

EFL Learner Reading Time Model for Evaluating Reading Proficiency

Katsunori Kotani^{1,2}, Takehiko Yoshimi^{3,2}, Takeshi Kutsumi⁴,
Ichiko Sata⁴, and Hitoshi Isahara²

¹ Kansai Gaidai University
16-1 Nakamiya Higashino-cho, Hirakata, Osaka, Japan
kat@khn.nict.go.jp

² National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan

³ Ryukoku University
1-5 Yokotani, Seta Oe-cho, Otsu, Shiga, Japan

⁴ Sharp Corporation
492 Minosho-cho, Yamatokoriyama, Nara, Japan

Abstract. We propose a reading time model for learners of English as a foreign language (EFL) that is based on a learner's reading proficiency and the linguistic properties of sentences. Reading proficiency here refers to a learner's reading score on the Test of English for International Communications (TOEIC), and the linguistic properties are the lexical, syntactic and discourse complexities of a sentence. We used natural language processing technology to automatically extract these linguistic properties, and developed a model using multiple regression analysis as a learning algorithm in combining the learner's proficiency and linguistic properties. Experimental results showed that our reading time model predicted sentence-reading time with a 22.9% error rate, which is lower than the models constructed based on linguistic properties proposed in previous studies.

1 Introduction

One of the critical issues in learning or teaching a foreign language is learners' individual differences in proficiency. Unlike first language acquisition, proficiencies in acquiring a foreign language vary greatly. Thus, a language teacher has to understand each learner's problems and help the learner contend with them. The learners' problems principally arise from lack of lexical or syntactic knowledge. For instance, if a learner encounters a lexical item the meaning of which the learner does not know, he or she has to guess the meaning based on contextual information. Reading such a sentence should take more time than reading a sentence without unknown lexical items. Given this, some learners' problems can be identified by measuring his or her reading time, because encountering unfamiliar lexical or syntactic items will interrupt the reading process [1].¹

¹ Reading process refers to a series of understanding tasks from word meaning to sentence/discourse meaning. It involves word recognition, syntactic parsing and semantic composition.

The reading process is typically measured with the following metrics: (i) reading time, (ii) eye-movement and (iii) brain activity [7]. Of these metrics, reading time is more easily applicable to language classrooms than the others from the viewpoint of cost. Reading time tests, which use reading time as an evaluation metric, seem to be unpopular as pedagogical tests. Recent research, however, has confirmed the reliability and validity of reading time tests as a metric of foreign language reading proficiency [15, 10].² Based on these findings, we decided to use reading time as a measure of a learner's reading problems.

In this paper, we present our reading time model (RT model), which functions as the baseline in identifying a sentence which might include reading problems. The RT model predicts the time required for learners of English as a foreign language (EFL) to read a sentence based on both the linguistic properties of a sentence and a learner's reading proficiency. In our approach, a learner's level of reading problems is identified by comparing a learner's actual reading time with the reading time predicted by the RT model. A remarkable difference between a learner's actual reading time and the predicted time might indicate reading problems. Without RT model, a teacher has to manually set up the reading time for a learner to read. This task is costly and time consuming, especially when there are a lot of sentences. The RT model automatically sets up the model reading time, thereby assisting a language teacher in identifying a learner's problems.

In our experiment, the RT model was derived with a multiple regression analysis, which took reading time as a dependent variable and linguistic and learners' properties as independent variables (discussed in § 2). This model was able to predict sentence-reading time with a 22.9% error rate.

2 Features

Sentence-reading time should vary depending on the linguistic properties of a sentence and a learner's reading proficiency, as previous linguistic/psycholinguistic studies have reported [2, 8]. Linguistic properties include lexical, syntactic and discourse factors. Of these factors, we picked linguistic factors, which can be automatically derived with state-of-the-art natural language processing tools, because the goal of this study is to implement the RT model into the Computer Assisted Language Learning (CALL) system. In the rest of this section, we will review the features used to construct the RT model.

2.1 Lexical Factors

The RT model uses word length and lexical difficulty as lexical factors that should affect sentence-reading time. It is supposed that the length of a word is positively correlated with its lexical difficulty, as research on readability often uses word length to determine readability [4, 16]. Based on this idea, we speculated that the length of a word should affect reading time and used word length as a lexical feature. We defined word length as the number of characters in a word.

