VLI Editorial Team

Editorial Board: Jeffrey Stewart, Raymond Stubbe, Aaron Gibson.


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

Copyright © 2020 Vocabulary Learning and Instruction, ISSN: Online 2187-2759; Print 2187-2767. All articles are copyrighted by their respective authors.
# Vocabulary Learning and Instruction

Volume 9, Number 2, August 2020
doi: http://dx.doi.org/10.7820/vli.v09.2.2187-2759

## Table of Contents

<table>
<thead>
<tr>
<th>Articles</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Letter from the Organizer</strong></td>
<td>iv</td>
</tr>
<tr>
<td><em>Jeffrey Stewart</em></td>
<td></td>
</tr>
</tbody>
</table>

### Session 1: Vocabulary and CALL

<table>
<thead>
<tr>
<th>Articles</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary Learning Using Student-Created Class Vocabulary Lists</td>
<td>1</td>
</tr>
<tr>
<td><em>G. Clint Denison and Imogen Custance</em></td>
<td></td>
</tr>
<tr>
<td>The Indirect Spaced Repetition Concept</td>
<td>9</td>
</tr>
<tr>
<td><em>Louis Lafleur</em></td>
<td></td>
</tr>
<tr>
<td>On Creating a Large-scale Corpus-based Academic Multi-word Unit Resource</td>
<td>17</td>
</tr>
<tr>
<td><em>James Rogers</em></td>
<td></td>
</tr>
<tr>
<td>Exploring the Effectiveness of Deliberate Computer-Assisted Language Learning</td>
<td>24</td>
</tr>
<tr>
<td><em>Andrew Obermeier</em></td>
<td></td>
</tr>
<tr>
<td>Vocabulary and Computer Technology: A Commentary on Four Studies</td>
<td>39</td>
</tr>
<tr>
<td>for JALT Vocabulary SIG</td>
<td></td>
</tr>
<tr>
<td><em>Tatsuya Nakata</em></td>
<td></td>
</tr>
</tbody>
</table>

### Session 2: Vocabulary Learning and Assessment

<table>
<thead>
<tr>
<th>Articles</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teaching Ideas for Improving Oral Performance through Formulaic Language Instruction</td>
<td>48</td>
</tr>
<tr>
<td><em>Chie Ogawa</em></td>
<td></td>
</tr>
<tr>
<td>Assessing Low-level Cognitive Processes of Word Recognition</td>
<td>55</td>
</tr>
<tr>
<td><em>Michael Holsworth</em></td>
<td></td>
</tr>
<tr>
<td>The Challenges of Measuring Multi-Word Expression Use in Conversation</td>
<td>63</td>
</tr>
<tr>
<td><em>Haidee Thomson</em></td>
<td></td>
</tr>
<tr>
<td>Deliberate Vocabulary Learning from Word Cards</td>
<td>69</td>
</tr>
<tr>
<td>Darrell Wilkinson</td>
<td></td>
</tr>
<tr>
<td>Vocabulary Learning and Assessment: A Commentary on Four Studies</td>
<td>75</td>
</tr>
<tr>
<td>for JALT Vocabulary SIG</td>
<td></td>
</tr>
<tr>
<td><em>Irina Elgort</em></td>
<td></td>
</tr>
</tbody>
</table>
Letter from the Organizer

Dear Readers,

The year 2019 has been a tumultuous one for the JALT Vocabulary Special Interest Group, much as 2020 has been a tumultuous year for the entire world. In September, a typhoon passed through Tokyo, forcing us to cancel the annual symposium where these papers were scheduled to be presented. Sessions were rescheduled for May and November 2020. However, the coronavirus disease 2019 (Covid-19) pandemic led to the cancellation of these events too.

Although we typically do not publish symposium papers until after the event, two years is simply too long to deprive the field of these fine contributions to L2 vocabulary acquisition. This year, we extend into the study of vocabulary in relation to computer-assisted language learning, with papers by Imogen Custance and Clint Denison, Louis Lafleur, James Rogers, and Andrew Obermeier, followed by commentary from Dr. Tatsuya Nakata of Rikkyo University. For our long-running learning and assessment symposium, we present commentary by Dr. Irina Elgort of Victoria University of Wellington on papers by Chie Ogawa, Michael Holsworth, Haidee Thomson, and Darrell Wilkinson.

We wish our readership the best in these difficult times. While it is satisfying to see these papers published despite the current state of the world, we hope to see you at a literal, rather than merely virtual, event in the near future.

