific Asia Conference on Language, Information and Computation*, 28-34.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices, *TESOL Quarterly*, 42(3), 475-493.

Flesch, R. (1948). A New Readability Yardstick, *Journal of Applied Psychology*, 32(3), 221-233.

Huang, C.-S., Kuo, W.-T., Lee, C.-L., Tsai, C.-C., & Liu, C.-L. (2010). Using linguistic features to classify texts for reading comprehension tests at the high school levels, *Proceedings of the Twenty Second Conference on Computational Linguistics and Speech Processing* (ROCLING XXIII), 98-112. (in Chinese)

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

Kuo, W.-T., Huang, C.-S., Lai, M.-H., Liu, C.-L., & Gao, Z.-M. (2009). 適用於中學英文閱讀測驗短文分類的特徵比較, *Proceedings of the Fourteenth Conference on Artificial Intelligence and Applications*. (in Chinese)

Lin, S.-Y., Su, C.-C., Lai, Y.-D., Yang, L.-C., & Hsieh, S.-K. (2009). Assessing text readability using hierarchical lexical relations retrieved from WordNet, *International Journal of Computational Linguistics and Chinese Language Processing*, 14(1), 45-84.

MOE. (2008). http://www.edu.tw/eje/content.aspx?site_content_sn=15326

Shih, R. H., Chiang, J. Y., & Tien, F. (2000). Part-of-speech sequences and distribution in a learner corpus of English, *Proceedings of Research on Computational Linguistics Conference* XIII (ROCLING XIII), 171-177.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Zhang, X. (2008). The effects of formal schema on reading comprehension – An experiment with Chinese EFL readers, *International Journal of Computational Linguistics and Chinese Language Processing*, 13(2), 197-214.