US-China Education Review B 1 (2012) 55-62
Earlier title: US-China Education Review, ISSN 1548-6613

# Use of a New Set of Linguistic Features to Improve Automatic Assessment of Text Readability

Takehiko Yoshimi
Ryukoku University, Shiga, Japan

Katsunori Kotani
Kansai Gaidai University, Osaka, Japan

Hitoshi Isahara
Toyohashi University of Technology, Aichi, Japan

The present paper proposes and evaluates a readability assessment method designed for Japanese learners of EFL (English as a foreign language). The proposed readability assessment method is constructed by a regression algorithm using a new set of linguistic features that were employed separately in previous studies. The results showed that the proposed readability assessment method, which used all the linguistic features employed in previous studies, yielded a lower error of assessment than readability assessment methods using only some of these linguistic features.

*Keywords:* computer-assisted language learning, EFL (English as a foreign language), readability, natural language processing

## Introduction

In foreign language teaching, providing learners with reading materials matched to their proficiency level is known to be effective. When selecting authentic reading materials of unknown readability, however, teachers must manually assess the readability of the materials, placing a heavy burden on teachers. One way to reduce this burden is to automate readability assessment. In order to achieve this goal, various readability assessment methods have been proposed (Flesch, 1948; Nagata, Masui, Kawai, & Siino, 2004; Schwarm & Ostendorf, 2005; Kotani, Yoshimi, & Isahara, 2011).

In constructing a readability assessment method using a regression algorithm, it is important to select linguistic features that significantly affect text readability. Previous readability assessment methods used linguistic features, such as average word length, average sentence length, number of nodes of a syntactic tree and grammatical constructions that are difficult for learners to comprehend. Although these methods yielded some valid results, readability assessment needs to be further improved for its practical use in computer-assisted language teaching.

The present paper proposes a readability assessment method using a new set of linguistic features that were employed separately in previous studies, which is expected to reduce the error of assessment. The proposed readability assessment method is constructed using a regression algorithm. The independent variables are various linguistic features and the dependent variable is the readability score for Japanese learners of EFL

Takehiko Yoshimi, associate professor, Department of Media Informatics, Faculty of Science and Technology, Ryukoku University.
Katsunori Kotani, associate professor, College of Foreign Studies, Kansai Gaidai University.
Hitoshi Isahara, professor, Information and Media Center, Toyohashi University of Technology.

(English as a foreign language). The proposed readability assessment method takes a text as input, extracts linguistic features of the text and estimates readability scores based on the extracted linguistic features. We report herein the experimental results of comparison between the readability assessment method using all the linguistic features and assessment methods using only some of these linguistic features.

## New Set of Linguistic Features

Of the linguistic features used in previous studies (Flesch, 1948; Nagata et al., 2004; Schwarm & Ostendorf, 2005; Kotani et al., 2011), the proposed readability assessment method employed all of the features as shown in Table 1.

Table 1

*Linguistic Features Employed in Previous Studies*

| | Lexical features | Syntactic features | Discourse features |
|---|---|---|---|
| Flesch (1948) | Average word length | Average sentence length | |
| Nagata et al. (2004) | Word difficulty score | Length of relative clauses, present-participle clauses and past-participle clauses | |
| Schwarm and Ostendorf (2005) | Average word length | Average sentence length, average number of noun phrases, verb phrases, and subordinate conjunctions | |
| Kotani et al. (2011) | Word difficulty score, number of un-registered words and word senses | Number of the nodes of a syntactic tree and number of nodes stored in short-term memory | Number of pronouns |

Average word length, used by Flesch (1948) and Schwarm and Ostendorf (2005), is computed as the ratio of the number of syllables divided by the number of words. Word difficulty scores, used by Nagata et al. (2004) and Kotani et al. (2011), represent the difficulty experienced by Japanese EFL learners in comprehending the words. Word difficulty scores were assigned based on a word level list containing more than 35,000 English words and providing difficulty scores for 11 levels (Someya, 2000). Number of un-registered words, used by Kotani et al. (2011), addresses the problem that authentic texts tend to contain words that are not registered in the word level list. Number of word senses, used by Kotani et al. (2011), addresses the problem that basic words in the word level list might be more difficult than expected. Number of word senses was counted using Word Net 2.0 (Fellbaum, 1998), a large lexical database of the English language.

