Vocabulary Learning and Instruction

Volume 10, Number 2,

December 2021
VLI Editorial Team

Editor-in-Chief:

Stuart McLean, Momoyama Gakuin University, Osaka, Japan

Senior Associate Editors:

Joseph P. Vitta, Kyushu University, Fukuoka, Japan
Christopher Nicklin, Rikkyo University, Tokyo, Japan

Editorial Board:

Ali H. Al-Hoorie, Royal Commission for Jubail and Yanbu, Saudi Arabia
Aaron Batty, Keio University, Tokyo, Japan
David Beglar, Temple University Japan, Osaka, Japan
Phil Bennett, University of Niigata Prefecture, Niigata, Japan
Luke Fryer, The University of Hong Kong, China
Dana Gablasova, Lancaster University, Lancashire, England
Aaron Gibson, Fukuoka University, Fukuoka, Japan
Phil Hiver, Florida State University, Tallahassee, USA
Tomoko Ishii, Metji Gakuin University, Tokyo, Japan
Brandon Kramer, Kwansei Gakuin University, Osaka, Japan
Geoffrey G. Pinchbeck, Carleton University, Ottawa, Canada
Tim Stoeckel, University of Niigata Prefecture, Niigata, Japan
Jeffrey Stewart, Tokyo University of Science, Tokyo, Japan
Tatsuya Nakata, Rikkyo University, Tokyo, Japan
Kiyomi Yoshizawa. Kansai University, Osaka, Japan
Raymond Stubbe, Kyushu Sangyo University, Fukuoka, Japan

The Editorial Team expresses a sincere thank you to Mana Ikawa, who designed the logo and the cover for the print version of VLI.

Copyright © 2021 Vocabulary Learning and Instruction, ISSN: Online 2187-2759. All articles are copyrighted by their respective authors.
# Vocabulary Learning and Instruction

Volume 10, Number 2, December 2021

## Table of Contents

<table>
<thead>
<tr>
<th>Articles</th>
<th>Page</th>
</tr>
</thead>
</table>
| Measuring Attrition of L2 Productive Vocabulary Knowledge Over the Summer Vacation  
*Brandon Kramer, Tohru Matsuo, Aaron C. Sponseller, Young Ae Kim, Suzuka Nishiyama, and Stuart McLean* | 1    |
| A Comparison of Textbook Vocabulary Load Analysis  
*Stuart Benson and Naheen Madarbakus-Ring* | 9    |
| The Contribution of High-frequency Multi-word Sequences to Speech Rate and Listening Perception Among EFL Learners  
*Michael McGuire and Jenifer Larson-Hall* | 18   |
| Comparisons of Word Lists on New Word Level Checker  
*Atsushi Mizumoto, Geoffrey G. Pinchbeck, and Stuart McLean* | 30   |
| Second Language Vocabulary Teaching and Learning: A Commentary on Four Studies for the JALT Vocabulary SIG  
*David Beglar* | 42   |
| Correlations of Modalities of Written Vocabulary Knowledge to Listening and Reading Proficiency: A Comparison  
*Jeffrey Stewart, Stuart McLean, and Aaron Olaf Batty* | 55   |
| The Internal Consistency and Accuracy of Automatically Scored Written Receptive Meaning-Recall Data: A Preliminary Study  
*Stuart McLean, Paul Raine, Geoffrey Pinchbeck, Laura Huston, Young Ae Kim, Suzuka Nishiyama, and Shotaro Ueno* | 64   |
| Considerations and Challenges in Longitudinal Studies of Lexical Features in L2 Writing  
*Minkyung Kim* | 82   |
| Developing a Measure of Proper Name Familiarity for Japanese University Students  
*Christopher Nicklin* | 91   |
| Discussion Paper: Using Statistics to Solve Practical Vocabulary Problems  
*Jenifer Larson-Hall* | 101  |
Letter from the Editors

Dear VLI Readers,

Welcome to the second VLI issue of 2021, which contains a set of studies written for the JALT Vocabulary SIG Symposium. The event was held in December 2021 at Doshisha University in Kyoto, and was a particularly special one due to the fact it was the first in-person conference that many of the presenters and attendees had experienced for quite a while due to the Covid-19 pandemic. Special thanks should be paid to Michael McGuire for the time and effort that he spent making the symposium the great success that it was.

The following pages comprise the fruits of the symposium, which was divided into two sessions. The morning session was dedicated to vocabulary learning and featured presentations by Brandon Kramer and Tohru Matsuo, Stuart Benson and Naheen Mandarbakus-Ring, Michael McGuire and Jenifer Larson-Hall, and Atsushi Mizumoto, with commentary provided by David Beglar. The focus of the afternoon session’s research was vocabulary assessment, with commentary provided by Jenifer Larson-Hall and presentations by Jeff Stewart and Aaron Batty, Stuart McLean, Minkyung Kim, and Christopher Nicklin.

We sincerely hope that you enjoy reading this set of studies as much as the presenters and attendees enjoyed the Kyoto symposium. Plans are already under way for the 2022 symposium, which we hope that many more of our readers will be able to attend. We hope to see you there!

Christopher Nicklin and Joseph P. Vitta

VLI Associate Editors
Measuring Attrition of L2 Productive Vocabulary Knowledge Over the Summer Vacation

Brandon Kramer\textsuperscript{a}, Tohru Matsuo\textsuperscript{b}, Aaron C. Sponseller\textsuperscript{b}, Young Ae Kim\textsuperscript{c}, Suzuka Nishiyama\textsuperscript{d}, and Stuart McLean\textsuperscript{d}
\textsuperscript{a}Kwansei Gakuin University; \textsuperscript{b}Osaka Jogakuin University & Junior College; \textsuperscript{c}Kyoto Seika University; \textsuperscript{d}Momoyama Gakuin University

Abstract

For many teachers and administrators, the degree to which attrition over summer vacation represents a threat to instructed language acquisition remains unclear. In a previous study, Kramer et al. (2019) looked at receptive vocabulary knowledge attrition over summer vacation, found no evidence of attrition using these measures, and called for future research to instead use tests of productive vocabulary knowledge which is more likely to be forgotten. Therefore, in this study we investigate the amount of summer attrition among Japanese university students ($N = 81$) and any mediation in that attrition attributable to digital paired-associate vocabulary studying, extensive reading, or experience travelling abroad. The results indicate that although there was no significant group difference in pre- and post-test productive vocabulary scores, a small but significant relationship was found between digital paired-associate vocabulary studying and vocabulary test score gains.

Keywords: Vocabulary, attrition, summer attrition, productive vocabulary knowledge

1 Background and Aim

The degree to which attrition over summer vacation represents a threat to instructed language acquisition remains unclear for many teachers and administrators. This study investigates that threat with regard to productive vocabulary knowledge studied through a coordinated vocabulary program. In a previous study, Kramer et al. (2019) found no evidence of receptive vocabulary knowledge attrition over summer vacation. They called for future research to instead use tests of productive vocabulary knowledge, which is more likely to be forgotten (Schmitt, 2010; Weltens & Grendel, 1993).

The following research questions were developed:

1. To what extent can systematic attrition be detected in Japanese university students’ L2 productive vocabulary knowledge after a 2-month summer break?
2. Is there a relationship between changes in L2 productive vocabulary knowledge and review using a digital paired-associate vocabulary study application, words read through extensive reading, or experience travelling abroad?
3. What patterns can be found in the attrition of partial vocabulary knowledge over a 2-month summer break?
2 Methodology

2.1 Participants

The participants ($N = 81$) for this study were first-year university students at a private women’s university in western Japan. The mean TOEIC score, calculated from those which were available was 351.4 ($SD = 73.7$, $n = 74$). All students studying at this institution participated in a school-wide vocabulary programme that covers the New General Service List (NGSL) (Browne et al., 2013) using a digital paired-associate flashcard and testing application called Vocabulary Builder (EnglishCentral, 2019). According to the university curriculum, all students study a little more than 500 words of the NGSL in their first semester, as shown in Table 1 below.

In addition to the vocabulary programme, all first-year students must read graded readers as part of an extensive reading programme using an online application called Xreading.com (Goldberg, 2019). The curricular goals of this extensive reading program are also presented in Table 1. Graded readers are written using easy and frequent vocabulary. Thus, they are thought to potentially have an effect on vocabulary retention over the summer vacation. The number of words which a student is counted as having read are only for those books for which they have scored over 60% on an associated follow-up comprehension test.

2.2 Instruments

For the measurement of productive vocabulary knowledge, an online vocabulary testing site called VocabLevelTest.org was used (McLean & Raine, 2019). Using this website, the participants were tested on their ability to accurately produce the English forms of words based on prompts that include the Japanese definition and an example sentence. An example prompt can be seen in Figure 1.

Students had 30 seconds to answer each question and the time remaining was indicated by a progress bar below the item prompt. They also had the option to skip the question if they had no knowledge of the target form. One complication of a form-recall test such as this is that there are sometimes multiple possible correct answers for a particular Japanese meaning and example sentence. Therefore, if they typed a word that was associated with the Japanese meaning, but was not the target word being tested, a prompt appeared asking the participant to try to produce the target word form again. Such a prompt is displayed in Figure 2.

<table>
<thead>
<tr>
<th>Program</th>
<th>First semester</th>
<th>Summer Vacation</th>
<th>Second Semester</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Central Vocabulary Builder</td>
<td>NGSL 501-1008</td>
<td>Voluntary review</td>
<td>NGSL 1001-1507</td>
</tr>
<tr>
<td>Xreading (ER)</td>
<td>180,000 words for full points</td>
<td>Voluntary reading which counts towards second semester goals</td>
<td>180,000 words for full points</td>
</tr>
</tbody>
</table>

Note. NGSL = New General Service List (Browne et al., 2013).
The target words tested were 30 randomly selected words each from bands 2 and 3 of the NGSL (frequency ranks 501-1507), for 60 words in total. The same words were tested on the pre- and post-tests. A list of the tested words is shown in Appendix A.

Although the website automatically scores the items dichotomously using a continually growing bank of acceptable answers, in order to measure partial knowledge the results were coded by two native speakers of Japanese who are highly proficient in English. The items were coded as belonging to one of the five categories as shown below ($\kappa = 0.75$; agreement = 84.6%).

A. Correct target word, with correct spelling and derivational form  
B. Correct target word and derivational form, but with incorrect spelling  
C. Correct target word and spelling, but with incorrect derivational form  
D. Spelling and derivational form are incorrect, but demonstrated some knowledge of the target word  
E. No demonstrated knowledge of the target word

In order to conduct the quantitative analyses for research questions 1 and 2, these categories were collapsed into a partial credit model where full points (2)
were assigned for perfect responses (category A only). Partial credit points (1) were assigned for responses which demonstrated any kind of partial knowledge (categories B, C, and D). Zero points were assigned for responses that were coded as containing no demonstrated knowledge of the target word (category E).

2.3 Procedures

The students completed the productive vocabulary pre-test on the final day of classes in the spring semester with instructions to study over the summer vacation in order to maintain their English proficiency. Specifically, all the students were instructed to continue reviewing vocabulary using Vocabulary Builder and to read as much as they could using Xreading.com. During the first class of the fall semester, the students completed the productive vocabulary post-test as well as a survey collecting background information such as TOEIC scores, summer study habits, and whether they travelled to any foreign countries during the semester break.

3 Results

To detect any productive vocabulary knowledge attrition between the pre- and post-test, a paired samples \( t \)-test was conducted using the collapsed categories described previously in JAMOVI (The JAMOVI Project, 2021). No statistically significant difference was found, \( t(80) = 0.12, p = 0.902, d = 0.01 \). Descriptive statistics for both the tests can be found in Table 2, with the distribution of score differences shown in Figure 3.

Over the summer vacation, only a limited number of students continued studying as recommended, either via Vocabulary Builder (\( n = 56; 69\% \) of the sample) or Xreading (\( n = 55; 68\% \) of the sample). See Table 3 for the descriptive statistics for these activities. Even fewer students visited foreign countries during the summer vacation (\( n = 18; 22\% \) of the sample). A multiple linear regression was conducted in JAMOVI (The JAMOVI Project, 2021) to predict the change in vocabulary scores based on the number of reviews using Vocabulary Builder, the number of running words read using Xreading, and international travel. The variables were added in the order of predicted importance (vocabulary review first, followed by words read in Xreading, and finally a dummy code indicating international travel), but only vocabulary review using Vocabulary Builder was retained in the model because of its significant relationship with gains from pre- to post-test, \( F(1,79) = 3.99, p = 0.049, R^2 = 4.9\% \).

Finally, the response data for all the participants were analysed using the original 5-category codes to learn exactly where any attrition was taking place.

Table 2. Descriptive Statistics for the Pre- and Post- Test Productive Vocabulary Scores

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Max</th>
<th>M</th>
<th>SEM</th>
<th>Median</th>
<th>SD</th>
<th>Skew</th>
<th>SES</th>
<th>Kurt</th>
<th>SEK</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-test</td>
<td>15</td>
<td>64</td>
<td>35.8</td>
<td>1.30</td>
<td>36</td>
<td>11.7</td>
<td>0.31</td>
<td>0.27</td>
<td>-0.45</td>
<td>0.53</td>
<td>0.807</td>
</tr>
<tr>
<td>Post-test</td>
<td>12</td>
<td>65</td>
<td>35.7</td>
<td>1.38</td>
<td>35</td>
<td>12.5</td>
<td>0.17</td>
<td>0.27</td>
<td>-0.72</td>
<td>0.53</td>
<td>0.828</td>
</tr>
</tbody>
</table>

Note. \( N = 81; \) SEM = Standard Error of Mean; SES = Standard Error of Skewness; SEK = Standard Error of Kurtosis.
The results are shown in Table 4. Of the total words answered by all participants, 84.60% (10,249 out of 12,115) showed no change between the pre- and post-tests. The two most common attrition patterns were A to C, in which target words were answered perfectly on the pre-test (A) but did not include the correct derivational form on the post-test (C), and B to E, in which the participants spelled target words
incorrectly on the pre-test (B) and demonstrated no knowledge on the post-test (E). Both the A to C and the B to E patterns made up 0.4% of the total responses. Noteworthy, however, is the extremely low amount of attrition as a percentage of all student responses, 1.63%.

4 Discussion

The first research question asked if there was any systematic attrition in the students’ L2 productive vocabulary knowledge after a 2-month summer break. We found no evidence of such attrition using a paired samples t-test of pre- and post-test scores on a vocabulary levels test targeting the NGSL. This is good news. It indicates attrition is difficult to detect over the summer vacation even when using a stringent form-recall vocabulary levels test as we did in this study.

The second research question asked if there is a relationship between changes in this vocabulary knowledge and review using a digital paired-associate vocabulary study application, words read through extensive reading, or experience travelling abroad. We found a significant but small relationship between digital paired-associate vocabulary review and gains on the post-test, with about 0.143 points predicted for every 100 vocabulary reviews. Students must therefore spend a considerable amount of time conducting many reviews if they wish to see noticeable gains on a test such as this over the summer vacation.

The final research question asked if any patterns could be found in the attrition of partial knowledge. With only 1.68% of all the items showing attrition from pre- to post-test, these results reaffirm the conclusion that attrition seems to be a minimal concern over summer vacation.

5 Limitations

The conclusions of this study are limited by weaknesses in the research design. Although students who reviewed vocabulary over the summer vacation studied the frequency bands which contain the target words on the pre- and post-tests, they might not have reviewed those exact words during the summer vacation, depending on the specific words which were brought up by the algorithm within Vocabulary Builder. Similarly, those students who engaged in extensive reading or travelled abroad might not have been exposed to the target words at all. Future research could make sure that these words were included in their summer review, and administer tests unique to each student based on the specific words which were reviewed.

Furthermore, while the target words on both the tests were chosen from the frequency range which was studied during the first year of university, many of these words were likely already well known by some of the participants, making them more resistant to decay, or never learned at all, making attrition impossible. Future studies could control this by measuring the knowledge of only words which were learned during the first semester.

Finally, the extent to which the students in our sample are representative of Japanese university students more generally is unknown. Our sample was drawn
from students at a university where English education is considered a central curricular pillar. Students at other universities may not study English nearly as much during the academic semesters or during summer vacations. Therefore, future research might look to replicate this study in a more typical Japanese university context in which English education is compulsory but not necessarily prioritized.

6 Conclusion

In a previous study, Kramer et al. (2019) looked at student attrition over the summer vacation with regard to the students’ ability to recognize and recall the meanings of L2 word forms that they studied during the school year. No systematic patterns of attrition were found, but it was theorized that if attrition were to take place over such a short period, tests of productive vocabulary knowledge, specifically requiring the recall of the L2 word forms, would be required to detect it. This study therefore measured the form-recall productive knowledge of students before and after summer vacation, but could not find any evidence of systematic attrition among the student sample. Furthermore, if the students wish to increase their vocabulary knowledge over this 2-month break, digital paired-associate vocabulary review (e.g., digital flashcard software) seems to be one way they could do this. But, based on the results of this study it will require a sizable effort to make any noticeable gains, with a 5-point gain in vocabulary test scores requiring an estimated 3,500 reviews.

References


The JAMOVI Project (2021). JAMOVI (Version 1.6) [Computer Software]. https://www.jamovi.org

Appendix A
List of NGSL Words Tested

<table>
<thead>
<tr>
<th>Band</th>
<th>Word</th>
<th>Rank</th>
<th>Band</th>
<th>Word</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>education</td>
<td>531</td>
<td>3</td>
<td>invite</td>
<td>1026</td>
</tr>
<tr>
<td>2</td>
<td>serve</td>
<td>534</td>
<td>3</td>
<td>copy</td>
<td>1035</td>
</tr>
<tr>
<td>2</td>
<td>range</td>
<td>516</td>
<td>3</td>
<td>handle</td>
<td>1039</td>
</tr>
<tr>
<td>2</td>
<td>standard</td>
<td>547</td>
<td>3</td>
<td>conduct</td>
<td>1052</td>
</tr>
<tr>
<td>2</td>
<td>similar</td>
<td>594</td>
<td>3</td>
<td>collect</td>
<td>1067</td>
</tr>
<tr>
<td>2</td>
<td>poor</td>
<td>618</td>
<td>3</td>
<td>appeal</td>
<td>1071</td>
</tr>
<tr>
<td>2</td>
<td>manager</td>
<td>661</td>
<td>3</td>
<td>collection</td>
<td>1082</td>
</tr>
<tr>
<td>2</td>
<td>necessary</td>
<td>676</td>
<td>3</td>
<td>alternative</td>
<td>1088</td>
</tr>
<tr>
<td>2</td>
<td>average</td>
<td>690</td>
<td>3</td>
<td>software</td>
<td>1089</td>
</tr>
<tr>
<td>2</td>
<td>yesterday</td>
<td>703</td>
<td>3</td>
<td>warning</td>
<td>1130</td>
</tr>
<tr>
<td>2</td>
<td>behaviour</td>
<td>706</td>
<td>3</td>
<td>concept</td>
<td>1211</td>
</tr>
<tr>
<td>2</td>
<td>population</td>
<td>710</td>
<td>3</td>
<td>refused</td>
<td>1207</td>
</tr>
<tr>
<td>2</td>
<td>review</td>
<td>730</td>
<td>3</td>
<td>reform</td>
<td>1213</td>
</tr>
<tr>
<td>2</td>
<td>despite</td>
<td>771</td>
<td>3</td>
<td>solve</td>
<td>1219</td>
</tr>
<tr>
<td>2</td>
<td>solution</td>
<td>788</td>
<td>3</td>
<td>reaction</td>
<td>1229</td>
</tr>
<tr>
<td>2</td>
<td>pressure</td>
<td>809</td>
<td>3</td>
<td>bottom</td>
<td>1240</td>
</tr>
<tr>
<td>2</td>
<td>modern</td>
<td>808</td>
<td>3</td>
<td>colleague</td>
<td>1249</td>
</tr>
<tr>
<td>2</td>
<td>avoid</td>
<td>850</td>
<td>3</td>
<td>attract</td>
<td>1296</td>
</tr>
<tr>
<td>2</td>
<td>refer</td>
<td>865</td>
<td>3</td>
<td>twice</td>
<td>1306</td>
</tr>
<tr>
<td>2</td>
<td>basis</td>
<td>908</td>
<td>3</td>
<td>relatively</td>
<td>1324</td>
</tr>
<tr>
<td>2</td>
<td>speech</td>
<td>893</td>
<td>3</td>
<td>actual</td>
<td>1333</td>
</tr>
<tr>
<td>2</td>
<td>encourage</td>
<td>911</td>
<td>3</td>
<td>opposite</td>
<td>1371</td>
</tr>
<tr>
<td>2</td>
<td>earn</td>
<td>925</td>
<td>3</td>
<td>stone</td>
<td>1372</td>
</tr>
<tr>
<td>2</td>
<td>professional</td>
<td>935</td>
<td>3</td>
<td>soldier</td>
<td>1376</td>
</tr>
<tr>
<td>2</td>
<td>memory</td>
<td>938</td>
<td>3</td>
<td>hate</td>
<td>1378</td>
</tr>
<tr>
<td>2</td>
<td>maintain</td>
<td>942</td>
<td>3</td>
<td>carefully</td>
<td>1384</td>
</tr>
<tr>
<td>2</td>
<td>daily</td>
<td>954</td>
<td>3</td>
<td>totally</td>
<td>1394</td>
</tr>
<tr>
<td>2</td>
<td>fix</td>
<td>972</td>
<td>3</td>
<td>army</td>
<td>1416</td>
</tr>
<tr>
<td>2</td>
<td>cross</td>
<td>974</td>
<td>3</td>
<td>familiar</td>
<td>1443</td>
</tr>
<tr>
<td>2</td>
<td>legal</td>
<td>977</td>
<td>3</td>
<td>arrangement</td>
<td>1458</td>
</tr>
</tbody>
</table>
A Comparison of Textbook Vocabulary Load Analysis

Stuart Benson\textsuperscript{a} and Naheen Madarbakus-Ring\textsuperscript{b}
\textsuperscript{a}University of Aizu; \textsuperscript{b}Nagoya University of Commerce and Business

Abstract

The popularity of using textbooks in second language programs in universities around the world continues to grow. Textbooks support teachers in their teaching by providing accessible materials and clear instruction. In addition, learners are guided by familiar lesson frameworks (e.g., beginning, middle, end) to guide their independent study (Swales, 1980). However, textbooks present many challenges. Learners’ difficulties include the range in lexical knowledge they must possess (Nation, 2006) and the different lexical and grammatical features that are found in written textbook registers (Biber et al., 1998). This study investigates and outlines the vocabulary load of two English for Academic Purpose textbooks, using the British National Corpus and Corpus of Contemporary American English (BNC/COCA) 25,000 (Nation, 2012) and JACET8000 (JACET, 2016) word lists. The results show that for each textbook, more lexical demands are needed for second language learners in the JACET8000 compared with the BNC/COCA 25,000 lists. Understanding the content in textbooks will inform of the vocabulary-level requirements needed when taught in tertiary-level programs. Using a general and a Japanese-specific word list to identify possible pedagogical priorities can help to determine textbook priorities for teachers that can be applied to teaching in the Japanese classroom.

1 Background

For many tertiary-level institutions, the textbook is an integral component of any English for Academic Purposes (EAP) program. Richards (2001) observed how textbooks provide integral structure and syllabus guidance for teachers and learners in a language program. Specifically, textbooks offer instruction, guidance, and activities for learners to practise language through using communicative (e.g., role play, discussion activities) and linguistic (e.g., vocabulary definitions, gap fill) components (O’Loughlin, 2012).

Commercially, textbooks range from beginner to advanced levels that aim to become more difficult with each subsequent unit to accommodate learners in their language learning (Mares, 2003). Therefore, one important consideration for both teachers and learners is vocabulary. Coxhead et al. (2010) observed how “the reciprocal relationship between vocabulary knowledge and textbooks is critical” (p. 4). Sun and Dang (2020) concurred, suggesting that learners pay attention to both the textbook content and vocabulary to optimise learning.

Despite these observations, textbooks may lack the vocabulary knowledge necessary for learners and could present several problems for practitioners. As Macalister and Nation (2019) suggested, a language course needs to include high-frequency language items to help learners achieve the best possible vocabulary coverage. As a result, research focusing on textbooks has now turned to investigating the vocabulary loads in textbook content. Researchers have used two forms of vocabulary load analysis: the number of high frequency words included in textbooks and the number of word families needed to reach 95% and 98% coverage (Sun & Dang, 2020).

## 2 Previous Research

Two main studies have shown the first 2,000 word families cover 80% of a written text and 95% of spoken language. Eldridge and Neufeld's (2009) study found 1,400 of the first 2,000 words in the first four levels of the *Success* series. Similarly, O’Loughlin’s (2012) study found 1,435 of the first 2,000 words in the first three levels of the *English File* series. These findings suggest that learners’ vocabulary knowledge of the first 2,000 high-frequency words would account for 75% of the written discourse used in textbooks (Sun & Dang, 2020).

Table 1 shows studies that analyzed the number of word family knowledge needed to reach 95% and 98% coverage, respectively. Four of the five studies (Coxhead et al., 2017; Hajiyeva, 2015; Sun & Dang, 2020; Yang & Coxhead, 2020) found the vocabulary load of different commercial textbooks (CTs) reached 95% coverage by the first 3,000 words (including supplementary lists). Matsuoka and Hirsh (2010) found the *New Headway* series reached 95% coverage by the first 2,000 words while Yang and Coxhead (2020) found a higher level 4 of the *Yilin* textbook needed 4,000 words. Therefore, most textbooks achieve 95% coverage if learners have vocabulary knowledge of the first 3,000 word families. However, the same studies concluded different vocabulary knowledge needed by learners when considering the reach for 98% coverage. The studies reported varying vocabulary loads between the first 5,000 words (for the *New Concept English* series [Yang & Coxhead, 2020]) to the first 9,000 words (for the various *University Textbook* titles [Hajiyeva, 2015] and *Yilin* [Sun & Dang, 2020]).

