A prediction model of foreign language reading proficiency based on reading time and text complexity

Katsunori Kotani[1], Takehiko Yoshimi[2], Hitoshi Isahara[3]

(1. School of English Language and Communication, Kansai Gaidai University, Osaka 573-1001, Japan;

2. Faculty of Science and Technology, Ryukoku University, Shiga 520-2194, Japan;

3. Information and Media Center, Toyohashi University of Technology, Aichi 441-8580, Japan)

Abstract: In textbooks, foreign (second) language reading proficiency is often evaluated through comprehension questions. In case, authentic texts are used as reading material, such questions should be prepared by teachers. However, preparing appropriate questions may be a very demanding task for teachers. This paper introduces a method for automatically evaluating proficiency, wherein comprehension questions are not required. This method assesses a learner's reading proficiency on the basis of the linguistic features of the text and the learner's reading time. A reading model following this method predicted reading proficiency with an ER (error rate) of 18.2%. This ER is lower than those of models proposed in previous studies. Furthermore, the ER of the authors' reading model for various learner groups classified by their RS (reading speeds) was examined. The result of this examination showed that the error rate was the lowest for the group of learners with fast RS.

Key words: computer-assisted language learning; English as a foreign language; learners' reading; natural language processing

1. Introduction

Computer-based evaluation plays an important role in foreign language learning and teaching. Various studies have investigated the methods of evaluating foreign language reading skills (Nagata, et al., 2002; Kotani, et al., 2007, 2008). These studies proposed statistical evaluation methods by using machine learning algorithms, such as SVMs (support vector machines) (Vapnik, 1998). Since these methods predict the optimal reading time, the authors can evaluate a learner's reading proficiency by comparing the optimal reading time with the learner's actual reading time.

Following the above-mentioned studies, the authors constructed a reading model that automatically evaluates a learner's reading proficiency on the basis of his/her reading time. The primary advantage of such evaluations is that they allow a learner's reading proficiency to be assessed without the use of comprehension questions. Although comprehension questions are useful for evaluating reading proficiency, preparing these questions by using authentic texts, which typically do not include such kinds of questions, is an ordeal for teachers. Moreover,

Katsunori Kotani, associate professor, School of English Language and Communication, Kansai Gaidai University; research fields: theoretical linguistics, computational linguistics.

Takehiko Yoshimi, associate professor, Faculty of Science and Technology, Ryukoku University; research field: computational linguistics.

Hitoshi Isahara, professor, Information and Media Center, Toyohashi University of Technology; research field: computational linguistics.

unlike in the case of comprehension questions, reading time can be measured by using a variety of texts, such as newspapers or magazine articles. It is evident that, since evaluations based on reading time do not require comprehension questions, teachers are relieved of an ordeal. In addition, if the time taken to read individual sentences is measured, a learner's strengths and weaknesses with respect to reading can be comprehensively analyzed for each sentence (Yoshimi, et al., 2009). Further, such sentence-by-sentence evaluation is a difficult task when methods based on comprehension questions are followed.

The authors shall briefly note other reasons for choosing reading time as an evaluation criterion. First, it is generally supposed that reading time is correlated with the rate of comprehension, as reported in Just and Carpenter (1987). Next, tests based on reading time are expected to have a pedagogical effect (Alderson, 2000; Bell, 2001). Finally, the reliability and validity of such tests have been acknowledged with respect to second language reading assessments (Shizuka, 1998; Naganuma & Wada, 2002; Kotani, et al., 2007; 2008).

Given the aforementioned properties of reading time, Kotani, et al. (2009a) proposed a reading model that estimates reading proficiency by examining a learner's reading time and the linguistic features of the text. Following Kotani, et al. (2009a), the authors derived a reading model that uses support vector regression based on reading time and linguistic features, as shown in Figure 1. In this model, reading proficiency refers to the score obtained for the reading section in the TOEIC (Test of English for International Communication). Linguistic features are classified into lexical, syntactic and discourse features.

