Creation and Validation of an Antonym Semantic Decision Task to Measure Japanese EFL Learners' Automaticity of Word Recognition

Tohru Matsuo

日本人の語彙認識の自動化を測定する反意語意味性判断テストの作成と妥当性の検証

松　尾　　徹

Abstract

The purpose of this study is to describe the creation of an Antonym Semantic Decision Task (ASDT) to measure Japanese EFL learners' automaticity of word recognition, in particular lexical meaning access speed and to validate the use of this instrument. To validate the instrument, this study focuses on whether reaction time and accuracy of visual word recognition in the ASDT differ according to the participants' proficiency levels defined by vocabulary size and by the word frequency level (1K, 2K, 3K, and 4K). The results of a one-way and two-way mixed analysis of variance (ANOVA) showed that reaction time and ASDT accuracy generally distinguish the participants' proficiency. Also, a general frequency effect for reaction time was found for both groups. However, Coefficient of Variance, an indication of the degree of automaticity, did not show frequency effects in either group.

Keywords: an antonym semantic decision task, reaction time, automaticity of word recognition

（Received September 19, 2019）

抄　　録

本研究では日本人英語学習者の語彙認識の自動化、特に語彙認識の構成要素の１つである単語の意味アクセスの正確さと流暢さを測定するための反意語意味性判断テスト作成手順を説明することとそのテストの妥当性を検証することを目的にしている。妥当性検証方法として、反意語を用いた意味性判断テストで測定する被験者の意味アクセスの語彙の反応速度と正確さが被験者の語彙サイズの差（英語能力）と語彙の頻度レベル（1000 語、2000 語、3000 語、4000 語）によって統計的有意差があるかを調査する。分散分析の結果から語彙の意味アクセスの反応速度と正確さは被験者の語彙サイズによって分けられたグループ間を区別することが判明した。また、全体的に頻度による語彙の反応速度の差が被験者グループ間で見られた。しかしながら、自動化の指標とされている Coefficient of Variance

（変動係数）では被験者のどちらのグループでも語彙の頻度レベルにおいて差が見られなかった。

キーワード：反意語意味性判断テスト、反応速度、語彙認識の自動化

（2019 年 9 月 19 日受理）

Introduction

This study describes the creation of an Antonym Semantic Decision Task (ASDT), a computerized test which is designed to measure automaticity of word recognition. This study also attempts to gather validation evidence for its use of ASDT as a measure of Japanese English as Foreign Language (EFL) learners' automaticity of word recognition, specifically lexical meaning access speed and accuracy. For the validation framework, Kojima (2010)'s framework using ASDT, which examined the proficiency effects and word frequency effects in test-takers' accuracy score and reaction time, is partially adopted.

Literature Review

In this section, the definition and components of word recognition, measures of automaticity of word recognition, and a previous study which employed the ASDT are reviewed; thereafter, the research hypothesis for validating the ASDT is presented.

Components of word recognition

Word recognition, which refers to how readily and automatically learners can recognize a word's written form and access the meaning of a word, is widely considered to be one of the most important processes contributing to skilled reading comprehension; therefore many researchers have been interested in measuring this skill (Grabe, 2009; Perfetti, 1999, 2007; Perfetti, Landi, & Oakhill, 2005).

Word recognition consists of several subcomponents, such as orthographic decoding, phonological processing, and lexical meaning access. Orthographic information refers to the visual recognition of word forms from the text; this has been considered as a key reading subskill (Cunningham, Perry, & Stanovich, 2001; Perfetti, 1999, 2007; Perfetti, Landi, & Oakhill, 2005). Phonological processing occurs when the link between the word form and phonological information is activated. This process is involved in accessing, storing, and manipulating phonological information (Torgesen & Burgess, 1998). Lexical access occurs when the link between the form of a word and its meaning in the reader's mental lexicon is activated.

Measuring automaticity of visual word recognition

In order to measure automaticity of visual word recognition, a Lexical Decision Task (LDT), in which test-takers are required to classify a stimulus as a word or nonword, is typically utilized. Even though this test is employed to specifically measure orthographic decoding speed, it can be also appropriate in first language (L1) studies such as measuring lexical meaning access speed of visual English words for native speakers of English as they cannot stop accessing the meaning of the word when they see it. In this way, LDT can measure both decoding and lexical meaning access, which are subcomponents of word recognition in L1 studies.

However, an LDT might not be appropriate to measure second language (L2) learners' lexical meaning access speed and accuracy. Grabe (2009) argued that it is possible for readers to initiate word recognition in the orthographic and phonological processing levels with access to little or no lexical content in contexts such as L2 reading. For example, beginner L2 readers can encounter many words whose form is recognized, and this recognition can activate the link between the word form and phonological information; however, no lexical access occurs because no lexical entry exists in the individual's mental lexicon. In such cases, word recognition occurs at orthographic and phonological levels, but no semantic information is available (Grave, 2009). Because word recognition involves not only decoding but also lexical meaning access, it is important to capture both components of word recognition.

Another important issue is how the degree of automaticity in word recognition can be quantified or measured. Segalowitz and Segalowitz (1993) claimed that one indication of automaticity is the coefficient of variance (CV), which is calculated as the mean standard deviation (SD) divided by the mean reaction time (RT). Segalowitz, Segalowitz, and Wood (1998) also proposed that the relationship between mean RT and CV can serve to discriminate between the mere speed-up and the development of automaticity in performance. However, Hulstijn, van Gelderen, and Schoonen (2009) questioned whether the distinction between faster performance and automaticity can be easily made by the CV. Hulstijn et. al. (2009) reviewed seven previous studies in which the CV was utilized and conducted two studies. The results provided minimal support for the proposal that CV reliably indicates development of automaticity. They argued that it is problematic to use CV as an operationalization of automaticity.

