Linguistic Variables Determining the Difficulty of Eiken Reading Passages

Akira HAMADA

日本言語テスト学会誌　第 18 号
JLTA Journal　Vol. 18
抜　刷
2015

Linguistic Variables Determining the Difficulty of Eiken Reading Passages

Akira HAMADA
Graduate School, University of Tsukuba/
The Japan Society for the Promotion of Science

Abstract

This study examined what linguistic variables affecting the cognitive process in reading comprehension determine the difficulty of Eiken reading passages. Using Coh-Metrix, a corpus analysis of Eiken first-grade to third-grade passages was run to compute lexical (word frequency and lexical diversity), syntactic (syntactic similarity), and meaning construction indices (argument overlap and occurrence of causal connectives and verbs). A stepwise discriminant function analysis showed that surface-level linguistic variables (i.e., lexical and syntactic indices) were stronger predictors in the discrimination of Eiken test grades than the linguistic variables affecting higher-level language processing. To verify whether these results corresponded with Japanese EFL learners' reading performance, Japanese university students completed recall tasks after reading second-grade and third-grade passages. A stepwise multiple regression analysis found that word frequency, lexical diversity, and syntactic similarity indices explained their recall productions. Consistent with the corpus analysis, the meaning construction indices did not explain the recall performances. These findings suggest that the difficulty of Eiken reading passages have probably been designed to measure learners' lower-level language processing abilities, such as word recognition and syntactic parsing.

Keywords: reading comprehension, linguistic variables, Coh-Metrix, corpus analysis, Eiken

1. Introduction

The most popular test for assessing the English four skills in Japan is the *Eiken Test in Practical English Proficiency*. This test has seven grades, enabling test-takers to choose the most appropriate grade according to their English proficiency. One outstanding feature of the Eiken proficiency test that distinguishes it from others (e.g., TOEIC® and TOEFL®) is its synchronization with the Japanese English language curriculum (Amma, 2010). For example, those who pass the third- and second-grade tests are considered equal to junior and senior high school students respectively in English proficiency. According to the Eiken Foundation of Japan (n.d.), Japanese junior and senior high school students account for 80% of all test-takers; consequently, numerous Japanese high schools and universities use the Eiken test when considering

students for admission.

The Eiken test considers positive washback effects, or not only aims to assess learners' English proficiency levels, but also to promote their listening, reading, speaking, and writing skills. The Eiken Foundation of Japan (n.d.) asserts that studying for the test also helps students prepare for the Japanese entrance examinations, since the Eiken test formats are compatible with the typical entrance examination tests used in Japan. Additionally, the results provide test-takers with can-do statements and learning advice to help them easily understand and diagnose their weak points among the English four skills. Thus, the Eiken test plays an essential role in Japanese English education for junior, senior, and university students, as well as adult non-students of English as a foreign language (EFL).

Reading passages and test items in the reading subsection of the Eiken test are designed to assess test-takers' reading comprehension skills, and promote their reading ability through washback effects. When Japanese EFL learners aim to develop their English reading skills by completing a certain grade of the Eiken test, teachers should provide appropriate reading passages to facilitate their learning. In particular, it is important to understand the characteristics of the Eiken test's reading passages to promote learners' reading skills effectively. As some researchers have noted (MacGregor, 1997; Miura & Beglar, 2002), to date, analyses of the test's characteristics, reliability, and validity have been sparse. Although Shimizu (2006) demonstrated the effects of question types (e.g., paraphrastic, inferential) on the test's difficulty, it is necessary to evaluate the reading passages' difficulty in isolation. When teachers adapt the test's reading passages to assess and promote their students' reading skills, understanding which linguistic variables make a text difficult aids selection of the most appropriate passages.

2. Literature Review

2.1 Linguistic Variables That Affect Text Difficulty

To comprehend a text successfully, readers must construct a well-organized mental representation of that text (Kintsch, 1998). The cognitive process of comprehending explicit textual information involves (a) matching semantic information to visual word input, (b) parsing syntactic structures, and (c) understanding each proposition described in the text (Grabe, 2009). Reader knowledge drives these cognitive processes. For example, lexical knowledge facilitates the retrieval of word meanings, grammatical knowledge aids sentence structure analysis, and background knowledge forms the whole comprehension of a text. Second language (L2) reading research has demonstrated that these cognitive processes are constrained by various linguistic variables in a text, which interact with learners' L2 knowledge (e.g., Jeon & Yamashita, 2014; Koda, 2005; Yamashita & Shiotsu, 2015).

At the word level, the most influential predictor of L2 text comprehension is

word frequency and *lexical diversity* (e.g., Crossley, Greenfield, & McNamara, 2008; Nation, 2013). When a reading passage contains various low-frequency words, EFL learners are more likely to encounter words unknown to them. As text comprehension deeply depends on the density of unknown words (Nation, 2013), comprehending a text requires readers to know as many words as possible (e.g., Jeon & Yamashita, 2014); otherwise, they may fail to grasp the explicit content of a text in passages that include various kinds of low-frequency words. In contrast, high-frequency words are processed more quickly and accurately than low-frequency words (Koda, 2005). Thus, texts that contain a greater proportion of high-frequency words can support the process of word identification, and contribute to L2 reading performance.

Syntactic structure complexity also affects text comprehension (Grabe, 2009). Although reading comprehension starts with recognizing information about each word, readers must also integrate the identified word meanings based on their knowledge of grammar rules such as word order (e.g., Jeon & Yamashita, 2014; Zhang, 2012; Yamashita & Shiotsu, 2015). The difficulty of parsing syntactic structures varies according to their complexity, a variable that can predict a text's difficulty. For example, Crossley et al. (2008) suggest that one of the criteria for evaluating syntactic complexity is the *syntactic similarity between adjacent sentences*, because Potter and Lombardi (1998) found that a prior syntactic structure could facilitate the recall of a successive sentence when both structures are similar.