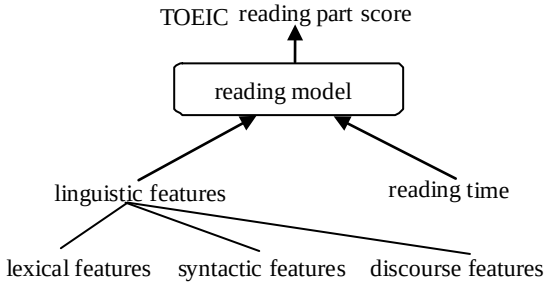

Figure 1 Reading model

Kotani, et al. (2009b) further investigated the relation between their reading model and learners' RS in order to determine whether any shortcomings still remained in the reading model. In this paper, the authors introduce new linguistic features aiming at decreasing the error rate. The experimental result showed that the error rate decreased from 18.5% to 18.2% owing to these features. The findings of this study will present a new direction with respect to the evaluation of foreign language reading proficiency under a computer-assisted language learning system. In light of this, the authors can use their reading model as an evaluation method for reading proficiency in a computer-based placement test.

2. Related studies

As shown in section 1, previous studies (Nagata, et al., 2002; Kotani, et al., 2007; 2008) have proposed the reading time model. In addition, another study (Schwarm, et al., 2005) has proposed the readability model. All these studies derived their reading models by using a statistical method. Nagata, et al. (2002) developed a word recognition time model using a neural network learning algorithm. Kotani, et al. (2007; 2008) conducted a multiple regression analysis, whereas Schwarm, et al. (2005) used SVMs.

Notwithstanding the above similarity, these models can be classified into 2 groups on the basis of their syntactic features. The reading model of Nagata, et al. (2002) comprises the syntactic features of specific grammatical constructions, such as relative clauses, participle clauses and to-infinitive clauses. The word recognition time is weighted for the words that appear in these constructions. According to Nagata, et al. (2002), it is difficult for Japanese learners of EFL (English as a foreign language) to comprehend such constructions. Similarly, a readability model developed by Schwarm, et al. (2005) is based on syntactic features involving the distribution of specific grammatical constructions such as noun phrases, verb phrases and subordinate clauses.

In contrast to these models, a reading model of Kotani, et al. (2007) uses syntactic features that are not limited to certain grammatical constructions. According to Kotani, et al. (2007), although the syntactic features of specific grammatical constructions undeniably affect reading time or readability, a reading model using these features should be able to tolerate the margin of technological errors related to natural language processing tools. When using a syntactic parser, the authors must consider the presence of technological errors, such as the incorrect labeling of syntactic nodes. For instance, a non-relative clause might be incorrectly labeled as a relative clause. Given this possibility, the authors have to reduce the error effect to the minimum.

Kotani, et al. (2007) solved this problem by using syntactic features that are available without labeling. Specifically, the syntactic features they used are the length of a sentence in terms of the number of words and the size of a sentence in terms of the number of syntactic branching nodes. Sentence length is commonly used as a syntactic feature for examining readability (Flesch, 1948; Smith & Kincaid, 1970). The number of branching nodes is believed to affect the reading time from the psycholinguistic perspective, for instance, through the garden-path effect (Frazier & Rayner, 1982). The reading model of Kotani, et al. (2007) exhibited higher prediction accuracy than a model having the labeling problem. Given this result, Kotani, et al. (2008) developed another reading model using not only syntactic features, but also lexical and discourse features.

3. Features of the reading model

Following the previous models (Kotani, et al., 2007; 2008), the authors develop a reading model that estimates a learner's reading proficiency in terms of his/her TOEIC score by examining the reading time of the learner and the linguistic features of the text. Hence, the authors' model differs from the previous models in that it evaluates a learner's reading proficiency, while the previous models identified sentences that were difficult to comprehend by a learner. In addition, this reading model has been developed by using support vector regression (Vapnik, 1998), whereas the previous models were constructed by using multiple regression analysis.

The authors' reading model examines both the reading time of the learner and the linguistic features of the text. Linguistic features encompass lexical, syntactic and discourse features. Since these features can be detected with state-of-the-art natural language processing tools, it is possible to implement this reading model in a computer-assisted language learning system. Next, the authors introduce certain linguistic features as well as a learner feature, namely, reading time.

3.1 Lexical features

Lexical features comprise word length and vocabulary difficulty for the reading model of Kotani et al. (2009b). Word length is defined as the number of characters in a word. It is well known that word length affects text readability and that readability formulas (Flesch, 1948; Smith & Kincaid, 1970) use word length as an independent variable for text readability. Hence, word length should be considered for evaluating a learner's

reading proficiency.