Jeffrey Stewart
JALT Vocabulary SIG Program Director
Vocabulary Learning Using Student-Created Class Vocabulary Lists

G. Clint Denison\textsuperscript{a} and Imogen Custance\textsuperscript{b}
\textsuperscript{a}Mukogawa Women's University; \textsuperscript{b}Kwansei Gakuin University

Abstract

In this article, we describe the pedagogical basis for class vocabulary lists (CVLs) and their implementation using Google Sheets. CVLs allow students to collaborate and build “notebooks” of vocabulary that they feel is important to learn. CVL choices of students ($N = 53$) in three classes of mixed non-English majors and one informatics class were compared against frequency-based lists (British National Corpus/Corpus of Contemporary American English Word Family Lists [BNC/COCA], New General Service List [NGSL], Test of English for International Communication [TOEIC] Service List [TSL]) using the Compleat Web Vocabulary Profiler (Web VP) to determine the usefulness of the selected vocabulary. An information technology keywords list, constructed using AntConc and AntCorGen, was compared against the informatics group’s CVL to determine if those students were choosing field-appropriate vocabulary. Results suggest that when given autonomy to choose vocabulary, students generally select useful and relevant words for their contexts (e.g., simulation, virtual, privacy, artificial, denuclearization, aftershock, heatstroke) and that CVLs supplement frequency-based lists in beneficial ways.

1. Background

In second and foreign language (L2) vocabulary research and instruction, it is generally accepted that learning will be more efficient if the most frequent words are studied first. To this end, much research has served to identify the most frequent English words, producing general purpose lists such as the BNC/COCA (Nation, 2012) and New General Service List (NGSL; Browne, Culligan, & Phillips, 2013) as well as more specific lists such as the TOEIC Service List (TSL; Browne & Culligan, 2016). In principle, such lists contain the words that should be learned first in a given context and are vital tools for improving the efficiency of instruction.

Although frequency-based lists answer the question of what to study, how to study must also be addressed. Successful vocabulary learning involves a combination of implicit and explicit instruction, meaningful engagement, and repeated meetings with words (Nation, 2013), facilitated by careful integration of words into a curriculum, systematic instruction, and repetition. However, such an intensive approach can be poorly implemented in many instructional contexts when important components of systematic instruction and curriculum integration are missing (Barker, 2007). Although instructors often refer to frequency-based lists when leveling materials and selecting words, the impact of such endeavors might be reduced...
when subsequent instructors adopt such an approach, attenuating the rate at which students can learn new words and expand coverage. In programs lacking sufficient coordination, the learners, not the materials or instructors, are the most stable elements in the system (Barker, 2007), and training learners to take responsibility for vocabulary development might be more sustainable.

For those studying an L2 to engage in a specific field, there are additional complications in vocabulary selection. The most frequent vocabulary between fields often differs, and language instructors might be limited in their knowledge of what vocabulary would be most useful in a particular field, a problem that is exacerbated as study becomes more specialized (Anthony, 2018). For specialized fields, developing students’ study skills might be more efficacious than relying on their instructors or using readily available frequency-based lists, which might not be specific to their fields of study.

One way to develop learners’ ability to study vocabulary is through the use of vocabulary notebooks, which have been shown to be effective pedagogical tools (e.g., Walters & Bozkurt, 2009). Although vocabulary notebooks are not novel, technology has allowed for new adaptations. Platforms such as Google Sheets (http://sheets.google.com/) present an opportunity for learners to collaborate and build a shared “notebook” to drive vocabulary learning. These student-created class vocabulary lists (CVLs) have the potential to increase motivation through a combination of collaborative and autonomous learning while simultaneously developing the ability to choose and study vocabulary. The shared format allows CVLs to be integrated into classroom activities and assessment, blending independent learning and classroom instruction.

2. Methodology

In this exploratory study, we examined CVLs and student word choice across two different classroom contexts using the Compleat Web Vocabulary Profiler (Web VP; Cobb, n.d.), AntConc (Anthony, 2019a), and AntCorGen (Anthony, 2019b). We hypothesized that students would generally choose relevant and useful words for their own learning contexts. However, it was unclear to what degree choices would overlap with publically available lists (i.e., NGSL, TSL, BNC/COCA) that contain the most useful words of English based on frequency, or to what degree students would choose words specifically relevant to their major.