<table>
<thead>
<tr>
<th>Study</th>
<th>Textbook</th>
<th>Running words</th>
<th>95% coverage</th>
<th>98% coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matsuoka and Hirsh (2010)</td>
<td>New Headway (Upper Intermediate)</td>
<td>44,877</td>
<td>95.5%–2,000</td>
<td>-</td>
</tr>
<tr>
<td>Hajiyeva (2015)</td>
<td>University Textbook (11 titles)</td>
<td>508,802</td>
<td>95%–3,500</td>
<td>98%–9,000</td>
</tr>
<tr>
<td>Coxhead et al. (2017)</td>
<td>English for Specific Purposes (ESP) Textbooks (15 titles)</td>
<td>380,078</td>
<td>95%–3,000</td>
<td>98%–7,000</td>
</tr>
<tr>
<td>Sun and Dang (2020)</td>
<td>Yilin (Four levels)</td>
<td>273,094</td>
<td>95.54%–3,000</td>
<td>98.02%–9,000</td>
</tr>
<tr>
<td>Yang and Coxhead (2020)</td>
<td>New Concept English (Two levels)</td>
<td>L3 -22,786</td>
<td>95.59%–3,000</td>
<td>98.08%–5,000</td>
</tr>
<tr>
<td></td>
<td>L4–18,109</td>
<td>96.51%–4,000</td>
<td>98.3%–6,000</td>
<td></td>
</tr>
</tbody>
</table>
These studies show that irrespective of the textbook title, the knowledge needed to reach 98% vocabulary coverage differs between textbooks. This suggests that further research is needed to analyze the vocabulary load in textbooks as the content may be too lexically demanding for the intended users.

3 Research Approach

We decided to investigate the vocabulary load of two EAP textbooks. The aim was to identify the vocabulary load of each textbook to determine its appropriateness to the student population it was being used for.

3.1 Materials

Two EAP textbooks were investigated for this study. The first textbook was an in-house textbook (IT), which was created by the teachers in an English department at a Japanese university. The book consists of eight units used over one academic year. The second was a commercial textbook (CT), which was developed by Cengage Learning, a reputable English language resources developer. The textbook consists of 12 units used over one academic year.

3.2 Context

Both books are considered appropriate to teach first year tertiary-level students based on their vocabulary and academic content. However, it is unclear about how appropriate these textbooks intended pre-intermediate levels correspond with the student populations’ current proficiency scores. This led to formulating the following research questions:

1. What is the vocabulary load of the in-house textbook (IT)?
2. What is the vocabulary load of the commercially published textbook (CT)?
3. How does the vocabulary load of each textbook compare when analyzed using a general word list and context-specific word list?

3.3 Method

The two chosen textbooks were scanned using Optical Character Recognition (OCR) Software. The scans were then cleaned by page and categorized by unit to prepare for the analysis. The data were run through the Range Program (Heatley et al., 2002) and the New Word Level Checker (Mizumoto, 2021). This study presents the preliminary results of the vocabulary load analysis of each textbook. The analysis used two word lists: a general word list and a context-specific word list. Nation’s (2012) BNC/COCA base words lists were utilized as the general word list. The BNC/COCA divides the most frequent 25,000 word families into 25, 1,000 word base lists, according to their frequency. In addition, Nation’s (2012) supplementary word lists (i.e., proper nouns, marginal words, transparent compounds, abbreviations) were also used. The New JACET8000 word list (JACET, 2016) was utilized as the context-specific word list. The New JACET8000 word list
is an 8,000 lemma educational word list, created for Japanese learners of English, specifically tertiary-level learners. The New Word Level Checker (Mizumoto, 2021) divides the New JACET8000 word list (JACET, 2016) into eight, 1,000 word base lists, according to their frequency. Nation’s (2012) BNC/COCA base word list was used due to its size, and with the Range program, the ability to analyze vocabulary at various frequency levels. The New JACET8000 word list (JACET, 2016) was used as it contains the most frequent vocabulary within the Japanese tertiary level, which is the context where each textbook is utilized.

4 Results

Table 2 presents the vocabulary load of the IT analyzed using Nation’s (2012) BNC/COCA word lists. Nation’s (2012) supplementary lists were included in the vocabulary load analysis when considering the 95% and 98% thresholds because Nation (2013) pointed out that once known, these words are not a burden to learners. In addition, due to their high coverage, the words in the supplementary lists were important to achieve 95% or 98% coverage.

As illustrated in Table 2, with the supplementary lists, the IT reached 95% coverage between the 2,000 and 4,000 base word families and reached 98% coverage between 4,000 and 7,000 base word families. Therefore, if learners know the first 7,000 word families plus the four supplementary lists, they could theoretically comprehend the IT. However, upon further analysis of Table 3, Unit 1 and 2 had varying levels of coverage that would affect comprehension of those units.

Table 3 shows the CT reached 95% coverage between the 2,000 and 3,000 base word families and reached 98% between 3,000 and 6,000 base word families. As with Table 2, Table 3 also shows some indiscretions in coverage within the textbook. For example, the vocabulary coverage in Unit 4 will be difficult for learners, compared with that of the other 11 units in the textbook. Compared with the results of the IT however, the ranges for each unit are similar, with few outliers.

Table 2. Cumulative Coverage of the IT by Nation’s (2012) 25,000 BNC/COCA Word Lists and Supplementary Lists

<table>
<thead>
<tr>
<th>Nation (2012) BNC/COCA + supplementary lists</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
<th>Unit 6</th>
<th>Unit 7</th>
<th>Unit 8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary lists (31–34)</td>
<td>2.28</td>
<td>1.46</td>
<td>1.8</td>
<td>1.45</td>
<td>8.31</td>
<td>2.83</td>
<td>1.96</td>
<td>0.99</td>
<td>2.36</td>
</tr>
<tr>
<td>1</td>
<td>87.55</td>
<td>81.61</td>
<td>83.41</td>
<td>87.96</td>
<td>84.38</td>
<td>79.12</td>
<td>81.85</td>
<td>83.54</td>
<td>83.78</td>
</tr>
<tr>
<td>2</td>
<td>95.16</td>
<td>91.86</td>
<td>90.15</td>
<td>93.83</td>
<td>93.74</td>
<td>92.65</td>
<td>90.96</td>
<td>93.05</td>
<td>92.72</td>
</tr>
<tr>
<td>3</td>
<td>98.44</td>
<td>94.76</td>
<td>95.16</td>
<td>98.39</td>
<td>96.36</td>
<td>97.40</td>
<td>96.85</td>
<td>96.76</td>
<td>96.71</td>
</tr>
<tr>
<td>4</td>
<td>98.86</td>
<td>95.51</td>
<td>96.54</td>
<td>98.71</td>
<td>97.13</td>
<td>98.40</td>
<td>97.91</td>
<td>98.13</td>
<td>97.54</td>
</tr>
<tr>
<td>5</td>
<td>99.39</td>
<td>96.23</td>
<td>98.12</td>
<td>99.95</td>
<td>98.39</td>
<td>99.06</td>
<td>98.62</td>
<td>98.98</td>
<td>98.39</td>
</tr>
<tr>
<td>6</td>
<td>99.69</td>
<td>97.41</td>
<td>98.68</td>
<td>99.70</td>
<td>98.67</td>
<td>99.42</td>
<td>99.26</td>
<td>99.39</td>
<td>98.93</td>
</tr>
<tr>
<td>7</td>
<td>99.76</td>
<td>98.66</td>
<td>98.94</td>
<td>99.88</td>
<td>98.98</td>
<td>99.52</td>
<td>99.42</td>
<td>99.45</td>
<td>99.31</td>
</tr>
</tbody>
</table>

Note: Bolded items are coverage thresholds.
Table 3. Cumulative coverage of the Commercial Textbook by Nation's (2012) 25,000 BNC/COCA Word Lists and Supplementary Lists

<table>
<thead>
<tr>
<th>Nation (2012)</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
<th>Unit 6</th>
<th>Unit 7</th>
<th>Unit 8</th>
<th>Unit 9</th>
<th>Unit 10</th>
<th>Unit 11</th>
<th>Unit 12</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC/COCA + supplementary lists (31–34)</td>
<td>3.87</td>
<td>5.51</td>
<td>6.87</td>
<td>4.27</td>
<td>4.91</td>
<td>4.71</td>
<td>4.17</td>
<td>4.44</td>
<td>6.85</td>
<td>4.61</td>
<td>2.94</td>
<td>3.97</td>
<td>4.62</td>
</tr>
<tr>
<td>1</td>
<td>84.89</td>
<td>88.5</td>
<td>85.9</td>
<td>83.74</td>
<td>85.54</td>
<td>87.48</td>
<td>88.28</td>
<td>84.23</td>
<td>84.48</td>
<td>86.28</td>
<td>84.95</td>
<td>84.74</td>
<td>86.00</td>
</tr>
<tr>
<td>2</td>
<td>95.19</td>
<td>96.89</td>
<td>93.43</td>
<td>93.76</td>
<td>94.34</td>
<td>94.81</td>
<td>95.33</td>
<td>92.00</td>
<td>92.98</td>
<td>94.48</td>
<td>95.06</td>
<td>92.69</td>
<td>94.47</td>
</tr>
<tr>
<td>3</td>
<td>97.24</td>
<td>98.64</td>
<td>97.61</td>
<td>96.27</td>
<td>97.85</td>
<td>97.84</td>
<td>97.41</td>
<td>96.1</td>
<td>97.17</td>
<td>97.19</td>
<td>98.22</td>
<td>97.35</td>
<td>97.49</td>
</tr>
<tr>
<td>4</td>
<td>97.71</td>
<td>99.12</td>
<td>98.41</td>
<td>97.24</td>
<td>98.25</td>
<td>98.65</td>
<td>98.31</td>
<td>97.35</td>
<td>97.96</td>
<td>98.00</td>
<td>98.62</td>
<td>98.53</td>
<td>98.23</td>
</tr>
<tr>
<td>5</td>
<td>98.62</td>
<td>99.63</td>
<td>99.27</td>
<td>97.97</td>
<td>99.04</td>
<td>99.11</td>
<td>98.89</td>
<td>98.33</td>
<td>98.58</td>
<td>99.24</td>
<td>99.15</td>
<td>98.87</td>
<td>98.91</td>
</tr>
</tbody>
</table>

Note: Bolded items are coverage thresholds.
Table 4 shows the vocabulary load analysis results of the IT using the New JACET8000 (JACET, 2016). The IT reached 95% coverage between 3,000 and 6,000 and reached 98% coverage between 4,000 and 8,000. Table 4 highlights Unit 2 and Unit 5 not reaching 98% coverage. Comparing these units alongside Units 4, 6, 7, and 8 indicates that a wide range of coverage will affect comprehension for Japanese tertiary-level learners.

Table 5 shows that the CT reached 95% coverage between the 2,000 and 5,000 base word lists and reached 98% coverage between 5,000 and 8,000 base word families. As with the IT, CT Units 5, 6, 8, 11, and 12 did not reach 98% coverage.

5 Discussion

When comparing the results of Tables 2 and 3 with Tables 4 and 5, it is clear that both textbooks are lexically demanding for Japanese tertiary-level learners. If the books were solely analyzed using a general word list, such as Nation’s (2012) BNC/COCA, the results would not illustrate this problem. Therefore, the preliminary results further highlight the need for using context-specific word lists.

Ideally, the lexical coverage in each unit of a textbook should be identical or show incremental increases. This is not the case for both; with the results highlighting indiscretions within each textbook that could affect the comprehension of certain units. This, however, is somewhat unsurprising in the IT, as it was created by several people and presumably, the vocabulary profile was not analyzed. This issue can be easily fixed as the materials can be redesigned following the results of this study. However, a university using the CT will need supplementary material, such as word lists, to assist learners when studying certain units.

6 Implications for Future Research

These preliminary results point toward further analyses in several areas. Firstly, the analysis could focus on other word types (e.g., academic vocabulary) to understand the different types of words that are included in the vocabulary load of each textbook. Secondly, an analysis of the incremental vocabulary increase per unit could indicate the changes in the level of difficulty between the
Table 5. Cumulative Coverage of the CT by the New JACET8000 (JACET, 2016)

<table>
<thead>
<tr>
<th>Base word lists</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
<th>Unit 6</th>
<th>Unit 7</th>
<th>Unit 8</th>
<th>Unit 9</th>
<th>Unit 10</th>
<th>Unit 11</th>
<th>Unit 12</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proper nouns</td>
<td>6.54</td>
<td>8.53</td>
<td>7.81</td>
<td>7.17</td>
<td>8.22</td>
<td>7.88</td>
<td>8.03</td>
<td>8.05</td>
<td>11.23</td>
<td>7.83</td>
<td>7.59</td>
<td>9.20</td>
<td>8.24</td>
</tr>
<tr>
<td>1</td>
<td>85.86</td>
<td>88.26</td>
<td>87.29</td>
<td>81.76</td>
<td>84.05</td>
<td>85.75</td>
<td>86.65</td>
<td>83.75</td>
<td>83.69</td>
<td>86.70</td>
<td>83.21</td>
<td>85.92</td>
<td>85.26</td>
</tr>
<tr>
<td>2</td>
<td>93.92</td>
<td><strong>95.58</strong></td>
<td>94.56</td>
<td>91.01</td>
<td>92.81</td>
<td>92.62</td>
<td>93.48</td>
<td>91.80</td>
<td>92.48</td>
<td>94.04</td>
<td>92.71</td>
<td>92.77</td>
<td>93.17</td>
</tr>
<tr>
<td>3</td>
<td><strong>96.35</strong></td>
<td>97.16</td>
<td><strong>96.50</strong></td>
<td>94.26</td>
<td><strong>96.09</strong></td>
<td><strong>95.57</strong></td>
<td><strong>95.83</strong></td>
<td>94.51</td>
<td>94.82</td>
<td><strong>96.89</strong></td>
<td><strong>95.49</strong></td>
<td><strong>96.04</strong></td>
<td><strong>95.82</strong></td>
</tr>
<tr>
<td>4</td>
<td>97.67</td>
<td>97.93</td>
<td>97.29</td>
<td>94.96</td>
<td>97.46</td>
<td>96.55</td>
<td>97.03</td>
<td><strong>97.11</strong></td>
<td><strong>95.93</strong></td>
<td>97.93</td>
<td>96.77</td>
<td>97.26</td>
<td>97.01</td>
</tr>
<tr>
<td>5</td>
<td>97.90</td>
<td><strong>98.05</strong></td>
<td>97.81</td>
<td><strong>95.70</strong></td>
<td>97.69</td>
<td>96.91</td>
<td>97.28</td>
<td>97.40</td>
<td>96.72</td>
<td><strong>98.43</strong></td>
<td>97.25</td>
<td>97.38</td>
<td>97.40</td>
</tr>
<tr>
<td>6</td>
<td><strong>98.09</strong></td>
<td>98.16</td>
<td><strong>98.15</strong></td>
<td>96.11</td>
<td>97.99</td>
<td>97.13</td>
<td>97.56</td>
<td>97.73</td>
<td>97.07</td>
<td>98.49</td>
<td>97.46</td>
<td>97.56</td>
<td>97.65</td>
</tr>
<tr>
<td>7</td>
<td>98.19</td>
<td>98.28</td>
<td>98.24</td>
<td>96.49</td>
<td>98.25</td>
<td>97.38</td>
<td><strong>98.21</strong></td>
<td>97.82</td>
<td>97.72</td>
<td>98.63</td>
<td>97.73</td>
<td>97.77</td>
<td>97.92</td>
</tr>
<tr>
<td>8</td>
<td>98.36</td>
<td>98.42</td>
<td>98.42</td>
<td>97.28</td>
<td>98.55</td>
<td>97.86</td>
<td>98.24</td>
<td>97.96</td>
<td>97.96</td>
<td>98.75</td>
<td>97.79</td>
<td>97.83</td>
<td><strong>98.15</strong></td>
</tr>
</tbody>
</table>

Note: Bolded items are coverage thresholds.
textbooks. These results could show if each unit increase is appropriate for learners using the textbook. Finally, using these results, supplementary materials could be designed to scaffold vocabulary learning for learners. In turn, these materials can help bridge any potential learning gaps between the textbook vocabulary and learners’ vocabulary knowledge.

7 Conclusion

This article presented the main vocabulary load findings from an IT and CT using a general and Japanese-specific word list. The initial results show that learners have higher lexical demands when considering vocabulary knowledge in JACET8000 compared with the BNC/COCA 25,000 general word list. These preliminary results will be analyzed further to determine the type of words, word families and lemmas used by skill and by unit in each textbook. Although textbooks are useful guidance in helping learners in second language classrooms, there is some evidence that there may be a discrepancy between the textbook context and learners’ vocabulary knowledge. Therefore, further analysis is needed to identify the textbook’s vocabulary load so that teachers can assist learners in reaching the vocabulary coverage.

References


Benson and Madarbakus-Ring: A comparison of textbook vocabulary load analysis


The Contribution of High-frequency Multi-word Sequences to Speech Rate and Listening Perception Among EFL Learners

Michael McGuire and Jenifer Larson-Hall

Doshisha University; University of Kitakyushu

Abstract

This experiment tested gains in spoken fluency and ability to complete a dictation listening task accurately among 33 Japanese L1 English language users. Both a control group (N = 17) and an experimental group (N = 16) studied Anki vocabulary cards each week for 10 weeks and described three picture stories that contained the vocabulary words every week. Both groups studied 10 common bigrams (such as take advantage) each week while the experimental group additionally studied sets of 10 reduced trigrams (how do you) and did narrow listening homework each week. The results for spoken storytelling fluency found a large advantage for the experimental group while fluency for the free speaking task showed a medium advantage for the experimental group that was not statistically significant. For the listening dictation task, both groups reduced their errors from pretest to posttest but neither group was statistically different from each other.

1 Introduction

Rather than building every sentence one word at a time, fluent speakers of a language assemble much of their speech from prefabricated, or formulaic, multi-word chunks of language (Wood, 2010). There are many terms used to describe this type of language and we have chosen the term multi-word sequences (MWS) for this article. Through repetition, the spoken production of high-frequency MWSs becomes automated, which allows for fluent speech with less cognitive load (Bybee, 2002a), and the proceduralization of these strings “could allow expression to occur fluently under the constraints of time which real-life speech entails” (Wood, 2010, p. 2).

It follows that language learners might benefit from learning and using more MWSs as well. We examine two types of frequent MWSs in our experiment: (1) Bigrams such as high probability or take advantage; and (2) The most frequent trigrams in spoken English, phrases such as and then I or how do you, which we call “reduced MWSs.” These phrases do not have much independent lexical content, and we label them as “reduced” since considerable phonological changes occur when they are uttered (Bybee, 2002b; Bybee & Hopper, 2001) (consider, for example, the reduction of how do you in fluent speech to [hawdʒə]). We are targeting these two types of MWSs because we believe their sheer frequency makes them valuable.
Previous research on the bigrams used here (Nguyen & Webb, 2017) found that English majors at a Vietnamese university did not have strong knowledge of the collocations in spite of the fact that all of them were composed from the 3,000 most frequent words of English (mean score 45% on average). The study did not try to teach participants the collocations; it simply tested their knowledge of them.

Although previous studies in the SLA literature have demonstrated that a focus on identifying and heightening awareness of the importance of formulaic sequences can lead to improved fluency measures in English users (McGuire & Larson-Hall, 2017; Wood, 2009), we only know of one study which targeted specific MWSs and then measured fluency on a speaking task. Thomson (2017) used a “fluency workshop” approach similar to Wood (2009) over 6 weeks and targeted 4-gram combinations like what would you like and I think I will. For productive knowledge of the 30 targeted MWSs in this study, Thomson found a marked difference for the experimental group over the control group (Cohen’s $d = 1.41$), which was perhaps to be expected given the control group had not seen the MWSs that the experimental group had seen in their materials. However, for productive fluency, Thomson used a restaurant role play of ordering and in speech rate the groups did not differ statistically, and effect size was negligible (eta-squared = 0.02). Examining the number of MWSs used the groups also did not differ but there was a small effect size (Cohen’s $d = 0.41$). Although this study was very carefully planned and carried out, the use of a very basic type of ritualized scenario for the pretest/posttest may have muted the effects of the experiment. The current study aims to improve on Thomson (2017) by using a more challenging task for testing (storytelling), adding a spontaneous task, and using a longer time frame for acquisition.

Considering our “reduced MWSs” category, previous studies which have looked at such collections of frequently occurring words have found that phonological reduction in these sequences is difficult for language learners. Henrichsen (1984) had 65 L2 students listen to recordings of English speech both with the presence and absence of connected speech processes (CSPs) and transcribe what they heard. Compared to native speakers, L2 listeners saw a large drop in accuracy when transcribing recordings that featured CSPs regardless of learners’ proficiency level.

To the best of our knowledge no major studies have examined the effect of reduced MWSs on spoken fluency. The present study combines previous work on teaching formulaic language along with a frequency-based corpus approach. We asked students to use flashcards to study 88 of the highest-frequency MWSs (which contain CSPs) and 100 bigrams involving high-frequency words. Given the basic finding in SLA that output encourages the acquisition of language (Swain, 1993), we designed a treatment that incorporated the production of the same MWSs weekly by describing picture stories. We hoped that this would encourage the participants to use the phrases in guided production and also produce MWSs spontaneously in unplanned speech, leading to greater overall fluency in their speech rate. This focus led us to the two main research questions:
1. Do students who learn and practice using common bigrams and high frequency reduced trigrams make improvements in overall speech rate when asked to describe picture stories incorporating the MWSs and when doing free speech?
2. Does this also improve listening perception of casual fluent speech that uses these formulaic sequences?

2 Methods

2.1 Participants

The study took place over 10 weeks in the spring semester with Japanese university students majoring in English at two universities. Because of attrition or failing to complete study assignments, a number of students either dropped out of the study or were dropped from the study. Two more students who completed all requirements were dropped from the study because they did not make any errors on the MWS portion of the dictation test, leaving no room for improvement. The number of participants in each group ended up being Control = 17 and Experimental = 16.

All study participants studied 10 collocations every week using the Anki flashcard program and described or listened to their classmates describing three picture stories every week that incorporated the collocations. The experimental group received practice with the reduced MWSs by (1) receiving 20 Anki flashcards per week with the 10 reduced MWSs on them; (2) watching a 5-minute introduction of the week’s MWSs and their CSPs; and (3) doing at least 15 minutes of narrow listening homework.

2.2 Selection of target language

- Collocations
  The 100 target collocations used in this study are found in Nguyen and Webb (2017) and contain randomly selected collocations of two-word adjective-noun and verb-noun collocations using only words from the first 3,000 most frequent words in English.

- Reduced MWSs
  The list of reduced MWSs used in this study (McGuire, 2020) was created for classroom use over the course of a semester in 10 weekly lessons after examining and categorizing the most frequent 3- and 4-gram MWSs in the spoken section of the Open American National Corpus (Fillmore et al., 1998). The MWSs that appear on the list are predominantly made up of function words and are not as psycholinguistically salient as fixed expressions or idioms, so students are not expected to memorize them like vocabulary. Rather, the list is intended to give students extensive practice with the most common MWS variations that feature CSPs.
2.3 Procedures

All participants loaded their list of MWSs every week into their Anki flashcard system (see Appendix A). Both groups had 10 collocations while the experimental group also had 10 reduced MWSs. All students were asked to study the cards at least three times a week. For the collocations, the Anki cards showed an English collocation, sentence, and audio on one side with a Japanese translation of the phrase and sentence on the other side. For the reduced MWSs, the cards showed the text on one side and an audio recording of the reduced MWS on the back. The participants were encouraged to repeat and shadow the reduced MWSs aloud when they turned over the card to the audio version in order to become familiar with the reduced pronunciation.

For the reduced MWS activities, the experimental group completed “Narrow listening” (Krashen, 1996) homework by choosing 15 minutes of YouTube videos of a person engaged in conversation and counted how many times they heard the weekly “reduced MWSs” in their listening. This activity allows students to become familiar with content and their selected person’s voice but left them free to revisit recordings in order to notice different linguistic features (Chang, 2017). Both groups saw three pictures per week that collectively incorporated all 10 weekly collocations. All stories contained a written description with the collocations included. The participants were given a few minutes to read the description previous to then looking at a version of the picture story without the written description. The students were asked to describe the picture in their own words but to not forget the collocations. Below is the story description that students read, while Figure 1 shows the page without the description but with the prompt words on it that participants used while telling the story.

A man is starting out building a new business. It is a fast-food restaurant called “Buddy’s”. He wants to do business in a part of town that used to have a lot of serious crime. However, in the past decade this area has been getting nicer. He reduces his risk of crime by installing an alarm system.

The participants took five tests before and after the experiment. Since we were concerned with fluency and listening perception, the two main tests we report here are two speaking tasks and a listening dictation. The pretest/posttest storytelling speaking task involved three multi-panel picture stories different from any used during the experiment. Similar to the practice conditions each week, the participants had 2 minutes to read a description of the story and then had to tell the story in their own words. Different from the weekly practice stories, there were no collocations listed on the story-less pictures, but there were some prompt words helpful for describing the story, possibly including one word of a collocation (for example, in one story which described a bank robbery, the collocation serious crime was not listed but crime was). A second speaking task asked the participants to freely describe what they would do in their ideal day. The participants had 1 minute to speak spontaneously. Audio recordings of these tasks were analyzed using the Syllable Nuclei v2 script (de Jong & Wempe, 2009) using PRAAT software (Boersma & Weenink, 2021) to determine speech rate (number of syllables divided by duration) as a quantitative measure of fluency.
In preparing these speech samples for analysis, the stories and free speaking were cut out of the recording stream when the participant began speaking and ended after they finished speaking. In some cases, there were interruptions in the storytelling and these interruptions were removed so that the monologue would sound natural.

The listening perception task was a dictation task consisting of a naturalistic conversation that participants wrote down word for word. It contained 10 collocations and 27 reduced MWSs. The participants heard the taped conversation twice at a normal native-speaker speed.

3 Results

For statistical reporting unless otherwise noted all of the data was found to be normally distributed using the Shapiro-Wilk normality test where the alternative hypothesis is that the distribution is not normal ($p \geq 0.05$; normality accepted) (see Larson-Hall, 2016, pp. 109–110). We declare an alpha level of 0.05 but focus discussion on the corresponding 95% confidence intervals (CIs) and effect sizes.

3.1 Storytelling task

Table 1 gives statistics for the storytelling speaking task. The number for each participant is a composite average of their speech rate from all three stories.
The speech rate measure divides the number of syllables by the duration (in seconds) of the speaking activity, so it includes the speed of speech as well as pauses and breakdowns; this measure is most similar to what one would think of as general fluency (Tavakoli et al., 2020). Both groups started with nearly identical speech rates, but the control group’s speech rate declined slightly from pretest to posttest while the experimental group’s speech rate increased. The statistics in the “Gain Score” column compare one group’s pre-test and post-test scores using the effect size for a paired-samples $t$-test. The statistics in the “Comparison” column compare the gain scores between the control and experimental group using an independent samples $t$-test.