Since word length cannot exhaustively explain the lexical effect, the authors include vocabulary difficulty as another lexical feature, following Kotani, et al. (2009b). It is reported that reading comprehension may be difficult for Japanese EFL learners even when only short words are used (Sano & Ino, 2000). Following Kotani, et al. (2009b), the authors have assigned vocabulary difficulty scores on the basis of heuristically determined vocabulary difficulty, which is summarized in the JACET 4,000 Basic Words List (JACET, 1993). Vocabulary difficulty is determined by English teachers working with Japanese EFL learners. This list provides the difficulty scores for 11 levels (Someya, 2000). Vocabulary difficulty is determined by summing up the scores of all the words in a given text.

Although the vocabulary difficulty list contains more than 35,000 words, authentic texts may contain words that are not registered in this list. Therefore, the reading model of Kotani, et al. (2009b) encounters a problem that the model cannot estimate the vocabulary difficulty for such words.

Since the vocabulary difficulty list is compiled mainly for EFL learners, it is supposed to cover vocabularies that they should learn. Hence, the vocabulary difficulty of unregistered words should be higher than the registered ones. Following this assumption, the authors can solve the problem of unregistered words by: (1) regarding unregistered words as more difficult than registered words; or (2) considering the number of unregistered words in a text as a lexical feature. The former solution is hardly feasible, because it is hard to precisely determine the vocabulary difficulty of unregistered words. Hence, the authors choose the latter solution in this paper.

While the vocabulary difficulty list covers basic vocabulary for EFL learners, some basic words might be more difficult than expected. For instance, the words, such as “get” and “make” are regarded as the least difficult words in the list. Since these words have various usages, their difficulty could depend on the context in which these words appear. The authors attempt to solve this problem by introducing a new lexical feature, i.e., the number of word meanings. The number of word meanings is measured by using WordNet2.0 (Fellbaum, 1998), which is a large lexical database of the English language. The number of word meanings of a text is determined by summing up the word meanings of each word in a text.

3.2 Syntactic features

As mentioned above, syntactic features comprise sentence length and the number of branching nodes in a syntactic tree. In addition to these features used by Kotani, et al. (2009b), the authors introduce a new feature, the number of syntactic nodes stored in short-term memory.

Here, the authors define sentence length as the number of words in a sequence of sentences. Similar to word length, it is generally supposed that sentence length negatively correlates with readability (Flesch, 1948; Smith & Kincaid, 1970). Given this property, sentence length should be regarded as a syntactic feature.

As sentence length is equivalent to the width of a syntactic tree, the height of a syntactic tree should also be considered. Since the number of syntactic nodes explains the size of a syntactic tree, the authors decided to use this quantificational information of syntactic nodes as another syntactic feature, following Kotani, et al. (2009b). In addition, it is confirmed that the number of branching nodes highly correlates with readability in the case of EFL learners (Kotani, et al., 2007). The garden-path effect is a similar branching node effect (Frazier & Rayner, 1982). Syntactic parsing is performed by using the Apple Pie Parser (Sekine & Grishman, 1995). The number of syntactic nodes in a text is determined by summing up those in all sentences in a text.

Since a syntactic tree represents a syntactic parsing result, it does not explain memory load in a syntactic parsing process. The authors proposed the number of syntactic nodes stored in short-term memory as a syntactic

feature that represents short-term memory load. Syntactic nodes stored in short-term memory refer to those stored in a stack when analyzing a sentence in a top-down fashion by using a push-down automaton. The number of nodes stored in a stack when parsing a text is determined by summing up those when parsing all the sentences in a text.

Figure 2 shows how the number of non-terminal symbols stored in a stack is determined for the sentence “The man saw the boy” in a push-down automaton (Yngve, 1960). When the first word “The” is inserted, the terminal symbol S is transformed into (NP, VP), and VP is memorized, that is, the symbol VP is stored in a stack. Next, the terminal symbol NP is transformed into (DT, N), and N is stored in a stack. Then, DT is rewritten as “The”. Therefore, the two non-terminal symbols N and VP are stored in a stack, while “The” is processed.