² Reading time-based evaluation should not exclude comprehension test-based evaluation, and these methods are fully compatible in identifying a learner's reading problems.

Word length was not the sole factor in lexical difficulty, because some short words are hard for EFL learners to understand [12]. Therefore, we added lexical difficulty scores heuristically derived based on JACET 4000 Basic Words [6], which classifies the difficulty of English vocabularies based on the empirical observations of English teachers working with Japanese EFL learners. Lexical difficulty was then rated using a ranking tool [17]. We measured lexical difficulty based on the scores derived with the tool [17], and summed the scores of words in a sentence as another lexical feature.

2.2 Syntactic Factor

The RT model uses sentence length and the number of branching nodes as syntactic factors that may affect sentence-reading time. Sentence length is supposed to be negatively correlated with readability [4, 16]. Therefore, we used sentence length as a syntactic feature in the RT model. A sentence's length was defined as the number of words it contained.

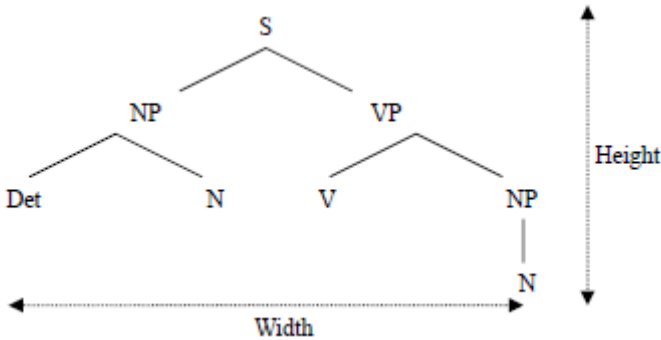


Fig. 1. Syntactic Tree

Sentence length is equivalent to the width of a syntactic tree, as shown in Figure 1. In addition to the width of the tree, we speculated that the height of a tree also indicates syntactic difficulty. To take into account both the width and height of a tree, we decided to use the number of branching nodes as another syntactic factor. Previous research [8] showed that the number of branching nodes was closely correlated with EFL learners' reading times. In addition to this empirical support, the number of branching nodes should be supported from the viewpoint of research on the garden-path sentence presented by [5]. We used the Apple Pie Parser to generate syntactic trees [14], and measured the number of branching nodes based on the trees.

2.3 Discourse Factor

In understanding the meaning of discourse, identifying anaphors is a crucial problem, and a pronoun is a typical anaphoric expression. Although there are other anaphoric

expressions, such as definite expressions, we decided to use the number of pronouns as our discourse feature because it is relatively easy to measure.

The RT model picked the number of pronouns as a discourse factor. In understanding a pronoun, a learner has to identify the referent of a pronoun. Hence, the number of pronouns should indicate discourse complexity.

2.4 Learner Factors

The RT model also takes into account a learner's reading proficiency. Among various metrics for measuring the reading proficiency of EFL learners, such as comprehension test scores, word recognition time or grammatical judgment time, we used a learner's reading score on the Test of English for International Communications (TOEIC) as a learner factor.

Learner factors should involve other factors, such as non-verbal factors, e.g., the degree of interest, motivation, and background knowledge. If one wants to construct a reading model employing these non-verbal factors, these factors have to be surveyed. In this study, we focus on learners' linguistic proficiency as a learner factor, so the RT model does not take non-verbal factors into account. Some research has also suggested that non-verbal factors less clearly affect the reading time of EFL learners [10]. We leave non-verbal factors for further studies.

3 Data Set

The RT model was built with the multiple regression analysis using linguistic and EFL learners' features (stated in §2) as dependent variables. The independent variable was the reading time required for Japanese EFL learners to read English passages from TOEIC textbooks.

EFL learners' reading times were collected in the following procedure. Reading materials were extracted from a TOEIC preparation textbook [9]. Eighty-four passages were chosen and classified into two test groups, consisting either of 7 or 14 passages.