Figure 2 shows a parallel coordinate plot for the speech rate of the control group and the experimental group separately. This plot charts the path from pretest to posttest with a separate line for each participant while the larger black line shows the average for the group.

For speech rate, the control group’s data was not normally distributed (Shapiro-Wilks $p = 0.046$) so Welch’s $t$-test was used. The control group’s decline was noticeable and the 95% confidence interval [0.10, 0.41] shows that the decline in the

Table 1. Statistics for the Storytelling Task

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th>Posttest</th>
<th>Gain score</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>1.66 (0.46)</td>
<td>1.41 (0.39)</td>
<td>0.25 (0.3)</td>
<td>95% CI: [0.10, 0.41]</td>
</tr>
<tr>
<td>(N = 17)</td>
<td></td>
<td></td>
<td>95% CI:</td>
<td>Welch's $t$-test: $t = -3.80$, $df = 30.6$, $p = 0.0006$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = 0.85$</td>
<td>95% CI: [0.28, 1.40]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$95% CI$:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[0.01, 0.41]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = 0.49$</td>
<td>95% CI: [0.04, 1.00]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$95% CI$:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[0.32, -0.01]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = 0.49$, 95%CI:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[0.32, -0.01]</td>
<td></td>
</tr>
<tr>
<td>Experiment</td>
<td>1.67 (0.43)</td>
<td>1.82 (0.39)</td>
<td>0.15 (0.31)</td>
<td>95% CI: [0.32, -0.01]</td>
</tr>
<tr>
<td>(N = 16)</td>
<td></td>
<td></td>
<td>95% CI:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = 0.49$, 95%CI:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[0.04, 1.00]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$95% CI$:</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[0.32, -0.01]</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Figures for the Storytelling Task Speech Rate Separated by Group.
population, which was -0.25 points in this sample, might be as small as 0.10 points or as large as 0.41 points. A Cohen's $d$ effect size for this decline is medium-to-large sized at $d = 0.85$ (but could be as small as 0.28). For the experimental group, the increase in speech rate is not statistical as the 95% CI passes through zero [0.32, -0.01]; the effect size is also small-to-medium ($d = 0.49$ on average but could be as little as zero). Comparing the gains of the two groups to each other using an independent samples $t$-test, the groups are statistically different with a possible difference in speech rate, with 95% confidence, from as little as 0.19 to as much as 0.63 faster. The effect size can be labelled large at $d = 1.3$, although the true effect size in the population might be as small as $d = 0.48$, which would only be a small-to-medium effect.

In summary, there were solid gains in speech rate in the experimental group, meaning that the time spent on learning collocations and reduced MWSs seemed to promote speaking fluency.

3.2 Free Speech Task

Table 2 gives statistics for the free speech task in which participants talked about what they would do on their ideal day (parallel coordinate plot in Figure 3).

Again, the control group slightly declined in their speech rate from pretest to posttest and the experimental group increased. But in both cases, the 95% CI goes through zero so neither change is statistical. However, the experimental group has a larger effect size than the control group. Statistically there is no difference between the control and experimental group, but the effect size shows that the difference between groups could be nothing or as large as $d = 1.4$. A larger number of participants could help narrow this confidence interval.

3.3 Dictation Listening Task

Table 3 shows descriptive statistics for errors on the dictation task. An error was any word that was incorrect or the absence of a word when one was needed. There were 177 words in the test, but some participants wrote in extra incorrect words, so the total number of errors is theoretically larger than 177. We examined the total number of errors (Figure 4) as well as errors due to reduced MWSs only (Figure 5).

Table 2. Statistics for the Free Speech Task.

<table>
<thead>
<tr>
<th></th>
<th>Pretest</th>
<th>Posttest</th>
<th>Gain score</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control ($N = 17$)</td>
<td>1.93 (0.45)</td>
<td>1.87 (0.42)</td>
<td>-0.06 (0.36)</td>
<td>Mean difference: 0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>95% CI: [-0.25, 0.12]</td>
<td>95% CI: [–0.01, 0.54]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$d = -0.18$</td>
<td>$d = 0.70$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>95% CI: [-0.65, 0.13]</td>
<td>95% CI: [-0.04, 1.42]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t = -2.00$, df = 29.8</td>
<td>$t = -2.00$, df = 29.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$p = 0.054$</td>
<td>$p = 0.054$</td>
</tr>
<tr>
<td>Experiment ($N = 16$)</td>
<td>1.84 (0.46)</td>
<td>2.05 (0.43)</td>
<td>0.21 (0.41)</td>
<td>$d = 0.50$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>95% CI: [-0.01, 0.42]</td>
<td>95% CI: [-0.03, 1.02]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$t = -2.00$, df = 28.8</td>
<td>(data skewed)</td>
</tr>
</tbody>
</table>

Vocabulary Learning and Instruction, 10(2), 18–29.
The descriptive statistics show that both groups improved by reducing their number of errors from the pretest to the posttest. Paired \( t \)-tests for each type of reduction in errors within each group (tests 1–4 in Table 4) were statistically significant and effect sizes were similar, as can be seen in Table 4. The experimental group declined in errors in reduced MWSs by a larger numerical amount than the control group, but the difference was not large enough to find statistical differences between the two groups when it came to gains in reducing errors in either the entire test or specifically just in MWSs (see independent sample \( t \)-test results 5–6 in Table 4). The average effect size for differences in the reduction of MWS errors was medium, and larger than the small average effect size for differences in the reduction of total errors.

It appears that both the control and experimental group were able to do much better on the dictation task by the end of the semester, thus reducing their errors both overall and specifically on MWSs by a large amount (looking at the average Cohen’s \( d \) for pretest versus posttest). This area does not show that the extra work done by the experimental group provided any important results.
Figure 4. Figures for the Listening Dictation Task (All Errors) Separated by Group.

Figure 5. Figures for the Listening Dictation Task (Only MWS Errors) Separated by Group.

Table 4. Inferential Tests for the Dictation Task

<table>
<thead>
<tr>
<th>Pretest versus posttest</th>
<th>t</th>
<th>Df</th>
<th>p</th>
<th>95% CI</th>
<th>Cohen's d</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Total errors Control</td>
<td>6.86</td>
<td>16</td>
<td>&lt; 0.0001</td>
<td>[10.69, 20.25]</td>
<td>1.66 [0.91, 2.40]</td>
</tr>
<tr>
<td>(2) Total errors Experimental</td>
<td>6.18</td>
<td>15</td>
<td>&lt; 0.0001</td>
<td>[12.12, 24.88]</td>
<td>1.55 [0.80, 2.27]</td>
</tr>
<tr>
<td>(3) MWS errors Control</td>
<td>5.95</td>
<td>16</td>
<td>&lt; 0.0001</td>
<td>[4.77, 10.05]</td>
<td>1.44 [0.75, 2.12]</td>
</tr>
<tr>
<td>(4) MWS errors Experimental</td>
<td>6.18</td>
<td>15</td>
<td>&lt; 0.0001</td>
<td>[7.53, 15.47]</td>
<td>1.54 [0.80, 2.27]</td>
</tr>
<tr>
<td>Control versus Experimental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) Total errors</td>
<td>0.82</td>
<td>31</td>
<td>= 0.42</td>
<td>[-4.55, 10.61]</td>
<td>0.28 [-0.41, 0.97]</td>
</tr>
<tr>
<td>(6) MWS errors</td>
<td>1.85</td>
<td>31</td>
<td>= 0.08</td>
<td>[-0.43, 8.61]</td>
<td>0.64 [-0.08, 1.35]</td>
</tr>
</tbody>
</table>
4 Discussion

We explored the question of whether targeted attention to reduced MWSs would improve our Japanese students’ speaking fluency so that their speech rate would increase when telling picture stories and speaking spontaneously. All participants memorized bigrams and practiced using them while telling stories in class, so it was expected that all participants might improve in their speaking fluency, but with special attention paid to very frequent 3-gram we thought the experimental group might improve even more. For speaking rate, we found that the control group declined in speech rate while the experimental group increased. Although the difference between groups was only statistical for the story task and not the free speaking task, effect sizes seemed to indicate that the experimental group did benefit in their speech rate from the additional work with reduced MWSs. We also expected the approximately 15 minutes of homework per week for the experimental class would have considerably improved the experimental group’s performance on a listening dictation test, especially when looking at errors due to reduced MWSs only. However, we did not find increased improvement for the experimental group. We plan to continue gathering more data next year using the same experiment in order to gain more power to examine the question of whether speaking fluency and listening ability can be enhanced through a study of reduced MWSs.

References


## Appendix

### Appendix A. Targeted Formulaic Sequences Used in Experiment.

<table>
<thead>
<tr>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
<th>Week 4</th>
<th>Week 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collocations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high drama</td>
<td>low income</td>
<td>fast food</td>
<td>save energy</td>
<td>difficult situation</td>
</tr>
<tr>
<td>large number</td>
<td>improve quality</td>
<td>possible exception</td>
<td>inner city</td>
<td>catch fish</td>
</tr>
<tr>
<td>work hard</td>
<td>main road</td>
<td>past decade</td>
<td>major concern</td>
<td>quick glance</td>
</tr>
<tr>
<td>long tradition</td>
<td>special occasion</td>
<td>public confidence</td>
<td>big surprise</td>
<td>change direction</td>
</tr>
<tr>
<td>strong desire</td>
<td>central figure</td>
<td>serious crime</td>
<td>take pleasure</td>
<td>free speech</td>
</tr>
<tr>
<td>make money</td>
<td>spend time</td>
<td>take pride</td>
<td>feel pain</td>
<td>take place</td>
</tr>
<tr>
<td>lose hope</td>
<td>give birth</td>
<td>express concern</td>
<td>lose weight</td>
<td>give voice</td>
</tr>
<tr>
<td>enter college</td>
<td>real life</td>
<td>take comfort</td>
<td>take notice</td>
<td>watch television</td>
</tr>
<tr>
<td>take advantage</td>
<td>pay attention</td>
<td>do business</td>
<td>make music</td>
<td>make progress</td>
</tr>
<tr>
<td>whole family</td>
<td>do research</td>
<td>reduce risk</td>
<td>build muscle</td>
<td>good reason</td>
</tr>
</tbody>
</table>

### Reduced MWSs

<table>
<thead>
<tr>
<th>a lot of</th>
<th>one of the</th>
<th>some of the</th>
<th>a couple of</th>
<th>that kind of*</th>
<th>a lot of people</th>
<th>kind of a</th>
<th>what kind of</th>
<th>out of the</th>
<th>lot of the</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don’t know</td>
<td>you know I</td>
<td>I think that*</td>
<td>I think that*</td>
<td>you know that*</td>
<td>you think they</td>
<td>I think the</td>
<td>that kind of*</td>
<td>but I think</td>
<td>I know that</td>
</tr>
<tr>
<td>you know you</td>
<td>I think it's</td>
<td>I think that's*</td>
<td>things like that</td>
<td></td>
<td>something like that</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you know</td>
<td>I think it's</td>
<td>I think that's*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you know the</td>
<td>I think it's</td>
<td>I think that's*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you know</td>
<td>I think it's</td>
<td>I think that's*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you know</td>
<td>I think it's</td>
<td>I think that's*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I don’t</td>
<td>you know</td>
<td>I think you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I don’t</td>
<td>you know</td>
<td>I think you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I don’t</td>
<td>you know</td>
<td>I think you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I don’t</td>
<td>you know</td>
<td>I think you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I don’t</td>
<td>you know</td>
<td>I think you</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Week 6</th>
<th>Week 7</th>
<th>Week 8</th>
<th>Week 9</th>
<th>Week 10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collocations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>become law</td>
<td>smell sweet</td>
<td>general population</td>
<td>important component</td>
<td>natural disaster</td>
</tr>
<tr>
<td>favorite subject</td>
<td>human nature</td>
<td>important feature</td>
<td>significant decrease</td>
<td>declare victory</td>
</tr>
<tr>
<td>small talk</td>
<td>open space</td>
<td>dark cloud</td>
<td>red cross</td>
<td>provide relief</td>
</tr>
<tr>
<td>natural beauty</td>
<td>wide range</td>
<td>top secret</td>
<td>private sector</td>
<td>have access</td>
</tr>
<tr>
<td>senior citizen</td>
<td>perfect match</td>
<td>gain knowledge</td>
<td>send word</td>
<td>lose faith</td>
</tr>
<tr>
<td>direct traffic</td>
<td>lose sight of</td>
<td>make noise</td>
<td>prime minister</td>
<td>take refuge</td>
</tr>
<tr>
<td>heavy rain</td>
<td>take note</td>
<td>commit murder</td>
<td>good fortune</td>
<td>gain admission</td>
</tr>
<tr>
<td>take issue</td>
<td>make sense</td>
<td>long silence</td>
<td>pay cash</td>
<td>feel sympathy</td>
</tr>
<tr>
<td>take flight</td>
<td>hard copy</td>
<td>high degree</td>
<td>basic rule</td>
<td>collect data</td>
</tr>
<tr>
<td>feel pressure</td>
<td>mass media</td>
<td>use technology</td>
<td>long term</td>
<td>give priority</td>
</tr>
</tbody>
</table>

### Reduced MWSs

<table>
<thead>
<tr>
<th>and you know</th>
<th>and it was*</th>
<th>what do you</th>
<th>to do it</th>
</tr>
</thead>
<tbody>
<tr>
<td>you have to</td>
<td>and it was*</td>
<td>what do you</td>
<td>to do it</td>
</tr>
<tr>
<td>and I think*</td>
<td>I have a</td>
<td>when I was</td>
<td>do you have*</td>
</tr>
<tr>
<td>and it was*</td>
<td>do you have*</td>
<td>it was a</td>
<td>do you think</td>
</tr>
<tr>
<td>and I don’t</td>
<td>we have a</td>
<td>it was just</td>
<td>do you do</td>
</tr>
<tr>
<td>and then I</td>
<td>I don’t have*</td>
<td>and I was*</td>
<td>how do you</td>
</tr>
<tr>
<td>and things like that</td>
<td>you have a</td>
<td>I was in</td>
<td>do you like</td>
</tr>
<tr>
<td>and I was*</td>
<td>to have a*</td>
<td>and that was</td>
<td>do you know</td>
</tr>
<tr>
<td>and all that</td>
<td>have a lot</td>
<td>there was a</td>
<td>do you feel</td>
</tr>
<tr>
<td>and so I</td>
<td>I have to*</td>
<td>was kind of</td>
<td>or do you</td>
</tr>
<tr>
<td>and then you</td>
<td>they have a</td>
<td>that was a</td>
<td>well do you</td>
</tr>
</tbody>
</table>

*Reduced MWSs that are repeated in multiple weeks.*
Comparisons of Word Lists on New Word Level Checker

Atsushi Mizumoto\textsuperscript{a}, Geoffrey G. Pinchbeck\textsuperscript{b}, and Stuart McLean\textsuperscript{c}
\textsuperscript{a}Kansai University; \textsuperscript{b}Carleton University; \textsuperscript{c}Momoyama Gakuin University

Abstract

This paper introduces a novel online vocabulary profiling application called the New Word Level Checker (https://nwlc.pythonanywhere.com/) and word list resources used by the application. First, the rationale for developing another web vocabulary profiler and the word lists included in the application are described. Next, the lexical units (i.e., how words are counted) and rules (e.g., case sensitivity, contractions, abbreviations with periods, hyphenated words, and compounds) employed in the application are explained. Then, the word lists adopted for the application are compared to show which lists are best used for different purposes. Pedagogical implications of the use of the application and word lists are discussed, especially focusing on matching learners with vocabulary-level appropriate tests.

Keywords: word list, word counting unit, vocabulary profiling, tool development, web application

1 Introduction

New Word Level Checker (NWLC) (https://nwlc.pythonanywhere.com/) is a web application for vocabulary profiling. Using the core concept of its predecessor, Word Level Checker (http://someya-net.com/wlc/), developed by Someya (2006), NWLC analyzes English texts submitted by users and produces a coverage profile based on the built-in, user-selected word lists.

NWLC is an online application specifically created to provide vocabulary profiling that has been optimized for Japanese English as a Foreign Language (EFL) learning contexts. It was necessary to develop NWLC because word counting methods in the word lists adopted in other vocabulary profiling applications such as VocabProfiler (https://www.lextutor.ca/vp/comp/) are not appropriate for Japanese EFL learners (e.g., McLean, 2018). Furthermore, vocabulary lists tailored for Japanese EFL learners are not included in other profilers. Another feature of NWLC is that it includes a knowledge-based word list, Scale of English Word Knowledge—Japanese (SEWK-J) (see below), in addition to conventional frequency-based word lists. Thus, it is possible, with NWLC, to match learners’ actual vocabulary knowledge with texts to be analyzed.

2 Word Lists Included in NWLC

As of September 2021, NWLC features five research-based, trustworthy word lists: SEWK-J, New JACET8000, SVL12000, the New General Service List...
It is advantageous to have these five word lists because each is built on different principles, and the number of words in the list differs from list to list. Users can select and compare the word lists depending on their purposes.

3 SEWK-J

The SEWK-J was developed to estimate the difficulty that the vocabulary in a text presents to Japanese learners of English. Thus, the SEWK-J list estimates the likelihood that a word is known to Japanese university students. The probability of knowledge of a word is based on a multiple regression performed by Pinchbeck (manuscript in preparation) using vocabulary test data of the 149-item New Vocabulary Levels Test (McLean & Kramer, 2016) administered to Japanese University EFL students as the criterion (dependent) variable. The regression formula includes the following predictive variables to provide estimates for about 75,000 lemma headwords:

1. English L2 vocabulary yes/no test data—Accuracy
2. English L2 vocabulary yes/no test data—Reaction time
3. English-Japanese loan words—Identity
5. Age of acquisition
6. Age of exposure
7. Word frequency in a large general corpus of English

4 New JACET8000

The New JACET List of Basic Words (New JACET8000) (JACET, 2016) is the updated version of JACET8000 (JACET, 2003), compiled by the Japan Association of College English Teachers (JACET). Based on the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), New JACET8000 serves as an educational word list for Japanese learners of English, especially university students. The list has 8,000 words, and for each 1,000 words, the level (i.e., from 1 to 8) is provided by the NWLC profiler. The New JACET8000 is available from Dr. Shin Ishikawa’s website (http://language.sakura.ne.jp/s/voc.html).

5 SVL12000

SVL12000 (Standard Vocabulary List 12000) was developed and published in 2001 by ALC Press, Inc. It is based on the BNC. Like New JACET8000, the word list is intended to be used for educational purposes, and many ALC-published materials use this list. As the name shows, the list has 12,000 words and can be divided into 12 levels for each 1,000-word band.

6 New General Service List

The NGSL was conceived as a modern update of the General Service List (West, 1953). It was created from the Cambridge English Corpus (CEC), and it
covers about 92% of general texts of English with a list of approximately 2,800 high frequency words (Browne et al., 2013). The NGSL is optimal for Japanese EFL learners because they learn approximately 3,000 words through junior and senior high school textbooks. Browne and his colleagues subsequently produced a secondary series of word lists for learners who have mastered the first 2,801 words with NGSL. These include the New Academic Word List (NAWL: 963 words), the TOEIC Service List (TSL: 1,259 words), and the Business Service List (BSL: 1,754 words).

The secondary word lists (i.e., NAWL, TSL, and BSL) exclude all words on the NGSL, but some words appear in more than one secondary list. For example, the word “impact” is in all three lists (NAWL, TSL, and BSL). Also, while the word “quit” is in both TSL and BSL, the word “syndicate” only appears in BSL. For this reason, NWLC considers the overlapping of words in the three lists and produces the output using the criteria shown below. In total, NWLC has 5,621 words for word profiling.

- Level 1: NGSL = 2,801 words
- Level 2: NAWL and TOEIC and BSL = 183 words
- Level 3: NAWL and TOEIC, NAWL and BSL, or TOEIC and BSL = 790 words
- Level 4: Only in NAWL, TOEIC, or BSL = 1,847 words

7 CEFR-J Wordlist

CEFR-J Wordlist was based on an English textbook corpus consisting of textbooks for primary and secondary schools in China, Korea, and Taiwan (Tono, 2019). The word levels were classified according to the CEFR levels, and they were aligned with the English Vocabulary Profile (https://www.englishprofile.org/wordlists). All headwords, as a result, have part-of-speech information and their corresponding CEFR(-J) levels as shown in Table 1. NWLC uses the CEFR-J Wordlist Version 1.5.

8 Word Counting Units and Rules

8.1 Proper Nouns and Numerals

In the NWLC, proper nouns and numerals (numbers) are first identified using an open-source Part of Speech tagger, spaCy (https://spacy.io/) and are then

Table 1. CEFR-J Level, Number of Headwords and School Levels in Japan

<table>
<thead>
<tr>
<th>CEFR-J Level</th>
<th>Number of Headwords</th>
<th>School Level in Japan</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1,164</td>
<td>Elementary</td>
</tr>
<tr>
<td>A2</td>
<td>1,411</td>
<td>Junior High</td>
</tr>
<tr>
<td>B1</td>
<td>2,446</td>
<td>Senior High</td>
</tr>
<tr>
<td>B2</td>
<td>2,778</td>
<td>University</td>
</tr>
<tr>
<td>Total</td>
<td>7,799</td>
<td></td>
</tr>
</tbody>
</table>

Vocabulary Learning and Instruction, 10(2), 30–41.
treated as possibly *known* words because they can be assumed to be understood by learners. The possessive’s (e.g., Todd’s dog) is also put into the *known* category in word profiler output. For the remaining words in the text, the following lemmatization (tokenization) rules are applied.

### 8.2 Lemmatization

A variety of word-counting methods have been used in previous studies (for a review, see McLean, 2018). In some word counting methods, “happy,” “happily,” “happiness,” and “unhappy” can be counted as one word family, with the headword “happy” (e.g., Nation, 2012). Four of the five word lists in NWLC adapt flemma counting (i.e., a base form as a headword and its inflected forms as a one word). For example, for the headword “study,” the following word forms are included and counted as one word: study, studies, studied, and studying. The flemma—a portmanteau of “family” and “lemma”—was first introduced by Pinchbeck (2014, 2017) to distinguish between word lists that include part of speech (POS) information and those that do not and is a recommended word counting method in the field of applied linguistics and in EFL teaching contexts (see McLean, 2018). Flemma counting combines inflections of lemma groups but does not distinguish the POS. That is, with flemma counting, the verb “study” and the noun “study” are both counted under the same headword “study.”

The distinction in the terms *flemma* and *lemma* is often not always made. For example, a type of resource used by the AntConc software that is termed a “lemma list” has no requirement for POS tag information. In fact, the resource labeled as, “lemma list,” as used in the NWLC (see below) is all based on *flemma*-grouped word lists.

In contrast to the family and flemma grouping methods, lemma counting can detect the POS differences. The CEFR-J Wordlist adopts lemma counting since the original CEFR word lists are also lemmatized. Thus, in the CEFR word lists, the *verb* “study” is classified as A1 and the *noun* “study” is A2.

When a user selects the New JACET8000 or the SVL12000, NWLC uses the flemma list, *AntBNC Lemma List* (https://www.laurenceanthony.net/software/antconc/), which is based on all words in the BNC corpus with a frequency greater than two, for flemmatization. Modifications were manually made to match the headwords of New JACET8000 and SVL12000. For example, the words “interesting” and “interested” are listed as two headwords in both New JACET8000 and SVL12000, so they were excluded from the flemma entry “interest” in these lists (i.e., interest = interest, interests). In addition, words with British spellings in New JACET8000 and SVL12000 are included in the revised flemma list (e.g., advertise = advertise, advertises, advertised, advertising, advertize, advertizes, advertized, advertizing).

For the NGSL, flemmatization is simple because all NGSL lists are provided as flemma groupings. NWLC uses the flemma lists available at the NGSL website. For the CEFR-J Wordlist, NWLC utilizes spaCy to assign a POS and its lemma form. For SEWK-J, NWLC refers to a flemma list developed by Pinchbeck (manuscript in preparation).
Figure 1 shows a flowchart representing the algorithm of (f)lemmatization in NWLC. As can be seen, the application does its best to provide, by referring to the selected word list, the vocabulary profile of the text uploaded by the users.

### 8.3 Capitalized Letters

As the headwords in New JACET8000, SVL12000, and CEFR-J Wordlists include capital letters (Table 2), they are treated as they are (e.g., “I” not “i”). However, when the NGSL or the SEWK-J is used, NWLC treats all words as lowercase because all headwords in these lists are lowercase. In other words, the New JACET8000, SVL12000, and CEFR-J profilers are case sensitive, whereas those using the NGSL and SEWK-J are case insensitive.
8.4 Contracted Forms

NWLC detects the contracted forms by using spaCy and reverts them to the uncontracted forms (e.g., I’m => I am). Note that, as contracted forms such as “s” as in “he’s” (he has, he is) and “d” as in “we’d” (we had or we would) are not distinguished by the spaCy tagger, NWLC regards “s” and “d” as they are. We consider this lack of precision acceptable because the headwords, “be,” “is,” “have,” “has,” “would,” etc. are all among the most frequent words. For users who need to accurately distinguish between these contracted forms, the results may need to be checked carefully.

The SVL12000, NGSL, and CEFR-J contain words with apostrophes (Table 3). If those words are in the input text, NWLC treats them as they are.

8.5 Abbreviations with Periods

The SVL12000 and CEFR-J include abbreviations that use periods. If those word lists are selected and the input text has those words in Table 4, NWLC treats them as they are.