Average sentence length, used by Flesch (1948) and Schwarm and Ostendorf (2005), is computed as the ratio of the number of words in the text divided by the number of sentences. Length of relative clauses, present-participle clauses and past-participle clauses, used by Nagata et al. (2004), focuses on grammatical constructions that are typically difficult for Japanese EFL learners to comprehend. Grammatical constructions (syntactic trees) were generated by Apple Pie Parser (Sekine & Grishman, 1995). Number of nodes of a syntactic tree takes into account the presence or absence of specific grammatical constructions that affect the reading comprehension of Japanese EFL learners. Number of nodes, stored in short-term memory (Yngve, 1960), explains memory load during psychological syntactic parsing.

Number of pronouns, used by Kotani et al. (2011), indicates the complexity of the discourse structure, as comprehension of a text requires identifying referents of pronouns during reading.

## Comprehension Rate Data Collection

In addition to the linguistic features reviewed above, readability scores were used as training data for regression in order to develop the proposed readability assessment method. Here, “readability score” refers to the comprehension rate, which is computed by dividing the number of correct answers by the number of comprehension questions on a text (ranging from 0.0 to 1.0).

Comprehension rate data were collected as follows. Sixty four paid participants were chosen on the basis of the following criteria that whose native language is Japanese and have taken the TOEIC (test of English for international communication, Retrieved from http://www.ets.org/toeic), a test of English language skills used in the workplace, within the previous one-year period.

We prepared test sets based on 84 texts extracted from the TOEIC preparation textbooks (Arbogast, Duke, Locke, Shearin, Bicknell, & Chauncey Group International, 2001; Lougheed, 2003). Each test set consisted of seven texts, and every test set contained different texts. Each text was accompanied by two to five multiple choice comprehension questions. We randomly provided participants with one or two test sets. Thirty one participants took one test set and 33 participants took two test sets.

Comprehension rate data were collected using a reading process recording tool (Yoshimi, Kotani, Kutsumi, Sata, & Isahara, 2005). This tool displays one sentence at a time. A sentence appears on the computer screen when the cursor is positioned over a reading icon, and it disappears when the cursor is moved away from the icon.

Participants used this tool while reading the text and answering the comprehension questions. When the cursor was positioned over a question icon, a comprehension question appeared. Participants answered the question by clicking on one of the answer icons.

After receiving instructions about the tool, participants practiced by reading several sample texts and answering sample comprehension questions. The participants were instructed first to read the text and then answer the comprehension questions. We also directed participants to attempt to understand the text well enough to correctly answer the comprehension questions. Since we did not impose time constraints, the participants could take as much time as they needed. In order to reduce the pressure on the participants, we did not inform them that the tool would be measuring their reading times.

We excluded comprehension rate data of four participants whose reading speed (WPM (words per minute)) was extremely fast or slow (> 200 WPM or < 70 WPM), as slow reading speed might have been the result of unnecessarily careful reading, and excessively fast reading speed could indicate that participants did not properly read the material (average reading speed of native English speakers is reported to be in the range of 200 WPM to 300 WPM (Carver, 1982)). We obtained 451 instances of comprehension data. One instance consists of the linguistic features of a text and the comprehension rate when a Japanese EFL learner reads the text. The mean age of the participants whose comprehension rate data were included in analysis was 29.8 years (S.D. (standard deviation): 9.5). Nine participants were males and 51 were females.

The distribution of the comprehension rate data is shown in Figure 1. The comprehension rate data comprise 10 values from 0.0 to 1.0. Each value refers to the comprehension rate calculated by dividing the number of correct answers (0 to 5) by the number of comprehension questions (2 to 5). The comprehension rate data showed a skewed distribution (plotted with a dotted line), because the instances of comprehension rate 1.0 comprised 59.6% of all instances (269 out of 451 instances). One reason for the high proportion of comprehension rate 1.0 could be the fact that the absence of time restriction in this experiment allowed the participants to spend as much time as they wanted to complete each question.