Contrary to the findings by Hulstijn et. al (2009), Lim and Godfroid (2014) argued that CV might be a valid measure of automatization at the sentence level. The authors investigated the development of automaticity in sentence processing and validated the use of the CV measure. They partially replicated Hulstijn, van Gelderen, and Schoonen's (2009) study, as they utilized the same analysis on a subset of the same computerized reaction time tasks. Forty Korean English as a Second Language (ESL) university students (20 intermediate and 20 advanced

proficiency learners) and 20 native speakers of English participated in the study. The results indicated that the CV in the sentence-level tasks decreased as the participants' proficiency level increased. The authors argued that no counterevidence against Segalowitz et al. (1998) was found.

Thus, even though the concept of CV is widely recognized among L2 automaticity researchers, the validation studies regarding the use of CV as a measure of the development of L2 automaticity is scant and inconclusive; however, it is impossible to completely disregard CV as a measure of automaticity based on the literature. Hence, in this study, I employ CV as a supplemental measure of automaticity in addition to RT.

A Previous study which employed an Antonym Semantic Decision Task

In order to ascertain the measurement of lexical meaning access in word recognition, a task in which test takers are forced to access the meaning of the word is imperative. For this task, a semantic decision task is usually used. Kojima (2010) employed an ASDT to measure decoding efficiency and lexical access. In the ASDT, test-takers decided whether the meaning of the target word was antonymous to the meaning of the prime word (e.g., the prime word *high* appears on the screen and a target word *low* is displayed after a specified time). The degree of automaticity was quantified using the coefficient of variation (CV) of reaction time. Kojima investigated the roles of word recognition speed, accuracy, and automaticity on Japanese EFL learners reading proficiency. Kojima also examined whether these three measures varied depending on word frequency. The participants were 44 Japanese English as a Foreign Language (EFL) undergraduate and graduate students and 22 native speakers of English. The students were divided into two groups based on their Test of English for International Communication (TOEIC) reading scores: 22 advanced readers (365-495), who were all English related majors, and 22 intermediate readers (210-340), whose majors varied.

The results indicated that word recognition accuracy and speed of recognition discriminated among the three reading proficiency groups. Generally, the native speaker group recognized the antonymous words more accurately and faster than the advanced Japanese L2 readers, who in turn, recognized the words more accurately and faster than the intermediate Japanese L2 readers. Moderate effects were observed for word recognition automaticity as measured by CV_{RT}. The effects of word recognition accuracy and speed became more prominent when word frequency decreased. That is, all participants generally responded to high frequency target words more accurately and faster than low frequency target words. On the contrary, such changes were not observed for word recognition automaticity; CV_{RT} was constant across the four frequency levels. Kojima indicated that the participants word recognition speed might not have been differentiated by the CV_{RT} measures because restructuring the underlying word recognition process takes more time and changes in CV_{RT} are

subtler compared to those of accuracy and reaction time. The author concluded that the more proficient people become in reading, the more quickly and accurately they recognize words, and these effects increase as word frequency increases.

The purpose of this study

Kojima's study (2010) was the only study that investigated whether ASDT distinguished learner's proficiency effects using TOEIC reading scores and word frequency levels. Her study needs to be partially replicated in order to examine to what degree similar patterns of ASDT accuracy rate and the mean reaction time of each word frequency level are consistent.

For the validation of the visual ASDT, the following hypotheses were used to examine the proficiency and word frequency effects on participants' mean accuracy rate and reaction time.

1. ASDT accuracy will improve as group proficiency and word frequency levels increase.
2. ASDT reaction time will decrease as group proficiency and word frequency levels increase.
3. Response stability, CV, in ASDT will decrease as group proficiency and word frequency levels increase.

Method

Participants

The participants were 166 Japanese law majors (124 male and 42 female students) attending a medium-ranked private university in western Japan. There were 94 (73 male and 21 female) first-year students and 73 (51 male and 21 female) second-year students, whose ages ranged from 18 to 21. The mean Institutional TOEIC scores of the first-year and second-year students were 290.74 (SD = 116.16) and 356 (SD = 148.63), respectively. They had studied English for six or seven years mainly through the Japanese secondary school system. None of the participants had studied English overseas although 12 students had been to the United States, Australia, or Canada for a short trip.

Instruments

The vocabulary size measure was based on Form 1 of the Vocabulary Size Test (Nation, 2008). The words included on the Vocabulary Size Test are based on twenty 1,000 British National Corpus (BNC) word lists developed by Nation (2006). The number of items on the original Vocabulary Size Test was truncated from 140 to 60 items. The new test was made up of 10 words per frequency level from the first to sixth 1,000-word levels. In a pilot test administered to 150 students, the items from the first 1,000 to the eighth 1,000-word frequency levels were used to estimate the participants' knowledge of written receptive vocabulary. The results of the pilot test indicated that most learners' vocabulary sizes were between 3,000 and 4,000 words

and that they rarely knew items beyond the sixth 1,000-word frequency level. Moreover, past research (Barrow, Nakanishi, & Ishino, 1999) indicated that it is quite unlikely that Japanese university students know words beyond the eighth 1,000-word frequency level.