Table 2. List of Words with Capitalized Letters

<table>
<thead>
<tr>
<th>New JACET8000 (20 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>April, August, December, English, February, Friday, I, January, Japanese, July, June, Monday, November, October, Saturday, September, Sunday, Thursday, Tuesday, Wednesday</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SVL12000 (44 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>April, August, Bible, Buddhism, Buddhist, Catholic, Christ, Christian, Christianity, Christmas, December, Easter, Fahrenheit, February, Friday, God, I, Islam, January, July, June, March, May, Messrs., Miss, Monday, Mr., Mrs., Ms., Muslim, November, OK, October, Protestant, Renaissance, Satan, Saturday, September, Shinto, Sunday, Thanksgiving, Thursday, Tuesday, Wednesday</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CEFR-J (72 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.M., AM, Apr, April, Atlantic, Aug, August, CD player, CD, CD-ROM, CV, Christian, DJ, DNA, DVD, Dec, December, Dr, Dr., E-mail, Englishman, Feb, February, Friday, HIV, I, ID card, ID, IT, Internet, Jan, January, July, June, MP3 player, Mar, March, May, Mediterranean, Miss, Monday, Mr, Mr., Mrs, Mrs., Ms, Ms., Nov, November, OK, OK, Oct, October, Olympia, Olympiad, Olympic, Olympics, PC, P.M., PM, Saturday, Sept, September, Shakespearean, Soviet, Sunday, T-shirt, TV, Thursday, Tuesday, Wednesday, X-ray</td>
</tr>
</tbody>
</table>

Table 3. List of Words with Apostrophes

<table>
<thead>
<tr>
<th>SVL12000 (1 word)</th>
</tr>
</thead>
<tbody>
<tr>
<td>o’clock</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NGWL (2 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>o’clock, ma’am</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CEFR-J (6 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>driver’s license, ’m, o’clock, ’re, ’s, ’d</td>
</tr>
</tbody>
</table>
8.6 Hyphenated Words

Hyphenated words are first divided into two words (e.g., “Osaka-based” is treated as two words, “Osaka” and “based”) in all word lists except for the cases where the selected list has hyphenated words as headwords (Table 5).

8.7 Compounds/Multi-Word Units

If a headword in the selected word list consists of more than one word (e.g., “bank account” and “mobile phone” in CEFR-J), it is counted as one word (unit), and NWLC returns the word profiling accordingly. NGSL and its three secondary lists (NAWL, TSL, and BSL) have only one compound, “ice cream,” while CEFR-J has 145 compounds (Table 6).

8.8 Comparisons of Word Lists: Which One to Use?

Table 7 summarizes the word counting units and rules, as described in the previous section, for the word lists adopted in NWLC.

Because all word lists in NWLC vary in terms of their rules and word definitions, direct comparisons are not straightforward. Therefore, a comparison of the coverage rate of each word list was made to provide a rule-of-thumb guideline for users to select which word list for what purposes.

Two English texts were analyzed with NWLC using all five word lists. The first English text was composed of (a) the Center Test in 2020 (6,309 words) and (b) the Kyotsu Test in 2021 (7,905 words), and both the reading and listening sections including the script for the listening section were used. The Kyotsu Test (Kyotsu means “common” in English), which was called the Center Test until 2020 with...
a different test format, is a standardized exam for students who intend to enter a national, public, or private university in Japan. Over 470,000 test takers sat for the Kyotsu Test in 2021. As the Center and Kyotsu test items reflect the Course of Study (national curriculum guidelines) in Japan, the test items from 2020 and 2021 used in this comparison can serve as a yardstick for evaluating the coverage rate, with each word list, of the standard English text that high-school graduates in Japan are likely to encounter.

The second English text analyzed with the word lists on NWLC was Test of English for International Communication (TOEIC). The text was collected and randomly sampled from materials such as TOEIC Official Test Preparation

---

### Table 6. List of Compounds

<table>
<thead>
<tr>
<th>NGWL (1 word)</th>
<th>CEFR-J (145 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ice cream</td>
<td>according to, air conditioning, air force, alarm clock, all right, bank account, because of, board game, bus station, bus stop, capital letter, car park, carbon dioxide, carbon footprint, carbon monoxide, CD player, central heating, chat show, check-in counter, check-in desk, chest of drawers, chewing gum, classical music, climate change, common sense, credit card, de facto, debit card, definite article, department store, digital camera, dining room, disc jockey, disk jockey, driver’s license, driving licence, each other, environmentally friendly, exchange rate, exclamation mark, extreme sports, face to face, fast food, fed up, fire brigade, fire station, first floor, first lady, first language, first name, first person, frying pan, full stop, global warming, good afternoon, good morning, good night, grocery store, ground floor, hard drive, have to, health care, heart attack, high school, hip hop, human rights, ice cream, ice hockey, ice skating, ID card, identity card, indefinite article, instead of, inverted commas, junk food, last minute, last name, living room, main course, martial art, message board, mineral water, mixing bowl, mobile phone, modal verb, MP3 player, native speaker, navy blue, net surfer, next door, next to, no one, olive oil, on to, ought to, out of, owing to, pen friend, per cent, petrol station, phrasal verb, pocket money, point of view, polar bear, police officer, police station, post office, primary school, prime minister, public transport, question mark, real estate, remote control, rush hour, science fiction, second person, secondary school, shop assistant, sitting room, soap opera, social networking, soft drink, sports center, sports centre, steering wheel, stock market, swimming costume, swimming pool, table tennis, text message, third person, tour guide, traffic jam, traffic light, travel agent, upside down, used to, vice president, video clip, video game, virtual reality, washing machine, weather forecast, web page, worn out</td>
</tr>
</tbody>
</table>

---

### Table 7. Summary of Word Counting Units and Rules

<table>
<thead>
<tr>
<th></th>
<th>SEWK-J</th>
<th>NewJ8</th>
<th>SVL</th>
<th>NGSL</th>
<th>CEFR-J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Words</td>
<td>74,810</td>
<td>8,000</td>
<td>12,000</td>
<td>5,621</td>
<td>7,799</td>
</tr>
<tr>
<td>Lemma or Flemma</td>
<td>Flemma</td>
<td>Flemma</td>
<td>Flemma</td>
<td>Flemma</td>
<td>Lemma</td>
</tr>
<tr>
<td>(F)lemma List</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Capitalized</td>
<td>0</td>
<td>20</td>
<td>44</td>
<td>0</td>
<td>72</td>
</tr>
<tr>
<td>Case</td>
<td>Insensitive</td>
<td>Sensitive</td>
<td>Sensitive</td>
<td>Insensitive</td>
<td>Sensitive</td>
</tr>
<tr>
<td>Apostrophe</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Abbreviation with Period</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Hyphen</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>62</td>
</tr>
<tr>
<td>Compound</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>145</td>
</tr>
</tbody>
</table>

---
Guide. Both listening and reading sections are included. Because many Japanese university students and businesspeople take the TOEIC test, it can be regarded as a representative text for EFL learners at those levels.

After ignoring question numbers, symbols, and Japanese text, the Center and Kyotsu Tests and the TOEIC comprised 14,214 and 104,299 words, respectively.

Figure 2 shows the results of the coverage rate of five word lists on NWLC. The level or category marked by the red boxes indicates 95% coverage of the text by each word list. As it is clear from the results, all five lists reach the 95% coverage threshold, which indicates that the vocabulary included in these word lists is within the range of practical use for these important, high-stake tests in Japan.

The blue boxes in the figure indicate 98% coverage of the text by each word list. This result demonstrates that, to reach the 98% coverage threshold which is

<table>
<thead>
<tr>
<th>Center &amp; Kyotsu tests (14,214 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEWK-J</td>
</tr>
<tr>
<td>New JACET8000</td>
</tr>
<tr>
<td>SVL12000</td>
</tr>
<tr>
<td>New general service list + α</td>
</tr>
<tr>
<td>CEFR-J</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TOEIC (104,299 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEWK-J</td>
</tr>
<tr>
<td>New JACET8000</td>
</tr>
<tr>
<td>SVL12000</td>
</tr>
<tr>
<td>New general service list + α</td>
</tr>
<tr>
<td>CEFR-J</td>
</tr>
</tbody>
</table>

95% 98%

Figure 2. Coverage rates of five word lists on NWLC.
Note. For New General Service List + α (three other lists), 3 indicates NAWL and TOEIC and BSL (183 words), 2 indicates NAWL and TOEIC, NAWL and BSL, or TOEIC and BSL (790 words), and 1 indicates that the word appears only in NAWL, TOEIC, or BSL (1,847 words). In this analysis, SEWK-J has only 10 levels, each with 1,000 words, for comparison with other lists. 

*Vocabulary Learning and Instruction, 10(2), 30–41.*
posited by vocabulary research literature (Schmitt et al., 2011), it requires several thousands more words in New JACET8000, SVL12000, and SEWK-J. In the case of the NGSL plus three secondary lists (NAWL, TOEIC, and BSL) and the CEFR-J, it was not possible to reach the 98% coverage threshold for either the Center and Kyotsu Tests or the TOEIC. This is not surprising given the fact that the number of words included in NGSL plus the three secondary lists is 5,621, and 7,815 for CEFR-J. It should be noted that when it comes to the coverage rate of text, CEFR-J has a disadvantage because it adopts the lemma counting unit, and the same word may have a different POS, which is not the case with the other four lists in NWLC; this means that many more lemmas are required to cover the same amount of text as compared to flemmas.

From the results of the comparisons of word lists used in NWLC, it can be claimed that all word lists could be used to examine the coverage rate of text because all the lists proved to have sufficient coverage of English texts, up to 95%. Although only the results of analyzing two English texts were reported in this article, the same tendency can be observed with other texts as well.

If the purpose of using the list is to learn the first 3,000 words in the lemma form, the NGSL may be a good choice. NGSL was reported to cover 92% of most general English texts (Browne et al., 2013). In this sense, the learning efficiency is much higher for learners. As CEFR-J distinguishes POS of the words, it could be used for assessing learners’ productive vocabulary use in writing or speaking, which in turn implies that CEFR-J could be used to measure the depth of vocabulary knowledge.

Among New JACET8000, SVL12000, and SEWK-J, New JACET8000 had better coverage than the other two in the current comparisons. As New JACET8000 is an updated word list integrating other word lists derived from large corpora, the coverage performance seems to be better than other word lists. If the user needs to learn more than 8,000 words, which is what the New JACET8000 covers, the SVL12000 or SEWK-J could be utilized. Especially, the SEWK-J list contains 74,810 headwords, and thus it is possible to inspect how many words learners need to know to reach the 98% coverage threshold, which cannot be accomplished with the other word lists in NWLC.

Although the degree of coverage of texts can be used as a measure of how efficiently a word list represents a target register, word lists can also be used for other purposes. Notably, the SEWK-J was not designed to increase the coverage performance; rather, it was intended to map the Japanese university students’ vocabulary knowledge on the word list. In this sense, it is unique in that the rank order of the headwords in SEWK-J represents the likelihood that a word in the list is known to the learners in the target population (i.e., Japanese university students). Thus, while frequency-based lists are of great value in directing learners towards the most frequent—and so valued—words in a list, when matching learners with texts, we should consider basing tests and lexical profilers on what learners do know (i.e., by using word lists based on learners’ actual vocabulary knowledge such as SEWK-J), rather than frequency-based lists, which indicate what learners should know (Paul Nation, personal communication, August 8, 2021). By matching learners with texts, vocabulary profiling will allow teachers
and materials developers to estimate how difficult candidate texts might actually be for learners at a given level of proficiency.

In terms of the assessment of vocabulary knowledge, word profilers and vocabulary tests should be based on the same word list scale to match the vocabulary level of a text with learners’ lexical knowledge. To this end, the Vocableveltest.org website (https://vocableveltest.org/) can be utilized because it facilitates the creation and administration of meaning-recall vocabulary levels tests, which better measure the type of vocabulary knowledge that can be employed when reading than meaning-recognition (multiple-choice or matching) tests (McLean et al., 2020; Stewart et al., 2021), on several of the same word lists that are used in the NWLC (e.g., SEWK-J, NGSL, NAWL, TSL, and New JACET 8000).

Acknowledgements

This research was made possible by a grant-in-aid for Scientific Research (21H00553 and 20K00792) from the Japan Society for the Promotion of Science.

References


Second Language Vocabulary Teaching and Learning: A Commentary on Four Studies for the JALT Vocabulary SIG

David Beglar
Temple University, Japan Campus

Abstract

Four papers by Brandon Kramer, Tohru Matsuo, Aaron Sponseller, Young Ae Kim, Suzuka Nishiyama, and Stuart McLean; Stuart Benson and Naheen Madarbakus-Ring; Michael McGuire and Jennifer Larson-Hall; and Atsushi Mizumoto, Geoffrey Pinchbeck, and Stuart McLean were presented in the morning session of the 2021 Vocabulary SIG Symposium held in Kyoto, Japan, on 04 December 2021. As discussant, it was my pleasure to comment on each of the papers. The four studies were investigations of the attrition of productive vocabulary knowledge; the lexical composition of two English for Academic Purposes (EAP) textbooks; the relationship between high-frequency multi-word sequences, speech rate, and listening perception; and the use of various word lists with the New Word Level Checker. After reviewing and summarizing each study in this paper, I present suggestions for further developing the studies and expanding the research agenda of each study.

Keywords: vocabulary, attrition, productive knowledge, lexical coverage, multi-word sequences, speech rate, oral fluency, listening perception, word list

1 Introduction

Second language vocabulary acquisition (SLVA) has received a tremendous amount of attention from researchers in the past three decades for good reason. Vocabulary is widely seen as the heart of multiple models of first (L1) and second language (L2) acquisition. For instance, Cutler and Clifton (2000) placed lexis at the heart of their first language listening model, while Rost (2016), a leading L2 listening authority, stated, “Recognizing words in fluent speech is the basis of spoken language comprehension…” (p. 28). Levelt and his colleagues (Levelt, 1989; Levelt et al., 1999) proposed a modular L1 speech model in which lexis occupies a crucial position because lexical items govern syntactic processing. Lexis also plays the key role in bilingual speech production models (e.g., see Chapter 4 in Kormos, 2006). In addition, word knowledge plays the central role in theories of L1 reading comprehension, such as the reading systems framework (Perfetti & Stafura, 2014). This line of thinking has been echoed by second language reading specialists such as Grabe and Stoller (2020), who stated, “The most fundamental requirement for fluent reading comprehension is rapid and automatic word recognition” (p. 16).
In this paper, I review and comment on four papers: Measuring Attrition of L2 Productive Vocabulary Knowledge Over the Summer Vacation (Kramer, Matsuo, Sponseller, Kim, Nishiyama, & McLean), A Comparison of Textbook Vocabulary Load Analysis (Benson & Madarbakus-Ring), The Contribution of High-Frequency Multi-Word Sequences to Speech Rate and Listening Perception (McGuire & Larson-Hall), and Comparisons of Word Lists on New Word Level Checker (Mizumoto, Pinchbeck, & McLean).

2 The Four Studies

2.1 Measuring Attrition of L2 Productive Vocabulary Knowledge Over the Summer Vacation (Kramer, Matsuo, Sponseller, Kim, Nishiyama, & McLean)

Relatively few second language vocabulary researchers have conducted studies of forgetting, and even studies of language attrition in the broader area of second language acquisition are rare; thus, little is known about this issue. For this reason, the study by Kramer et al. is a welcome addition to the SLVA literature. The primary purpose was to investigate the attrition of productive lexical knowledge over summer vacation by 81 first-year Japanese university students. The learners had taken part in a vocabulary instructional programme that included the use of a digital paired-associate learning application designed to develop both receptive and productive lexical knowledge and had also engaged in an extensive reading programme. The learners studied high-frequency vocabulary from the New General Service List (NGSL) (Browne et al., 2013) that covered the 501–1,507 word-frequency range and were asked to read 180,000 tokens of graded readers during the academic year. The participants’ productive vocabulary knowledge of 60 randomly selected words from the targeted frequency range was assessed before and after summer break with an online vocabulary testing site, VocabLevelTest.org (McLean & Raine, 2019). The researchers’ focus on productive vocabulary set a high bar, as learning productive vocabulary is more difficult than learning receptive vocabulary (Mondria & Weirsma, 2004, p. 79). A paired-samples t-test of the pre-test and post-test results showed no significant change in lexical knowledge ($p = 0.90$, Cohen’s $d = 0.01$). The researchers also conducted a multiple regression to investigate determinants of the changes in the productive test scores. Only vocabulary review with the digital application, which accounted for 4.90% of the variance, was a significant predictor. Extensive reading and overseas experience did not make significant contributions to the regression model.

Kramer et al. also used a more detailed approach to investigating possible attrition by using a five-part system for categorising lexical knowledge that ranged from No demonstrated knowledge of the target word to Correct target word with correct spelling and derivational form. The authors reported that 84.60% of the 12,115 total words tested did not change from pre-test to post-test, while the greatest amount of attrition at any level of the five-part system was 0.40%; thus, attrition was negligible. Taken together, the results indicated that the participants’ productive vocabulary knowledge was generally stable over the summer break. This finding might have occurred for two reasons. First, the majority of the target items should have been known by the learners before entering the university; thus, the
instructional programme might have primarily served to consolidate rather than teach knowledge of the target words. Second, the digital application required productive practice; thus, it involved transfer-appropriate processing, which states that memory retrieval is easier when the cognitive processes engaged in when materials are encoded match those used during encoding (e.g., Graf & Ryan, 1990). In other words, the digital application matched the test format. Although admittedly difficult to determine, an effort needs to be made to operationalise the degree to which the learners used the receptive and productive parts of the digital application. This information could shed light on the effectiveness of the application and thereby provide usage guidelines for teachers and learners.

In addition, the fact that extensive reading was not a significant predictor in the regression model was reasonable given that the participants read little \((M = 14,813 \text{ words read})\) during summer vacation and it only provided receptive encounters with vocabulary. Even if overseas experience had been a significant predictor, it would have been uninterpretable without knowledge of the interactions the learners had while overseas. For instance, it is important to know whether the participants visited an English-speaking country, who they lived with (e.g., another Japanese student or an English-speaking family), how much English they were exposed to, and how much English they produced.

As the authors acknowledged, one problem with the study was the use of different words on the pre-test and post-test. This approach made direct comparisons impossible; instead, the researchers were forced to assume that the words on the two tests were approximately equal in difficulty. However, there was no indication that the words on the tests were controlled for issues such as frequency, concreteness, or part of speech. One way to make two or more tests directly comparable is to use Rasch item anchoring, as this technique allows for direct comparisons in person ability when a number of common items are included on each test form (see Chapter 5 in Bond et al., [2021] for further information).

This study raises interesting questions which, to the best of my knowledge, have not been investigated, and as is usual in many types of research, the most interesting questions revolve around the reasons why a phenomenon does or does not occur. For instance, does the degree to which a lexical item has been automated influence its stability in long-term memory? The degree of automatisation can be assessed using reaction time measures, which can assess aspects of words such as orthographic form, phonological form, and semantic meaning (e.g., see Matsuo [2017] and Shimono [2019] for studies that included these types of reaction time measures). If automaticity is a determinant of stability in long-term memory, then it would be a candidate for becoming an instructional objective in foreign language curricula.

A second possibility concerns depth of lexical knowledge. If depth of knowledge is defined as knowledge of collocations and associations, then it means that a target word has been placed in a lexical network in the learners’ mental lexicon. It is plausible that the development of this network consolidates knowledge of target words and provides multiple mental pathways to access them. A somewhat different version of this idea was proposed by Meara (2009, see Chapters 3 and 4), who suggested that receptive lexical knowledge develops.
into more cognitively demanding productive knowledge as words are embedded into a lexical network. If it is shown that embedding target vocabulary in such a network improves long-term retention, then it would also be a candidate as an instructional objective.

Future researchers can also investigate part of speech and attrition. For instance, whereas adjective and adverbs generally have one form and nouns can only add two inflectional morphemes, plural -s and possessive -’s, verbs are highly complex. Verbs can take four inflectional morphemes, irregular forms are commonplace and are formed in various ways, verb tenses combine to form a complex tense-aspect system, and many verbs convey abstract meanings, which raises yet another line of research: The possibility that attrition is related to the degree of concreteness of a word’s meaning. This phenomenon, which is known as the concreteness effect (e.g., Pexman et al., 2007), states that concrete words are better remembered (Schwanenflugel et al., 1992), better recognised (Fliessbach et al., 2006), read faster (Schwanenflugel & Shoben, 1983), and acquired more quickly than more abstract words (Mestres-Missé et al., 2014).

### 2.2 A Comparison of Textbook Vocabulary Load Analysis (Benson & Madarbakus-Ring)

Benson and Madarbakus-Ring conducted an investigation of the vocabulary load of an in-house textbook and a commercially produced textbook to determine the appropriateness of the two textbooks for first-year Japanese university students. The textbooks were analysed with Range (Heatley et al., 2002) and the New Word Level Checker (Mizumoto, 2012). Nation’s British National Corpus/Corpus of Contemporary American English (BNC/COCA) base word lists, which were accompanied by his supplementary lists, were used as a general word list, and the New JACET8000 word list (JACET, 2016) was used as a context-specific word list, as it was created for tertiary-level Japanese learners of English. Using the conventional targets of 95% and 98% lexical coverage of a written text (Nation, 2013, pp. 206–208), the authors first analysed the inhouse textbook and reported that knowledge of the first 3,000 words on the BNC/COCA list and the supplementary list would enable learners to know 95% of the tokens in the textbook, while the first 5,000 words and the supplementary list were needed to meet the 98% criterion. The analysis of the commercial textbook produced similar results: Knowledge of the first 3,000 words was needed to meet the 95% criterion and knowledge of the first 4,000 words was necessary to meet the 98% criterion.

When the in-house and commercial textbooks were analysed with the New JACET8000 word list, they yielded the same results: The first 3,000 words provided 95% coverage, while knowledge of 8,000 words was needed to meet the 98% criterion.

One way to interpret the above results is to use the 3,715-word figure for written receptive vocabulary knowledge of Japanese university students provided by McLean et al. (2014). When matched with the results of this paper, average students would know slightly more than 95% of the tokens in both the in-house
and commercial textbooks, a figure I would argue is sufficient if the students are engaged in intensive reading. My position raises a question that has not been sufficiently addressed in the SLVA literature: Should the 95% and 98% criteria be applied equally to all types of reading, or should SLVA researchers distinguish different purposes for reading as reading specialists have done for decades? It is possible that different degrees of lexical coverage work effectively for different reading purposes. It is also important to acknowledge that not all unknown words are equal, as one might occur once in a text, while another occurs multiple times. Moreover, the helpfulness of the contexts in which they are embedded can vary greatly (see Nation [2013], pp. 363–364 for a list of 14 variables that influence an unknown word’s guessability in a reading text).

Extensive reading and intensive reading usually differ in a number of important ways. Students typically engage in extensive reading without the benefit of pre-reading activities and with little or no teacher support. In such cases, the 98% criterion is defensible. However, intensive reading often involves pre-reading tasks, including vocabulary activities; the texts are far shorter than extensive reading texts; glossing can clarify the meaning of low-frequency or technical vocabulary (Nation, 2013, pp. 238–247), learners can use a dictionary (Nation, 2013, pp. 414–436), and learners have some ability to guess unknown words from context (Nation, 2013, pp. 348–388). In addition, it is generally agreed that some unknown vocabulary should be included in intensive reading texts given that without such lexical challenges, vocabulary growth would be unnecessarily limited. What this means where lexical coverage is concerned is unknown and therefore needs to be investigated, but there is a possibility that even 95% coverage is not required for intensive reading. Finally, vocabulary coverage figures might not provide an accurate picture of text comprehension because of what Hagoort (2005) called unification. During unification, word-level semantic and syntactic information is integrated into the developing context. Learners struggling with unification could have comprehension difficulties even when they know 100% of the tokens in a reading text because they fail to understand how individual tokens fit together to create meaning. Empirical research can shed light on this issue.

This study raises other interesting issues where the 95% and 98% criteria are concerned. It is possible that the two figures are interpreted too inflexibly in the field. For example, in Table 5 in Benson and Madarbakus-Ring’s paper, knowledge of 8,000 words is needed to achieve 98% lexical coverage of the commercial textbook using the New JACET8000 list. However, knowledge of 4,000, 5,000, 6,000, and 7,000 words produced coverage figures of 97.01%, 97.40%, 97.65%, and 97.92%, respectively. Thus, knowledge of 4,000 words provides 97.01% coverage and knowledge of 8,000 words provides 98.15% coverage. This result raises two points. First, there is a massive difference between knowing 4,000 and 8,000 words in a foreign language. The McLean et al. (2014) study included 3,427 participants, yet the maximum vocabulary size estimate was 7,400 words; thus, while none of their participants appeared to have an 8,000-word receptive vocabulary, the mean of 3,715 was close to knowledge of 4,000 words. In contrast, the coverage difference between knowledge of 4,000 and 8,000 words in Table 5 was 1.14% (98.15 – 97.01 = 1.14). This small difference raises an important question: Can the inflexible

Vocabulary Learning and Instruction, 10(2), 42–54.
use of the 95% and 98% coverage criteria lead researchers and classroom teachers astray in some instances?

Researchers can answer this question by investigating small differences in lexical coverage with learners at different proficiency levels using texts written at different levels of difficulty (e.g., general texts vs. academic texts) and that are used for different purposes (e.g., intensive reading, narrow reading, timed reading, and extensive reading).

Another issue to consider is that both word lists used in this study divide words into 1,000-word frequency bands. Although this division is conventional, it is imprecise and therefore of limited usefulness. If for instance, the texts were analysed using 500- or 100-word categories, it might be found that knowledge of 2,500 words rather than 3,000 words provides 98% coverage of a text. This 500-word difference is large in terms of percentage difference (i.e., 18%), and according to the result obtained from McLean et al. (2014) study, the Japanese learners acquired an average of 619 words per year (3,715 words ÷ 6 years of study = 619) if they were first-year university students and less if they were in higher grades; thus, 500 words represents nearly 1 year of study. More detailed vocabulary coverage analyses are possible with lextutor.com and the Scale of Word Knowledge–Japanese (SEWK-J) list (Stuart McLean, personal communication, November 29, 2021).

Another concern with this type of analysis is that it assumes that knowledge of single words is an adequate predictor of text comprehension. There is an abundance of evidence that reading comprehension is based on multiple factors including knowledge of collocations (Matsuo, 2017), multi-word expressions (e.g., Siyanova-Chanturia & Van Lancker Sidtis, 2018), syntactic parsing (Jeon & Yamashita, 2014), background knowledge (Kintsch, 2012), and inferencing skills (van den Broek et al., 2015). Vocabulary researchers too rarely acknowledge that skilled reading is based on a number of distinct component skills, of which lexical knowledge is just one.