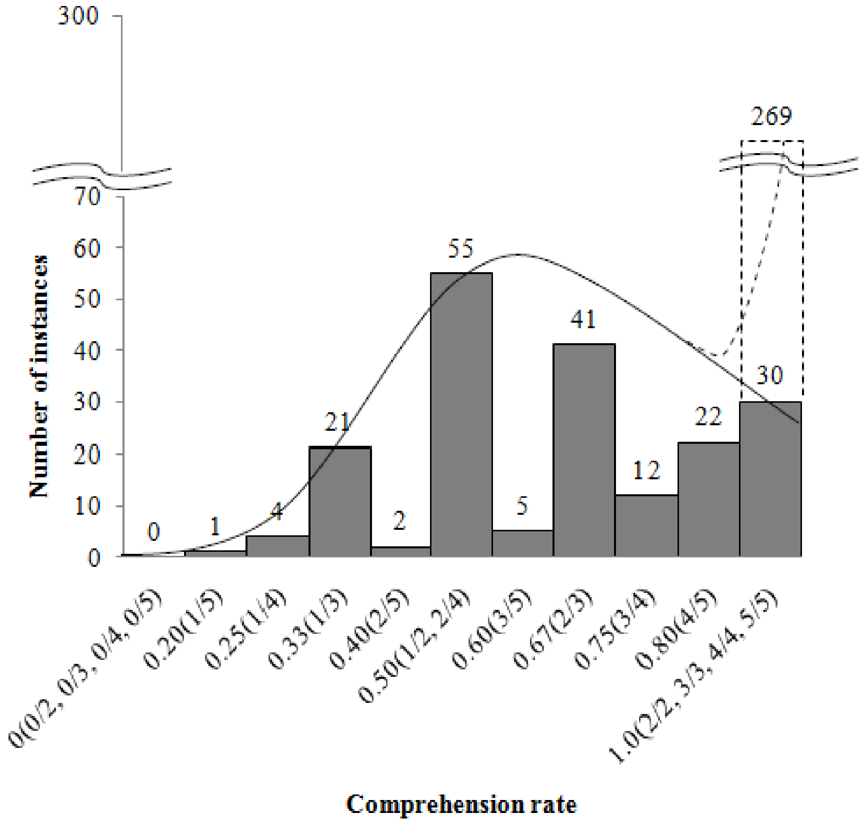

*Figure 1.* Graph of comprehension rate data.

As the target of this readability assessment method was texts intended for use in norm-referenced tests, the text readability scores should follow a normal distribution. Since a readability assessment method trained with skewed data estimates skewed values, it is highly likely that a method trained with skewed data cannot properly assess the readability of texts intended for use in norm-referenced tests. To address this problem, we modified the distribution of the comprehension rate data by randomly selecting 30 instances of comprehension rate 1.0. The modified comprehension rate (plotted with an actual line) included 193 instances. We considered this to be a roughly normal distribution.

## Evaluation Experiment

In this section, we describe experiments for the evaluation of the proposed readability assessment method.

### Experimental Method

The readability assessment methods were evaluated using the 193 instances of comprehension rate data described in above. The evaluation was performed using five-fold cross-validation tests.

Support vector regression (Vapnik, 1998) was carried out using an algorithm implemented in mySVM (support vector machine) software (Retrieved from http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html). The $d$-th polynomial kernel function ($d$ = 1, 2, 3 and 4) was selected and the other settings remained at the default values.

The performance of the readability assessment methods was examined in terms of the absolute error (the absolute value of the difference between the estimated value and the observed value). The estimated values refer to the readability scores calculated with the readability assessment methods, and the observed value indicates the Japanese EFL learners' actual comprehension rate obtained in the data collection described in above. The absolute error shows the degree to which the readability assessment methods correctly indicate the readability scores.

The proposed readability assessment method was constructed with all the linguistic features (used by Flesch (1948), Kotani et al. (2011), Nagata et al. (2004) and Schwarm and Ostendorf (2005)). In order to examine the appropriateness of the proposed readability assessment method, we compared the absolute error of the proposed readability assessment method with that of readability assessment methods constructed using only some of the linguistic features. Hereafter, the readability assessment methods are referred to using the study authors' initials. A readability assessment method using features of Flesch (1948) is referred to as F method, a method using features of Kotani et al. (2011) as K method, a method using features of Nagata et al. (2004) as N method, and a method using features of Schwarm and Ostendorf (2005) as S method. The name of a method using features of multiple methods is from the combination of the names of the relevant methods. For instance, a method using features of F method, K method and S method is referred to as FKS method.