Regarding the understanding of text proposition, many researchers have proposed various reading models. For example, Kintsch (1998) has suggested that readers achieve textbase comprehension by linking concepts described repeatedly throughout a passage. This assumption is known as *argument overlap*, and it provides evidence that meaning construction is predictable if the main verbs share common arguments between sentences (Bohn-Gettler, Rapp, van den Broek, Kendeou, & White, 2011). Additionally, readers need to build a situation model of a text to achieve a level of comprehension deeper than textbase (Kintsch, 1998). Trabasso and Sperry (1985) have established a *causal network model*, which assumes that the strong factor predicting situation model construction is the understanding of causal relatedness between propositions. For example, Linderholm et al. (2000) demonstrated that the insertion of causal connectives (e.g., *therefore*, *thus*, and *because*) and causal verbs (e.g., *cause*, *result*, and *lead*) into a less-cohesive text facilitates the construction of situation models for less-skilled readers.

One way to analyze the linguistic variables involved in text difficulty is to examine how they differ between each text based on a corpus analysis (e.g., Crossley et al., 2008; Nagata, Iguchi, Masui, & Kawai, 2005). For example, Nagata et al. (2005) simulated the classification model of the Eiken reading passages based on their test grades. Their model used two linguistic variables: word frequency and the number of post-modifications, such as relative clauses and participles. Although the two variables

contributed to the classification of the Eiken reading passages, their model considered only the surface-level linguistic factors of language processing (i.e., the lexical and syntactic levels). Moreover, there are no data explaining how these linguistic variables influence EFL learners' reading performance.

2.2 Coh-Metrix and its Applicability

To discover how linguistic variables affect reading comprehension, recent natural language processing research has shown interest in Coh-Metrix, an online tool developed by a research group at the University of Memphis to evaluate text readability based on various linguistic variables. In particular, it focuses on analyzing textual coherence at the word, syntactic, discourse, and conceptual levels (Graesser, McNamara, Louwerse, & Cai, 2004; Graesser & McNamara, 2011). Therefore, Coh-Metrix makes it "possible to computationally investigate various measures of text and language comprehension that supersede surface components of language and instead explore deeper, more global attributes of language" (Crossley et al., 2008, p. 480). The central feature of Coh-Metrix is its ability to compute the numerous linguistic variables involved in the cognitive process of reading comprehension (Crossley et al., 2008; Graesser et al., 2004; Graesser & McNamara, 2011).

Given that the word frequency effect facilitates or inhibits reading comprehension, this study employs Coh-Metrix to calculate word frequency across multiple Eiken reading passages. The word frequency refers to the CELEX Database, which is composed of 17.9 million COBUILD corpus data (The CELEX Lexical Database, n.d.). Coh-Metrix computes the word frequency scores for all content words (Graesser et al., 2004). The scores indicate the mean logarithm of word frequency in a particular text, ranging from 0 to 6, in which 0 indicates the most infrequent words while 6 indicates the most common words used in English. Additionally, Coh-Metrix computes the vocd-D value in its assessment of lexical diversity of a text. According to McCarthy and Jarvis (2010), the vocd-D value is strictly adjusted by text length. Generally, higher vocd-D values indicate that a text contains a more varied mixture of word types; subsequently, EFL learners would require a large vocabulary to read such texts.

As for the effect of syntactic structures, Coh-Metrix calculates the proportion of intersection tree nodes between adjacent sentences as the syntactic structure similarity (Graesser et al., 2004). Specifically, the words embedded into a particular sentence are parsed according to a grammar rule and arranged in nodes, such as NP, VP, and propositional phrase, to create a tree structure. Then, the proportion of the number of intersection tree nodes to be shared by adjacent sentences is evaluated. Whereas Nagata et al. (2005) showed that the number of embedded clauses can be used to discriminate the text difficulty (i.e., test grades) of Eiken reading passages, Coh-Metrix cannot report the proportion of embedded clauses. However, it is assumed that if the

texts contain more embedded clauses, the syntactic structure similarity will be diverse.

Finally, three variables that affect the construction of a text's mental representation are computed. Coh-Metrix evaluates the degree of argument overlap, which occurs when there is a noun in one sentence and the same noun or corresponding pronoun in its adjacent sentence (Graesser et al., 2004). The degree of argument overlap represents the average number of sentences in a particular text that have argument overlap between adjacent sentences. Additionally, the average number of causal verbs and connectives in each text are also accounted for to estimate the difficulty of constructing a situation model based on a text's causality. The indices of causal verbs (i.e., a verb that represents causing something to happen, such as *kill* causing an animate being to die) and causal connectives (e.g., *since*, *so that*, *because*, *the cause of*, and *as a consequence*) represent its frequency of occurrence in a particular passage (Graesser et al., 2004).

The main goal of this study is to determine whether the difficulty of the Eiken test's reading passages is based upon the linguistic variables involved in the cognitive process of reading comprehension. A corpus analysis investigated which linguistic variables reflect the difficulty (i.e., the test grade) of the reading passages using discriminant function analysis. Subsequently, an experimental test was conducted to determine if the linguistic variables related to the text's difficulty affect Japanese EFL learners' reading performance.

3. Research 1: Corpus Analysis

The purpose of the corpus analysis was to reveal what kinds of linguistic variables predict the difficulty of the Eiken reading passages. The linguistic variables automatically produced by Coh-Metrix are involved in the cognitive process of reading comprehension. A multiple discriminant function analysis was employed to test which variables relatively contributed to discriminating the text grades as an evaluation index of the difficulty of the Eiken reading passages. The first research question (RQ) is summarized as follows:

RQ1: How well do the lexical, syntactic, and meaning construction indices discriminate the Eiken reading text grades?

3.1 Corpus Collection

A corpus database was compiled for use with Coh-Metrix to analyze the difficulty of the reading passages. The passages were used in multiple-choice matching questions between 1998 and 2011, collected from the first, pre-first, second, pre-second, and third grades. Some passages in the first grade were obviously longer (over 800 words) than any other passage in the same grade and were excluded from the database. The text genres were narrative, expository, and essay. Table 1 shows tokens,

words per text, and readability, as calculated by Coh-Metrix version 3.0.

3.2 Procedure

Although Coh-Metrix 3.0 linguistic index banks have 11 categories and 108 variables to evaluate text features, this study selected six variables to discriminate the text difficulty based on prior studies.