The following is a sample item.

soldier: He is a **soldier.**

a. person in a business
b. student
c. person who uses metal
d. person in the army

Antonym Semantic Decision Task

When conducting a semantic decision task in reaction time research, a two-word judgment task is often used (Jiang, 2012). In this task, two words are simultaneously presented to a participant who must decide whether the two words are synonyms or not. Many researchers have utilized this type of task to examine lexical representations and processing (e.g., Azuma, Williams, & Davie, 2004; Morita & Matsuda, 2000; Perfetti & Zhang, 1995). However, a two-word judgment task was not utilized in this study. As the primary purpose of the semantic decision task was to measure the meaning access component of word recognition, only the reaction time of the target word should be measured, and not the reaction time to both words. Therefore, the semantic priming method was used in this study. McDonough and Trofimovich (2009) defined semantic priming as a facilitation in the speed or accuracy of processing a word (e.g., nurse) when it is preceded by a semantically related word (e.g., doctor) relative to when it is preceded by a semantically unrelated word (e.g., butter). Even though semantic priming effects are susceptible to strategic influences (i.e., training participants to expect the association of the priming word), they are largely automatic and often precede conscious attention or awareness. As the meaning access component of word recognition should be largely automatic for fluent readers, the priming method is a suitable way to measure this construct.

Word pairs that are synonymous or related in meaning are typically utilized as test materials in the semantic priming method. However, in this study, antonymous word pairs were employed. The justification of using antonymous word pairs is that antonymous words pairs were easier to create than synonymous ones for noun pairs (e.g., teacher-student, father-mother, boy-girl). In addition, several researchers (Kojima, 2010; Shiotsu, 2009, 2010; Yamashita, 2013) have employed antonymous word pairs to measure lexical meaning access during word recognition and they reported that the approach functioned well. In the Antonym Sematic Decision Task, a prime word (e.g., strange) appears on the computer screen, and after

a specified duration of 1,000 ms, a target word (e.g., familiar) appears on the screen. Test-takers decide whether the target word is antonymous to the prime word as quickly as possible by pressing keyboard buttons.

Creation of the items for Antonym Semantic Decision Task

Items created by Kojima (2010) were adapted for use in the ASDT. The original version contained 128 pairs of items, all of which were content words (i.e., nouns, verbs, adjectives, and adverbs). Half of the 128 pairs were antonyms in which the stimulus word and the target word were related, and half were not semantically related. All the prime words were selected from the 2,000 high-frequency words of the Japan Association of College English Teachers (JACET) 8,000 words list (Aizawa, Ishikawa, & Murata, 2005). Thirty-two corresponding antonymous target words were selected from the first four 1,000-word frequency levels on the JACET 8,000 list (32 words × 4 frequency levels = 128 total items). Half of the target words (64 items) were matched with unrelated words that had the same number of letters and were from the same word frequency levels. Thus, each 1,000-word frequency level consisted of 16 antonymous pairs and 16 unrelated pairs. In order to control the number of letters in the target words at each frequency level, the 32 stimulus words at each level consisted of eight words with four or five letters, 19 words with six to eight letters, and five words with nine to 11 letters. However, the parts of speech of the target words in each level were not equally distributed.

The following changes were made to the stimulus words. First, adverbs were excluded because no adverbs were included on the Lexical Decision Task in other studies. Second, the word frequency level of each stimulus word was examined with Vocabprofilers BNC-20 (Cobb, 2013) and each word was categorized based on its frequency in the BNC Corpus because the word frequency level of the JACET 8,000 is not identical to the frequency level of the BNC. Thereafter, 18 target words for each of the first four 1,000-word frequency levels (18 items × 4 word frequency levels = 72 total items) were constructed using the following procedure. Forty-six of Kojima's original stimulus words were used, and 26 stimulus words were newly added.

Next, the 72 corresponding prime words were selected. Forty-seven prime words were adopted from Kojima's original prime words, and 25 prime words were newly added. These 25 antonymous items were mostly from the first 1,000 to the third 1,000 word families in the BNC, which were selected from the JACET 8,000 list. The meanings of the antonymous words corresponding to the target words were checked using the Thesaurus.com webpage. Third, the target words' lexical properties—the number of letters and syllables—were controlled by selecting six sets of three stimulus words, each of which consisted of four, five, and six letters at each word frequency level. Besides controlling for the number of letters and syllables, the part of speech of the words was also controlled. Among the six sets of three stimulus words,

two sets were adjectives, two were nouns, and two were verbs. Fourth, as filler items, another 72 prime words were constructed from Kojima's list and words from the JACET 8,000 list. Most of these prime words were in the first 1,000 or second 1,000 word frequency levels in the BNC. Finally, 72 words unrelated to the prime words were constructed. These 72 unrelated words had the same number of letters, part of speech, and frequency level as the 72 antonymous target words. All stimulus words were checked by three native speakers of English who hold doctoral degrees in the field of education (TESOL) and who teach English at a university.

Procedure

The Vocabulary Size Measure was administered during a regular class period. The participants completed the test in about 30 minutes. The Antonym Semantic Decision Task (ASDT) was conducted with SuperLab version 5 (2011). SuperLab is an experiment generator package used to design and administer many types of psychometric experiments that require presenting stimuli on the screen or auditory stimuli via speakers. SuperLab has been utilized by many psycholinguistic researchers to examine lexical representation and processing.