2.1. Participants

The study involved participants from two groups (N = 53). The first group involved a special advanced English program for non-English majors at a women’s university in Japan. Twenty-six students in second-year (n = 12), third-year (n = 10), and fourth-year (n = 4) cohorts participated over two semesters. The second group involved a required English communication course for first-year students in the science and technology department at a coeducational university in Japan. One class of first-year informatics students (n = 27) participated over one semester.
2.2. Student-Created Class Vocabulary Lists

A CVL for each class—three in the first group, one in the second group—was created using Google Sheets. Editing permission was granted to students, allowing for simultaneous access during study, homework, and in-class activities. Students had access to their class’ sheet only. Based on the vocabulary notebook recommendations of Schmitt and Schmitt (1995), the spreadsheet included columns for the English word, the Japanese translation, a pronunciation code, the part of speech, a usage example, and related words. The pronunciation code (B. Teaman, personal communication, February, 2018) consisted of two digits: the first being the number of syllables in the word and the second being the stressed syllable. The code was intended to direct students’ attention to the phonology of words in addition to the orthography. Students were asked to include an example sentence that clearly showed the meaning of the word. The related words column was intended to help integrate new vocabulary into existing lexical networks as “the best way to remember new words is to incorporate them into language that is already known” (Schmitt & Schmitt, 1995, p. 133).

To build the CVLs, students were required to choose words that they deemed useful or important and complete each column in the spreadsheet. Although students were given autonomy, they were instructed on how to make good choices following Barker’s (2007) guiding questions. Before choosing words, students were encouraged to ask themselves, “Have I seen it more than once?” “Would I use the translation of this word in my language?” and “Do I have a special reason for wanting to know this word?” among others (see Barker’s guidelines for a full list). Instructors worked as curators, correcting errors when necessary and possible.

In the first group, participants added three words every week for two semesters. To direct students’ attention to common words, NGSL and TSL words were identified using the Web VP (Cobb n.d.) and highlighted by the instructor. CVLs were used to make quizzes and activities for each class. A partial CVL created by second-year participants in the first context is shown in Figure 1.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>Japanese</td>
<td>Pronunciation</td>
<td>Part of Speech</td>
<td>Related words</td>
<td>Example Sentence (show meaning clearly)</td>
</tr>
<tr>
<td>421</td>
<td>prominently</td>
<td>明るく</td>
<td>4-1</td>
<td>adjective</td>
<td>stand out, outstandingly, marked, noticeable</td>
<td>Many designers display logo prominently on clothes and bags.</td>
</tr>
<tr>
<td>422</td>
<td>engagement</td>
<td>関係</td>
<td>3-2</td>
<td>noun</td>
<td>involvement, participation</td>
<td>Japan continued political and economic engagement with Asia countries.</td>
</tr>
<tr>
<td>433</td>
<td>subtle</td>
<td>薄い</td>
<td>2-1</td>
<td>adjective</td>
<td>almost, nearly, slightly</td>
<td>There are subtle differences between the two.</td>
</tr>
<tr>
<td>441</td>
<td>scope</td>
<td>幅広い</td>
<td>1-1</td>
<td>noun</td>
<td>visual field, view</td>
<td>The teachers have to secure the wide activity scope for children.</td>
</tr>
<tr>
<td>442</td>
<td>bleak</td>
<td>寒い</td>
<td>1-1</td>
<td>adjective</td>
<td>cold, chilly</td>
<td>Hokkaido always blows a bleak wind in winter.</td>
</tr>
<tr>
<td>443</td>
<td>vaccine</td>
<td>ワクチン</td>
<td>2-2</td>
<td>noun</td>
<td>inject some vaccine</td>
<td>Children has to take vaccine for preventing some diseases.</td>
</tr>
<tr>
<td>444</td>
<td>ban</td>
<td>訴える</td>
<td>1-1</td>
<td>verb</td>
<td>Suppression, Oppose</td>
<td>He was banned from driving.</td>
</tr>
<tr>
<td>445</td>
<td>criticism</td>
<td>批判</td>
<td>4-1</td>
<td>noun</td>
<td>referee, judgment</td>
<td>I need literary criticism.</td>
</tr>
<tr>
<td>446</td>
<td>prevail</td>
<td>有利</td>
<td>2-2</td>
<td>verb</td>
<td>advantageous, to be in the lead</td>
<td>Right will prevail in the end.</td>
</tr>
<tr>
<td>447</td>
<td>forbidden</td>
<td>禁止する</td>
<td>3-2</td>
<td>verb</td>
<td>give praise</td>
<td>My supervisor complemented me on my work.</td>
</tr>
<tr>
<td>448</td>
<td>enterprise</td>
<td>事業</td>
<td>3-1</td>
<td>noun</td>
<td>prohibited, banned</td>
<td>It is forbidden to come in here.</td>
</tr>
<tr>
<td>449</td>
<td>settlement</td>
<td>解消,解決</td>
<td>3-1</td>
<td>noun</td>
<td>agree, solve</td>
<td>I reach a settlement.</td>
</tr>
<tr>
<td>450</td>
<td>dine</td>
<td>食事をする</td>
<td>1-1</td>
<td>verb</td>
<td>eat, eat</td>
<td>He dines out once a week.</td>
</tr>
<tr>
<td>451</td>
<td>avenue</td>
<td>大通り</td>
<td>3-1</td>
<td>noun</td>
<td>road, big</td>
<td>There is an avenue of poplars.</td>
</tr>
<tr>
<td>452</td>
<td>obvious</td>
<td>明らか</td>
<td>3-1</td>
<td>adjective</td>
<td>open, overt</td>
<td>I had some obvious errors in my test.</td>
</tr>
</tbody>
</table>