(1) Two lexical indices: mean word frequency and lexical diversity (vocd-D)
(2) A syntactic complexity index: the degree of syntactic structure similarity
(3) Three meaning construction indices: the degree of argument overlap, and average number of causal verbs and causal connectives

These independent variables were involved in lower-level language processing (i.e., word recognition and syntactic parsing) and higher-level language processing (i.e., situation model construction) since many studies in psycholinguistics have suggested the discourse comprehension models (Crossley et al., 2008; Grabe, 2009; Kintsch, 1998; Koda, 2005; Trabasso & Sperry, 1985).

Table 1
Lexical Features and Traditional Readability of the Passages

			Words per text		FKGL		FRE	
Grades	*n*	Tokens	*M*	*SD*	*M*	*SD*	*M*	*SD*
First	92	45,883	498.73	50.69	12.68	1.70	42.13	8.50
Pre-first	96	40,066	417.35	72.05	11.87	1.33	45.95	7.07
Second	74	26,264	354.92	22.11	9.32	1.09	59.98	6.19
Pre-second	54	15,668	290.15	18.28	7.94	1.00	67.25	5.53
Third	40	10,212	255.30	12.96	5.99	0.96	75.15	5.29

Note. *n* indicates the number of passages. FKGL = Flesch-Kincaid Grade Level, FRE = Flesch Reading Ease (calculated in Coh-Metrix 3.0 program).

Coh-Metrix 3.0 was run for the numeric conversion of each Eiken text feature in terms of the selected linguistic variables. Words were eliminated from the analysis when they were not listed in the CELEX database referenced by Coh-Metrix. To answer RQ1, the selected independent variables were submitted into a stepwise discriminant function analysis, which showed an estimate of relative importance for each variable to predict text difficulties (see Tabachnick & Fidell, 2014).

3.3 Results

Table 2 displays the descriptive statistics of the six linguistic variables for each Eiken test grade (see also Figure 1). Pearson correlations showed no strong relationships among the variables ($rs < .70$; Table 3 and Figure 2). This ensured that multicollinearities did not affect a subsequent discriminant function analysis

(Tabachnick & Fidell, 2014). However, it should be noted that the variance-covariance matrices were not homogeneous (Box's M test, $p < .001$). Accordingly, Pillai's criterion was used in the interpretation of statistical significance. To examine which individual variables best discriminated among five test grades, all six variables were submitted into a stepwise discriminant function analysis.

Table 2

Means and Standard Deviations of the Linguistic Variables for the Five Grade Levels

		vocd-D		Frequency		Syntactic similarity		Argument overlap		Causal connectives		Causal verbs	
Grades	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
First	92	115.74	19.56	2.05	0.10	.08	.01	.38	.14	24.27	6.56	25.06	5.59
Pre-first	96	111.79	19.46	2.10	0.12	.08	.02	.42	.13	26.86	8.44	27.43	6.34
Second	74	88.64	15.54	2.30	0.13	.10	.02	.48	.13	30.56	11.01	30.90	7.71
Pre-second	54	86.24	13.18	2.39	0.12	.11	.02	.54	.14	32.27	11.96	33.80	8.17
Third	40	68.00	11.98	2.49	0.11	.14	.02	.57	.13	25.37	9.84	39.59	8.18

Table 3

*Means, SDs, and Correlations Among Linguistic Variables (*N = *356)*

Variables	*M*	*SD*	1	2	3	4	5	6	7
1. Grade	NA	NA	—						
2. vocd-D	99.20	23.73	-.68	—					
3. Frequency	2.21	0.20	.79	-.54	—				
4. Syntactic similarity	.09	.03	.64	.45	.53	—			
5. Argument overlap	.46	.15	.42	.44	.29	.29	—		
6. Causal connectives	27.61	9.79	.18	<u>-.07</u>	.12	.13	.13	—	
7. Causal verbs	29.87	8.32	.52	-.31	.37	.54	.14	<u>.05</u>	—

Note. Correlation coefficients between Grade and the linguistic variables were calculated by Spearman's method; the others were in Pearson's method. Insignificant correlations were underlined ($p > .05$).

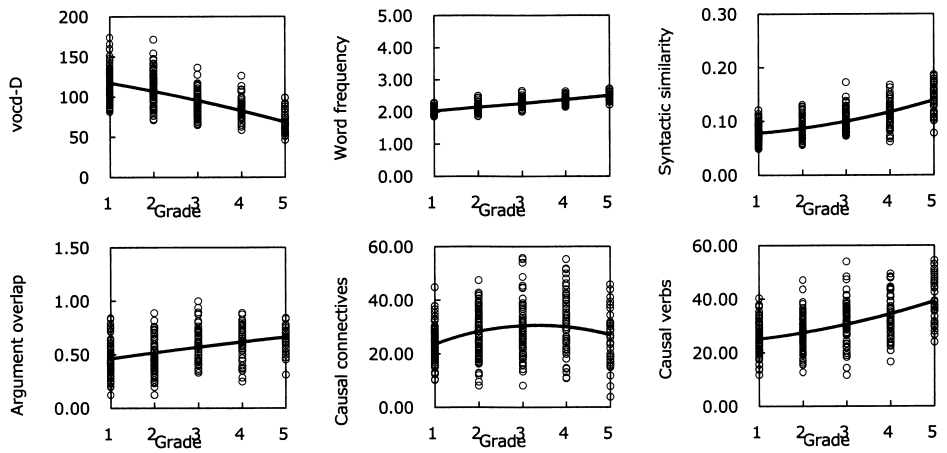

Figure 1. Scatterplots between Eiken Grades and linguistic variables with an approximate curve (N = 356). 1 = first grade, 2 = pre-first grade, 3 = second grade, 4 = pre-second grade, and 5 = third grade.