The participants took the ASDT individually on a laptop computer in a quiet room. The room was reserved for the reaction time tests so that interruptions were completely avoided; each of two computers was set on a separate table so that each student was able to focus on the computerized test. Before the main trial, they listened to oral instructions in Japanese regarding the concept of antonymous words and a description of the task. The participants had to decide whether the meaning of the target word was antonymous to the meaning of the prime word as quickly as possible. After completing 15 practice items with an oral explanation by the researcher, they began the test. The 144 word pairs were divided into four sets of trials, with each trial consisting of 36 word pairs (18 word pairs were antonymous and 18 were unrelated). The frequency levels of the target words were distributed equally among the four trials. On each trial, the +++ sign, which indicated the focal point, appeared on the screen for 1,000 ms, and then after a blank screen for 100 ms, the prime word appeared on the screen. The target word appeared on the screen after 1,000 ms. The interval between the prime and the target word was set at 1,000 ms to ensure that the obtained prime effects were due to automatic processes without being affected by task expectation. The participants responded Yes (i.e., It is antonymous to the meaning of the prime word) by pressing B, and No (i.e., It is not antonymous to the meaning of the prime word) by pressing N on the keyboard. No feedback concerning correctness was given. Most participants completed the test in 10 to 15 minutes.

Results

The results section consists of two parts. In the first section, an overview of the results is

provided and the outlying responses are discussed. In the second part, each of the hypotheses is examined.

Initial analysis

A search for outliers was made using (a) overall mean accuracy, (b) target accuracy rate, (c) false alarm rate, and (d) mean reaction time for the correctly identified target words. Outliers were detected using a z-score of ±3.29 (Tabachnick & Fidell, 2007). One participant was identified as an outlier. Participant 48 had a false alarm rate of 56%; 40 out of 72 non-antonymous words were wrongly identified as antonymous words. This was exceptionally high, as the mean false alarm rate for 166 students was 12%, which implied that this participant did not take the tests seriously. Therefore, participant 48 was excluded from further data analysis. The 164 participants' overall mean accuracy rate on the antonym semantic decision task was 71.38% with accuracy rates ranging from 55% to 93%.

Test scores for correctly identified antonymous words were calculated for each word frequency level and for overall performance. Outliers were first defined using reaction times shorter than 300 ms and longer than 3,000 ms; 166 out of 6,636 reactions were identified and were excluded from the further analysis because they were not regarded as a reflection of true reaction time. This change affected 2.5% of the target items. Second, the criterion of 2.5 standard deviations (*SDs)* from the mean reaction time was utilized, and responses more than 2.5 *SD*s beyond individual mean RTs were replaced with a value at the 2.5 *SD* point. This change affected 1.8% of the data across all participants. Only correctly identified real words were included in the final reaction time analyses.

In order to examine the effect of vocabulary proficiency on accuracy and reaction time in the lexical decision task, the participants were divided into two groups using the Rasch person measures (logits) on the Vocabulary Size Measure. The first group was made up of 82 higher proficiency students and the second was made up of 82 lower proficiency students. Table 1 shows the descriptive statistics for the Vocabulary Size Measure for the two groups. The mean of the high proficiency students, .75, was much higher than that of the low proficiency students, -2.4. In addition, the 95% confidence interval means for the two groups did not overlap, which indicated that they were significantly different. The skewness and kurtosis statistics were converted into z-scores to examine the normality of the distributions. A z-score of 3.29 (Tabachnick & Fidell, 2007) was used as the cut point. The kurtosis of the Low Proficiency Group was normal as the z-score was 1.44, skewness was significantly non-normal (z-score = 3.73). The negative skew for the Low Proficiency Group indicated that many of these participants had high scores in the group. The z-scores for skewness and kurtosis for the High Proficiency Group were 4.92 and 4.94, respectively, which indicated that both were significantly non-normal. The positive skewness indicated that more than half of the scores

were below the mean and the positive kurtosis indicated that the data distribution was taller than a standard normal distribution.

Table 1. Descriptive Statistics for the Vocabulary Size Measure by Groups

	Low proficiency group	High proficiency group
M	-0.24	0.75
SE	0.05	0.03
95% CI	[-0.33, -0.14]	[0.68, 0.82]
SD	0.42	0.31
Skewness	-0.99	1.31
SES	0.27	0.27
Kurtosis	0.76	2.59
SEK	0.53	0.53

Note. All statistics are based on Rasch logits.

Primary Analysis

Hypothesis 1 stated that ASDT accuracy would improve as group proficiency level and word frequency level increase. Table 2 shows the descriptive statistics for the overall accuracy rate (both correctly identified antonymous words and non-antonymous words) by group. The High Proficiency Group's accuracy rate, 75.54% was higher than that of the Low Proficiency Group, 67.23%. For the Low Proficiency Group, the z-scores for skewness and kurtosis were .52 and 1.36, respectively, and for the High Proficiency Group, .03 and 1.57, respectively; thus the skewness and kurtosis for both groups were regarded as normal.

Table 2. Descriptive Statistics for Overall Accuracy Performance Rate by Group

	Low proficiency group	High proficiency group
M	67.23	75.54
SE	.70	.61
95% CI	[65.83, 68.63]	[74.32, 76.75]
SD	6.37	5.54
Skewness	.14	-.01
SES	.27	.27
Kurtosis	-.72	.83
SEK	.53	.53

Note. The unit of the overall accuracy score is percentile.