### Experimental Results

Table 2 shows the median absolute errors of the readability assessment methods in ascending order. The name of each method is followed by a bracketed number that indicates the *d*-th polynomial kernel function showing the lowest median absolute error.

Table 2

*Results of Readability Assessment Methods*

| Ranking | Assessment method | Median absolute error | Range | Difference from FKNS method | F | K | N | S |
|---|---|---|---|---|---|---|---|---|
| 1 | FKS method (*d* = 1) | 0.094 | 0.669 | -0.010 | ✓ | ✓ | | ✓ |
| 2 | KNS method (*d* = 1) | 0.097 | 0.615 | -0.006 | | ✓ | ✓ | ✓ |
| 3 | KS method (*d* = 1) | 0.099 | 0.584 | -0.004 | | ✓ | | ✓ |
| 4 | FKNS method (*d* = 1) | 0.104 | 0.666 | 0.000 | ✓ | ✓ | ✓ | ✓ |
| 5 | FK method (*d* = 1) | 0.124 | 0.609 | 0.020 | ✓ | ✓ | | |
| 6 | S method (*d* = 1) | 0.124 | 0.582 | 0.020 | | | | ✓ |
| 7 | FNS method (*d* = 1) | 0.126 | 0.685 | 0.022 | ✓ | | ✓ | ✓ |
| 8 | FKN method (*d* = 1) | 0.127 | 0.590 | 0.024 | ✓ | ✓ | ✓ | |
| 9 | FS method (*d* = 1) | 0.128 | 0.520 | 0.024 | ✓ | | | ✓ |
| 10 | F method (*d* = 2) | 0.132 | 0.471 | 0.028 | ✓ | | | |
| 11 | N method (*d* = 2) | 0.133 | 0.474 | 0.029 | | | ✓ | |
| 12 | KN method (*d* = 2) | 0.134 | 0.725 | 0.030 | | ✓ | ✓ | |
| 13 | K method (*d* = 2) | 0.135 | 0.818 | 0.031 | | ✓ | | |
| 14 | FN method (*d* = 2) | 0.139 | 0.508 | 0.035 | ✓ | | ✓ | |
| 15 | NS method (*d* = 4) | 0.143 | 0.637 | 0.039 | | | ✓ | ✓ |
| Sum of ranking frequency | | | | | 58 | 48 | 73 | 47 |

Although the proposed readability assessment method (FKNS method) achieved the fourth lowest median absolute error (0.104), this was relatively low, as the difference between this error and the lowest median absolute error (FKS method, 0.0904) is only 0.010. The significance of the difference between the median absolute error of the FKNS method and that of the FKS method was examined using the Wilcoxon pair matched rank sum test. A significant difference was not found ($p = 0.78$). In addition, the median absolute error of the FKNS method was lower than that of the S method, which had the lowest error among the F, K, N and S methods. The significance of the difference between the median absolute error of the FKNS method and that of the S method was examined using the Wilcoxon pair matched rank sum test. A significant difference was found ($p < 0.05$). These results suggest the effectiveness of the FKNS method, which employs all of the linguistic features used in the F, K, N and S methods.

Check marks on the right side of Table 2 indicate features used to construct the readability assessment method. For instance, as the features used to construct the FKS method were those from the F, K and S methods, a check mark appears in the F, K and S methods. The sum of ranking frequency is defined as the sum of ranking points marked with a check mark in each method. For instance, because methods using features of the F method are ranked first, fourth, fifth, seventh to 10th and 14th, the sum of ranking frequency of the F method (58) was computed by adding these ranking points. More accurate readability assessment methods have a lower sum of ranking frequency. The sums of ranking frequency of the K and S methods (48 and 47, respectively) were less than those of F and N methods (58 and 73, respectively). This suggests the effectiveness of the linguistic features of the K and S methods.

Figures 2, 3 and 4 show the results of the FKS method (the method with the lowest median absolute error), FKNS method (the proposed readability assessment method) and NS method (the method with the highest median absolute error), respectively. The graph in Figure 2 (FKS method) and that in Figure 3 (FKNS method) do not differ significantly. On the other hand, the graph in Figure 3 (FKNS method) and that in Figure 4 (NS method) differ markedly, in that the FKNS method shows a steeper initial rise in the cumulative relative frequency as compared to the NS method. That is, the FKNS method has a cumulative frequency of 47.7% in the absolute error range from 0.0 to 0.1, and 71.5% in the range from 0.1 to 0.2, whereas the NS method has a cumulative frequency of 38.9% in the absolute error range from 0.0 to 0.1 and 59.1% in the range from 0.1 to 0.2. This suggests the validity of the FKNS method.