The participants were recruited from a job information website on the conditions that (i) the participant submit a TOEIC score sheet, (ii) the participant was an English learner, and (iii) the participant should live close to the site of the experiment. Of those who responded, 64 participants were chosen based on their having taken a TOEIC test in the year preceding the experiment. Each participant was randomly provided with either a 7-passage text set or a 14-passage text. 31 participants took the seven-passage text test, and 33 participants took the 14-passage text test.

Sentence-reading time was recorded with a reading process recording tool [19]. Participants read sentences displayed on a computer monitor and answered comprehension questions after reading a passage. The rate of correct answers to comprehension questions was used to determine outliers.

The reading process recording tool measures sentence-reading time in 10-millisecond units. As shown in Figure 2, a sentence appears when a participant puts the cursor on the corresponding numbered icon. A comprehension question appears



Fig. 2. Screenshot of a Reading Process Recording Tool

when the cursor is put over a Q (question) icon. In answering a comprehension question, a participant has only to click on one of the four answer icons ((A) through (D)).

Before the experiment, participants were instructed (i) to read a passage first and then answer comprehension questions, (ii) to try to understand a passage well enough to answer the comprehension questions correctly, (iii) to take as long as they needed (there was no time restriction) and (iv) to practice the reading process recording tool with several practice passages and questions.

We eliminated outliers from the collected reading time data. The data were excluded (i) if the rate of correct answers to comprehension questions was less than 70%, and (ii) if the participant's reading speed (in terms of words read per minute (WPM)) was faster than 200 WPM or slower than 70 WPM. We decided to omit the low comprehension rate data because the RT model should predict the sentence-reading time of EFL learners with a reasonable level of comprehension.³ We excluded slow reading speed data because of the possibility of irregular reading, i.e., that a learner might have read too carefully in order to correctly understand passages. The fast reading speed data were also regarded as irregular reading for EFL learners because 200 WPM is as fast as native English speakers.⁴ Hence, we supposed that such fast reading speed did not appropriately represent EFL learners' reading speed.

³ However, there is still the problem whether the 70% correct rate was adequate for further study.

⁴ The average speed for native English speakers is reported to be 200-300 WPM [3].

As a result, a total of 1807 reading times were obtained. The data consisted of 80 passages, 448 sentences, read by 61 participants. Mean age of the participants was 29.8 years old (S.D. 9.5). Of the participants, 8 were male and 53 were female. Table 1 presents the participants' TOEIC reading score (SCR) distribution. The mean TOEIC reading score of participants is 318.0, which is higher than the mean score of Japanese EFL learners, i.e., 254 to 270, according to the TOEIC technical manual [18].

Table 1. Frequency Distribution of Participants' TOEIC Reading Scores

Intervals	Frequency
$0 \leq \text{SCR} \leq 50$	0
$50 \leq \text{SCR} \leq 100$	0
$100 \leq \text{SCR} \leq 150$	3
$150 \leq \text{SCR} \leq 200$	4
$200 \leq \text{SCR} \leq 250$	11
$250 \leq \text{SCR} \leq 300$	10
$300 \leq \text{SCR} \leq 350$	6
$350 \leq \text{SCR} \leq 400$	11
$400 \leq \text{SCR} \leq 450$	11
$450 \leq \text{SCR} \leq 500$	5

4 Experiment

The RT model was built using multiple regression analysis based on all the features discussed in § 2. This section discusses the methods and results of our experiment, which tested the effectiveness of the RT model.

4.1 Methods

Of the 1807 reading times, 1627 were used as learning data for the multiple regression analysis to develop the RT model, and 180 were used as test data to verify the model. Multiple regression analysis was carried out for reading times with all factors shown in § 2 entered simultaneously.

The RT model was evaluated based on the error rate derived from Formula (1). In Formula (1), predicted value refers to reading time calculated with the RT model, and observed value is defined as actual reading time measured with the tool shown in § 3.

$$E(\text{error})R(\text{ate}) = \frac{| \text{predicted value} - \text{observed value} |}{\text{observed value}} \times 100\% \quad (1)$$

First, we evaluated the RT model derived from all of the features in § 4.2, and then we compared the accuracy of the RT models built with different linguistic feature combinations in § 4.3. Finally, we compared our RT model with other models that use syntactic features as in previous studies [11] and [13] in § 4.4.