An independent sample *t*-test was conducted to examine the difference between the groups' accuracy performance. Levene's test was not significant ($p = .06$). Therefore, the equal variance was assumed. The test was significant, $t(162) = -8.91$, $p < .001$, which indicated that the

two groups' overall accuracy rate differed significantly.

Table 3 shows the proportion of the mean accuracy rate (i.e., 1 is perfect) and the *SDs* for correctly identified real words by frequency and group. The accuracy means discriminate between Low Proficiency Group and High Proficiency Group across the word frequency levels, although the high standard deviation of both groups implies considerable individual variance in their responses.

Table 3. Accuracy Rate for Correctly Identified Antonymous Words by Frequency Level and Group

Word		Accuracy rate	
frequency	Proficiency group	*M*	*SD*
1K	Low	.76	.10
	High	.83	. 08
2K	Low	.58	.16
	High	.72	.14
3K	Low	.38	.16
	High	.47	.16
4K	Low	.27	.16
	High	.33	.17
Overall	Low	.50	.12
	High	.59	.11

Note. 1K = the first 1,000 word frequency; 2K = the second 1,000 word frequency; 3K = the third 1,000 word frequency ; 4K = the fourth 1,000 word frequency.

In Figure 1, both the low and high proficiency groups clearly show frequency effects on accuracy for correctly responding to antonymous words because their accuracy decreased consistently as word frequency level decreased.

The accuracy scores were analyzed using a mixed ANOVA. Group was the between-subjects factor (two levels: Low Proficiency Group × High Proficiency Group) and frequency level was the repeated-measures factor (four levels: 1K × 2K × 3K × 4K). The sphericity assumption, which hypothesizes that the variances of the data taken from the same participants are equal, was met. There was a significant main effect of word frequency level on accuracy measure, $F(3, 486) = 756.29$, $p < .001$, partial $\eta^2 = .82$. Tests of within-subjects contrasts showed a linear relationship. There was also a significant main effect of proficiency on accuracy (between subjects), $F(1, 162) = 1.43$, $p < .001$, partial $\eta^2 = .14$. Moreover, there were significant interactions between word frequency and group, which indicated that the accuracy score of each word frequency level differed between the two groups, $F(1, 486) = 4.77$, $p = .003$, partial $\eta^2 = .03$.

As a post-hoc analysis, all the pairwise comparisons for mean accuracy scores by

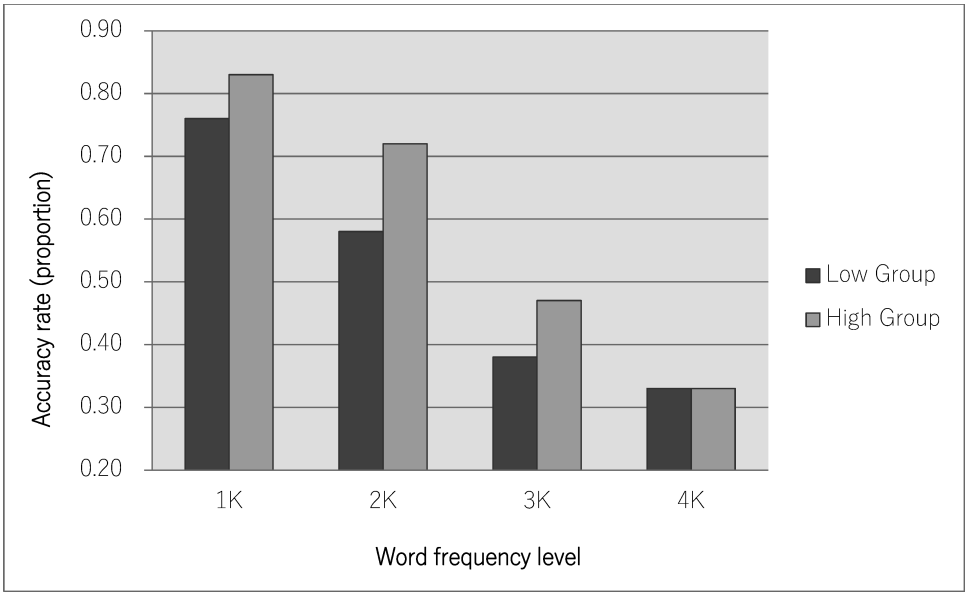

Figure 1. Mean accuracy scores by group proficiency level and word frequency level. Low Group = Low Proficiency Group; High Group = High Proficiency Group.

frequency were conducted separately for each group using a mixed ANOVA. The assumption of sphericity was met. For the Low Proficiency Group, $F(3, 243) = 384.50$, $p < .001$, partial $\eta^2 = .83$, significant differences were found for the all levels comparisons ($p < .01$, Bonferroni adjustment for multiple comparisons). For the High Proficiency Group, $F(3, 243) = 376.93$, $p < .001$, partial $\eta^2 = .82$, significant differences were found for all comparisons. The results indicated that the differences in accuracy rate by word frequency level shown in Figure 1 were statistically significant. Hence, Hypothesis 1 was fully supported.