Figure 1. Example CVL accessible through a shared Google spreadsheet.
In the second group, students considered topics they wanted to discuss in English, either related to their major or more general. Once topics were chosen (e.g., What are the good points of virtual reality?, What do you think about the Internet of things, etc.), students conferred in Japanese to decide what vocabulary would be necessary for those topics, and whether they knew those words in English. Unknown English words deemed necessary were added to the CVL. Each week students discussed one topic. After discussions, the CVL was reviewed and further words retrospectively deemed useful were added. Each student was asked to add three words to the CVL during the first six weeks of the semester. Students were free to focus on a single topic, however, they were encouraged to include words that could be used in a range of topics. Words on the CVL were tested as part of mid and end of term exams.

2.3. Analysis

To determine if students were choosing frequently occurring words, choices were compared against the NGSL, TSL, and BNC/COCA lists using the Web VP (Cobb, n.d.). These lists contain the most frequently used words in English (i.e., NGSL and BNC/COCA) and words needed for qualification testing (i.e., TSL), both considered important by the participants. Each CVL was analyzed to determine the degree of overlap between CVLs and the above-mentioned lists. Student choices not accounted for on the lists (i.e., off-list words) were also examined. Word types rather than tokens were examined to control for repetition.

For the second group, an additional comparison was made using a keyword list generated from an information technology (IT) corpus to determine how many words specific to the field of informatics had been selected. The corpus of 5,190 journal articles was created using AntCorGen (Anthony, 2019b) and included a subset of articles with computer and information sciences tags in the PLOS-One database. Articles with these tags but from PLOS publications other than the PLOS-One journal (e.g., PLOS Pathogens) were excluded. A keyword list for this corpus was created in AntConc (Anthony, 2019a) using the BNC word frequency list as the reference. Keyness values were calculated using a log-likelihood (four terms) keyword statistic with a \( p < 0.0001 \) Bonferroni-corrected significance threshold. An odds ratio (OR) was calculated for effect size. Both keyness and OR were examined following Pojanapunya and Watson Todd (2018), who suggested that log-likelihood keyness highlights words of general use, while OR highlights more specialized words. Only words with keyness values greater than 1,000 were compared against the CVL.

3. Results and Discussion

In the first group, second-, third-, and fourth-year students chose 504, 490, and 208 word items, respectively, while informatics students in the second group chose 111.

The distribution of word types by percentage is shown in Tables 1 and 2, while visualizations are shown in Figures 2 and 3.

Results suggest that students largely chose words considered useful in general. The majority of choices made by second-year (64.09%) and third-year
(66.95%) students in the first group and informatics majors (77.4%) in the second group were covered by NGSL and TSL. Informatics students chose a greater proportion of high-frequency words, possibly due to the focus on speaking tasks in their course. The exception was fourth-year students in the first group whose off-list choices constituted 54.33% of their CVL. This difference can partially be explained by the advanced level of those students and the content of the fourth-year discussion course, which focused on authentic materials from English news sources. Off-list choices such as denuclearization, noncompliance, and statesman were all chosen from news articles. Similarly, aftershock and heatstroke were selected by third-year students due to their relative importance in Japan, as were

Table 1. Percentage of Total Tokens for Word Types by List (NGSL/TSL)