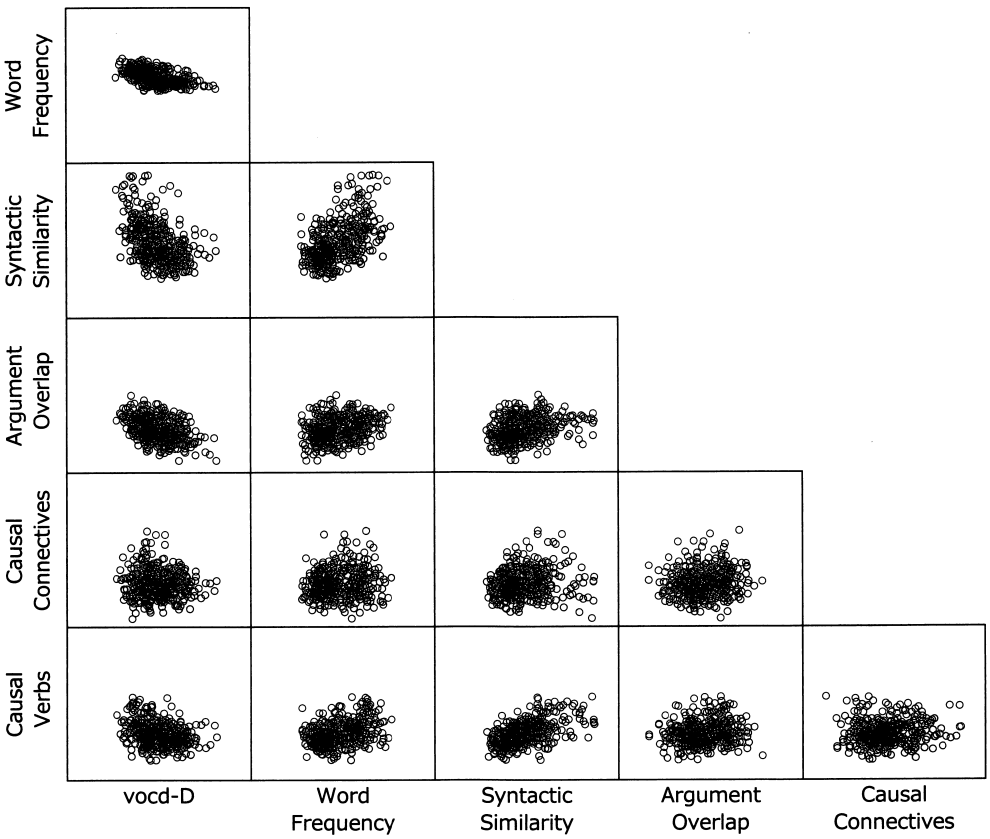

Figure 2. Scatterplots among linguistic variables (N = 356).

In the discriminant function analysis, eigenvalues, Wilks's Lambda, standardized discriminant function coefficients (DFCs), and classification results are the main focus (Tabachnick & Fidell, 2014). The eigenvalues are the percentage of variance explained by each discriminant function. The Wilks's Lambda determines whether the functions are meaningful. As this value approximates zero, each function model should fit the observed data. The standardized DFCs measure the extent to which each independent variable succeeds in classifying a dependent variable. The higher the standardized DFCs are, the more effectively the corresponding linguistic variables contribute to discrimination of the test grades.

The first function explained 96.9% of the variances and the canonical correlation was .91 (the eigenvalue was 4.86). The Wilks's Lambda for the first function was significant, $\Lambda = .15$, $\chi^2(24) = 669.79$, $p < .001$, indicating that the first function corresponded better to the observed data. The second function incrementally explained 99.6% of the variances, and the Wilks's Lambda was also significant, $\Lambda = .86$, $\chi^2(15) = 51.79$, $p < .001$. The third and fourth functions did not significantly contribute to the discrimination ($ps > .05$). Considering that the second function also did not relatively fit the observed data, subsequent discussion will focus on the first function. Table 4 and Figure 3 show the classification results. Although the cross-validated accuracy of the classifications was 60.7%, the classification accuracies of the pre-first and pre-second grades were less than 50.0%. When these two grades were excluded from the discriminant function analysis, the classification accuracy reached 95.8%.

All six variables significantly discriminated the Eiken grade levels (see Table 5). Regarding the standardized DFCs, the analysis demonstrated that the word frequency index was the best predictor of test grades (.75). Word frequency was followed by lexical diversity (-.49), causal verbs (.37), syntactic similarity (.35), and argument overlap (.22). The worst predictor was causal connectives (.09).

Table 4

Cross-Validated Classification Results

		Predicted grades									
		1st		Pre-1st		2nd		Pre-2nd		3rd	
Actual grades	*n*	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
1st	92	66	71.7	24	26.1	2	2.2	0	0.0	0	0.0
Pre-1st	96	39	40.6	44	45.8	12	12.5	1	1.1	0	0.0
2nd	74	0	0.0	3	4.1	46	62.2	24	32.4	1	1.3
Pre-2nd	54	0	0.0	0	0.0	21	38.9	25	46.3	8	14.8
3rd	40	0	0.0	0	0.0	0	0.0	5	12.5	35	87.5

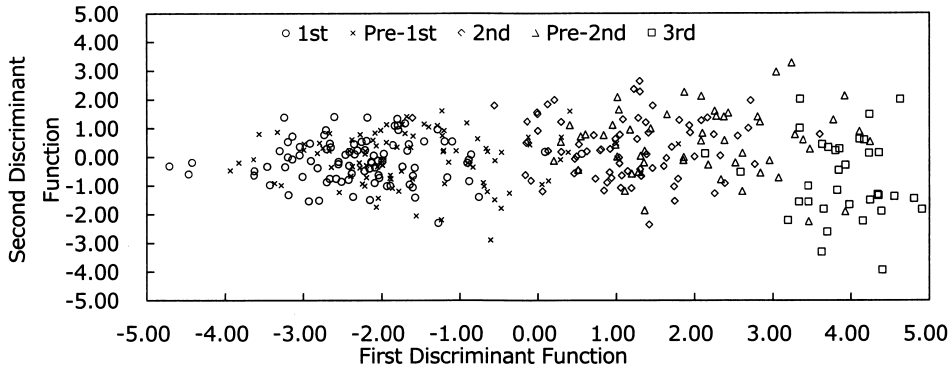

Figure 3. Scatterplots of five Eiken grades on two discriminant functions derived from word frequency, vocd-D, syntactic similarity, argument overlap, causal verbs, and causal connectives.