2.3 The Contribution of High-Frequency Multi-Word Sequences to Speech Rate and Listening Perception (McGuire & Larson-Hall)

Relatively few studies have been focused on teaching vocabulary for productive use, so the study by McGuire and Larson-Hall addresses an important gap in the field. The researchers used Anki, a digital vocabulary learning application, to teach two types of multi-word sequences — bigrams such as high probability and trigrams such as and then I. They then investigated the degree to which knowledge of the multi-word sequences affected speech rate, one aspect of oral fluency, and performance on a listening dictation task. All 33 participants studied 10 collocations each week using the Anki app and described or listened to classmates describe three picture stories that incorporated the taught collocations. The experimental group (n = 16) also practiced 10 of the multi-word sequences each week using Anki and watched a 5-minute introduction about them. The participants described three multi-panel pictures and talked about their ideal day on the speaking pre-test and post-test. Audio recordings were analysed with PRAAT (Boersma & Weenink, 2021), a computer program for analyzing, synthesizing, and
manipulating recorded speech samples, to ascertain speech rate, which was operationalised as the number of syllables produced divided by the duration of the speech sample. In the listening dictation task, the participants transcribed a naturalistic conversation containing 27 of the multi-word sequences while listening twice at normal speed. The pre-test and post-test results are shown in Table 1. As shown in the Table, the experimental group improved to a small degree on the speaking tests, as indicated by the Cohen’s $d$ effect sizes, while both groups made large improvements on the dictation tests. These findings indicated that receptive knowledge of the form of multi-word sequences was acquired better than productive knowledge of the sequences.

The overall results were promising, as they showed that L2 learners are capable of acquiring multi-word sequences for both receptive and productive use. Larger $n$-sizes would have produced more robust results and likely solved the statistical issues that arose with the 95% confidence intervals passing through zero for the experimental group. However, the authors are to be commended for using 95% confidence intervals—a practice that is all too rare in the field of SLVA—as they show the degree to which variables are measured precisely. In addition, a second experimental group would add more nuance to the study and make the results more informative. For instance, in the study, the participants were asked to engage in homework in which they counted the number of times they heard the reduced multi-word sequences. However, this type of focus on linguistic form can negate meaningful processing (VanPatten, 2020). Therefore, including an additional experimental group that receives input in which the focus is on meaning rather than form might provide more nuanced results.

An additional possibility is to assess oral fluency in a more detailed way. Oral fluency is often divided into three categories: breakdown fluency, speed fluency, and repair fluency. Breakdown fluency is assessed using silence and pausing.
behaviour, and can be operationalised using silent pauses at AS-boundaries (AS = Analysis of speech unit; “…an independent clause or subclausal unit, together with any subordinate clause(s) associated with it” Foster et al., [2000], p. 365), silent pauses mid-clause, filled pauses at AS-boundaries, and filled pauses mid-clause (Pang & Skehan, 2021). Speed fluency can be assessed by calculating pruned (i.e., repetitions, restarts, and self-repairs are deleted) and unpruned speech rates. Moreover, speech rate can be divided into sub-components such as lexical retrieval speed, speed of articulation, and sentence building speed (De Jong et al., 2012). However, speech rate must be interpreted carefully, given that some native speakers of a language speak relatively slowly. This issue can be addressed by measuring research participants’ L1 speech rates as suggested by Segalowitz (2010, p. 2). Repair fluency can be assessed using false starts/100 words, self-repairs/100 words, and repetitions/100 words. A final possibility is to gather qualitative judgements of oral fluency using human raters. In this way, it is possible to determine whether changes in oral fluency can be detected by the human ear. Human rater data can be profitably analysed using software such as FACETS (Linacre, 2021), which provides interval-level person ability estimates, fit statistics for all facets in the model (e.g., participants, raters, tasks, and evaluation criteria), and bias analyses.

In addition to a more detailed account of changes in oral fluency, when teaching vocabulary for productive purposes, it is useful to investigate changes in lexical complexity, which can be operationalised as lexical frequency (i.e., the number of tokens in particular word-frequency bands), lexical diversity (i.e., the ratio of the number of unique word types to the total number of tokens), and lexical density (i.e., the proportion of content words to the total number of tokens) (Révész et al., 2016). Lexical complexity is one form of linguistic development that would be reasonably expected to occur in a study such as this one (Baese-Berk et al., 2021).

A final possibility is to investigate the contribution of foreign language learning aptitude (e.g., Linck et al., 2013) to lexical acquisition in the context of an oral proficiency study such as this one. Two related constructs that could be investigated are working memory and rote learning ability (i.e., associative memory), as empirical research has suggested that they have direct relationships to vocabulary learning (e.g., Nagata et al., 1999).

2.4 Comparisons of Word Lists on New Word Level Checker (Mizumoto, Pinchbeck, & McLean)

Analysing the lexical composition of listening and reading texts is a useful skill for foreign language instructors, materials developers, and curriculum specialists. For this reason, the paper by Mizumoto, Pinchbeck, and McLean provides useful data for each of the above audiences. The researchers presented an investigation of an online vocabulary profiling application, the New Word Level Checker (https://nwic.pythonanywhere.com/), in which they discussed the rationale for developing the application and compared the lexical coverage results of five word lists with two English-language examinations. The five word lists that make up the New Word Level Checker are the SEWK-J, the New JACET8000
Table 2. 95% and 98% Coverage Results for the Five Word Lists

<table>
<thead>
<tr>
<th>Word list</th>
<th>Center and Kyotsu Tests</th>
<th>TOEIC Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95% coverage</td>
<td>98% coverage</td>
</tr>
<tr>
<td>SEWK–J</td>
<td>4,000</td>
<td>—</td>
</tr>
<tr>
<td>New JACET8000</td>
<td>3,000</td>
<td>8,000</td>
</tr>
<tr>
<td>SVL 12000</td>
<td>4,000</td>
<td>9,000</td>
</tr>
<tr>
<td>NGSL</td>
<td>NGSL, NAWL, TOEIC</td>
<td>—</td>
</tr>
</tbody>
</table>


(JACET, 2016), the Standard Vocabulary List 12000 (SVL12000), the NGSL (Browne et al., 2013), and CEFR–J (Tono, 2019). The authors described the word counting units and rules applied with each corpus, including how proper nouns and numerals were treated, lemmatisation, words with capital letters, contractions, hyphenated words, compounds, and multi-word units.

In the final part of the paper, the five word lists were used to analyse (a) the 2020 Center Test and 2021 Kyotsu Test as well as (b) a TOEIC test. The authors analysed the tests using the five word lists in order to investigate how much vocabulary knowledge was needed to reach the 95% and 98% text coverage criteria. The results are shown in Table 2. As indicated in the Table, using the SEWK-J, New JACET8000, and SVL 12000 word lists, knowledge of 3,000–4,000 words was needed to achieve the 95% criterion on the Center and Kyotsu examinations, while knowledge of 8,000–9,000 words was required to meet the 98% criterion. Importantly, the 98% criterion could not be determined using the SEWK-J, NGSL, and CEFR-J lists. The TOEIC test results showed that 95% coverage could be achieved with knowledge of 4,000–5,000 words. However, the results for the 98% criterion varied widely with the SEWK-J list indicating that knowledge of 10,000 words was required and the New JACET8000 list indicating that knowledge of 6,000 words was needed.

As noted above, in some instances, three of the lists were unable to determine 98% coverage due to limitations in the lists themselves. In addition, an anomalous result is that the SEWK-J list could not provide 98% coverage of the Center and Kyotsu tests even though it could for the more difficult TOEIC test. Furthermore, the results for the NGSL are not comparable to the other word lists, as it includes specialised lists such as the New Academic Word List and TOEIC list, which are made up of words from a variety of frequency levels.

The paper can be clarified and extended in a number of ways. First, the selection criteria for the five lists should be specified in order to justify why these lists are preferable to other better-known word lists such as the BNC and COCA. Second, as suggested by Laufer (2014), it would be helpful to show the amount of overlap among the five lists at the various word frequency levels, as this would clarify where they do and do not differ, and with the use of a more detailed analysis of individual words, how they differ. Third, the idea that proper nouns “...can be assumed to be
understood by learners” has not been properly researched and should therefore not be assumed, particularly where learners at lower proficiency levels are concerned.

The authors also used a multiple regression to estimate the probability of a word being known on the SEWK-J list by Japanese university students. The regression analysis could be improved in a number of ways, including checking the assumptions of the analysis, justifying the use of one vocabulary test to predict scores on another vocabulary test, using a vocabulary test that requires evidence of semantic knowledge rather than a yes-no test, ensuring that reaction time data are gathered with proper equipment and software, and treating loan word status in a nuanced way. A large-scale study of loanwords currently underway (Edelman, 2021) shows that knowledge of loanwords varies substantially among Japanese university students, and should not be treated in a monolithic way.

The authors cited Nation (personal communication) when stating that tests and lexical profilers should be based on what learners know rather than what they should know. However, justification is needed for this statement, as it cannot be generalised to all situations. Although this proposal might work when selecting a course textbook, there are good reasons for teaching learners words they do not know, but should know—a position the authors do not appear to support—particularly in educational contexts in which the learners have no specific need for the English language. In such cases, using frequency-based lists to determine learning targets often provides the best cost-benefit ratio (Nation, 2013, p. 38).

The researchers also concluded that all five word lists can be used for text analysis because they “…have sufficient coverage of English texts, up to 95%” despite mounting evidence that 98% is the more valid figure where unassisted reading is concerned. In addition, they stated that the CEFR-J list distinguishes part of speech and can be used for assessing learners’ productive vocabulary, “which in turn implies that CEFR-J could be used to measure the depth of vocabulary knowledge.” The reasoning behind this statement needs to be clarified, as the term depth of vocabulary knowledge has multiple meanings and it is unclear how a vocabulary list can be used to measure that construct, however it is defined.

Finally, the suggestions made above in connection to the Benson and Madarbakus-Ring paper also apply here: (a) Using 1,000-word frequency bands produces imprecise estimates and should therefore be abandoned, (b) assuming that knowledge of single words is an adequate predictor of text comprehension flies in the face of a large amount of L1 and L2 research, and (c) if learners are engaged in intensive reading, the 95% and 98% vocabulary coverage figures are potentially too strict. As noted above, each of these issues requires further study.

3 Conclusion

The four papers reviewed above provide a great deal of food for thought about how to move the field of SLVA forward. I enjoyed reading the papers and found them stimulating in a number of ways, as they touch upon a number of important issues where vocabulary learning and teaching are concerned: the stability of learners’ lexical knowledge, matching learners with textbooks, teaching productive vocabulary, and analysing the lexical composition of listening and reading texts. As researchers

Vocabulary Learning and Instruction, 10(2), 42–54.
shed further light on the learning processes involved in foreign language vocabulary acquisition, teachers and students will benefit greatly, as vocabulary knowledge is the linguistic foundation of second language acquisition.

References


Laufer, B. (2014). Vocabulary in a second language: Selection, acquisition, and testing: A commentary on four studies for JALT vocabulary SIG. *Vocabulary Learning and Instruction, 3*(2), 38–46. https://doi.org/10.7820/vli.v03.2.laufer


Correlations of Modalities of Written Vocabulary Knowledge to Listening and Reading Proficiency: A Comparison

Jeffrey Stewart\textsuperscript{a}, Stuart McLean\textsuperscript{b}, and Aaron Olaf Batty\textsuperscript{c}
\textsuperscript{a}Tokyo University of Science; \textsuperscript{b}Momoyama Gakuin University; \textsuperscript{c}Keio University

Abstract

In recent years, there has been increasing debate and research regarding which modality of vocabulary knowledge has the strongest correlation to reading, with particular focus on distinctions between testing L2 form and L2 meaning, and between recall of answers from memory and recognition of answers from fixed options. However, relatively little attention has been paid to find out which modality has the strongest correlation to listening ability. A recent meta-analysis by Zhang and Zhang (2020) indicated that meaning recall was the superior predictor of reading proficiency. Although their results showed that form recall had the highest correlation to listening, the difference between form recall and meaning recall was statistically insignificant. The present study uses data from McLean et al. (2020) of learner responses to 1000-item vocabulary tests employing written tests of meaning recall, form recall, meaning recognition and Yes/No modalities, sampling them with replacement to create thousands of 100-item tests using a bootstrapping approach. The test scores were then correlated to measures of listening and reading proficiency for comparison. The results indicated that for written tests, meaning recall, form recall, meaning recognition and form recognition had the strongest correlations to both reading and listening, in descending order. All comparisons were statistically significant.

Keywords: Vocabulary, vocabulary testing, meaning recall, listening, reading

1 Background

In recent years, there has been increasing debate in the field of vocabulary learning and instruction regarding which modality of vocabulary knowledge has the strongest correlation to reading proficiency (Stewart et al., 2021; Stoeckel et al., 2021; Webb, 2021). A meta-analysis by Zhang and Zhang (2020) indicated that meaning recall, wherein learners recall and write or type L1 word meaning from memory after encountering the L2 written form, had a stronger correlation than both form recall (see the meaning, produce the L2 word form from memory) and the commonly used meaning recognition format (see the L2 word form, select a correct definition of the word from a list of fixed options, e.g., multiple-choice tests).
Shortly before Zhang and Zhang’s meta-analysis was released, McLean et al. (2020) published an additional study of the relationship between modalities of vocabulary and reading proficiency using a bootstrapping approach. Bootstrapping involves sampling a population with replacement in order to reach better estimates of confidence intervals and determine how the estimates from replicate experiments could be distributed (Kulesa et al. 2015). In the McLean et al. (2020) study, learners took 1,000-item tests in each modality, which were continually sampled with replacement to produce thousands of test forms and correlated to scores on the Reading section of the Test of English for International Communication (TOEIC®) test (https://www.ets.org/toeic). The study found further evidence that meaning recall was superior to meaning recognition as a predictor of reading ability. Written meaning recall was also found to have statistically and significantly higher correlations to reading when compared to form recall and form recognition modalities of vocabulary knowledge, as measured by L1-L2 multiple choice tests and written form recognition Yes/No tests, respectively.

Relatively less conclusive findings have been reported regarding written modalities of vocabulary knowledge as they relate to L2 listening ability. As listening is essentially a receptive aspect of language proficiency, theoretically meaning recall vocabulary knowledge could also be a strong predictor of listening ability. Despite this, Zhang and Zhang’s meta-analysis more tentatively suggested that Form Recall\(1\) could possibly have a higher correlation to listening \( [r = 0.63 (95\% \text{ CI} = 0.53 – 0.72)] \) when compared to meaning recognition \( [r = 0.50 (95\% \text{ CI} = 0.41 – 0.58)] \).

Although it seems plausible that an L2 to L1 modality of vocabulary knowledge such as meaning recall could have stronger correlations to the receptive forms of language proficiency such as listening, Zhang and Zhang’s finding that form recall could potentially be the superior predictor of listening ability was in alignment with a competing theory proposed by Laufer and Goldstein (2004) and Laufer et al. (2004). These studies found that written form recall was the most difficult form of vocabulary knowledge. Laufer et al. (2004) combined these four forms of vocabulary knowledge into a single unidimensional scale under the Rasch model. Under this framework, the mastery of “stronger” (i.e., more difficult) forms of vocabulary knowledge would imply mastery of “weaker” (i.e., less difficult) forms of vocabulary knowledge. This means that stronger forms of knowledge could have stronger correlations to other aspects of language proficiency more generally, even in cases where weaker forms of knowledge may appear to have stronger theoretical links to them, such as receptive vocabulary knowledge and its relationship to reading. However, Zhang and Zhang’s results were inconclusive. This was because although the difference between meaning recall and meaning recognition was statistically significant, the difference between form recall \( [r = 0.63 (95\% \text{ CI} = 0.53 – 0.72)] \) and meaning recall \( [r = 0.58 (95\% \text{ CI} = 0.54 – 0.62)] \) was not. Therefore, there is still doubt as to whether written receptive form recall or written receptive meaning recall is the better predictor of listening.
An important caveat must be mentioned when discussing modalities of vocabulary knowledge in relation to listening ability. It is very likely that an auditory vocabulary test using spoken forms of words would have the strongest correlation to listening ability; evidence suggests that learners’ written and spoken receptive vocabulary knowledge differ significantly (Masrai, 2020; Milton et al., 2010, 2013; Milton & Hopkins, 2006; Mizumoto & Shimamoto, 2008; Uchihara & Harada, 2018). Summarizing the results of such research, Zhang and Zhang’s (2020) meta-analysis found that auditory-modality vocabulary tests had an average correlation of 0.6 to listening, compared to 0.52 for orthographic (written) modality tests. Although the difference was statistically insignificant, this lends further support to the hypothesis.

However, the existing spoken receptive levels tests have their limitations, most notably in their item formats. The Aural Lex (Milton & Hopkins, 2006) utilizes a spoken Yes/No format that does not require learners to demonstrate knowledge of the form-meaning link. The spoken receptive meaning recognition (multiple-choice) Listening Vocabulary Levels Test (LVLT) is only available for Japanese (McLean et al., 2015), Chinese (Zhang & Graham, 2020) and Vietnamese (Ha, 2021) learners, as tests that expect to measure spoken receptive meaning-recognition lexical knowledge should present answer options in the learners’ L1 so as not to confound L2 written receptive and spoken receptive ability. However, while future research should strive to utilize appropriate tests, it is not always possible.

Cautions about using written tests of vocabulary knowledge to predict forms of proficiency that involve spoken vocabulary should be heeded. As Beglar (2010) advised, using written meaning recognition format tests such as the VST to measure listening vocabulary size “is not recommended as reading and listening vocabulary sizes can vary considerably” (p. 114). However, there may be value in determining if there is a single vocabulary test type that has, on average, the highest correlation to a wide range of L2 competencies, including listening, reading, writing and speaking. While it is highly unlikely that one test format will be superior in all situations, it may be possible to identify the best trade-off in terms of practicality and efficacy in situations where researchers wish to measure proficiency on multiple skills, but only have limited time to test vocabulary knowledge. Perhaps such an item format could be identified as an ideal overall measure of general L2 vocabulary knowledge.

In this paper, we will re-examine the data from Mclean et al. (2020) by correlating the aforementioned aspects of written vocabulary knowledge (meaning recall, form recall, meaning recognition and Yes/No) to listening proficiency as measured by the Listening section of the TOEIC test, using the same bootstrapping method as the original paper. A major limitation of this study is that the examined data set does not include the testing of spoken forms of words. Nor does it include measures of the productive skills of spoken and written proficiency. However, the results may provide evidence that undermines or lends further support to some of Zhang and Zhang’s statistically insignificant findings and, at a minimum, shed some light on how written modalities of form, meaning, recall and recognition can affect predictions of listening and reading ability and confirm if differences exist depending on whether reading or listening is
tested. Future studies can examine if the findings related to these modalities are generalizable to the same modalities as applied to auditory tests of spoken word forms.

2 Method

One hundred and three learners took four 1,000-item vocabulary tests of the third 1,000 most frequent words according to the New General Service List (NGSL; Browne et al., 2013). Each of the four tests examined a different modality of vocabulary knowledge: meaning recall, meaning recognition, form recall and form recognition (as measured by a Yes/No test). The 1,000 items were sampled with replacement to create 1,000 100-item tests, the scores of which were correlated to the learners’ scores on the Listening section of the TOEIC test. For further details on these tests, the tested sample of learners, and test administration, please refer to McLean et al. (2020).

3 Analysis

One thousand 100-item tests were generated by sampling with replacement from the McLean et al. (2020) data set for each of the modalities in question: written meaning recall, written meaning recognition, written form recall and written Yes/No. These test scores were then correlated to learners’ TOEIC Listening scores, resulting in 4,000 correlations. This process was then repeated in order to compare the same conditions of correlations to TOEIC Reading scores. Following the procedure of McLean et al. (2020), a factorial $2 \times 4$ analysis of variance (ANOVA) was then conducted with the predictor variables “Channel” (listening or reading proficiency) and “Modality” (vocabulary knowledge type). The results of this ANOVA, descriptive statistics and post-hoc tests of the conditions can be seen below in Tables 1–3.

Channel (listening or reading proficiency) had a statistically significant effect on correlations, with correlations to listening being slightly lower than correlations to reading. However, this difference was very slight, with average correlations only 0.012 higher for reading, and a partial eta-squared of 0.119, indicating a negligible effect size. The effect of modality (type of vocabulary knowledge) on correlations to reading and listening proficiency was substantially higher, with a large partial eta-squared of 0.892. The interaction effect between channel and modality was statistically significant, but, as with channel, very slight in terms of effect size.

The examined forms of vocabulary knowledge displayed the same ranks of correlational strength to both listening and reading proficiency. As can be seen in Table 2, the written meaning recall had the strongest correlation to both listening and reading, at $r = 0.765$ and 0.778 respectively. For both forms of proficiency, meaning recall was followed by form recall, meaning recognition and Yes/No, in that order.

The differences between correlations, where each $r$ value was a case, were significant for all conditions, as can be seen in Table 3. This includes
comparisons between written form recall and written meaning recall and listening proficiency. Written meaning recall’s correlation to listening was 0.765 compared to 0.745 for form recall, with a Cohen’s $d$ effect size of 1.41.

4 Discussion

The findings show that correlations of forms of written vocabulary knowledge to listening skills are very similar to their correlations to reading proficiency, albeit slightly lower. In contrast to Zhang and Zhang’s (2020) findings, meaning recall, rather than form recall, was the best predictor of listening proficiency. Furthermore, unlike in Zhang and Zhang’s study, the difference between the two was statistically significant ($p < 0.001$), with a Cohen’s $d$ effect size of 1.41. Both Zhang and Zhang and McLean et al. (2020) found that meaning recall was superior to meaning recognition as a predictor of reading ability. Although Zhang and Zhang reported the same relationship with regard to listening ability, the result was statistically insignificant. The current study reconfirms the finding to a statistically significant degree, at least with regard to written tests. A caveat to this claim is that Zhang and Zhang considered meta-analytic data while this current study presents correlations where each acts as a case in parametric testing.
<table>
<thead>
<tr>
<th></th>
<th>Mean difference</th>
<th>SE</th>
<th>t</th>
<th>P Tukey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening, Form Recall</td>
<td>Reading, Form Recall</td>
<td>0.003</td>
<td>7.175e-4</td>
<td>4.056</td>
</tr>
<tr>
<td></td>
<td>Listening, Form Recognition</td>
<td>0.095</td>
<td>7.175e-4</td>
<td>132.818</td>
</tr>
<tr>
<td></td>
<td>Reading, Form Recognition</td>
<td>0.086</td>
<td>7.175e-4</td>
<td>120.260</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recall</td>
<td>−0.020</td>
<td>7.175e-4</td>
<td>−27.260</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recall</td>
<td>−0.032</td>
<td>7.175e-4</td>
<td>−45.044</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recognition</td>
<td>0.074</td>
<td>7.175e-4</td>
<td>103.216</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>0.046</td>
<td>7.175e-4</td>
<td>63.872</td>
</tr>
<tr>
<td></td>
<td>Listening, Form, Recognition</td>
<td>0.092</td>
<td>7.175e-4</td>
<td>128.762</td>
</tr>
<tr>
<td></td>
<td>Reading, Form Recognition</td>
<td>0.083</td>
<td>7.175e-4</td>
<td>116.205</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recall</td>
<td>−0.022</td>
<td>7.175e-4</td>
<td>−31.316</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recall</td>
<td>−0.035</td>
<td>7.175e-4</td>
<td>−49.099</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recognition</td>
<td>0.071</td>
<td>7.175e-4</td>
<td>99.160</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>0.043</td>
<td>7.175e-4</td>
<td>59.817</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recall</td>
<td>−0.009</td>
<td>7.175e-4</td>
<td>−12.557</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recall</td>
<td>−0.106</td>
<td>7.175e-4</td>
<td>−147.521</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recognition</td>
<td>−0.128</td>
<td>7.175e-4</td>
<td>−177.861</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>−0.021</td>
<td>7.175e-4</td>
<td>−29.602</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>−0.049</td>
<td>7.175e-4</td>
<td>−68.945</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recall</td>
<td>−0.106</td>
<td>7.175e-4</td>
<td>−147.521</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recall</td>
<td>−0.119</td>
<td>7.175e-4</td>
<td>−165.304</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recognition</td>
<td>−0.012</td>
<td>7.175e-4</td>
<td>−17.045</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>−0.040</td>
<td>7.175e-4</td>
<td>−56.388</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recall</td>
<td>−0.013</td>
<td>7.175e-4</td>
<td>−17.783</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>0.094</td>
<td>7.175e-4</td>
<td>130.476</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>0.065</td>
<td>7.175e-4</td>
<td>91.133</td>
</tr>
<tr>
<td></td>
<td>Listening, Meaning Recognition</td>
<td>0.106</td>
<td>7.175e-4</td>
<td>148.259</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>0.078</td>
<td>7.175e-4</td>
<td>108.916</td>
</tr>
<tr>
<td></td>
<td>Reading, Meaning Recognition</td>
<td>−0.028</td>
<td>7.175e-4</td>
<td>−39.344</td>
</tr>
</tbody>
</table>

Note. *p*-value adjusted for comparing a family of 8. All vocabulary knowledge modalities are written form.

McLean et al. (2020) proposed two theories for why meaning recall could outperform meaning recognition as a correlate of reading, which contrasted with the arguments from some researchers that meaning recognition could have more applicability to the skills required for reading (Laufer & Aviad-Levitzky, 2017). First, fixed options could introduce a source of error not contained in recall item formats due to guessing effects. Second, working under a continuum/cline theory of vocabulary strength (Laufer et al., 2004; Stewart et al., 2012), stronger forms of vocabulary knowledge could subsume weaker forms, meaning a learner who has mastery over form recall would also have mastery over meaning recall and meaning recognition, meaning form recall could therefore...
have equal or better correlations not only to productive language skills but also receptive language skills such as reading and listening.

However, the present findings indicate that even if accurate with regard to meaning recall and meaning recognition, this theory does not appear to extend more broadly to form recall modalities. When placed in comparison to one another, recall of meaning outperforms recall of form, and both recall modalities examined in this study outperform both recognition modalities, regardless of whether form or meaning is tested. This implies that the distinction between recall and recognition exerts a stronger influence on correlations to other aspects of language proficiency than the distinction between form and meaning.

The results suggest that when possible, recall format tests are preferable measures of vocabulary knowledge when attempting to relate vocabulary knowledge to other language proficiency skills, particularly for research that must rely on sensitive measurement instruments. Although the differences in correlation between formats may appear to be small, these differences can become important in hypothesis testing, particularly in experiments that are underpowered, a common occurrence in the field of second language acquisition (e.g., Nicklin & Vitta, 2021). More sensitive and reliable measures can help to guard against Type II error, where researchers falsely reject correct hypotheses due to a lack of statistical significance.