4.2 Prediction Performance of Our Model

Table 2 shows the error rate of the RT model. From the table 2 we found that most of the 180 test data showed low error rates. This tendency is clearly observed, because the relative frequency is the highest at the interval between 0% and 10%. Moreover, as the right tail is longer, the distribution of the error rate is positively skewed. The normality of the error rate was examined using the Kolmogorov-Smirnov test, and it was found that the error rate did not follow the normal distribution. The median error rate was 22.9%, and the range was 99.2.

Table 2. Error Rate (ER) Frequency Distribution Table

Intervals	Frequency	Relative Frequency (%)	Cumulative Frequency	Cumulative Relative Frequency (%)
$0\% \leq ER \leq 10\%$	43	23.9	43	23.9
$10\% \leq ER \leq 20\%$	34	18.9	77	42.8
$20\% \leq ER \leq 30\%$	37	20.6	114	63.3
$30\% \leq ER \leq 40\%$	19	10.6	133	73.9
$40\% \leq ER \leq 50\%$	20	11.1	153	85.0
$50\% \leq ER \leq 60\%$	10	5.6	163	90.6
$60\% \leq ER \leq 70\%$	9	5.0	172	95.6
$70\% \leq ER \leq 80\%$	5	2.8	177	98.3
$80\% \leq ER \leq 90\%$	1	0.6	178	98.9
$90\% \leq ER \leq 100\%$	2	1.1	180	100.0

4.3 Prediction Performance and Features of a Reading Time Prediction Model

The RT model predicts sentence-reading time based on linguistic and learner factors. To examine the linguistic effects on the RT model, we constructed RT models that used different combinations of linguistic features. All models employed weighted learner factors equally.

Table 3. Constituent Features and Error Rates of Reading Models

RT Model	Constituent Features	Error Rate (%)
RT Model 1	All Features	22.9
RT Model 2	Lexical & Learner Features	25.7
RT Model 3	Syntactic & Learner Features	24.6
RT Model 4	Discourse & Learner Features	37.2
RT Model 5	Lexical & Syntactic & Learner Features	24.2
RT Model 6	Lexical & Discourse & Learner Features	27.3
RT Model 7	Syntactic & Discourse & Learner Features	24.1

Table 3 lists the error rates of each RT model as defined by the median error rate of the model. RT Model 1, which is derived from all of the features, has the lowest error rate. The error rate seems to increase when a model lacks syntactic features, e.g., RT Models 2, 4 and 6. Given this, we suppose that syntactic features affect RT models more strongly than any other features.

4.4 Comparison of Our Reading Model with other Reading Models Built with Syntactic Features

RT models account for syntactic complexity as indicated by sentence length and the number of branching nodes. Previous studies defined syntactic complexity using other syntactic factors. A previous study [13] developed a reading model based on (i) the height of a syntactic tree, (ii) the number of noun phrases, (iii) the number of verb phrases and (iv) the number of subordinate conjunctions in a sentence. Another study [11] built a reading model using the presence or absence in a sentence of (v) relative clauses, (vi) participle clauses and (vii) *to*-infinitive clauses. We constructed two reading models using the syntactic features proposed by these previous studies [11, 13], and compared our RT model with these models regarding prediction accuracy.⁵

We found that the RT model based on features (i)-(iv) [13] had an error rate of 23.9%, and the RT model using features (v)-(vii) [11] had an error rate of 24.9%. The error rate of our RT model was 22.9%. Thus, the error rate of our RT model was 4.2% ($= \frac{23.9 - 22.9}{23.9} \times 100\%$) lower than that of the model [13], and 8.0% ($= \frac{24.9 - 22.9}{24.9} \times 100\%$) lower than that of Nagata et al. [11].

5 Related Work

Our study shares a problem on features with a study for predicting EFL learners' reading time [11]. A previous study [11] developed an RT model that computed text reading time by summing up the word recognition time of each word in a passage. Word recognition time is weighted for words appearing in particular constructions such as relative clauses, participle clauses and *to*-infinitive clauses. This is based on an assumption that these constructions are more difficult for EFL learners than other constructions. This RT model is derived with a neural network learning algorithm.