Table 5

Predictors in Stepwise Discriminant Function Analysis

	Wilks's	Equiv.	Approx.		Standardized DFCs	
Step: predictors	Λ	*F*	*F*	*p*	Function 1	Function 2
1. Frequency	.34	173.01		< .001	.75	.29
2. vocd-D	.24	89.89		< .001	-.49	.06
3. Syntactic similarity	.19		68.55	< .001	.35	-.61
4. Causal verbs	.17		52.80	< .001	.37	-.01
5. Causal connectives	.15		43.70	< .001	.09	.76
6. Argument overlap	.15		36.86	< .001	.22	.22

3.4 Discussion

The results of the discriminant function analysis indicate that the linguistic variables involved in the cognitive process during text comprehension predicted the difficulty of the Eiken reading passages. However, the classification accuracy was not sufficient even when the discriminant model used all six variables. Specifically, the model did not accurately distinguish between the pre-first and pre-second grades. These findings suggest that the reading passages' difficulty between the first and pre-first and between the second and pre-second grades reflects the surface level of readability, that is, FKGL, FRE, and text length, but not the linguistic variables affecting the cognitive process of reading comprehension. In relation to these findings, subsequent discussion focuses on the importance of each predictor in discriminating difficulty.

The two lexical indices, word frequency and vocd-D (i.e., lexical diversity), were the strongest predictors of text difficulty. The results are consistent with Nagata et al. (2005), implying that text difficulty is mainly manipulated by word level

variables. In particular, higher-grade texts are likely to contain various kinds of low-frequency words. As many researchers have suggested (e.g., Nation, 2013), test-takers must possess a large vocabulary to comprehend such reading passages.

The degree of syntactic similarity between adjacent sentences also had predictive power in classifying the test grades. The descriptive statistics showed that the syntactic structures were diverse in the higher-grade texts (see Table 2 and Figure 1). This suggests that the difficult reading passages use various grammar items. In fact, the grammar items served in the third-grade passages are virtually restricted to those that Japanese junior high school students learn. Nagata et al. (2005) found that a certain grammar item (e.g., post-modification by a relative clause) becomes an indicator in determining the difficulty of Eiken reading passages. In addition to this, the corpus analysis by Coh-Metrix showed that syntactic similarity is another indicator of text difficulty in Eiken tests.

At the meaning construction level, the argument overlap index was a significant but relatively small predictor. The degree of argument overlap consistently increased from the first- to third-grade passages (see Figure 1). Because this index reflects the ease of constructing textbase representations (Bohn-Gettler et al., 2011; Grabe, 2009; Graesser & McNamara, 2011; Graesser et al., 2004; Kintsch, 1998), it may be difficult for Japanese EFL learners to understand textbase statements in higher-grade texts. However, it should be noted that the argument overlap index had limited influence, and that the relative effect on reading comprehension was expected to be minor. Next, the causal verb index made a relatively strong contribution to the difficulty classification. The occurrence of causal verbs decreased in the higher-grade texts (see Figure 1). In contrast, the causal connective index was the weakest predictor, because the relationship between the test grades and the occurrence of causal connectives was incongruent, particularly in the third grade (see Figure 1). When the causal relatedness between statements is implicit, readers must infer the causality to construct a well-organized situation model (Trabasso & Sperry, 1985). Therefore, the lower occurrence of causal verbs and connectives requires them to infer the causal relatedness to construct a situation model (Linderholm et al., 2000). Given that making inferences depends on L2 learners' reading proficiency (Grabe, 2009; Koda, 2005), it is reasonable that higher-grade texts should include fewer causal verbs and connectives as necessary, in order to require the learners to generate inferences.

However, the classification results suggest that the reading passages in some grades were not wholly based on the linguistic variables involved in text meaning construction. In other words, the difficulty of the reading passages may be discriminated by only surface-level linguistic variables such as lexical and syntactic difficulties (Nagata et al., 2005). On the other hand, it is possible that the test grades do not strictly reflect the difficulty of the reading passages, because the time limit of the test and the question types also affect test difficulty (Shimizu, 2006). Therefore, in a

subsequent experimental test, this study examines whether the linguistic variables present in the Eiken reading passages exert influence on Japanese EFL learners' reading comprehension.

4. Research 2: Experimental Test

The corpus analysis by Coh-Metrix demonstrated that the lexical, syntactic, and meaning construction indices discriminated the test grades of the Eiken reading passages. The experimental test then examined whether these variables corresponded with Japanese EFL learners' reading performance. The second research question addressed herein is as follows:

RQ2: Do the linguistic variables involved in the Eiken reading passages predict Japanese EFL learners' reading performance?

4.1 Participants

The participants included 51 Japanese university students majoring in philosophy, linguistics, economics, and education. However, the data from 10 participants were excluded because those participants were absent from either a vocabulary size test or an experimental session. All participants had studied English as a foreign language for a minimum of six years, but none had studied in an English-speaking country. The 2,000 to 6,000 word-level from Version 3 of the Mochizuki vocabulary size test (Aizawa & Mochizuki, 2010) was used to estimate participants' English proficiency. The results indicated that the average vocabulary size ranged between 3,538 and 5,962 words (M = 4,880, SD = 617, Cronbach's α = .95).

4.2 Materials

Regarding the reading materials, 40 kinds of booklet were prepared to examine the relationship between the linguistic variables and text comprehension. The booklets included two Eiken-grade reading passages. Considering the range of participants' vocabulary sizes, one passage was selected from a second-grade test, and the other from a third-grade test; both texts were in use between 1998 and 2011, compiled for the corpus analysis in Research 1. The second- and third-grade passages were a part of Question 4-B and 4-C, respectively.

The two-tailed paired t tests were used to verify whether the 40 passage sets were represented in the corpus analysis results. Given that the insertion of causal connectives facilitates reading comprehension (Linderholm et al., 2000), the causal connective index was excluded from the subsequent analysis; this is because the results of the t test revealed that their occurrence reversed significantly between difficult (second-grade) and easy (third-grade) passages. Table 6 shows the mean number of words, sentences, readability of passages, and characteristics of the linguistic variables.