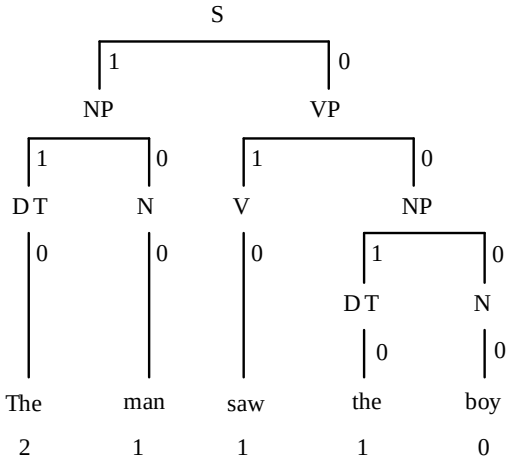

Figure 2 The number of nodes stored in short-term memory

The number of nodes stored in a stack is measured as shown in Figure 2 (Yngve, 1960), beginning from 0, a number is assigned to each branch from the right to left direction. The sum of the numbers in the path from S to a word indicates the number of symbols stored in a stack for that word. The following numbers are assigned to each word as the number of nodes stored in a stack for the sentence “The man saw the boy”: 2, 1, 1, 1 and 0.

Murata,et al. (2001) modified this number assignment procedure in certain aspects, for instance, NP that has no post-modifier will not be transformed. Thus, as NPs in the sentence “The man saw the boy” have no post-modifier, the numbers of nodes in a stack are 1, 1 and 0. The number of nodes in a stack is determined by following this revised procedure. Thus, the sum of the numbers of nodes in a sentence is regarded as a syntactic feature.

3.3 Discourse features

The authors use the number of pronouns as a discourse feature, following Kotani, et al. (2009b). While reading a text, references to pronouns should be identified. This identification of pronominal references requires comprehension of the discourse structure. Hence, the authors suppose that the number of pronouns indicates the complexity of a discourse structure.

Although there are other anaphoric expressions, such as definite expressions, these are not included as a discourse feature. This exclusion is also because of the technological error effect (see section 2). The authors consider that the detection of pronouns involves fewer technical problems.

3.4 Learner feature

In contrast to other models that generally use comprehension questions, the authors’ reading model evaluates a learner’s reading proficiency on the basis of reading time. As mentioned above, this is because that evaluations based on reading time allow assessments using authentic texts. However, the authors’ reading model can be

implemented by using comprehension questions. Further, in the authors' model, they can consider not only reading time, but also the effective RS—a complex measure of the RS and comprehension rate (Jackson & McClelland, 1979). In this paper, the authors have developed an initial reading time model. An effective RS model will be investigated in a future study.

In principle, there are other learner factors, such as text interest, reading motivation and background knowledge of a topic. However, these non-verbal learner features are not included in their reading model, because non-verbal learner features appear to have little impact on reading proficiency (Naganuma & Wada, 2002). The authors will examine non-verbal learner features in a future study.

4. Reading time data collection

The construction of the reading model requires the reading time data of EFL learners. These data were collected in the following manner. The participants in the authors' data collection process were recruited from a job information website. The participants were chosen on the basis of the following criteria: (1) They had taken the TOEIC and could submit the score sheet; (2) They were EFL learners; and (3) They lived near the site of the data collection process. From the participants who responded, 64 took part in the data collection process. The native language of all the participants was Japanese.

In this data collection process, the authors prepared test sets based on 84 texts from TOEIC preparation textbooks (Arbogast, et al., 2001; Logheed, 2003). Each test consisted of 7 texts, and every test set contained different texts. Each text was accompanied by several multiple-choice comprehension questions. The authors randomly provided 1 or 2 test sets to the participants. As a result, 31 participants took 1 test set and 33 participants took 2 test sets.

Reading time data were collected using a reading process recording tool (Yoshimi, et al., 2005). This tool measures reading time on a 10-millisecond time scale. It 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.

The participants used this tool not only for reading the text, but also for answering comprehension questions. When the cursor was positioned on a question icon, a comprehension question appeared. The participants answered a question by clicking on one of the answer icons. Even though the tool recorded the reading time for comprehension questions, the authors excluded it from their reading time data.

After receiving instructions about the tool, participants practiced by reading several texts and answering comprehension questions. The participants were instructed to first read a text and then answer the comprehension questions. The authors also directed participants to understand the text sufficiently well to correctly answer the comprehension questions. Since the authors did not impose time constraints, the participants could take as much time as they needed. In order to reduce the pressure on the participants, the authors did not inform them that the tool would be measuring their reading times. The participants were misled to believe that the goal of the data collection was to examine the comprehension scores for TOEIC comprehension questions on a sentence-by-sentence reading basis.