With regard to testing learner vocabulary in pedagogical contexts, the tradeoff of the higher predictive validity and internal reliability (McLean et al., 2020) of recall item formats must be balanced with the relative difficulty in administering and scoring recall-format tests relative to recognition-format tests. Browser-based vocabulary tests such as Vocableveltest.org (McLean & Raine, 2018), which can be completed at computer terminals or on learners’ smartphones, can help mitigate many of these issues. In addition to compiling learners’ answers in digital format, the software can automatically score whitelisted and blacklisted answers and flag novel responses for manual grading. With each successive administration of the test, fewer and fewer responses require manual scoring. The results of the current study also suggest that form recall tests represent a tradeoff that further simplifies the marking of recall responses while maintaining a relatively strong correlation to receptive language proficiency, as form recall also outperforms meaning recognition measures as a correlate to receptive language proficiency. Scoring can be further simplified if learners can provide the L2 word form rather than a meaning as it greatly limits the number of possible answers. It is also possible that, when used as a correlate to general language proficiency, form recall’s somewhat lower correlations to receptive language proficiency could be balanced by a higher correlation to productive language skills. Future research should attempt to replicate these results regarding form, meaning, recall and recognition modalities using an auditory test format of spoken word forms, as the modalities examined in the current study were all used written forms. It could also be beneficial for future studies to examine the correlations of meaning recall and form recall in relation to speaking and writing, in order to examine and compare their overall predictive power with regard to general language proficiency.
Note:

1. In Zhang and Zhang (2020) form, recall includes dictation modalities where learners have to dictate the word they hear. Dictation item formats do not require learners to demonstrate knowledge of the spoken form meaning link.

References


Stewart et al.: Correlations of vocabulary to listening and reading


Vocabulary Learning and Instruction, 10(2), 55–63.
The Internal Consistency and Accuracy of Automatically Scored Written Receptive Meaning-Recall Data: A Preliminary Study

Stuart McLeana, Paul Raineb, Geoffrey Pinchbecket, Laura Hustond, Young Ae Kime, Suzuka Nishiyamaa, and Shotaro Uenof

aMomoyama Gakuin University; bKeio University; cCarleton University; dJosai International University; eKyoto Seika University; fHirakata Junior High School

Abstract

Vocabularytest.org is a testing platform on which users can create online self-marking meaning-recall (reading or listening) and form-recall (typing) tests that address a number of limitations of the existing vocabulary level tests and vocabulary size tests. A major limitation of many existing vocabulary tests is the written receptive meaning-recognition (multiple-choice or matching) format which is associated with increased error due to guessing and decreased power to measure the type of vocabulary knowledge suitable for reading practice (McLean et al., 2020; Stewart et al., 2021a; Stoeckel et al., 2021), despite being designed for this purpose (Nation, 2012; Schmitt et al., 2020; Webb et al., 2017). Conversely, scoring meaning-recall tests by hand is labour-intensive, and the internal consistency and accuracy of automatically marked data are unknown. Thus, this study investigated the internal consistency and accuracy of automatically marked responses of 98 words from the fifth 100 most frequent words of English. This study tested for knowledge of high-frequency words as a more robust test of the marking system, as these words possess multiple-meaning senses, making their automatic marking problematic. Furthermore, the predicted limited range of learners’ knowledge of these 98 words was expected to result in data of a low internal consistency. However, the automatically marked data had a high internal consistency (Cronbach’s α = 0.868) and was 98% similar to human marked meaning-recall responses.

Keywords: meaning-recall, automatic marking, accuracy

1 Background: Addressing the Limitations of Existing Vocabulary Levels and Size Tests

Many of the most commonly used vocabulary levels and size tests (hereafter levels tests) are based on word families, and sample between 5 and 30 meaning-recognition (multiple-choice or matching format) items to represent 1,000 words. Levels test design has seen few innovations and has been the focus of recent critical scrutiny (Kremmel, 2016; Schmitt et al., 2020; Stewart et al., 2021a; Stoeckel et al., 2021b; Stoeckel et al., 2021). Vocabularytest.org (McLean & Raine, 2018)
addresses some of the known limitations of levels tests and allows teachers and researchers (hereafter teachers) to create online self-marking levels tests to their required parameters. Once tests are created, teachers are provided with web addresses and QR codes that are shared with learners so that they can complete levels tests. Learners can be provided with feedback, and teachers can download actually typed responses, dichotomously scored responses, and the time taken to complete each response.

1.1 Item Format

When a receptive levels test is administered through the written receptive modality (as opposed to spoken), the conventional assumption has been that the test does in fact measure the examinee’s knowledge of vocabulary required for reading, rather than listening (Beglar, 2010; McLean & Stoeckel, 2021; Nation, 2012). Figure 1 illustrates the four main vocabulary test item types. Both theory and research support the use of written receptive meaning-recall items to measure the type of lexical knowledge that can be employed when reading (Aviad-Levitzky et al., 2019; McLean et al., 2015a, 2020; Stoeckel et al., 2019; Zhang & Zhang, 2020).

![Figure 1. Form-Meaning Link Vocabulary Item Types.](image)
Vocabulary Learning and Instruction, 10(2), 64–81.

Vocableveltest.org allows teachers to create written (Figure 2) and spoken (Figure 3) receptive meaning-recall tests, as well as written productive form-recall tests (Figure 4). Vocableveltest.org automatically marks learners’ responses using an extensive bank of possible valid responses collected from dictionaries and the inspection of learners’ responses through two methods (Figure 5). In the first method, incorrect responses are inspected and valid responses are added to the answer bank. In the second method, teachers give feedback on completed tests with automatically scored responses. Teachers’ suggestions for test bank changes are stored and presented to site administrators, who can supplement and edit the answer bank. The written and spoken versions of the receptive meaning-
recall tests can now be completed in Japanese, Vietnamese, French, Chinese, Dutch, and Arabic.

1.2 Word Lists

Existing levels tests have been based on a limited number of word lists. When matching learners with vocabulary-level-appropriate materials (McLean,
2014), the use of knowledge-based word lists is preferable to a frequency list (Paul Nation, personal communication, August 8, 2021; Schmitt et al., 2021). Knowledge-based lists rank words according to how well words are known within a given population. Vocableveltest.org facilitates the creation of levels tests from nine different word lists (Figure 6). The Scale of English Word Knowledge—Japan (SEWK-J) is a list derived from a predictive model of English word knowledge for native Japanese speakers. Please see Mizumoto et al. (2021) for a description of the parallel text profiler that uses the same word list to estimate the difficulty of candidate-text vocabulary.

1.3 Word Counting Unit

In L2 English research, the lexical units most often discussed are as follows: (a) the “type,” any specific orthographic form (e.g., use); (b) the “lemma,” comprised of a base word of a particular part of speech (POS) and its inflectional forms (useverb, usedverb, usesverb, and usingverb); (c) the “flemma,” a base word form and inflectional forms, regardless of POS (useverb, usedverb, usedadjective, usesverb, usingverb, useadjective, usedadjective, usesverb, usingverb, useadjective, usefuladjective, usefullyadverb, uselessnessnoun, uselessadjective, uselessadverb, usernoun, usersnoun, usesverb, and usingverb). “Flemma” and “lemma” are terms that have sometimes been used to refer to the same thing.

Views differ on the appropriateness of different lexical units with different learners for different purposes, with some supporting the use of WF6 (Laufer & Cobb, 2020; Laufer et al., 2021). Others the flemma or lemma (Brown et al., 2020, 2021; Kremmel & Schmitt, 2016; McLean, 2018, 2021; McLean & Stoeckel, 2021; Mochizuki & Aizawa, 2000; Stewart et al, 2021a; Stoeckel...
et al., 2020, 2021; Ward & Chuenjundaeng, 2009). While the majority of the evidence supports the use of the flemma or lemma with some EFL and ESL learners, the WF6 is appropriate with native English speakers and in some EFL and ESL settings (e.g., Northern Europe). Thus, Vocableveltest.org allows teachers to select lists based on various lexical units (Figure 6).

### 1.4 Band Sizes and Number of Bands

Levels tests have traditionally been based on 1,000-word bands, a practice for which the rationale has not been explained. Kremmel (2016) and McLean (2021) argue for the adoption of 500-word bands for high-frequency words as these words provide such a great deal of coverage. Furthermore, for beginning learners with gaps in their knowledge of high-frequency words, the most frequent 1,000 words might never be mastered. In a survey of 3,427 Japanese learners, McLean et al. (2014) found that even learners from university departments with a hensachi of 61 and over (a high rank in Japan, based on average scores for standardised academic tests) did not demonstrate mastery of the first 1,000 words of English (Figure 7). Vocableveltest.org facilitates the creation of levels tests at 100-, 250-, 500-, and 1,000-word band sizes (Figure 8). Teachers can specify which bands they want their levels test to cover. For example, a teacher who wants to check if speed-reading materials written at the 1,000-word level are lexically appropriate only needs to test learners’ knowledge up to the 1,000-word level. If the teacher believes that their learners have already mastered the first 500 words of English, it would not be necessary to test learners’ knowledge of the first 500 words of English. If learners’ knowledge of only a few bands is tested, more items can be deployed in the test to better represent those target bands.

![Figure 7. A Scatter Plot of VST Item Accuracy at Each 1,000-Word Frequency Level for the Three hensachi Groups with Best-Fit Lines.](image)

*Vocabulary Learning and Instruction, 10*(2), 64–81.
1.5 Sample Size

In practice, teachers and researchers cannot require learners to complete thousands or tens of thousands of items. Thus, levels tests present learners with samples of between 5 and 30 items which represent target word bands of 1,000 or 560 words. While it is clear that the number of items used to represent a word band affects how representative a sample is, in the case of levels tests that sample from word bands, sampling also influences the reliability, accuracy, and construct validity of the test.

Reliability is defined as the “consistency of measurement” (Bachman & Palmer, 1996, p. 19). Thus, a reliable sample from 1,000 words is a number of items that, if reselected, does not result in a significantly different estimate of a learner’s vocabulary knowledge. For example, if 30 items can represent a reliable sample, no two sets of 30 randomly sampled items (or sampled in a stratified way) will result in significantly different test scores from the same learner. Figures 9 and 10 show data from a learner who correctly answered 750 of 1,000 items representing the third 1,000-word band of English. From this data, samples of 5, 10, 20, 50, and 100 items from the third 1,000-word band were selected. The frequency (Y-axis) of the different resulting knowledge estimates (X-axis) are shown in Figures 9 and 10. The accuracy of the test is therefore a function of the sample size of words that represent the target band, and the limited representativeness, accuracy, and reliability of a sample size reduce the construct validity of both a sample size and a test. Figure 11 suggests that even samples of 100 or 200 items can occasionally result in inaccurate estimates. However, the degree of inaccuracy or value of adding more items declines significantly from 40-item tests. Vocableveltest.org allows users to select the number of items that is most appropriate for their setting (Figure 12).
Figure 9. Monte Carlo Study of Vocabulary Size Estimates Using Tests of 5, 10, and 100 Items (Adapted from Gyllstad, McLean, & Stewart, 2021).

Note: The true number of words known by this learner is 750 (black line).

Figure 10. Monte Carlo Study of Vocabulary Size Estimates Using Tests of 20, 50, and 100 Items (Adapted from Gyllstad et al., 2021).

Note: The true number of words known by this learner is 750 (black line).
Figure 11. Mean Difference in Scores Between Learner’s True Score on 1,000 Item Tests and Estimates from Bootstrapped Samples (adapted from Gyllstad et al., 2021).

Note: The data presented in this figure was from 103 participants. Thus, each data point represents the mean inaccuracy of 103,000 vocabulary knowledge estimates relative to a learner’s true score.

Figure 12. A Screenshot Showing Sample Size Selection.

1.6 Customised Level Tests, Pretests, and Posttests

Usually, Vocableveltest.org will randomly select items within the parameters selected by the teachers. However, teachers can opt to select which items will be present in tests they create from over 7,000 items. Thus, teachers can customise level tests, pretests, and posttests.

Vocabulary Learning and Instruction, 10(2), 64–81.
1.7 Feedback

1.7.1 Feedback for students

The existing levels tests do not automatically provide feedback to learners. Upon completion of levels tests on Vocableveltest.org, if teachers select the feedback option, students are provided with item-level feedback (Figure 13), and the percentage of correct responses at each band (Figure 14).

1.7.2 Feedback for teachers

Teachers can view the same feedback that students view, for all of their learners (Figure 13). Teachers can view mean scores for all learners who have completed a single test (Figures 15), thereby helping teachers quickly and simply estimate a class’s level of lexical mastery. Teachers can download an Excel sheet of the learners’ (a) typed responses, (b) dichotomously-marked responses, (c) the time taken to complete each response, and (d) class name and standardised test scores.

![Feedback](image)

Figure 13. An Image of Feedback for Learners.
To reduce possible cheating, Vocableveltest.org has the four following features. Firstly, the text within the site cannot be automatically translated by computer or smartphone browsers. Secondly, the text within the test item stems cannot be copied. Thirdly, text cannot be pasted into answer boxes. Fourthly, the learners must complete the items within 20 seconds.
2 Methods

The present study asks the following research questions.

1. What is the internal consistency of responses that are automatically marked by Vocableveltest.org relative to human markers?
2. Can Vocableveltest.org automatically mark written receptive meaning-recall data accurately?

2.1 Participants

The participants were 78 female Japanese university students at a private university in Western Japan. The participants came from two English classes and elected this English class from several available English classes. The participants were within a range of English proficiencies. Test of English for International Communication (TOEIC®) test scores ranged between 300 and 700. All participants agreed to the use of their data in this study.

2.2 Instruments

2.2.1 Target word selection

The participants completed 98 items from the fifth 100 words of the New JACET8000 list. At the time of data collection, 98 of the fifth 100 words (401–500) of New JACET8000 were words that could be tested through Vocableveltest.org (excluding meeting and following, as at the time of data collection these words did not have item stems).

The internal consistency of vocabulary test data is often the product of the number of items tested and the range of knowledge of items among the participants. Thus, a limited and ecologically-valid number of items was tested: 98 items is 60% the length of commonly used tests (McLean et al., 2015b; McLean & Kramer, 2015, 2016). Secondly, as all 98 items were from the fifth 100-word band, it was expected that the learners would be familiar with these items and would therefore be homogenous in their knowledge of items, reducing the variance in the data and its internal consistency. The high mean scores and limited standard deviations (Table 1) support these two assumptions.

We wanted to rigorously test the automatic scoring accuracy against human raters. Skipped items are unambiguously incorrect. Thus, if this item sample included a large number of low-frequency words which would likely be unknown to students and therefore skipped, it would result in an artificially high similarity of marking between the Vocableveltest.org automatic scoring and the human raters. Instead, high-frequency words, which were less likely to be skipped, were selected. Furthermore, a second reason for using high-frequency words in this study was that they are often associated with multiple possible meanings, leading to multiple valid typed L1 responses for each test item. This serves as a robust test of Vocableveltest.org’s ability to accurately mark meaning-recall levels test responses.
2.2.2 Item presentation

The meaning-recall items were completed on Vocableveltest.org. The website presents learners with a non-defining context sentence with the target word bolded and underlined (Figure 1). Before completing the test, test-takers read instructions and complete questions that encourage learners to consider and express the part of speech and affixes within the target forms.

2.3 Procedures

Each week the participants completed target items within each 100-word band of the NEW JACET 8000 with feedback on answers. The participants submitted a screenshot of the scores, and wrote unknown words in lexical journals which were submitted as homework and used when conducting writing tasks to encourage recycling of previously unknown words. The first week of the semester, the participants completed the target items.

2.4 Marking

The responses from the 78 participants were downloaded from Vocableveltest.org, and the automatically marked dichotomous data was used. The participant-typed responses were presented to two native Japanese speakers, Marker 1 and Marker 2, teachers of English, who dichotomously scored the responses. The two markers were instructed to score responses that demonstrated knowledge of the target word including any affixes and any meaning-senses for the target word as correct.

3 Results and Discussion

3.1 Internal Consistency

The internal consistency of the hand-marked data by Marker 1, Marker 2, and Vocableveltest.org was Cronbach’s $\alpha = 0.863$, 0.858, and 0.869, respectively. Under Nunnally’s (1978) guidelines, an $\alpha = $ value of 0.80 is required for tests used in basic research, and a value of at least 0.90 is advisable for applied settings, although a value of 0.95 or higher is ideal. Considering the limited number of participants, the number of items, and the high degree of homogeneity of learners’ knowledge of the items, the computer marking yielded reasonably high internal consistency, which is slightly higher than the human markers.
3.2. Marking Accuracy

Table 2 shows that the inter-rater reliabilities among the two markers and automatically marked data were sufficient for research purposes. Tables 2 to 5 show the degree of similarity in marking between the three marking methods. This degree of similarity provides evidence that Vocableveltest.org can mark data similar to human markers. The discrepancies between marking are due to two main causes. Firstly, the participants added particles to nouns. For example, in response to the stem, “He is in a hospital”, some learners added に after locations (e.g., 病院, hospital), or を after object nouns, which the human markers marked as correct, but Vocableveltest.org marked as incorrect. Secondly, Vocableveltest.org’s answer bank included some responses that the human markers scored as incorrect. For example, in response to the target word best (stem: He is the best

<table>
<thead>
<tr>
<th>Marker</th>
<th>Marker 1</th>
<th>Vocableveltest.org</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker 1</td>
<td>0.874</td>
<td>0.959</td>
</tr>
<tr>
<td>Marker 2</td>
<td>0.853</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Inter-rater Reliability (Kappa) Figures

<table>
<thead>
<tr>
<th>Vocableveltest.org</th>
<th>Incorrect</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker 1 Incorrect</td>
<td>518 (6.777%)</td>
<td>61 (0.798%)</td>
</tr>
<tr>
<td>Correct</td>
<td>76 (0.994%)</td>
<td>6,989 (91.431%)</td>
</tr>
</tbody>
</table>

Table 3. Degree of Agreement between the First Marker and Automatic Marking

<table>
<thead>
<tr>
<th>Vocableveltest.org</th>
<th>Incorrect</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker 2 Incorrect</td>
<td>512 (6.698%)</td>
<td>79 (1.033%)</td>
</tr>
<tr>
<td>Correct</td>
<td>82 (1.073%)</td>
<td>6,971 (91.196%)</td>
</tr>
</tbody>
</table>

Table 4. Degree of Agreement between the Second Marker and Automatic Marking

<table>
<thead>
<tr>
<th>Vocableveltest.org</th>
<th>Incorrect</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker 1 Incorrect</td>
<td>563 (7.365%)</td>
<td>16 (0.209%)</td>
</tr>
<tr>
<td>Correct</td>
<td>28 (0.366%)</td>
<td>7,037 (92.059%)</td>
</tr>
</tbody>
</table>

Table 5. Degree of Agreement between the First and Second Marker

Vocabulary Learning and Instruction, 10(2), 64–81.
guitar player) the responses 良い (good) and 優れた (excellent) were scored as correct by Vocableveltest.org and incorrect by both of the human markers.

As shown in Table 5, the human raters scored more similarly to each other (98.8% agreement) than to the automatic scoring system. The automatic system was nevertheless very similar to the human raters. Of 7,644 responses automatically marked, 7,483 (97.894%) and 7,507 (98.208%) were scored in the same way as Marker 1 and Marker 2, respectively. The discrepancies between automatic marking and human raters are the result of inconsistencies between the answer bank and the human raters’ decisions, which can largely be resolved by ongoing updates to the answer bank, and/or by providing marking instructions and/or calibration training to raters. It is also important to note that the advantages of recall tests over recognition tests outweigh the small differences observed between the human and automatic rating, particularly when vocabulary tests are being used as a proxy for reading proficiency. Stewart et al. (2021a) found that the Pearson’s correlation between data from 30 written receptive meaning-recall items and TOEIC reading ($r = 0.74$) was significantly stronger ($p \leq 0.001, d = -3.622$) than that of 30 written receptive meaning-recognition items and TOEIC reading ($r = 0.65$). Thus, while the limitations of automatically marked data are salient, they seem small relative to the implicit limitations of meaning-recognition items, which offer sub-optimal construct validity. Thus, we would argue that the initial investment required to produce this platform has been worthwhile.

4 Conclusion

This article introduces Vocableveltest.org and explains how it addresses a number of limitations of the existing levels tests. In this study, student participants were tested on their knowledge of high-frequency words, and these responses were then used to evaluate the automatic scoring system, which was compared to human markers. In this preliminary study, it was found that automatically marked data was found to have high internal consistency ($\alpha = 0.869$), which was slightly higher than two human markers ($\alpha = 0.858$ and 0.863). The 7,644 automatically marked responses were found to agree 97.894% (7,483 responses) and 98.208% (responses) of the time with the two human markers. We will continue to work with teachers and other researchers to provide a user-friendly vocabulary testing platform that continues to be improved and optimised for different groups of learners.

Acknowledgments

This research was made possible by a Grant-in-aid for Scientific Research (20K00792 and 20K00898) from the Japan Society for the Promotion of Science.

Note:

1. This is the updated version of JACET8000 list (JACET, 2003), compiled by the Japan Association of College English Teachers (JACET). The New JACET8000 can be downloaded from Dr. Shin Ishikawa’s website. JACET reserves the copyright to the list (JACET, 2016).
References


JACET Kihongo Kaitei Iinkai (JACET, Committee for Revision of the JACET Wordlist). (2003). JACET list of 8000 basic words. JACET.


McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalised. *Reading in a Foreign Language, 33*(1), 126–140. https://nflrc.hawaii.edu/rfl/item/528


*Vocabulary Learning and Instruction, 10*(2), 64–81.


Considerations and Challenges in Longitudinal Studies of Lexical Features in L2 Writing

Minkyung Kim
Nagoya University of Commerce and Business

Abstract
Exploring the longitudinal development of second language (L2) lexical use has been one of the important topics in L2 vocabulary research. One approach to examining longitudinal changes in L2 lexical use is to capture changes in lexical features as found in learner production, such as L2 writing, over time. To further facilitate this approach, the purpose of this article is to discuss considerations and challenges for conducting longitudinal studies on lexical features in L2 writing. The article first provides a summary of relevant previous studies, followed by the promise of longitudinal studies on lexical features in L2 writing. It then presents considerations and challenges in longitudinal studies of lexical features in L2 writing in terms of the data collection and analysis and the choice of lexical measures. More research on the longitudinal changes in lexical features in L2 learner production seems warranted. Ultimately, more longitudinal research in lexical features in L2 learner production will help us have a deeper understanding of L2 lexical development and design better vocabulary intervention in L2 classrooms.

Keywords: lexical features; lexical richness; longitudinal study; L2 writing

1 Introduction
In second language (i.e., language other than mother tongues; L2) vocabulary research, one of the important topics is to examine longitudinal developmental patterns of L2 vocabulary (lexical) use. One approach to exploring it is to track changes in lexical features as found in learner production, such as L2 writing, over time under the assumption that lexical use can be indicative of the underlying L2 vocabulary ability. An increasing body of research has examined developmental patterns of L2 lexical use in L2 writing (e.g., Gené-Gil, Juan-Garau, & Salazar-Noguera, 2015; Knoch et al., 2015; Mazgutova & Kormos, 2015; Polio & Shea, 2014; Verspoor et al., 2008; Zheng, 2016). The purpose of this article is to discuss considerations and challenges for conducting longitudinal studies of lexical features in L2 writing. In the following section, I first present previous studies that examined lexical features in L2 writing.

2 Previous studies on lexical features in L2 writing
While there are various conceptualisations of lexical features, this article focuses on Read’s (2000) conceptualisation of lexical richness that includes...
Kim: Considerations and Challenges in Longitudinal Studies

Vocabulary Learning and Instruction, 10(2), 82–90.

lexical density (i.e., proportion of content words [i.e., verbs, nouns, adjectives, and adverbs] as compared to function words [e.g., articles and prepositions]), lexical diversity (i.e., use of different types of words), lexical sophistication (i.e., use of advanced and difficult words), and lexical accuracy (i.e., errors in the use of words). Much cross-sectional research has investigated the relationships between lexical richness and L2 writing quality. While lexical density has shown little relation with L2 writing quality (Engber, 1995), lexical diversity, sophistication, and accuracy have shown close relations with L2 writing quality, such that higher-rated L2 texts tend to include more diverse words (e.g., Crossley & McNamara, 2012), more sophisticated words (e.g., Kim et al., 2018; Laufer & Nation, 1995), and fewer lexical errors (Crossley et al., 2019).

On the other hand, the longitudinal changes in lexical features in L2 writing remain more open. With respect to lexical density, a few L2 writing studies examined longitudinal changes (Zheng, 2016). Zheng examined lexical density over a year in Chinese upper-intermediate-level learners’ English argumentative writing and reported no significant change in lexical density. Similarly, a few L2 writing studies explored longitudinal changes in lexical accuracy (Llach, 2011; Polio & Shea, 2014). Polio and Shea (2014) examined lexical errors produced in English as a second language (ESL) university-aged students’ descriptive writing over a semester and reported no significant change in lexical errors. Llach (2011) explored lexical errors in the English letter writing of fourth grade Spanish EFL learners over 2 years and reported that the number of lexical errors, particularly for form-related errors, decreased.

On the other hand, changes in lexical diversity and sophistication over time have been increasingly examined in the L2 writing literature. Previous studies have found that lexical diversity in L2 writing tends to increase when investigation durations were more than 1 year (e.g., Gené-Gil et al., 2015; Verspoor et al., 2008; Zheng, 2016). For example, Gené-Gil et al. (2015) examined the English narrative writing of Spanish-Catalan secondary EFL learners over 3 years and reported an increasing pattern of lexical diversity over time. In a similar vein, developing patterns in lexical sophistication over time have been reported, such that L2 writers tend to use more advanced words over time (e.g., Mazgutova & Kormos, 2015; Storch & Tapper, 2009; Zheng, 2016). For example, Mazgutova and Kormos (2015) found that after 1-month-long intensive English course, undergraduate ESL students’ argumentative essays included more advanced content words (i.e., less frequently used content words). However, some other L2 writing studies have reported no changes in lexical diversity or sophistication over time when university-aged L2 learners’ writing was analysed for a shorter span of time (e.g., one semester; Bulté & Housen, 2014; Yoon & Polio, 2017) and when university-aged L2 learners did not participate in language-focused programmes during data collection periods (e.g., Knoch et al., 2015). These findings indicate that to observe L2 writers’ developmental patterns in lexical diversity and sophistication, researchers may need to investigate L2 writing for a longer span of time (at least 1 year) or involve L2-focused instruction.

In sum, while previous longitudinal L2 writing studies on lexical features have provided important information about the developmental patterns of L2
lexical use, it still seems that we do not know much of L2 lexical development, and more research is needed.