Table 6

Linguistic Variables of the Eiken Texts Used in the Experimental Test

Variables	2nd (k = 40)		3rd (k = 40)				Cohen's
	M	*SD*	*M*	*SD*	*t*	*p*	*d*
Tokens	351.00	23.82	255.30	12.96	-22.32	< .001	4.99
Sentences	19.68	1.87	19.93	2.27	0.54	.592	0.12
FKGL	9.46	1.18	5.99	0.96	11.82	< .001	3.23
FRE	59.23	6.67	75.15	5.29	-14.46	< .001	2.65
vocd-D	89.75	16.03	68.00	11.98	-6.87	< .001	1.54
Word frequency	2.31	0.13	2.49	0.11	6.57	< .001	1.50
Syntactic similarity	.10	.02	.14	.03	6.72	< .001	1.57
Argument overlap	.49	.13	.65	.11	6.14	< .001	1.33
Causal connectives	31.29	8.18	39.59	8.18	4.34	< .001	0.97
Causal verbs	30.44	10.53	25.37	9.84	-2.93	.004	0.66

Note. FKGL = Flesch-Kincaid Grade Level, FRE = Flesch Reading Ease (calculated in the Coh-Metrix program).

4.3 Procedure

This study adopted a recall test to assess the participants' reading comprehension to avoid the effects of question types on reading performance. Participants were tested during a regular English class or individually. Before starting the experimental session, they were notified of the study's general purpose and it was explained how the data would be used. After completion of a vocabulary size test within 15 minutes, participants randomly received one of the 40 booklets and instructions on how to complete a recall test. The time allotted for reading was five minutes per passage. After the participants had finished each passage, they were asked to write everything that they could comprehend from the passage in Japanese. A second reading and recall test was then conducted in the same way as the first test. The reading order of the second- and third-grade passages was counterbalanced among participants.

4.4 Scoring and Data Analysis

For scoring of recall data, the passages were parsed into a set of idea units (IUs) by the author based on Ikeno's (1996) criteria. Two weeks later, the same procedure was conducted to ensure intra-rater reliability, resulting in 95.4% agreement. Disagreements were resolved by referring to the criteria once more. Each IU in the recall protocols was allotted one point if the literal or paraphrased information was reproduced. Scoring of the recall protocols for each passage was repeated twice by the author, resulting in a high agreement ratio of 92.5%. All disagreements were resolved by re-scoring the data.

Before the statistical analysis, the arcsine transformation was performed on the total score because each passage differed in the number of IUs. To answer RQ2, two lexical indices (word frequency and vocd-D), a syntax index (syntactic similarity), and two meaning construction indices (argument overlap and causal verbs) were submitted into a stepwise multiple regression to predict the recall production.

4.5 Results

Table 7 shows the correlation matrix between recall production rates and each linguistic variable. There was a moderate correlation between recall production and word frequency (.51). Word frequency was followed by syntactic similarity (.45), vocd-D (-.45), and argument overlap (.44). The causal verb index showed a significant but extremely weak correlation with the recall production (.24). These correlation coefficients indicate that recall production decreased when (a) low-frequency and diverse words were used in the passages, (b) the syntactic structures were not parallel, and (c) the texts had few causal verbs, as visualized in Figure 4.

Table 7

Means, SDs, and Pearson Correlations Among Variables (N = *82)*

Variables	*M*	*SD*	1	2	3	4	5
Independent variable: Recall	.46	.22	-.45	.51	.45	.44	.24
Predictors:							
1. vocd-D	79.40	18.09	—	-.22	-.33	-.56	-.18
2. Frequency	2.40	0.15		—	.25	.23	.17
3. Syntactic similarity	.12	.03			—	.29	.37
4. Argument overlap	.57	.14				—	.24
5. Causal verbs	34.99	10.32					—

Note. Insignificant correlations were underlined ($p > .05$).

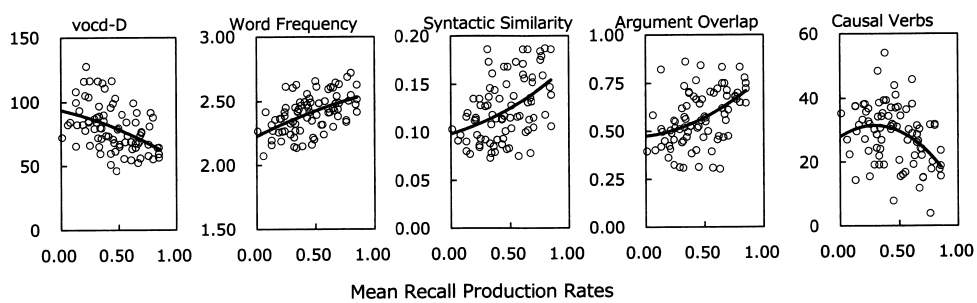

Figure 4. Scatterplots between recall production rates (x-axis) and linguistic variables (y-axis) with an approximate curve ($N = 82$).

To determine which linguistic variables affected recall production, a stepwise

multiple regression analysis was conducted. Requirements for performing a multiple regression analysis were confirmed as accurately as possible (for a review, see Hirai, 2012; Tabachnick & Fidell, 2014):

- Multicollinearity: None of the variables correlated strongly with each other (r < .70), and the tolerance values of each factor were not less than .86. These suggested that there were no multicollinearities among them.
- Independence of residuals: The result of the Durbin-Watson statistic was 1.53 (not less than 1.00 or more than 3.00). This showed that there were no correlations among any combinations of variables' residuals.
- Outliers: A leverage method was used to find any outliers of the data set; the maximum value of a leverage was .14, which was less than a criterion of .17 (= 2 × {6 [the number of predictors] + 1} / 82 [a sample size]).
- Normality, homoscedasticity, and linearity of residuals: Although some residuals of the variables submitted into the regression model were not homogeneous, the normality and linearity of the residuals were regarded as good by a visual inspection of corresponding plots.