After collecting the reading times of all the participants, the authors excluded the erroneous reading time data. First, the authors assigned restrictions on the RS, which was measured in terms of WPM (words per minute). The authors excluded the data in the cases where the RS were extremely fast or slow; reading time data was regarded

as improper data if the RS was more than 200 WPM or less than 70 WPM. A slow RS might have been the result of unnecessarily careful reading. Fast RS were also judged as improper data because the average RS of native English speakers is reported to be in the range of 200-300 WPM (Carver, 1982).

The authors' reading time data comprise 451 instances, i.e., the reading times of 60 participants for 84 texts. The mean age of the participants was 29.8 years (SD: 9.5). Nine participants were males and 51 were females.

5. Validity of the reading model

In this section, the authors describe the experimental method for assessing their reading model and report the corresponding results.

5.1 Experimental method

The authors constructed their reading model by using support vector regression (Vapnik, 1998) with the TOEIC reading section scores as a dependent variable and all the linguistic features and the learner features shown in section 3 as independent variables. Support vector regression was performed by using an algorithm implemented in the mySVM software (Rüping, 2000). The first order polynomial was set as a type of kernel function, and the other settings were retained as the default ones. The authors evaluated the prediction performance of the reading models by using 5-fold cross-validation tests using the 451 instances in the reading time data described in section 4.

5.2 Performance of the reading model

The authors report the performance of their reading model in terms of the ER (error rate) computed by using the following formula. The ER shows the degree to which the reading model correctly predicted a learner's TOEIC reading section scores. The predicted value refers to a learner's score in the TOEIC reading section as calculated using the reading model, and the observed value indicates the learner's actual score in the TOEIC reading section taken in the data collection described in section 4.

$$\mathrm{ER} = \frac{|\text{Predicted value} - \text{Observed value}|}{\text{Observed value}} \times 100\%$$

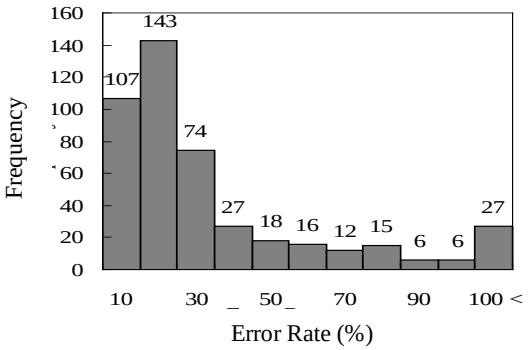

Figure 3 Histogram of the ER

The distribution of the ERs for the authors' reading model is shown in Figure 3. The median ER was 18.2% and the range was 247.6. The distribution of ERs indicates that a lower error rate is more frequently observed. The frequency is the highest at the interval of 10%-20%. Moreover, the distribution of the ER is positively skewed.

5.3 Comparison with other models

In order to validate the authors' reading model, they compare it with other reading models that were derived using the syntactic features proposed by previous studies (Kotani, et al., 2009b; Nagata, et al., 2002; Schwarm, et al., 2005).

The authors' reading model basically follows the reading model of Kotani, et al. (2009b) (henceforth, K-model). What differs between these models are lexical features and a syntactic feature, as introduced in section 3. The syntactic feature of Nagata's model (henceforth, N-model) comprises specific grammatical constructions, i.e., ones including a relative clause, participle clause and to-infinitive clause. The syntactic feature of Schwarm's model (henceforth, S-model) comprises the height of a syntactic tree, number of noun phrases, number of verb phrases and number of subordinate conjunctions.

The ER of the K-model was 18.5%, that of the N-model was 19.1% and that of the S-model was 19.2%. The error reduction rate against the K-model was 1.6% (=(18.5-18.2)/18.5×100), that against the N-model was 4.7% (=(19.1-18.2)/19.1×100) and that against the S-model was 5.2% (=(19.2-18.2)/19.2×100). This result indicates the superiority of the authors' reading model.

5.4 Analysis of ER

The results of the experiment revealed that the authors' reading model had an ER of 18.2%, using the linguistic features stated in section 3. Improvements in the authors' reading model can be achieved by using different linguistic features. Further, the authors examined the ER of their reading model with respect to 3 learner groups categorized according to RS. If the ER is higher for a particular group, the authors can examine the linguistic features that greatly influence the learners in the group in order to improve the model's prediction performance.