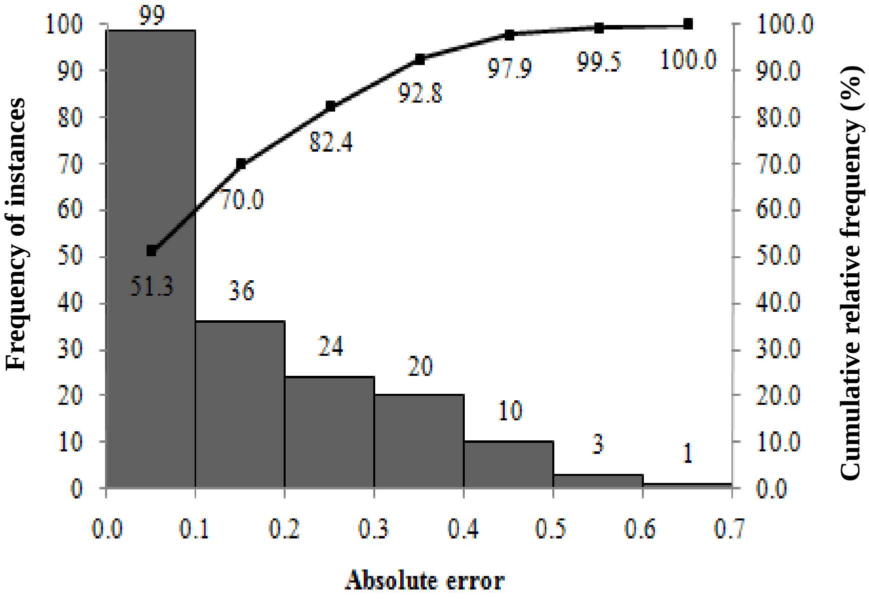

*Figure 2.* Graph of absolute error of the FKS method.

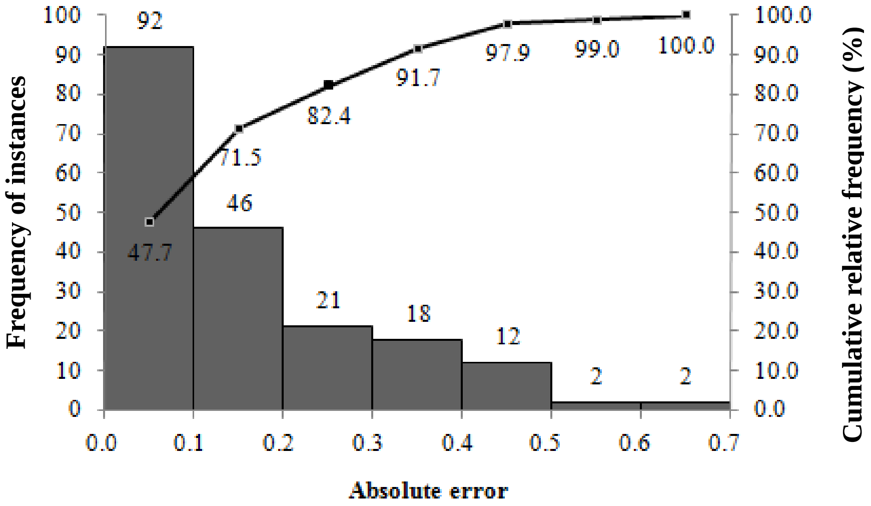

*Figure 3.* Graph of absolute error of the FKNS method.