Table 8

Summary of Multiple Regression Analysis for Variables Predicting Recall Production

Predictors	*B*	95% CI	*SE B*	β	*t*	*p*
(Invariable)	-0.85	[-1.51, -0.18]	0.33		-2.54	.013
Frequency	0.56	[0.31, 0.82]	0.13	0.39	4.37	< .001
vocd-D	0.00	[-0.01, 0.00]	0.00	-0.28	-3.03	.003
Syntactic similarity	1.71	[0.51, 2.90]	0.60	0.26	2.84	.006
Argument overlap					1.85	.069
Causal verbs					-1.03	.308

Note. adjusted R^2 = .42 (N = 82, p < .001). CI = confidence interval for B.

All five variables were submitted into the regression analysis. A regression model was the most suitable for the observed data when the three variables (i.e., word frequency, vocd-D, and syntactic similarity) were used, $F(3, 78) = 20.34$, $p < .001$, resulting in adjusted R^2 of .42. The meaning construction indices (i.e., argument overlap and causal verbs) did not explain the recall production (ps > .05). Table 8 shows a summary of the stepwise multiple regression analysis.

4.6 Discussion

The results of the multiple regression analysis demonstrated that lexical (i.e., word frequency and vocd-D) and syntactic variables affected Japanese EFL learners' recall performance. These show that text comprehension suffers when the texts contain

various types of low-frequency words, and when their syntactic structures differ between adjacent sentences, consistent with Nagata et al. (2014). Additionally, these findings are congruent with prior research suggesting that surface-level linguistic variables strongly affect L2 learners' text comprehension (Crossley et al., 2008; Grabe, 2009; Jeon & Yamashita, 2014; Koda, 2005; Nation, 2013; Yamashita & Shiotsu, 2015).

The meaning construction indices did not predict the participants' reading performance, although the argument overlap variable correlated with the recall production rates on the same level with lexical diversity and syntactic similarity. Because the argument overlap variable also correlated with the other variables, it is possible that the simple predictive power of the argument overlap was reduced (see also Tabachnick & Fidell, 2014). In this study, the sample size of the multiple regression analysis was relatively small in order to determine a significant predictor (see Hirai, 2012). Therefore, the inconsistent results with the reading model of argument overlap (Kintsch, 1998) may be tentative. In contrast, the insignificance of the causal verb index can be explained by past research, which suggested that situation model construction in L2 reading is often difficult due to certain constraints on the learners' cognitive processes (Grabe, 2009; Koda, 2005). Moreover, Linderholm et al. (2000) showed that even L1 readers had difficulty constructing situation models representing the causal relatedness of a text. Thus, EFL learners might not be sensitive enough to text causality in reading, and so the causal verb index would not affect recall performance.

5. General Discussion

This research examined whether linguistic variables automatically evaluated by Coh-Metrix could successfully classify Eiken test grades and affect EFL learners' reading comprehension. In the corpus analysis, the lexical, syntactic, and meaning construction indices correlated with the text difficulty classification, but the classification accuracy was not high enough. Specifically, word frequency and lexical diversity made a more effective contribution to discrimination than other linguistic variables. In the experimental test, the degree of text comprehension differed according to the lexical and syntactic indices, but not the meaning construction index. Taken together, these findings provide evidence that surface-level linguistic variables have a stronger influence on the difficulty of the Eiken reading passages.

The corpus analysis considered the six linguistic variables that affect the cognitive processes employed in reading. It produced results indicating that the lexical and syntactic indices can be used to classify Eiken test grades, which is fully consistent with Nagata et al. (2005). Generally, lexical difficulty (e.g., the density of low-frequency words) and syntactic complexity constrain the reading comprehension process (Grabe, 2009; Koda, 2005; Nation, 2013). Therefore, the results of the

discriminant function analysis suggest that the difficulty of the Eiken reading passages can be manipulated by increasing the number of low-frequency words and complicating the sentences' syntactic structures.

More importantly, the degree of argument overlap, the number of causal verbs, and the number of causal connectives made significant contributions to classifying the grades. As expected from some theoretical reading models, for example the argument overlap model in Kintsch (1998) and the causal network model in Trabasso and Sperry (1985), the higher-grade reading passages were likely to contain less argument overlap, causal verbs, and causal connectives. Although this result suggests the higher-grade passages are designed to require test-takers to engage in higher-level language processing, it should be noted that the classification accuracy was not high. Specifically, the discriminant function analysis erroneously classified pre-first- as first-grade passages, and pre-second- as second-grade passages. These findings suggest that the text difficulty of these adjacent grades might only differ minutely at the surface-level of readability (i.e., FKGL, FRE, and text length).

Consistent with the above prediction, the results of the recall test showed that the surface-level linguistic variables (i.e., lexical and syntactic indices) could determine the difficulty of Eiken reading passages. In particular, whereas the causal verb index as a deeper-level linguistic variable was a strong predictor of text discrimination, it did not predict recall production. In contrast, the regression analysis found that word frequency, lexical diversity, and syntactic similarity explained 42% of recall production in the second- and third-grade passages. These findings expand upon prior research (Nagata et al., 2005) by demonstrating that the discrimination of Eiken test grades from the computed linguistic variables reflects Japanese EFL learners' reading comprehension. Although the meaning construction indices did not become significant predictors in the present study, Crossley et al. (2008) also showed that the influence of linguistic variables on reading comprehension was smaller in the meaning construction indices than in the surface-level variables.

Although the present study provides a better understanding of the Eiken reading passages' linguistic features, some limitations constrain the generalizability of the findings. In the experimental test, the sample size of participants was limited; therefore, the regression analysis did not ensure the independence of the recall data. This might reduce the reliability of the regression analysis results. Additionally, the findings reported in this article were obtained from a very limited range of English reading proficiencies (i.e., the participants were university-level students). To address these issues, further research should conduct a larger-scale replication.

Finally, this study implies that teachers might need to use Eiken reading passages for instructing and testing reading skills with great care. Considering that the linguistic and syntactic variables were determiners of the difficulty of Eiken reading passages, they can be used to promote lower-level language processing skills (i.e.,

word recognition and syntactic parsing). However, if teachers try to use the Eiken reading passages to develop their students' reading abilities, they should complement them with appropriate reading strategies to facilitate a deeper-level of text comprehension. For example, Grabe (2009) suggested teaching strategies such as making inferences, using background knowledge, and understanding discourse structure.