<table>
<thead>
<tr>
<th>Frequency level</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Informatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGSL 1</td>
<td>13.10</td>
<td>16.12</td>
<td>12.50</td>
<td>30.6</td>
</tr>
<tr>
<td>NGSL 2</td>
<td>17.06</td>
<td>17.76</td>
<td>14.42</td>
<td>26.1</td>
</tr>
<tr>
<td>NGSL 3</td>
<td>13.10</td>
<td>14.29</td>
<td>12.02</td>
<td>11.7</td>
</tr>
<tr>
<td>TSL</td>
<td>20.83</td>
<td>18.78</td>
<td>6.25</td>
<td>9.0</td>
</tr>
<tr>
<td>Total on-list</td>
<td>64.09</td>
<td>66.95</td>
<td>45.19</td>
<td>77.4</td>
</tr>
<tr>
<td>Off-list</td>
<td>35.71</td>
<td>32.86</td>
<td>54.33</td>
<td>22.52</td>
</tr>
</tbody>
</table>

Note: Due to rounding, percentage totals might not equal to 100.

Table 2. Percentage of Total Tokens for Word Types by List (BNC/COCA)

<table>
<thead>
<tr>
<th>Frequency level</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Informatics</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1</td>
<td>15.87</td>
<td>14.08</td>
<td>12.98</td>
<td>27.0</td>
</tr>
<tr>
<td>K2</td>
<td>17.26</td>
<td>20.61</td>
<td>17.79</td>
<td>34.2</td>
</tr>
<tr>
<td>K3</td>
<td>29.37</td>
<td>29.18</td>
<td>27.88</td>
<td>22.5</td>
</tr>
<tr>
<td>K4</td>
<td>11.90</td>
<td>10.20</td>
<td>11.54</td>
<td>7.2</td>
</tr>
<tr>
<td>K5</td>
<td>11.71</td>
<td>6.33</td>
<td>6.73</td>
<td>0.9</td>
</tr>
<tr>
<td>K6</td>
<td>4.56</td>
<td>4.08</td>
<td>7.69</td>
<td>—</td>
</tr>
<tr>
<td>K7</td>
<td>3.37</td>
<td>4.08</td>
<td>2.40</td>
<td>1.8</td>
</tr>
<tr>
<td>K8</td>
<td>1.19</td>
<td>2.04</td>
<td>1.44</td>
<td>—</td>
</tr>
<tr>
<td>K9</td>
<td>0.40</td>
<td>1.43</td>
<td>2.88</td>
<td>1.8</td>
</tr>
<tr>
<td>K10</td>
<td>0.20</td>
<td>0.61</td>
<td>1.44</td>
<td>—</td>
</tr>
<tr>
<td>K11</td>
<td>0.40</td>
<td>0.61</td>
<td>2.40</td>
<td>0.9</td>
</tr>
<tr>
<td>K12</td>
<td>—</td>
<td>0.61</td>
<td>1.44</td>
<td>—</td>
</tr>
<tr>
<td>K13</td>
<td>0.20</td>
<td>0.61</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>K14</td>
<td>0.40</td>
<td>—</td>
<td>0.48</td>
<td>—</td>
</tr>
<tr>
<td>K15</td>
<td>0.20</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>K16</td>
<td>0.40</td>
<td>—</td>
<td>0.48</td>
<td>—</td>
</tr>
<tr>
<td>K17</td>
<td>0.20</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>K18</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.9</td>
</tr>
<tr>
<td>K19</td>
<td>0.20</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Off-list</td>
<td>1.98</td>
<td>5.31</td>
<td>1.92</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Note: Due to rounding, percentage totals might not equal to 100. Words beyond the K19 band were not selected.
landslide and blackout, by the informatics students. These choices suggest that there was often a clear learning goal when off-list words were selected.

As for words selected by informatics students, 24 (21.6%) also appeared on the IT keyword list (see Table 3), indicating that students selected several words that were also relevant to their major. Of these items, five are not present on the NGSL or TSL lists, suggesting that they would not have been targeted for instruction were only these frequency-based lists considered.

Figure 2. Percentage of total for word types by list (NGSL/TSL) for each class.

Figure 3. Percentage of total for word types by list (BNC/COCA) for each class.
4. Conclusion

Frequency-based lists have aided instructors, administrators, and materials writers in deciding what vocabulary should be targeted for instruction; however, the results of this study indicate that CVLs also have potential benefits. One advantage is that learners using CVLs choose relevant vocabulary at appropriate times, which is important in English for specific purposes (ESPs) contexts. Although a keyword list for a particular major could be constructed, its utility would be limited by the extent to which underlying concepts had already been established (Nation, 2008). Using CVLs mitigates this issue as students choose vocabulary based on known or recently encountered concepts, particularly useful when instructors are not experts in learners’ fields.