The 451 instances of the reading time data were divided into 3 groups according to learners' RS (slow: 70 WPM≤RS<100 WPM, intermediate: 100 WPM≤RS<130 WPM, fast: 130 WPM≤ RS), and the median ERs were calculated for each group. As shown in Figure 4, the distribution of the ERs differs between slow and intermediate RS groups (19.8% and 18.6%) and fast RS group (15.4%). Therefore, the authors' reading model should be modified by examining the reading behavior of slow and intermediate EFL learners. For instance, the authors' reading model still does not include phrase-level features, such as idiomatic expressions. Since these features appear to make reading difficult for slow and intermediate RS learners, but not for fast RS learners, the authors' reading model should include these features in order to reduce the ER. The authors will further address this problem in future research.

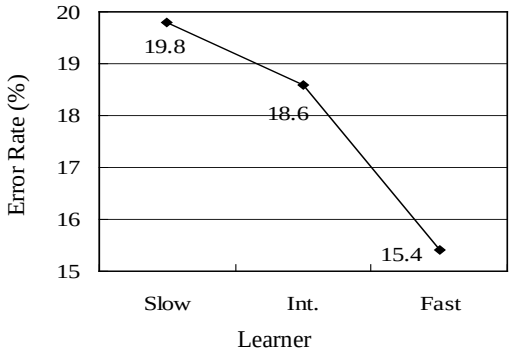

Figure 4 Graph of the ER for the learner groups

6. Conclusion

The authors proposed a method for evaluating the reading proficiency of learners. The model was developed so that it could predict learners' scores in the TOEIC reading section by examining their reading times and the linguistic features of the text. It predicted the scores with an ER of 18.2%. This ER was lower than that of the reading model of Kotani, et al. (2009b). Furthermore, the authors found that the predictions of their model were more accurate than those of the reading models proposed in previous studies. Finally, the authors confirmed that the ER of their reading model was the lowest for the group of learners with fast RS.

References:

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Arbogast, B., Ashmore, E., Chauncey Group International, Peterson's, Duke, T., Jerris, K. & Locke, M. (2001). *TOEIC official test-preparation guide: Test of English for international communication*. NJ: Peterson's, Lawrenceville.

Bell, T. (2001). Extensive reading: Speed and comprehension. *The Reading Matrix, 1*(1).

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

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

Frazier, L. & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, (14), 178-210.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.

JACET. (1993). *JACET 4,000 basic words*. Tokyo: The Japan Association of College English Teachers.

Jackson, M. D. & McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology, 108*, 151-181.

Just, M. A. & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Newton, MA: Allyn and Bacon.

Kotani, K., Yoshimi, T., Kutsumi, T., Sata, I. & Isahara, H. (2007). Effects of syntactic factors on EFL learners' reading time. *Information Technology Letters, 6*, 457-460.

Kotani, K., Yoshimi, T., Kutsumi, T., Sata, I. & Isahara, H. (2008). EFL learner reading time model for evaluating reading proficiency. *Lecture Notes in Computer Science, 4919*, 655-644. Springer Berlin: Heidelberg.

Kotani, K., Yoshimi, T., Kutsumi, T., Sata, I. & Isahara, H. (2009a). Predicting foreign language learners' reading proficiency based on reading time and text complexity. Proceedings of *International Technology, Education and Development Conference*, 3040-3049.

Kotani, K., Yoshimi, T. & Isahara, H. (2009b). Automatic evaluation of foreign language reading proficiency based on reading time and linguistic features. Proceedings of *Conference of the Pacific Association for Computational Linguistics*, 35-40.

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

Murata, M., Uchimoto, K., MA, Q. & Isahara, H. (2001). Magical number seven plus or minus two: Syntactic structure recognition in Japanese and English sentences: Computational linguistics and intelli-gent text processing. *Lecture Notes in Computer Science, 2004*, 43-52. Springer Berlin: Heidelberg.

Naganuma, N. & Wada, T. (2002). Measurement of English reading ability by reading speed and text readability. *JLTA Journal, 5*, 34-52.

Nagata, R., Masui, F., Kawai, A. & Siino, T. (2002). A method of rating English reading skill automatically: Rating English reading skill using reading speed. *Computer & Education, 12*, 99-103.

Rüping, S. (2000). *MySVM-Manual.* University of Dortmund, Lehrstuhl Informatik 8. Retrieved September, 12, 2008 from http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/.

Sano, H. & Ino, M. (2000). Measurement 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*, 523-530.

Sekine, S. & Grishman, A. (1995). A corpus-based probabilistic grammar with only two non-terminals. Proceedings of *the 4th International Workshop on Parsing Technologies*, 216-223.

(to be continued on Page 41)