Both Nagata et al.'s model [11] and our RT model encounter an empirical problem that the prediction performance depends on the performance of natural language processing techniques because both models use syntactic features derived with a syntactic parser, which is not free from technical error. The error effect of parsing should be limited as much as possible. From this viewpoint, our syntactic features should involve fewer errors than those of Nagata et al.'s model [11]. Since our syntactic features are sentence length and the number of branching nodes, our model does not need to label syntactic nodes. By contrast, the syntactic features of Nagata et al.'s model [11] have to undergo labeling, e.g.,

⁵ Note that the reading model [13] does not predict sentence-reading time but text readability, and the reading model [11] predict text reading time by summing up the word recognition time in a sentence which is weighted for words appearing in particular constructions.

relatives, participles and *to*-infinitives. This creates the possibility of technical errors due to labeling. For instance, a relative clause might be parsed as a non-relative clause, or vice versa. This kind of parsing error might degrade the prediction accuracy of the RT model. Regarding the technical errors for the labeling, our RT model should also be more robust than the model using labeled features to be.

6 Conclusion

We presented an RT model that predicts EFL learners' sentence-reading times based on linguistic properties of a sentence and a learner's reading proficiency. This is the first step toward identification of sentences that have lexical or syntactic problems that are challenging to a learner. The results of our experiment show that the RT model predicts a learner's sentence-reading time with an error rate of 22.9%.

We must still examine the prediction performance of the RT model in more detail. Then, we will pursue a more accurate reading time model.

References

1. Alderson, J.C.: *Assessing Reading*. Cambridge University Press, Cambridge (2000)
2. Bell, T.: Extensive reading: Speed and comprehension. *The Reading Matrix*, 1(1) (2001)
3. Carver, R.P.: Optimal rate of reading prose. *Reading Research Quarterly* 18(1), 56–88 (1982)
4. Flesch, R.: A new readability yardstick. *Journal of Applied Psychology* 32, 221–233 (1948)
5. Frazier, L., Rayner, K.: Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology* 14, 178–210 (1982)
6. JACET. JACET 4000 Basic Words. The Japan Association of College English Teachers, Tokyo (1993)
7. Just, M.A., Carpenter, P.A.: *The Psychology of Reading and Language Comprehension*. Allyn and Bacon, Newton (1987)
8. Kotani, K., et al.: Effects of syntactic factors on EFL learners' reading time. *Information Technology Letters* 6, 457–460 (2007)
9. Lougheed, L.: *How to Prepare for the TOEIC Test: Test of English for International Communication*. Barron's Educational Series, Inc., Hauppauge, New York (2003)
10. Naganuma, N., Wada, T.T.: Measurement of English reading ability by reading speed and text readability. *JLTA Journal* 5, 34–52 (2002)
11. Nagata, R., et al.: A method of rating English reading skill automatically: Rating English reading skill using reading speed. *Computer & Education* 12, 99–103 (2002)
12. Sano, H., Ino, M.: Measurement of difficulty on English grammar and automatic analysis. *IPSI SIG Notes* 117, 5–12 (2000)
13. Schwarm, S.E., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 523–530 (2005)

14. Sekine, S., Grishman, A.: A corpus-based probabilistic grammar with only two non-terminals. In: Proceedings of the 4th International Workshop on Parsing Technologies, pp. 216–223 (1995)
15. Shizuka, T.: The effects of stimulus presentation mode, question type, and reading speed incorporation on the reliability/validity of a computer-based sentence reading test. *JACET Bulletin* 29, 155–172 (1998)
16. Smith, E.A., Kincaid, P.: Derivation and validation of the automated readability index for use with technical materials. *Human Factors* 12, 457–464 (1970)
17. Someya, Y.: Word Level Checker: Vocabulary Profiling Program by AWK, Ver. 1.5 (2000) (consulted November 6, 2006), http://www1.kamakuranet.ne.jp/someya/wlc/wlc_manual.html
18. The Chauncery Group International, Ltd., TOEIC Technical Manual, The Chauncery Group International, Ltd., Princeton, NJ (1998)
19. Yoshimi, T., et al.: 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 (2005)