Hypothesis 2 stated that reaction time in the ASDT would decrease as group proficiency level and word frequency level increase. Table 4 shows the descriptive statistics for the reaction times for correctly identified antonymous words by group. The mean reaction time of the High Proficiency Group, 1,061.94 ms was slightly faster than that of the Low Proficiency Group, 1,080.85 ms. However, the 95% confidence intervals overlapped, which indicated that the mean reaction times were not significantly different. For the Low Proficiency Group, *z*-scores of skewness and kurtosis were .03 and 1.34, respectively, and for the High Proficiency Group, 1.56 and .16, respectively; thus the distributions were acceptably normal.

Table 4. Descriptive Statistics for the Mean Reaction times for Correctly Identified Antonymous Words by Group

	Low proficiency group	High proficiency group
M	1080.85	1061.94
SE	24.09	21.89
95% CI	[1032.91, 1128.79]	[1018.39, 1105.49]
SD	218.19	198.20
Skewness	-.01	.42
SES	.27	.27
Kurtosis	-.73	.09
SEK	.53	.53

Note. The unit of reaction time is in milliseconds; CI = Confidence interval.

An independent sample *t*-test was conducted to examine group differences of overall reaction time when correctly responding to antonymous words. Levene's test was not significant ($p = .25$). Therefore, the equality-of-variance assumption was met. The t-test was not significant, $t(162) = .63$ $p = .53$. The result indicated that the two groups' overall mean reaction times were not statistically different.

Table 5 shows the means and standard deviations for the reaction times for correct responses to antonymous words by group and word frequency level. It was important to note that seven students in the low proficiency group did not contribute to the mean reaction time at the fourth 1,000 word level because they did not correctly identify any of the target words in

Table 5. Means and Standard Deviations for Reaction Time by Group and Word Frequency Levels

Word frequency	Proficiency group	Reaction time (msec)	
		M	*SD*
1K	Low	993	214
	High	954	181
2K	Low	1019	195
	High	1015	201
3K	Low	1243	326
	High	1216	253
4K	Low	1243	392
	High	1224	332
Overall	Low	1081	214
	High	1062	198

Note. 1K = the first 1,000 word frequency; 2K = the second 1,000 word frequency; 3K = the third 1,000 word frequency; 4K = fourth 1,000 word frequency.

4K word frequency levels. Therefore, the sample size of the low proficiency group at 4K word frequency level was 75. As Table 5 shows, the reaction times for the Low Proficiency Group ranged from 993 ms at the 1K level to 1243 ms at the 4K level. For the High Proficiency Group, mean reaction times increased from 954 ms at the 1K level to 1224 at the 4K level.

The scores were analyzed using a one-way mixed ANOVA. Group was the between subjects factor (Low Proficiency Group × High Proficiency Group) and Word frequency levels was the repeated measure factor (1K × 2K × 3K × 4K). The sphericity assumption, which states that the variances of the data taken from the same participants are equal, was violated; hence, the results are reported using the Greenhouse-Geiser correction. Frequency effects were significant, $F(1.75, 284.24) = 61.51$, $p < .001$, partial $\eta^2 = .28$. Moreover, tests of within-subjects contrasts was significant ($p < .001$), indicating a linear relationship among the four frequency bands. There was no significant interaction between word frequency level and proficiency, which indicated that the performance of the two groups at each word frequency level was not drastically different. Figure 2, which shows the mean reaction times by group and word frequency, graphically displays this linear relationship.

In order to investigate the differences in reaction time among the four frequency bands, pairwise comparisons were conducted separately for each group using one-way repeated measure. Because the sphericity assumption was violated, all the results are reported with Greenhouse-Geisser correction. For the low proficiency students group, the results were

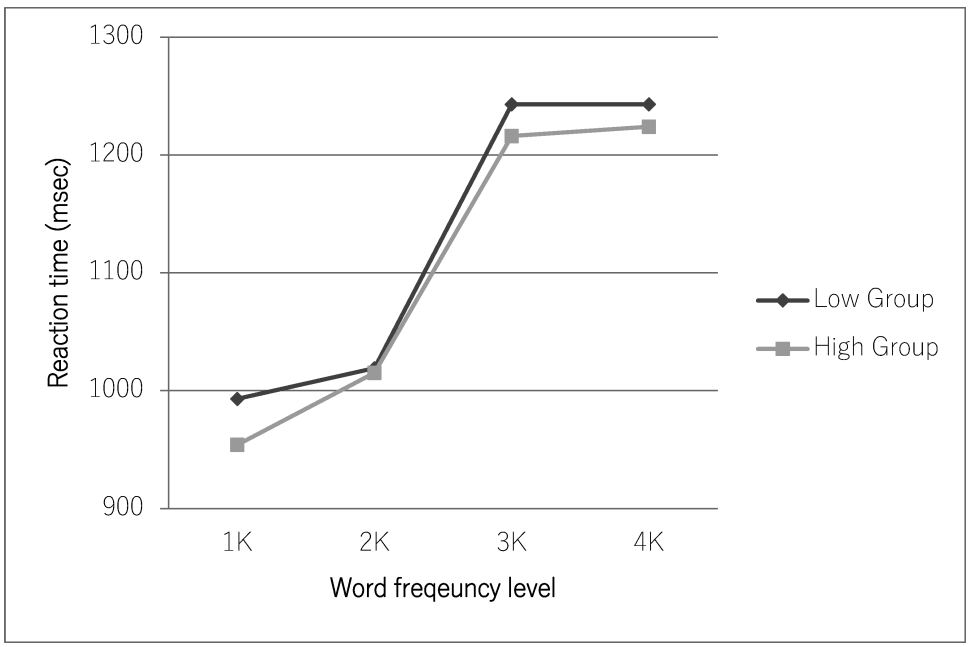

Figure 2. Mean reaction time by group proficiency level and word frequency level. Low Group = Low Proficiency Group; High Group = High Proficiency Group.

significant, $F(1.68, 136.06) = 21.47$ $p < .001$, partial $\eta^2 = .21$. Significant differences were observed in all the pairwise comparisons except for the 1K-2K and 3K-4K pairs. As for High Proficiency Students Group, $F(1.93, 156.68) = 59.26$, $p < .001$, partial $\eta^2 = .42$. Frequency effects on reaction time were observed in all the pairs except for the 3K-4K pairs. Table 6 shows the pairwise comparisons for reaction time by word frequency levels and group proficiency levels.