Testing higher-level language processing such as inferences may be limited in using Eiken passages. Although inferential questions can complement the testing, Shimizu (2006) showed that the number of this type of question was small in the Eiken tests regardless of test grades. Accordingly, adopting an inferential question like TOEFL® ("What can be inferred from the passage?") is necessary. More importantly, creating a passage that taps test-takers for making inferences is also effective (Linderholm et al., 2000) because the results of the present study imply that the Eiken reading passages are not written for that purpose. These modifications regarding reading passages and test items should allow assessment of the various aspects of test-takers' reading comprehension skills, and lead to development of their reading ability through washback effects.

Acknowledgements

This research was partly supported by a Grant-in-Aid for Scientific Research (No. 25・487) from the Japan Society for the Promotion of Science, and Eiken Research Aid from the Eiken Foundation of Japan in 2012. I would like to express my deep gratitude to Professor Yuji Ushiro and his colleagues for their valuable suggestions.

References

Aizawa, K., & Mochizuki, M. (2010). *Eigo goishidou no jissen idea shu: Katsudourei kara test sakusei made* [Practical handbook for English vocabulary teaching]. Tokyo, Japan: Taishukan Shoten.

Amma, K. (2011). English proficiency tests and their practical application. In. S. Ishikawa, T. Nishida, & C. Saida (Eds.), *Language testing and assessment: Approach to four-skill measurement and university entrance examination* (pp. 144–172). Tokyo, Japan: Taishukan Shoten.

Bohn-Gettler, C. M., Rapp, D. N., van den Broek, P., Kendeou, P., & White, M. J. (2011). Adults' and children's monitoring of story events in the service of comprehension. *Memory & Cognition, 39*, 992–1011. doi:10.3758/s13421-011-0085-0

Chall, J., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula.* Cambridge, MA: Brookline Books.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability

using cognitively based indices. *TESOL Quarterly*, *42*, 475–493. doi:10.1002/j.1545-7249.2008.tb00142.x

Eiken Foundation of Japan. (n.d.). *EIKEN test in practical English proficiency.* Retrieved from http://www.eiken.or.jp/eiken/

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge University Press.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*, 371–398. doi:10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, *36*, 193–202. doi:10.3758/BF03195564

Hirai, A. (2012). *Kyouiku shinri kei kenkyu no tameno data bunseki nyumon: Riron to jissen kara manabu SPSS katsuyouhou* [An introduction to data analysis for educational/psychological research: Using SPSS based on theory and practice]. Tokyo, Japan: Tokyo Shoseki.

Ikeno, O. (1996). The effects of text-structure-guiding questions on comprehension of texts with varying linguistic difficulties. *JACET Bulletin*, *27*, 51–68. Retrieved from http://ci.nii.ac.jp/naid/110003814657

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*, 160–212. doi:10.1111/lang.12034

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge University Press.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* Cambridge University Press.

Linderholm, T., Everson, M. G., van den Broek, P., Mischinski, M., Crittenden, A., & Samuels, J. (2000). Effects of causal text revisions on more- and less- skilled readers' comprehension of easy and difficult texts. *Cognition and Instruction*, *18*, 525–556. doi:10.1207/S1532690XCI1804_

MacGregor, L. (1997). The Eiken test: An investigation. *JALT Journal*, *19*, 24–42. Retrieved from http://jalt-publications.org/files/pdf/jalt_journal/jj-19.1.pdf

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. doi:10.3758/BRM.42.2.381

Miura, T., & Beglar, D. (2002). The Eiken vocabulary section: An analysis and recommendations for change. *JALT Journal*, *24*, 107–129. Retrieved from http://jalt-publications.org/files/pdf/jalt_journal/2002b_jj.pdf

Nagata, R., Iguchi, T., Masui, F., & Kawai, A. (2005). A method for rating English texts by reading level for Japanese learners of English. *Systems and Computers in Japan*, *36*, 1–13. doi:10.1002/scj.20326

using cognitively based indices. *TESOL Quarterly*, *42*, 475–493. doi:10.1002/j.1545-7249.2008.tb00142.x

Eiken Foundation of Japan. (n.d.). *EIKEN test in practical English proficiency.* Retrieved from http://www.eiken.or.jp/eiken/

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* Cambridge University Press.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*, 371–398. doi:10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, *36*, 193–202. doi:10.3758/BF03195564

Hirai, A. (2012). *Kyouiku shinri kei kenkyu no tameno data bunseki nyumon: Riron to jissen kara manabu SPSS katsuyouhou* [An introduction to data analysis for educational/psychological research: Using SPSS based on theory and practice]. Tokyo, Japan: Tokyo Shoseki.

Ikeno, O. (1996). The effects of text-structure-guiding questions on comprehension of texts with varying linguistic difficulties. *JACET Bulletin*, *27*, 51–68. Retrieved from http://ci.nii.ac.jp/naid/110003814657

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*, 160–212. doi:10.1111/lang.12034

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge University Press.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* Cambridge University Press.

Linderholm, T., Everson, M. G., van den Broek, P., Mischinski, M., Crittenden, A., & Samuels, J. (2000). Effects of causal text revisions on more- and less- skilled readers' comprehension of easy and difficult texts. *Cognition and Instruction*, *18*, 525–556. doi:10.1207/S1532690XCI1804_

MacGregor, L. (1997). The Eiken test: An investigation. *JALT Journal*, *19*, 24–42. Retrieved from http://jalt-publications.org/files/pdf/jalt_journal/jj-19.1.pdf

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. doi:10.3758/BRM.42.2.381

Miura, T., & Beglar, D. (2002). The Eiken vocabulary section: An analysis and recommendations for change. *JALT Journal*, *24*, 107–129. Retrieved from http://jalt-publications.org/files/pdf/jalt_journal/2002b_jj.pdf

Nagata, R., Iguchi, T., Masui, F., & Kawai, A. (2005). A method for rating English texts by reading level for Japanese learners of English. *Systems and Computers in Japan*, *36*, 1–13. doi:10.1002/scj.20326