### 3 The promise of longitudinal studies on lexical features in L2 writing

Considering that more longitudinal studies are needed, to help researchers think about their potential research on longitudinal changes in lexical features, in this section, I present four main research domains.

First, more research can describe longitudinal changes in lexical features over time. While previous studies have examined longitudinal changes in lexical features, more research is needed to better understand developmental patterns of L2 lexical use. Researchers can focus on various levels of L2 proficiency, learners from various first language (L1) backgrounds, and learners from different school levels. In addition, when describing longitudinal changes, researchers should understand the importance of examining longitudinal development at not only group levels through group statistics but also individual levels through individual case studies (Lowie & Verspoor, 2019).

Second, research can explore longitudinal interrelationships among linguistic features. Changes in lexical features do not occur in isolation. Instead, they likely take place in relation to changes in other linguistic features, such as syntactic and cohesive features. For example, when writers make greater use of more information-dense, advanced words in a clause, the length of that clause may become shorter not to hold too much information in it. Studies from a Dynamic Systems Theory perspective (Larsen-Freeman, 2006; Verspoor et al., 2008) have focused on developmental relationships of linguistic features, suggesting four types of longitudinal relationships among linguistic features: support (i.e., jointly developing), competition (i.e., developing in an opposite direction), precursor (growth of one linguistic feature precedes that of another), and asymmetry (i.e., relations between linguistic features varying over time; Verspoor & van Dijk, 2011). Such relationships have not yet examined thoroughly, and more research is needed to shed light on the complex nature of longitudinal interrelationships among linguistic features.

Third, research can also examine longitudinal interrelationships between lexical features and L2 writing quality. Over time, it is ideal for L2 learners to develop both their L2 vocabulary ability and their L2 writing ability. In addition, L2 writing quality (indicative of L2 writing ability) tends to be influenced by linguistic features, including lexical features (indicative of L2 vocabulary ability), found in L2 writing samples. Thus, linking lexical features and L2 writing quality over time may be promising to answer a question of what lexical features predict L2 writing quality longitudinally. If some lexical features consistently predict L2 writing quality over time, they can be the focus of vocabulary intervention in L2 writing classrooms.

Lastly, after identifying longitudinal developmental patterns in lexical features, the next important step that researchers can take is to explain what factors lead to such developmental patterns. Researchers can examine learner-related
factors, such as cognitive strategies and motivation, and contextual factors, such as exposure to L2s and instruction. Essentially, explaining longitudinal changes in lexical features should be an important goal to figure out how to design and implement intervention that considers learner and contextual factors.

4 Considerations and challenges in longitudinal studies on lexical features in L2 writing

To conduct longitudinal studies on lexical features in L2 writing, there are many basic questions about the research design, which also impose some challenges. These questions should be answered prior to collecting and analysing longitudinal data. The answers to these answers will help guide the research design and impact the degree to which the results of the study can generalise to other settings and populations. In this section, I present a set of questions along with considerations and challenges associated with key research design decisions that need to be made. Of course, each research design decision will be guided by the research questions of the study.

Who is the target learner population? The learners’ age ranges and proficiency levels will depend on the research questions (e.g., adolescent or adult learners, beginning-level or advanced learners). The target learner population should be decided first to create writing tasks appropriate to that learner population.

How long is the research duration? The research duration will depend on the research questions of the study (e.g., whether interested in weekly, monthly, or yearly changing trends). If the duration is too short, it may be difficult to detect changes. On the other hand, longer research designs in duration may lead to higher attrition rates.

How frequently do learners write? The rates for how frequently learners produce writing samples will also depend on the research questions of the study. If researchers focus on learner variation over time, they can collect samples frequently (e.g., weekly), which may enable them to zoom in on various changing patterns including gradual changes, sudden spurts, and stagnation (Lowie & Verspoor, 2015). On the other hand, if researchers focus on learner development, they can collect samples less frequently (e.g., bi-monthly) but with a longer research design in duration because L2 development does not tend to take place in a short term. To track both learner variation and learner development, longer research durations with high frequency samples may be ideal but may not be always feasible.

What are writing tasks? It is important to consider the appropriateness of the writing tasks to the research questions. Choosing writing tasks must be meticulous because the choice of writing tasks (e.g., genre types, topics) influence learners’ lexical use (Crossley, 2020). For longitudinal data collection, if researchers use the same writing prompt for each data collection occasion, it may result in the confounding effect of the repeated practice on learner performance. That is, learners may perform better over time not because of their development, but because of their repeated practice. On the other hand, if researchers use different writing prompts across data collection occasions, they should make several key decisions.
First, the genre of the writing tasks must be consistent throughout the research duration because different genres (e.g., narratives vs. persuasion) will lead to different lexical use. Second, it is ideal to have at least two different prompts per data collection occasion to counterbalance and reduce the potential prompt effects. Lastly, it is important to be cautious when interpreting the results and comparing writing samples produced in response to different prompts (Read, 2000).

**Is there instruction during the data collection period?** The amount, intensity, and quality of L2 instruction during the data collection will influence the rates of changes in lexical features in L2 writing. Information about instruction needs to be collected, and instructional factors need to be considered when interpreting the results.

**What is the basic unit of lexical analysis?** For data analysis, basic units of lexical analysis can be tokens, lemmas (i.e., base forms of words, such as confirm for confirmed and confirms), and word families (i.e., word forms that share a common meaning, such as confirm for both inflected forms [e.g., confirmed] and derived forms [e.g., confirmation]). Generally, a widely used basic unit of lexical analysis is lemma, so that inflected forms and their base forms are counted as instances of the same lemma (Read, 2000).

**Are function words included for the analysis?** When measuring L2 vocabulary knowledge in receptive forms, in most cases, the focus of analysis is on content words. However, when examining lexical use in L2 writing, many previous studies included function words for analysis. Researchers need to decide whether to include function words for their analysis.

**Are words related to prompts included for the analysis?** An important consideration in analysing lexical features in written texts is whether to include words related to prompts. If learners use words that appear in the prompt in their texts, those words may not represent learners’ vocabulary ability, but rather reflect the prompt wording. Thus, while some previous studies did not control for prompt-based words, it may be advisable to remove content words that appear in the prompt prior to analysis.

## 5 Considerations and challenges in choosing measures of lexical features

In addition to the questions provided above, an important, but challenging, task is to choose measures for lexical features. Various measures have been used in assessing lexical features. In this section, I present brief explanations and considerations for the use of measures of the four main lexical features: lexical density, diversity, sophistication, and accuracy.

Lexical density is calculated by the proportion of content words to function words in the text. Content words tend to have specific semantic content, belonging to open classes of words (i.e., new members can be added), whereas FWs tend to have a non-conceptual meaning, belonging to closed classes of words (i.e., new members are rarely added; Corver & Van Riemsdijk 2001). Different approaches to defining function words have been suggested. For example, Halliday (1985) suggested that phrasal verbs which consist of one verb and one or more prepositions
Lexical diversity indicates how much the text includes unique words. It is traditionally related to a type-token ratio (TTR) which calculates the number of unique words divided by the total number of words in the text. However, TTR has a limitation that it is influenced by text length. Beyond TTR, various measures not influenced by text length have been suggested including the measure of textual lexical diversity (MTLD; the mean length of sequential word strings in a text that maintain a given TTR value, such as .720; McCarthy & Jarvis, 2010) and hyper-geometric distribution diversity (HD-D; the probability of encountering any of its type in a random sub-sample drawn from the text; McCarthy & Jarvis, 2010). When measuring lexical diversity, a caveat is that it is difficult to be convinced that lexical diversity purely assesses writers’ vocabulary ability because there is a possibility that lower levels of lexical diversity may be due to writers’ purposeful repetition of the same words to build cohesion by lexical repetition.

Lexical sophistication is considered as the use of advanced words in a text (Read, 2000). Advanced words were originally based on word frequency (i.e., how frequently a word is used) with less frequently used words being more advanced. Recently, using various computational indices, beyond word frequency, the operationalisation of lexical sophistication has been expanded, such that advanced words include academic words, words whose meanings are more specific, words that are more abstract and less familiar, and words acquired at a later age (Kim et al., 2018). However, the measure that may best represent lexical sophistication has not been agreed upon and may differ depending on L2 proficiency and writing tasks. Thus, researchers should carefully select lexical sophistication features for their analysis. In addition, given the richness of lexical sophistication measures and to avoid redundancy, if researchers use various lexical sophistication measures, they should consider construct distinctiveness, so that each measure targets different sub-constructs of lexical sophistication.

To measure lexical accuracy, various classification schemes of lexical errors have been suggested (e.g., Engber, 1995; Llach, 2011). For example, Engber (1995) distinguished errors resulting from lexical choice (i.e., meaning-related lexical errors) from those from lexical forms. One of the challenges in examining lexical accuracy lies in the subjective nature of lexical error detection, and thus, it is difficult to establish inter-rater reliability. Indeed, when examining various types of errors (e.g., preposition errors, verb phrase errors) in L2 writing, the inter-rater reliability of lexical errors was the lowest (0.54; Polio & Shea, 2014). In addition, beyond the number of lexical errors, types of lexical errors also need to be considered because while the number of lexical errors may not change over time, the specific types of lexical errors may change as L2 proficiency increases. For example, lower proficiency learners tend to produce L1-influenced errors, while higher proficiency learners tend to produce more target-language-oriented errors (Llach, 2011).

Lexical features presented above can be coded and calculated manually, but it will be labour-intensive and time-consuming. With the advent of natural language processing (NLP) tools, automated indices of lexical features can be
calculated quickly, flexibly, and reliably. I present some freely accessible automated tools for lexical features. The Coh-Metrix calculates lexical diversity and sophistication indices (Graesser et al., 2004). The Lexical Complexity Analyzer measures lexical density, variation, and sophistication (Lu, 2012). The Tool for the Automatic Analysis of Lexical Diversity (TAALED; Kyle et al., 2021) and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES, Kyle & Crossley, 2015) provide measures of lexical diversity and lexical sophistication, respectively. The Grammar and Mechanics Error Tool (GAMET; Crossley et al., 2019) provide several measures related to lexical errors.

6 Conclusion

In this article, I provided a summary of previous studies on longitudinal studies on lexical features in L2 writing, and considerations and challenges associated with capturing longitudinal changes in lexical features in L2 writing. To disentangle the unpredictability and nonlinearity of L2 lexical development and to clearly distinguish development (progressing towards more target-like behaviour), which we aim at understanding, from mere change (becoming different over time), it seems clear that more research on longitudinal changes in lexical features in L2 learner production is warranted. Ultimately, more longitudinal research in lexical features found in L2 learner production will help us have a deeper understanding of L2 lexical development and implement better vocabulary intervention in L2 classrooms.

References


Gené-Gil, M., Juan-Garau, M., & Salazar-Noguera, J. (2015). Development of EFL writing over three years in secondary education: CLIL and


Developing a Measure of Proper Name Familiarity for Japanese University Students

Christopher Nicklin
Rikkyo University

Abstract

In this study, an instrument for measuring proper name (PN) familiarity was developed for a psycholinguistic experiment investigating the effect of PNs on Japanese university students’ English reading fluency. Familiarity has previously been operationalized in disparate ways, producing contradictory results. Furthermore, authors of previous studies did not conduct validation analyses on their familiarity instruments. To address this issue, a four-point Likert-type scale instrument was constructed to assess Japanese university students’ familiarity with a set of 100, two-syllable PNs. The responses of 216 participants from 2 Japanese universities were subjected to Rasch analysis with the rating scale model to determine whether the resulting data fit the expectations of the model. The results suggested that a dichotomous response instrument was more appropriate than the scale-based instruments utilized in previous studies.

Keywords: Proper nouns, Familiarity, Rasch analysis

1 Background

Proper names (PNs) constitute a lexical class comprising the names of people, places, facilities and institutions, objects, and works of art that might be considered unique (Valentine et al., 1996). Despite accounting for approximately 1 and 5% of text in novels and newspapers, respectively, (Nation, 2006), PNs are essentially ignored in second language (L2) research. Because they are distinguished through the use of initial capital letters, it is assumed that learners will know PNs and they are thus included in text coverage counts as known items (e.g., Nation, 2006; Schmitt, 2008; Webb & Chang, 2015; Webb & Macalister, 2013). This assumption contrasts with research suggesting that advanced English learners’ reading and listening comprehension is disrupted by unfamiliar PNs (e.g., Erten & Razi, 2009; Kobeleva, 2012).

The omission of PNs from text coverage counts can be legitimately questioned when considering that L2 researchers and pedagogists traditionally quantify the likelihood that readers will know a word, and in turn lexical ‘difficulty’, through corpus-based frequency. Proper names such as Holden and Hartright, with merely 3,135 and 32 appearances in the Corpus of Contemporary American English (COCA; Davies, 2008), respectively, are assumed to be unproblematic.
because they are capitalized. This assumption becomes more problematic with the case of extensive reading, which involves learners reading massive amounts of simplified text with high speed and comprehensibility (Waring & McLean, 2015). One of the main benefits of extensive reading is its propensity to develop learner’s reading fluency (e.g., Beglar & Hunt, 2014). However, this quality is potentially hindered if the assumption regarding PNs’ unproblematic nature is inaccurate. To determine whether the assumption holds true, the effect of PNs on L2 learner’s reading fluency warrants research explicitly addressing the issue.

One way to approach the issue involves reliance on the traditional corpus frequency-based approach. However, it is debatable as to how pertinent the material gathered in COCA are to a population such as Japanese English as a foreign language (EFL) university students. Another way would be to assess how familiar a sample of the target population are with the target PNs. Familiarity has been operationalized by L1 and L2 vocabulary researchers in disparate ways, such as through the use of post-experimental interviews (e.g., Schmitt & Underwood, 2004), and varying question prompts along with five- (e.g., Libben & Titone, 2008; Titone et al., 2019) or seven-point scales (e.g., Carroll & Conklin, 2020; Valentine et al., 1991). Furthermore, the results produced with the instruments did not undergo any validation.

The present study reports on the development of PN familiarity ratings, measured in Rasch logits, for inclusion as an independent variable in an experiment conducted to assess the effect of PNs on Japanese EFL university students’ English reading fluency. The familiarity variable will eventually be included in linear-mixed effects model as a predictor of reading times in a self-paced reading experiment. With this in mind, the following research question was addressed.

1. To what extent are the Rasch model expectations met by an instrument designed to measure how familiar a set of English proper nouns are to a group of Japanese university students?

2 Method

In accordance with recent calls for multisite samples (Vitta et al., 2021), participants (\(N = 263\)) were recruited from six intact English classes at two Japanese universities, Site A (\(n = 207\)) and Site B (\(n = 56\)). Permission was granted to conduct the research from both universities and all participants signed a consent form. Both universities prohibited reporting standardized proficiency measures, which constitutes a limitation of the study. However, all participants had received at least 6-years of pre-university classroom English instruction.

The target PNs were extracted from a small 256,956-word corpus constructed from 15 Oxford Bookworms graded readers. Three books were randomly selected from each of Stages 2 through 6 of the Bookworms series, and each of the 15 books was converted into .txt file, tagged with TagAnt (Anthony, 2015), and analyzed with AntConc (Anthony, 2014). In total, 127 two-syllable PNs that appeared five or more times were extracted from the corpus. Twenty-seven PNs were omitted based upon graded-reader corpus frequency (i.e., the PNs with fewest occurrences were removed), which left 100 target PNs remaining.
In addition to the 100 target PNs, 65 control items were included on the instrument, consisting of the 33 most common two-syllable Japanese family names as of 2009\(^1\) and 32 non-names. The non-names were constructed from a list of 32 two-syllable location PNs (e.g., London) that were extracted from the graded-reader corpus and had the first letter (or sound) of each item substituted for next consonant or vowel in the alphabet, relative to the letter being substituted. For instance, London became Mondon, and Iran became Oran. The 165 items were presented to participants in random order on a Google Form. Test takers were instructed in Japanese with the following Japanese prompt:

"世界中から集められた名前が表示されます。例えば、英語、日本語、その他の言語のものがあります。それぞれの名前に馴染みがあるかどうか、1〜4で評価してください。

※1 = この名前に馴染みがない 4 = この名前に非常に馴染みがある"

[You will see a group of names from around the world. For example, some will be English, some will be Japanese, some will be from other languages. Please rate on a scale of 1 to 4 how familiar each name is to you: 1 = Not familiar at all and 4 = Very familiar”]

Test takers were prompted (in Japanese) to answer [lit.] How familiar do you feel with this name and were asked to respond by selecting one of the following options: I’m not familiar with this name, I’m a little familiar with this name, I’m familiar with this name, and I’m very familiar with this name. All responses were automatically recorded on a Google spreadsheet for analysis.

The data preparation process comprised three stages. Firstly, the responses were recoded as numerical values ranging from 1 to 4, where 1 = I’m not familiar with this name and 4 = I’m very familiar with this name. Secondly, items and participants with high false-alarm (FA) rates were removed from the dataset. False-alarm rate related to the number of times one of the 32 non-names, such as Zorkshire, was responded to as being very familiar, familiar, or a little familiar, while person FA related to the number of times a participant responded to a non-name as being very familiar, familiar, or a little familiar. A FA rate cut-off of 10% was utilized to ensure that participants were not overestimating their PN knowledge and to exclude participants who were perhaps not concentrating.

The FA-rate check revealed that five items received 27 or more a little familiar, familiar, or very familiar responses and were removed from further analysis for being too endorsable. Based on responses to the remaining 27 non-names, 47 participants with an FA rate > 10% (i.e., three or more [(27/100)*10 = 2.7] a little familiar, familiar, or very familiar responses to non-names) were excluded from further analysis. Finally, a spreadsheet was constructed that contained only the responses to the PNs. The non-names were removed because their only function was for FA-rate calculations, and the Japanese PNs were removed because they functioned purely as filler items.

To address the research question, the responses of the 100 target items and remaining 216 persons were subjected to Rasch analysis (Rasch, 1960) with the rating scale model (RSM; Andrich, 1978), which allows for polytomous data to be

---

\(^1\) Retrieved from https://www.japantimes.co.jp/life/2009/10/11/lifestyle/japans-top-100-most-common-family-names/
The Rasch misfit statistics for each item and person were analyzed to determine whether they were conducive to measurement. Misfit was determined via Wright and Linacre’s (1994) thresholds for productive measurement, whereby infit and outfit mean square (MNSQ) values outside of 0.50 to 1.50, and t-scores outside of −2.00 to 2.00 are considered detrimental to measurement. Items and persons with fit statistics below 0.50 were considered acceptable because they are unlikely to have practical implications in human science research (Bond et al., 2020). Misfitting items and persons were examined individually to assess why they failed to conform to the model’s expectations. For instance, an item might misfit the model’s expectations because of an error by a participant who produced several unusual answers (e.g., responded not only to several low familiarity target items as Very familiar but also responding to several high familiarity items as Not familiar). In this instance, it might be better to remove the person as opposed to the item and then re-run the analysis. The result of this iterative process was a set of logits representing a measure of the familiarity of each target item.

3 Results and Discussion

Despite the precedent in the literature for measuring familiarity on a scale, the results from the four-point scale utilized in the present study failed to meet the expectations of the RSM. When misfitting persons were removed, large numbers of items were also required to be removed due to insufficient responses to all four category levels. The middlemost categories failed to distinguish between Familiar and A little familiar across the sample suggesting that the meaning of each category was not invariant across the participants. However, the Guttman plot (see Figure 1) did suggest that a binary choice of either Familiar or Not familiar, might be more successful. This was based upon the observation that the top-left corner cells were generally all lighter than the bottom-right cells and that a diagonal line was visible stretching from the bottom-left to top-right corner. Consequently, the Very familiar, Familiar, and Slightly familiar categories were collapsed into a single category representing simply Familiar (to some degree). This resulted in a recoded, binary dataset, whereby 0 = I’m not familiar with this name and 1 = I’m familiar with this name (to some degree), which was fit with a dichotomous Rasch model using TAM.

Although collapsing categories might seem controversial because the new dataset represents answers that were incongruent with the response categories presented to the participants, it is an acceptable practice when ascribing to a school of thought that considers the model as subject to exploration (Wright & Linacre, 1992). Under this school of thought, the analyst is responsible for extracting the maximum amount of meaning from the observed responses. Wright and Linacre explained that collapsing categories with RSM frequently results in equivalent fit and results, and that such behavior is acceptable provided the decision can be
The initial iteration of the dichotomous model resulted in misfitting items with outfit mean square (MNSQ) values > 1.5 and outfit and infit t-scores > 2.00. Thus, the initial analysis was followed-up with nine iterations until all remaining items and persons satisfactorily fit the expectations of the Rasch model. The final iteration comprised 92 items and 185 persons, all displaying MNSQ fit statistics within Wright and Linacre’s thresholds, but with three items (Baby, Hatta, Sunset) and two persons displaying infit t-scores above 2.00. However, this was deemed acceptable because 5% of the items or persons are expected to misfit by chance (Beglar, 2010). Reliability was measured with expected a posteriori (EAP) reliability, which is a measure that, although not the same, can be interpreted in the same way as Cronbach’s α (Neumann et al., 2011). The EAP reliability of 0.91 was high, indicating that the results produced by the reduced sample were reliable. Figure 2 illustrates that the TAM person parameters’ distribution was Gaussian, which is an assumption of the marginal maximum likelihood estimation equation utilized in the TAM analysis. The R script containing details of each iteration (and also the Supplementary Materials) is available online at https://github.com/nicklin/vli2022. Supplementary Materials A contains a brief report of the confirmatory eRm analysis, which produced almost identical results. The descriptive statistics for the final iteration are displayed in Table 1.
Figure 2. Distribution of the Person Parameters.

Table 1. Descriptive Statistics for the Final Iteration of the Dichotomous Rasch Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Measure</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>Skew</th>
<th>Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>Raw</td>
<td>60.61</td>
<td>60.00</td>
<td>1.00</td>
<td>177.00</td>
<td>0.67</td>
<td>-1.13</td>
</tr>
<tr>
<td></td>
<td>Logit</td>
<td>1.48</td>
<td>2.53</td>
<td>-3.65</td>
<td>5.76</td>
<td>-0.30</td>
<td>-1.04</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.31</td>
<td>0.18</td>
<td>0.16</td>
<td>1.01</td>
<td>1.82</td>
<td>3.36</td>
</tr>
<tr>
<td></td>
<td>Infit MNSQ</td>
<td>1.00</td>
<td>0.07</td>
<td>0.87</td>
<td>1.24</td>
<td>1.17</td>
<td>2.28</td>
</tr>
<tr>
<td></td>
<td>Outfit MNSQ</td>
<td>0.89</td>
<td>0.25</td>
<td>0.29</td>
<td>1.38</td>
<td>-0.32</td>
<td>-0.40</td>
</tr>
<tr>
<td></td>
<td>Infit-t</td>
<td>0.14</td>
<td>0.70</td>
<td>-1.68</td>
<td>2.93</td>
<td>1.58</td>
<td>4.71</td>
</tr>
<tr>
<td></td>
<td>Outfit-t</td>
<td>0.01</td>
<td>0.76</td>
<td>-1.36</td>
<td>2.79</td>
<td>0.89</td>
<td>1.24</td>
</tr>
<tr>
<td>Persons</td>
<td>Raw</td>
<td>30.14</td>
<td>10.37</td>
<td>7.00</td>
<td>55.00</td>
<td>0.06</td>
<td>-0.50</td>
</tr>
<tr>
<td>(N = 185)</td>
<td>Logit</td>
<td>1.18</td>
<td>1.17</td>
<td>-1.31</td>
<td>4.35</td>
<td>0.28</td>
<td>-0.32</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.34</td>
<td>0.03</td>
<td>0.30</td>
<td>0.47</td>
<td>1.59</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>Infit MNSQ</td>
<td>0.99</td>
<td>0.19</td>
<td>0.56</td>
<td>1.45</td>
<td>0.09</td>
<td>-0.71</td>
</tr>
<tr>
<td></td>
<td>Outfit MNSQ</td>
<td>0.80</td>
<td>0.31</td>
<td>0.24</td>
<td>1.44</td>
<td>0.28</td>
<td>-0.99</td>
</tr>
<tr>
<td></td>
<td>Infit-t</td>
<td>-0.03</td>
<td>1.00</td>
<td>-2.68</td>
<td>2.13</td>
<td>-0.15</td>
<td>-0.63</td>
</tr>
<tr>
<td></td>
<td>Outfit-t</td>
<td>-0.12</td>
<td>0.60</td>
<td>-1.51</td>
<td>0.99</td>
<td>-0.33</td>
<td>-0.80</td>
</tr>
</tbody>
</table>

Note: Min, Minimum; Max, Maximum.

The final selection process involved selecting 30 PNs for the psycholinguistic experiment assessing the effect of PNs on Japanese EFL university students’ L2 English reading fluency. The TAM-derived logits were reverse signed for ease of interpretability (i.e., the logit for Tony [-1.96] became 1.96, thus larger values represented more familiar according to the target group), resulting in a spread of 92 PNs from the least familiar PN, Halcombe (-5.76) to the most familiar, William (3.65). The 15 least familiar PNs with logits < -4.00 were removed from the list, leaving a spread from approximately -4.00 through 4.00. The 15 excluded PNs were also the items with the largest standard errors (SEs), indicating that they were estimated with the least precision. Furthermore, all PNs with alternative meanings, which constituted potential confounds, were removed (Baby, Rosso, and Sunset). From the remaining 74 PNs, 15 PNs with logits above and below zero were selected, with the aim of achieving an even spread of values from -4.00 to 4.00. The final 30 PNs...
are plotted by familiarity logit in Figure 3, while the logit, SE, and fit statistics for each PN are presented in Supplementary Materials B.

With regard to the research question, the results of the analysis suggest that the results of the proper noun familiarity instrument met the expectations of the Rasch model. The final 30 items comprise a spread of logits ranging from −3.91 to 3.65, with a maximum SE of 0.42. The MNSQ fit statistics indicated that the items fit the expectations of the Rasch model well, with all values within Wright and Linacre’s (1994) 0.50 to 1.50 thresholds for productive measurement. Furthermore, the order of familiarity ratings makes logical sense. For instance, it is understandable that William and Alice are the most familiar names from the list and that they are more familiar to Japanese university students than Tony and Sophia, who are located several places lower. William is the name of an English Prince who features regularly in the Japanese news, while the name Alice is embedded in popular culture through the Lewis Carroll novels, Alice in Wonderland and Through the Looking Glass, and their Disney interpretations. It is also understandable that Holden, Dobbin, and Hartright are the least familiar names, as these can hardly be said to be typical names.