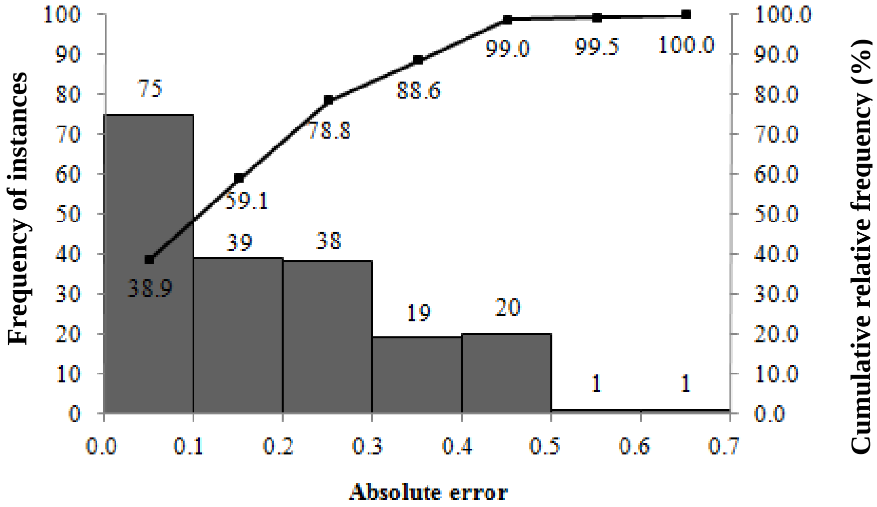

*Figure 4.* Graph of absolute error for NS method.

## Conclusions

We proposed a readability assessment method for Japanese EFL learners using a new set of linguistic features that were used separately in previous studies. Although the median absolute error value of the proposed readability assessment method was not the lowest among the methods tested, the difference between the median absolute error of the proposed method and the lowest median absolute error was not statistically significant. In contrast, a statistically significant difference ($p$ < 0.05) was found between the median absolute error of the proposed readability assessment method and that of the previous readability assessment methods (F, K, N and S methods). These experimental results indicate that the proposed readability assessment method can assess text readability more effectively than previous readability assessment methods.

The present paper leaves several problems unresolved. First, the absolute error of the proposed readability assessment method should be further reduced. Second, the possibility of incorporating the features of learners, such as the scores of TOEIC into the proposed readability assessment method, should be investigated.

## References

Arbogast, B., Duke, T., Locke, M., Shearin, R., Bicknell, J., & Chauncey Group International. (2001). *TOEIC official test-preparation guide: Test of English for international communication*. Lawrenceville, N. J.: Peterson's.

Carver, R. P. (1982). Optimal rate of reading prose. *Reading Research Quarterly, 18*(1), 56-88.

Fellbaum, C. (1998). *Word Net: An electronic lexical database*. Cambridge, M. A: The MIT Press.

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

JACET. (1993). *JACET 4000 basic words*. Tokyo: The Japan Association of College English Teachers.

Kotani, K., Yoshimi, T., & Isahara, H. (2011). A machine learning approach to assessment of text readability for EFL learners using various linguistic features. *US-China Education Review, 1*(6), 767-777.

Lougheed, L. (2003). *How to prepare for the TOEIC test: Test of English for international communication*. Hauppanuge, N. Y.: Barron's Educational Series, Inc..

Nagata, R., Masui, F., Kawai, A., & Siino, T. (2004). A method of rating English texts by reading level for Japanese learners of English. *The Transactions of the Institute of Electronics, Information and Communication Engineers. J-87-D-II*(6), 1329-1338.

Sano, H., & Ino, M. (2000). Assessment of difficulty on English grammar and automatic analysis. *IPSJ SIG Notes, 117*, 5-12.

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. Proceedings of *the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 523-530). June 25-30, University of Michigan, USA.

Sekine, S., & Grishman, A. (1995). A corpusbased probabilistic grammar with only two non-terminals. Proceedings of *the 4th International Workshop on Parsing Technologies* (pp. 216-223). September 20-24, Prague and Karlovy Vary, Czech Republic.

Someya, Y. (2000). *Word level checker: Vocabulary profiling program by AWK, 1.5.* Retrieved from http://www1.kamakuranet.ne.jp/someya/wlc/wlc_manual.html

Vapnik, V. (1998). *Statistical learning theory*. N. Y.: Wiley-Interscience.

Yngve, V. H. (1960). A model and a hypothesis for language structure. *The American Philosophical Society, 104*(5), 444-466.

Yoshimi, T., Kotani, K., Kutsumi, T., Sata, I., & Isahara, H. (2005). A method of measuring reading time for assessing EFL learners' reading ability. *Transactions of Japanese Society for Information and Systems in Education, 22*(1), 24-29.