In sum, even though the effect of group proficiency on reaction time was not statistically significant, the tendency that the mean reaction time of the high proficiency group was faster than that of low proficiency students was observed in the raw data. In addition, general word frequency effects were found in both groups. Hence, these results generally supported Hypothesis 2.

Table 6. Pairwise Comparisons for Reaction Time by Frequency and Groups

Level	Reaction time	
differences	Low group	High group
1K-2K	ns	*
1K-3K	*	*
1K-4K	*	*
2K-3K	*	*
2K-4K	*	*
3K-4K	ns	ns

Note. * = Difference significant at <.05, Bonferroni adjusted for multiple comparisons. Low Group = Low Proficiency Group; High Group = High Proficiency Group.

Hypothesis 3: Response stability, CV in the antonym semantic decision task will decrease as group proficiency and word frequency level increase.

For the calculation of CV_{RT} (*SD* divided by the mean reaction time), 3 participants in the low proficiency group were excluded at the 3K level because they identified only 1 item of the 3K word frequency level correctly. Therefore, no standard deviation was calculated. In addition, 11 students in the low proficiency group and 3 students in the high proficiency group were excluded at the 4K level because they either missed all the items of 4K words or correctly identified only 1 target item at the 4K level.

As the overall column in Table 7 shows, the mean CV_{RT} of the High Proficiency Group and the Low Proficiency Group did not show any differences. Therefore, a proficiency effect on CV_{RT} was not presented. Regarding the effect of word frequency on CV_{RT}, it seems that CV_{RT} increased as the word frequency level decreased, which is opposite from the hypothesized result as smaller value of CV_{RT} is considered to be more stable. Howerver, the changes were subtle across the four frequency levels. A one-way repeated ANOVA was conducted for each

group to find all the pairwise differences. The ANOVA was not significant for either group, which indicated that CV_{RT} did not discriminate either proficiency of group or word frequency levels. Hence, Hypothesis 3 was not supported.

Table 7. Means and Standard Deviations for Coefficient of Variation by Group and Word Frequency Levels

		Coefficient of variation	
Word frequency	Proficiency group	*M*	*SD*
1K	Low	.38	.10
	High	.37	.08
2K	Low	.34	.11
	High	.36	.12
3K	Low	.31	.12
	High	.32	.10
4K	Low	.31	.13
	High	.35	.13
Overall	Low	.37	.08
	High	.37	.08

Note. 1K = the first 1,000 word frequency; 2K = the second 1,000 word frequency; 3K = the third 1,000 word frequency; 4K = fourth 1,000 word frequency.

Discussion

Overall accuracy (correctly identified antonymous words and correctly rejected unrelated words) generally improved as the participants' proficiency increased. Moreover, the standard deviation of the accuracy scores decreased as group proficiency increased, which indicated that the low proficiency group defined by vocabulary size responded less consistently than the high proficiency group.

Frequency effects on accuracy for correctly responding to antonymous words were significant for both the high and the low proficiency groups because their accuracy decreased consistently as word frequency level decreased as it was shown in Figure 1. This suggested that the ASDT validly distinguished learners' proficiency levels and word frequency levels for lexical meaning accuracy.

Frequency effects on mean reaction time with correctly identified antonymous words were generally observed in both high and low proficiency groups, as both groups tended to respond more quickly to high frequency words than to low frequency words. The lack of a significant difference in mean reaction times at the 3K-4K levels suggests that the 3K and 4K words were equally unfamiliar to them.

Contrary to frequency effects on mean reaction time, overall reaction time was not

significantly different between the high and low proficiency groups even though overall mean reaction time for the high proficiency group was faster than that of the low proficiency group in the raw data. This implies that the ASDT did not distinguish proficiency levels for lexical meaning access speed in this study. However, this does not automatically mean that ASDT is not a valid measure for the lexical meaning access speed as it clearly showed overall word frequency effects in both groups for reaction time. Close examination of mean reaction time by group and word frequency levels in Figure 2 indicated that even though there seemed to be a difference between the high proficiency group and low proficiency group for the mean reaction time at the 1K level, the mean reaction time for 2K, 3K, and 4K word frequency levels were almost the same. This suggests that neither the high proficiency nor low proficiency group have developed automaticity of word recognition beyond 2K word frequency levels; although, their vocabulary sizes were different. This result partially supports Laufer and Nation (2001), who argued that development of fluency lags behind increases in overall vocabulary size.