The result of this short study has one main implication for the use of familiarity as a variable in L2 research. Although familiarity has been utilized in previous research, the authors of those studies operationalized the variable in various guises, such as five- and seven-point Likert scales. The present study is the only one of these studies that has investigated a scale-derived approach to familiarity with Rasch analysis, and the result suggested that such an approach might be inappropriate for L2 learners. The participants failed to use the categories of the scale in a consistent manner, thus the categories failed to separate. Collapsing the categories and constructing a dichotomous model solved this issue, thus a dichotomous instrument should be utilized if this process is replicated with another set of PNs. It is possible that this conclusion is relevant for familiarity measures in
L1 psycholinguistic research, but Rasch analysis of such instruments is required for confirmation.

4 Conclusion

In the present study, the development of a PN familiarity measure for inclusion as a variable in a model of Japanese EFL university students’ reading fluency was reported upon. Although previous researchers have utilized familiarity variables by collecting information from the target population, none have reported the results of a validation analysis to determine whether the instruments were measuring what they were designed to measure. Rasch analysis with the rating scale model indicated that the scale-derived approach to familiarity adhered to in previous L1 research might be inappropriate for L2 research because the meaning of each category was not invariant across the participants. It is possible that alternative wording of the instrument could result in more consistent responses, but further research would be required. However, once the categories were collapsed to allow a binary, “Yes” or “No” approach to familiarity, the model fit the expectations of the dichotomous Rasch model. The main implication of this short study is that the construction of a familiarity measure in L2 research, and perhaps even L1 research, should involve an instrument with a dichotomous response as opposed to four-, five-, or seven-point scales, because the difference between the categories are unlikely to be consistent across all participants.

References


---

*Vocabulary Learning and Instruction, 10*(2), 91–100.


Discussion Paper: Using Statistics to Solve Practical Vocabulary Problems

Jenifer Larson-Hall

University of Kitakyushu

Most of us who do research on language acquisition have had to use statistics to evaluate the results of experiments. Some may use only the statistical procedures they learned in graduate school and may thus miss out on new advances in statistics that might shed light on some problems in a more straightforward way. The three papers that conduct empirical studies that I will discuss today have used statistical procedures that you may not be very familiar with—bootstrapping, Monte Carlo simulations, and Rasch (or item response theory [IRT]) analysis. Their use of these procedures, however, means that they are able to give quite precise and interesting answers to the questions that they have asked. The fourth paper I will discuss is not an empirical study but a review of studies and call for future research going forward.

I'd like to start with the paper by Stewart, McLean, and Batty (2021, issue 10.2) entitled “Correlations of modalities of written vocabulary knowledge to listening and reading proficiency: A comparison.” In this article, the authors basically used the data that were described in more detail in a separate article published by all three authors in Language Testing (McLean et al., 2020). In that 2020 study, they examined which of the four modalities of vocabulary tests best correlated to reading ability as measured by Test of English for International Communication (TOEIC) Reading section scores. In the present (2021) study, the authors used the data to examine the correlations to listening ability, as measured by the TOEIC Listening section scores and compared that to the reading correlations. Putting aside the question of whether the TOEIC tests accurately measure reading or listening ability, I would like to examine the specific way in which both papers by these authors answered the research questions they set out.

Both of these studies used interesting methodological and statistical methods to significantly increase their power to give interesting answers to their respective research questions. The McLean, Stewart & Batty (2020) data (I will name the data MSB based on the authors’ last name order in the 2020 article) consists of 4,000 data points per person, with 103 participants. Each participant gave their answers regarding the third most frequent band of 1,000 words from the New General Service List (Browne et al., 2013), which includes words like supplier, personnel, and stimulus. The test takers answered 1,000 questions about their knowledge for each of the four modalities of tests that have been used to measure vocabulary knowledge (see Table 1).

The first innovative methodology of the MSB data is that it is complete, at least for one 1,000-word vocabulary band. Because the range of possible
vocabulary words is so large, most studies simply sample from the vocabulary range that they are interested in. This is the first study I have heard of which asks the participants to judge every word. The authors used their complete information to create samples of each 1,000-item dataset in differing lengths to answer the question of how long a sample had to be to accurately reflect the true score of the participants. Because they had not sampled the participants but instead exhaustively tested them, they knew their true score and were thus able to look at what sample lengths were accurate.

The second innovative step was to use bootstrapping with this data. This was not necessary to answer their research question. For a sample of length 100, for example, the authors could have just randomly drawn out 100 numbers from the 1,000 available numbers for each participant to create an appropriate sample. They could have done this for all 103 participants and then correlated the average score of each sample to the own participant’s TOEIC reading and listening scores.

However, the authors noted that to find very small effect sizes they would need to sample several hundred students (McLean et al., 2020). Obviously asking even 103 participants to judge 4,000 items must have been a Herculean task, so there is no way to criticize the authors for not having more participants. They thus decided to use another method to increase the generalizability of their data: bootstrapping. Bootstrapping is one recent statistical tool that has been introduced in the SLA field (Larson-Hall & Herrington, 2009) as a way to deal with smaller samples and also a way to overcome the problem of the fact that for the small sample sizes (< 50) used in our field it is essentially impossible to determine whether the data are in fact normally distributed. However, to perform parametric statistics one must assume that the data are exactly normally distributed and not slightly heavier in the tails of the empirical (sampled) distribution than the normal distribution, which is called a contaminated normal distribution and which research has shown can cause type II errors (Tukey, 1960; Wilcox, 2001). This results in a classification of a difference or relationship as non-statistical when it is in fact statistical.

Table 1. Four modalities to test vocabulary as cross-tabulated by meaning vs. form and recognition vs recall

<table>
<thead>
<tr>
<th>Modality</th>
<th>Example for native English speaker learning Japanese</th>
<th>Example for native Japanese speaker learning English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning recall</td>
<td>See L2 form, write L1 word</td>
<td>親切な学生です。</td>
</tr>
<tr>
<td>Form recall</td>
<td>See L1 form, write L2 word</td>
<td>進行は遅い。</td>
</tr>
<tr>
<td>Meaning recognition</td>
<td>See L2 form, choose L1 meaning</td>
<td>楽曲</td>
</tr>
<tr>
<td></td>
<td>a) game</td>
<td>a) 遊び</td>
</tr>
<tr>
<td></td>
<td>b) movie</td>
<td>b) 映画</td>
</tr>
<tr>
<td></td>
<td>c) song</td>
<td>c) 歌</td>
</tr>
<tr>
<td></td>
<td>d) yard</td>
<td>d) 庭</td>
</tr>
<tr>
<td>Form recognition</td>
<td>See L1 form, say whether you know it</td>
<td>知っている単語をチェック：</td>
</tr>
<tr>
<td></td>
<td>□ 遊び</td>
<td>□ game</td>
</tr>
<tr>
<td></td>
<td>□ 映画</td>
<td>□ movie</td>
</tr>
<tr>
<td></td>
<td>□ 歌</td>
<td>□ song</td>
</tr>
<tr>
<td></td>
<td>□ 庭</td>
<td>□ yard</td>
</tr>
</tbody>
</table>
Bootstrapping treats the data of each participant as a pool of data to be randomly sampled from and then creates an empirical distribution from this random sampling of the data. In other words, bootstrapping does what the researchers themselves would like to do: repeat the experiment. In this case, there was quite a large pool to sample from given that each participant had 1,000 data points. Thus, for the length of 100 items, the authors randomly sampled 100 items from those 1,000 data points, and they did that 1,000 times. Each time a number was randomly picked from the pile of 1,000 data points, it was tallied and then dropped back into the pile of 1,000 numbers--this is called sampling with replacement. Each of those 1,000 samples of length 100 was then reduced to its average number. This was repeated for all 1,000 samples. The authors now had an average score for a sample of length 100 for each participant that was itself the average of 1,000 samples of size 100 from the original 1,000 data points. If this seems confusing, perhaps an illustration will help (see Figure 1). This average, then, was correlated with the TOEIC listening score.

In the McLean et al. (2020) article they had 21 different lengths of tests, and they created 1,000 bootstrap samples from each test (21 × 1,000 = 21,000) for the 4 different types of tests (4 × 21,000=84,000 tests) for each person. There were 103 participants so the total number of samples they had was (84,000 × 103 = 8,652,000). They could then calculate a mean score of the samples for each modality of test at each of the 21 lengths of test. This number represented the scores of the 103 participants but with 1,000 samples from each, meaning it was much more accurate than simply taking a random sample of that length from each participant would have been. Essentially it was like surveying 1,000 × 103 = 103,000 participants for each length in a certain modality.

Now there was nothing special about the MSB data that meant they needed to use bootstrapping. The point I want to bring up here is that anyone can use bootstrapping with their data as a way of generating more resilient and accurate samples.

![Figure 1. Bootstrapping, illustrated for one participant in one modality with samples of length 100 and 1,000 bootstrapping samples taken.](image-url)
The current paper used exactly the same methodology as the McLean et al. (2020) article for dealing with the data, the only difference being that the scores were correlated with the TOEIC listening test.

This paper did not show the data graphically so I just wanted to show what that correlation between the vocabulary levels test at length 100 and the TOEIC listening data looked like (Figures 2 and 3). By the way, to plot Figure 2 I used Mizumoto’s langtest.jp site (http://langtest.jp/shiny/cor/) and the ellipse shows the robust part of the correlation. I included Figure 3 because it has a LOESS line which follows the pattern of the data instead of just drawing the straight least-squares line on the data.

The strong correlation in this data is clear. The point is that this data looks like every other scatterplot. There’s nothing strange about the MSB data even though it was bootstrapped.

How does one perform this bootstrapping? McLean et al. (2020) said they used an Excel formula. The formula would have sampled with replacement X number of items from the pool of 1,000 and then calculated an average for that one sample, then repeated that 999 more times. Simple bootstrapping is not difficult. Bootstrapping can be used together with other robust procedures, however. One discussed in Larson-Hall and Herrington (2009) is trimming. This removes a certain percentage of the data on both ends of a distribution, thus excluding outliers in a principled manner and being more robust to single extreme values than the mean score but keeping more of the data than the median, which throws away all but one or two values.

The fact of the matter is that I cannot stop praising this paper for its use of larger samples to begin with and then for using bootstrapping to ensure that the
samples would be highly representative of the population from which they are drawn. Tversky and Kahneman (1971) point out that most people have wrong intuitions about samples and think that all samples are similar to their population. For example, if a researcher found a statistical correlation of $r = 0.30$ in a sample of $N = 40$ participants, most researchers think that testing $N = 20$ participants would similarly result in finding a statistical correlation of about the same magnitude. Tversky and Kahneman call this a “belief in the law of small numbers” when what we can actually believe in is the representativeness of large numbers and large samples only.

The McLean et al. (2020) paper also showed pretty convincingly that meaning recall tests (see L2 word form and write L1 word) are the best types of tests to use when one wants to test vocabulary ability. Meaning recall tests had the highest correlation to reading proficiency, reaching $r = 0.76$ at a test length of 40 items, indicating that vocabulary ability as measured by meaning-recall explains almost 50% of the variance in the TOEIC reading portion. The superiority of the meaning recall tests to the other three types of tests still held true in a comparison involving the same amount of time spent on the test (by the participant). And it turned out this result also applies to correlations with vocabulary skills and the TOEIC listening tests as well.
Of course, meaning-recall involves more work for the person scoring the test, but McLean has been working on reducing that burden by creating a website where the computer can learn what answers you favor and ask you to judge unorthodox answers. This is of course the vocableveltest.org site described in McLean, Huston, Raine, Kim, Ueno, Pinchbeck and Nishiyama called “The internal consistency and accuracy of automatically scored written receptive meaning-recall data: A preliminary study.”

To me it seems that Stuart McLean has organized a fruitful and practical line of research that he is methodically pursuing in collaboration with many others. With strong research showing that meaning-recall tests (see L2 word form and write L1 word) are the best way to measure vocabulary that is useful for reading and listening abilities, he has set about trying to make such tests more user friendly. I had known about McLean’s vocableveltest.org site before, but what I hadn’t realized before was how it related to best practices for vocabulary testing as uncovered by research.

I’m struck by the research result, quoted in the paper as coming from McLean et al. (2014), that even Japanese students at more exclusive schools (with a hensachi over 61) did not demonstrate mastery of the first 1,000 words of English. It seems clear that at least in Japan, the vocableveltest.org might be fruitfully used by almost all university professors who are teaching English as a way to discover where their students might have gaps in their basic knowledge of English. Since bands as small as 100 can be used, teachers have an easy way to test their students’ knowledge of foundational English vocabulary a little at a time.

I was also impressed with the measures implemented at this site to prevent cheating. I think that is an important consideration these days when online dictionaries can be easily accessed during testing and when some students may be doing testing at home, not under our watchful eyes.

Now the McLean et al. paper does something very similar to the bootstrapping in the Stewart, McLean, and Batty paper: it uses Monte Carlo simulations to determine how many items will be sufficient to accurately represent a learner’s knowledge of the vocabulary in a 1,000-word band. Basically, Monte Carlo simulations use the same process as bootstrapping, although instead of gathering data from live participants, a computer randomly simulates a distribution whose mean is specified in advance. Thus, for Figure 9 of McLean et al. (2021, issue 10.2), the computer creates a normal distribution of numbers whose mean is $X = 750$ with 1,000 samples of the distribution. The researcher then randomly samples some number of 5, 10, 20, and 100 item samples from that distribution. From Figure 9, it looks like 1,000 samples were taken at each of those test size lengths. The average score of that sample is then calculated, and that is plotted for the number of times each of those possibilities is chosen. Thus we can see that when we actually know the true score if we only sample five items from the entire distribution (and this would be equivalent to testing five items from a 1,000-band vocabulary level), we could end up with an average score of 200 points occasionally (maybe 10–15 times). Most often we would get a score of around 800 (400 times) but basically with so few points, we could end up almost anywhere along the continuum! Five points are therefore not very likely to get us close to the real score. Obviously 100
sampled items get us within a much closer range to the real score although it may slightly under- or over-estimate the real score. Figure 10 shows that sampling 50 items is pretty close to 100 items although not quite as accurate. Of course, since the McLean et al. (2020) paper went through the same process with actually sampled data, meaning we knew each participant’s real score, it was already shown that a sample size of 40 was almost as reliable as a sample size of 100 and the size that the authors seemed to consider “good enough.”

Does this mean that the McLean et al. (2021) study did not actually require the Monte Carlo simulation to prove its point? I’m pretty sure it does! I think it could have used the MSB data set. However, does this mean that the McClean, Stewart, and Batty (2020) as well as the Stewart, McLean, and Batty paper in this volume did not actually need to sample students and could have just as well used a Monte Carlo simulation? Well no, because the authors did want to find out exactly how the four test modalities would compare to each other, and they couldn’t tell the computer what the mean scores for each modality were in advance the way the Monte Carlo simulation could.

I do quibble a bit with the McLean et al. (2021) paper when it asserts that “Figure 11 suggests that even samples of 100 or 200 items can occasionally result in inaccurate estimates.” I assume this refers to the fact that although the true mean is 750, the sampled average with 100 items could be as low as 630 or as high as 870. Statistics are probabilistic and whenever we calculate some statistic from a sample, be it the vocabulary knowledge of a 1,000-band set of words or the effect size of a $t$-test, our statistics may return a point value but any point value is only an estimation and will never be the true score for the population. If that point is understood, then confidence intervals (CIs) become much more valuable. CIs will give an interval around the point value that could plausibly contain the true value, with 95% confidence. Of course, sample sizes of 100 or 200 will have smaller CIs than sample sizes of 20 or 50 but that doesn’t mean they are inaccurate. No statistic is accurate, meaning exactly right. As Crawley (2012) says, all models are wrong, but some models are better than others (p. 403). Statistics can give us the best wrong model, but it will always still be wrong.

I have to say that I have heard Stuart discuss his vocableveltest.org site several times and because of this paper I actually went and tried it out myself. I am doing vocabulary activities with several classes I teach, but I’m afraid the website isn’t useful to me as is. I am using Gardner and Davies’ (2014) Academic Vocabulary List at my university, and that list is not yet available in vocableveltest.org. Another class I teach has vocabulary activities with the words in the textbook, and I don’t see any way to enter your own list of words into vocableveltest.org yet. Thus, unless I want to test students’ abilities with basic vocabulary using the NGSL I wouldn’t use this tool yet. I also tested out the site and found that some of the sentences are problematic. For example, I used the site to test out my Japanese ability by answering form recall items (see L1 form, write L2 word; see Figure 4). As you can see, my Japanese ability in this test was not great, but notice the second sentence, “Are they pair?” This is not correct English and if I got this strange sentence with a sample of only five sentences it makes me wonder what percentage of sentences might be problematic.
However, I have to say I am extremely glad to see that this website has been created. I really enjoyed the format where I could see my incorrect answer and then see possible correct answers too. I do think the site could be useful for large numbers of researchers. I imagine, since the last line of the paper states it, that the website will continue to be improved and that McLean is open to adding new types of data to the website so researchers like me could use our vocabulary lists too.

The review paper by Kim, entitled “Considerations and challenges in longitudinal studies of lexical features in L2 writing (2021, issue 10.2),” indicates that this author also has an ambitious agenda that could help researchers. It would be great if there were eventually a website where researchers could upload a piece of L2 writing or transcribed speaking and get information about the lexical sophistication, diversity, density and accuracy of the vocabulary used in that writing. Actually, as Kim notes, for English that day may not be far away, as the Crossley, Kyle, and colleagues’ tools (TAALED, TAALES, GAMET) can implement algorithms to check for many of these things.

I have examined these tools in my own research, however, and I have found that they are not as helpful when the target language is not English. I had a case study with five L1 English speakers learning Japanese. I tested the participants over 3 years and transcribed their storytelling of Japanese picture books. Previous research of this type has simply reported on the type/token ratio of such utterances, but I thought that with the more sophisticated tools available today I would be able to examine lexical density or sophistication and track how that changed over time. However, the tools are not created for dealing with Japanese and although I was corresponding with Scott Jarvis for a few months as I tried to jury-rig some of the tools to work for Japanese, including learning how to manipulate Jarvis’ Python codes, it was slow work and I eventually gave up. This, of course, is an additional challenge if we assume that the L2 writing Kim alludes to includes any L2 other than English.

Beyond this problem, as Kim notes, there is still controversy surrounding the best ways to measure and compute various ideas like lexical sophistication, lexical errors, or any of the other measures of L2 writing. For example, on the
Kim notes that if vocabulary development is investigated through writing (or I would add, transcribed speaking measures), then it would be better to look at writing over spans longer than 1 year where participants are in a situation where they are likely to get large amounts of input. Since language acquisition is a rather slow process if we are really thinking about real productive or receptive abilities and not just scores on a test where time is available for explicit analytical abilities to come into play, then this recommendation is probably a great one for any researcher who wants to study development at all! Frankly, I’ve come to think that studies that test out a teaching technique over a short period (like an hour!) and look at the results are usually fairly worthless. We don’t have any evidence that groups that might perform better on an immediate posttest, or even at a delayed posttest of several weeks later, will still retain any of that information better over the long run. I know that longitudinal research of the type Kim is advocating and which I am seconding takes a lot of time and there’s always a lot of attrition from participants, but better to struggle with that than waste our resources with short-term studies that aren’t worth the paper they’re printed on!

The issue I see with Kim’s proposal, however, is that this area of research looks enormous and I am not sure it is at a state where many concrete recommendations can be made. Obviously I agree that examining vocabulary development over the long run when participants are getting lots of input is a great idea. However, Kim herself states, “we do not know much of L2 lexical development, and more research is needed” (p. 3). The papers cited in the literature review certainly do not paint any kind of vivid picture of what is going on with lexical development. When researchers in the field are still investigating what lexical features to measure and how to best measure them, and when in fact very little research has actually been done longitudinally, this call for more research sounds too broad to my ears.

Let’s take just one area that Kim mentions. She says that a consideration in the research design is what the basic unit of lexical analysis is and lists possible units as being tokens, lemmas, or word families. However, this area alone is quite controversial and could probably encompass a substantial research agenda. The unit of flemma, explained by the McLean et al. paper given in this session as “a base word form and inflectional forms, regardless of POS” (p. 6), differs from the lemma by not separating words with different parts of speech that are identical in form into separate categories (e.g., use_N and use_V are separate lemmas but one flemma). However, the question of whether the word family, lemma, or flemma is more appropriate for use with English learners is something that has, to my mind, still not been resolved and has been addressed by several researchers from our vocabulary SIG (McLean, 2017; Stoeckel et al., 2020). I want to turn back to the safety of the other papers here, which I think address very small but practical and solvable issues.
Therefore, I would like to mention Nicklin’s paper, called “Developing a measure of proper name familiarity for Japanese university students (2021, issue 10.2).” This very small and discrete study seems to be one brick in an agenda determined to understand whether proper nouns disrupt comprehension abilities. I have to admit that before reading this paper I would have definitely said that they did not, but it appears there is at least as much support for the conclusion that proper names make reading English difficult as there is for the finding that studying vocabulary in semantic sets is worse than studying it in thematic sets. That very specific vocabulary finding based on only a few studies got a whole chapter in Folse’s Vocabulary Myths book (Folse & Briggs, 2004), while I have never heard of proper noun problems before, so I am glad Nicklin has drawn my attention to it.

Nicklin does not yet provide a paper that establishes whether proper nouns impede comprehension but rather provides an example of a tool that he will use in his investigation of this question. He thus provides an example of how to proceed in other endeavors where new tools are created. These examples are needed! I remember in my early research I used a grammaticality judgment test that had been used in critical period studies. In my own research, I cited the reliability statistics of that test from the previous research papers as if the high reliability numbers found in someone else’s paper meant that the test was reliable without any further demonstration on my part. What I didn’t understand was that reliability is a function of how a particular sample performs on a particular test, so that the only reliability statistics I should have provided were those for my own sample.

Nicklin here uses Rasch modeling, or what is known as item response theory (IRT), to validate what ultimately becomes a range of 30 items that span a continuum of familiarity for Japanese university-level learners of English. Again, just as the first two papers I mentioned used some innovative statistical methods, this paper also uses a statistical method that may not be very familiar to some readers. Although IRT is by no means a new procedure it is not available in the base version of SPSS and possibly because of this is not widely understood.

Paolillo (2000) notes that those who ignore methodological concerns do so at the peril of misinterpreting their data and making false conclusions about their experiments. I think Nicklin has clearly shown that this could have easily happened in his study if he had not investigated the ability of his Likert-scale familiarity questionnaire to distinguish four levels of familiarity. Other researchers often use Likert-scale questions without any analysis of whether the scale is actually reliable or valid, and again, that is why this paper is quite exemplary.

IRT improves upon classical test theory. The R packages that Nicklin used to conduct his analysis were called TAM (test analysis modules) and eRm (extended Rasch measurement). Using such packages one can call for the descriptive statistics that are calculated in a classical test analysis approach to the data: item facility (which calculates how many people got the item “correct” out of the people who took it), item discrimination (a measure that lets the researcher see whether those who scored highly overall on the test scored highly overall on a particular item), and the biserial correlation with the item excluded. While classical
test analysis basically examines how difficult items are for the test takers, IRT gives information about the skill levels of the test takers themselves. While classical test analysis cannot talk about reliability beyond the scores of the test takers themselves and cannot generalize test scores beyond the sample and items tested, IRT’s advantage is that it is parameter invariant, meaning that “item statistics that are obtained from the application of IRT models are independent of the sample of examinees to which a test is administered” (McKinley, 1989). It is also an advantage that it can score individuals according to their ability levels and give error measurements for individual items in the test. I’ve always been intrigued by the promise of IRT for doing adaptive testing. Because it can take the responses of a test taker and see what scores they receive on items at different difficulty levels it is able to quickly discriminate what a test-taker’s ability level is.

Nicklin’s analysis was more interested in the items of the test and which items would span a range from very unfamiliar through very familiar proper nouns. I checked out Nicklin’s supplementary materials and while I cannot claim that I could follow the entire analysis, one great thing about R is that all of the R commands that Nicklin used, including his data and commands for drawing the graphics found in the paper, are reported online and available to anyone who wants to check them or use them to do something similar to what Nicklin did. If you are not familiar with the free R statistical program yet, I cannot recommend highly enough that you become so. It is the statistics of the future.

In conclusion, three of the studies presented today resemble each other in that they have used lesser-known but powerful types of statistical analysis to make their conclusions more valid and reliable. Although it may seem that such analysis is too sophisticated or advanced for some readers, with an initial investment of time to learn the basics of the R statistical language they are not beyond the abilities of normal researchers. I do want to acknowledge that Kim is no stranger to statistical analysis herself, as shown by her careful work in Kim et al. (2017), where she used Principal Components Analysis to find 12 components of a lexical sophistication measure and then used correlations and regression between the components and a measure of writing proficiency to discover which components could successfully predict writing proficiency. From these papers, we can see the benefit of learning about the latest statistical methods and employing them carefully in our own research.

Notes:

1. Nicholson (2015) says that although the TOEIC scores have been found to be reliable some researchers question the premise of using the TOEIC test in such situations at all since the test was developed to measure how well English L2 users can communicate in a business workplace. Nicholson says that although the TOEIC test is fairly prestigious in Asia and is used by many employers, the lack of independent verification that the TOEIC measures actual reading or listening skills is “startlingly low” (p. 225).
2. I used the document entitled “NGSL + 1.01 + with + SFI” and the words listed between 2001 and 3000 to obtain these words. McLean, Stewart, and Batty (2020) said that they used flemmas as sorted by the words’ SFIs (standard frequency indices), but the Excel file column itself specifies that these are lemmas so I do not know whether anything was done to the original document to change to a lemma distinction in words.

3. Note that the authors provided me with the data upon request. My belief is that this should be the norm for any published paper. One step further is to make your data publicly available on a site such as github (as Nicklin did), Open Science Framework repository (osf.io), or the IRIS database (https://www.iris-database.org/iris/app/home/index;jsessionid=9B0A42048B0D992898D00BC52905D123).

4. The Japanese term hensachi when used in connection with universities refers to the average score of students entering any university based on a national test whose scores are normed to a scale where 50 represents the mean score of all students who took the test. Students whose scores were one standard deviation above the mean would receive a 60. By the 68–95–99 rule for a normal distribution then, 68% of all students would fall within one standard deviation of the mean, so only 16% of students (32%/2) would have scored above 60.

References


*Vocabulary Learning and Instruction, 10*(2), 101–113.