Regarding the use of CV, it did not distinguish proficiency levels as the mean CV_{RT} of the High Proficiency Group and Low Proficiency Group did not show any differences. In addition, no significant differences were observed pairwise comparison of word frequency levels. Moreover, as Table 6 shows, it seems that CV_{RT} increased as the word frequency level decreased, which is opposite to hypothesized result as the smaller value of CV_{RT} is considered to signify more automatized word recognition. However, these results aligned with Kojima's study (2010) in which the CV_{RT} scores of low frequency words (level 4) were smaller than those of words with high frequency words (level 1). Kojima (2010) hypothesized that the accuracy deteriorated as the word frequency levels decreased, especially at frequency level 4. Due to the elimination of wrong responses, the proportion of *SD* did not increase more than the corresponding proportional increase in *RT,* which led to the smaller value of CV_{RT}. This study supported her hypothesis as the mean ASDT accuracy for 4K words levels was 33 percent for the high proficiency group and 27 percent for the low proficiency group. The results implied that CV might not be sensitive enough to measure relatively low proficiency of EFL learners' word recognition.

Conclusion

This study validated the Antonym Sematic Decision Task, which was designed to measure relatively low proficient Japanese L1 EFL learner's automaticity of word recognition, specifically lexical meaning access speed and accuracy. The results showed that ASDT accuracy generally distinguished proficiency and word frequency. Moreover, a general frequency effect for reaction time was found for both high and low proficiency groups. However, overall mean reaction time did not significantly distinguish between high and low

proficiency groups, which implied that the participants have not developed automaticity of word recognition even though their vocabulary sizes are different. In addition, the score of CV_{RT} did not distinguish word frequency levels for the participants in this study, which implied that CV might be less sensitive than RT when the participants' lexical proficiency is relatively low.

References

Aizawa, K., Ishikawa, S., & Murata, T. (Eds.). (2005). JACET 8000 words. Tokyo, Japan: Kiriharashoten.

Azuma, T., Williams, E. J., & Davie, J. E. (2004). Paws + cause = pause? Memory load and memory blends in homophone recognition. *Psychonomic Bulletin & Review, 11*(4), 723-728. doi:10.3758/BF03196626

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, 27, 223- 247. doi:10.1016/S0346-251X(99)00018-4

Carrell, P. L., & Grabe, W. (2002). Reading. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 216-231). London, England: Arnold.

Cunningham, A., Perry, K., & Stanovich, K. (2001). Converging evidence for the concept of orthographic processing. *Reading and Writing, 14*(5), 549-568. doi:10.1023/A:1011100226798

Cobb, T. (2013). Vocabprofilers BNC-20 in Compleat Lexical Tutor [Online computer software]. Retrieved from www.lextutor.ca/vp/

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

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading*. New York, NY: Longman.

Harrington, M. (2006). The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA Yearbook, 6*, 147-168. doi:10.1075/eurosla.6.10har

Hulstijn, J. H., van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us about the language acquisition device? *Applied Psycholinguistics, 30*(4), 555-582. doi:10.1017.S0272263102002115

Jiang, N. (2012). *Conducting reaction time research in second language studies*. New York, NY: Routledge.

Koda, K. (2004). *Insights into second language reading*. New York, NY: Cambridge University Press.

Kojima, M. (2010). Effects of word recognition speed, accuracy, and automaticity on reading ability. *Annual Review of English Language Education in Japan, 21*, 151-160.

Lim, H., & Godfroid, A. (2014). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijin, Van Gelderen and Schoonen's 2009 study. *Applied Psycholinguistics*, (5), 1-36. doi:10.1017/S0142716414000137

McDonough, K., & Trofimovich, P. (2009). *Using priming methods in second language research*. New York, NY: Routledge.

Morita, A., & Matsuda, F. (2000). Phonological and semantic activation in reading two- kanji compound words. *Applied Psycholinguistics, 21*(4), 487-503.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59-82.

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston, MA: Heinle.

Perfetti, C. (1999). Comprehending written language: A blueprint for the reader. In C. Brown & P. Hagoort

(Eds.), *Neurocognition of language* (pp. 167-208). Oxford: Oxford University Press.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357-383. doi:10.1080/10888430701530730

Perfetti, C., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. Snowling & C. Hulme (Eds.), *The science of reading* (pp. 227-247). Malden, MA: Blackwell.

Perfetti, C. A., & Zhang, S. (1995). Very early phonological activation in Chinese reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 24-33.

Segalowitz, N., & Segalowitz, S. J. (1993). Skilled performance, practice and differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics, 13*(3), 369-385. doi:10.1017/S0142716400010845

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics, 19*(1), 53-67. doi:10.1017/S0142716400010572

Shiotsu, T. (2009). Reading ability and components of word recognition speed: The case of L1 Japanese EFL learners. In H. Han & N. J. Anderson (Eds.), *Second language reading research and instruction* (pp. 15-37). Ann Arbor, MI: University of Michigan Press.

Shiotsu, T. (2010). *Components of L2 reading: Linguistic and processing factors in the reading test performances of Japanese EFL learners*. Cambridge: Cambridge University Press.

SuperLab (version 5.0) [Computer software]. San Pedro, CA: Cedrus Corporation.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.) Boston, MA: Pearson Education Inc.

Torgesen, J. K., & Burgess, S. R. (1998). Consistency of reading-related phonological processes throughout early childhood from longitudinal-correlation and instructional studies. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3-40). Mahwah, NJ: Erlbaum.

Yamashita, J. (2013). Word recognition subcomponents and passage level reading in a foreign language. *Reading in a Foreign Language, 25*(1), 52-71.