In general contexts, a learner-centered approach has the potential to increase relevance and engagement, thereby enhancing the learning environment while using CVLs to develop learners’ ability to choose and study vocabulary independently might be more effective in the long-term than teacher-fronted approaches that cannot assist learners beyond graduation. The fact that the majority of vocabulary selected was on-list (NGSL/TSL) for all except the fourth-year students also indicates that learners are capable of selecting words that are useful long term. Even for fourth-year students, off-list choices were reasonable and useful considering the learning context and represent a shift toward less frequent vocabulary as learning progresses.

Although we certainly do not argue against the use of frequency-based lists, CVLs have the potential to help develop not only vocabulary, but also important language learning skills that facilitate sustainable learning. There is clear potential to improve vocabulary instruction through further research examining effective ways in which frequency-based lists and CVLs can be used together.
References


The Indirect Spaced Repetition Concept

Louis Lafleur
Ritsumeikan University

Abstract
The main goal of this research is to systemize, build, and test prototype software to demonstrate Indirect Spaced Repetition (ISR) as a viable concept for Second Language Vocabulary Acquisition (SLVA). ISR is designed around well-founded spaced repetition and SLVA principles. Most importantly, it is based on Nation’s (2001) recommendation to consider all three tiers of word knowledge (meaning, form, and function/use) and subsequent 18 aspects of word knowledge for a more balanced approach in teaching and learning vocabulary. ISR prototype software was achieved in the conceptual phase of the research. The resulting prototype flashcard software was given an in-depth trial for a period of 2 weeks by seven university students. Participants were given a post-project survey to evaluate ISR software (ISRS) under four categories: enjoyment, usefulness, usability, and general consideration. Post-test survey findings showed above-average satisfaction and consideration to use such software in the future. However, these findings also revealed that some areas could be further improved, such as addressing some hardware/software issues (e.g., IT infrastructure problematics and lag) and integrating gamification elements (e.g., performance feedback/reports).

Keywords: Vocabulary learning, (Indirect) Spaced Repetition, (Spaced) Interleaving, 18 aspects of word knowledge, Computer Assisted Language Learning (CALL)

1 Background
Spaced Repetition is often mistaken as a new concept as the term is often associated with recently published study software and applications. In many cases, these programs fail to give credit to the founders of the spaced repetition system (SRS). First, Hermann Ebbinghaus (1885/1964), a cognitive psychologist, established the spacing effect which accounts for why learners have better memory retention when they engage in spaced learning (multiple but short study sessions) compared to when they engage in massed learning (a single or very few long study sessions). Second, Pimsleur (1967) suggested an exponentially expanded spaced review schedule called graduated interval recall. Finally, Sebastian Leitner (1972) systemized such ideas into a spaced-interval-based box/compartment flashcard study system capable of organizing review cards/items across multiple intervals while allowing the addition of new items into the mix at one’s convenience.
Although digitized in recent years, the concept remains largely the same. The learner is presented with a question or a prompt first (traditionally, flashcard side A) for which they must try to recall or guess an appropriate answer before confirming it (traditionally, flashcard side B). If the learner’s recall is successful, then that item’s study/review interval spacing/length will be increased. After a number of consecutive successful recalls/reviews, the spacing of reviews will move from shorter (e.g., daily) to longer (→ weekly, etc.). However, if one’s recall fails, this would signify the need for a shorter study/review interval, and therefore, the interval should be reduced considerably (back to the previous or first interval, e.g., daily) before being increased again through successful recall.

In terms of study efficiency, there is an agreement that massed learning (or cramming) is not as efficient as spaced learning. However, there are two different schools of thought about how to spread out study intervals within spaced learning, expanded or uniform (see Figure 1).

- In most comparative studies, uniform interval lengths are longer than expanded intervals at first.
- A good research practice is to have an equivalent number of intervals and total period for fair comparison.

A number of studies found no statistical difference in efficiency between uniform and expanded intervals when tested shortly after their set study periods (Balota, Duchek, & Logan, 2007; Carpenter & DeLosh, 2005; Landauer & Bjork, 1978). However, longer delayed studies such as Schuetze and Weimer-Stuckmann (2010, 2011) showed that the uniform group outperformed the expanded group at 83% retention versus 59% on a 9-month post-test.

The decision to implement expanded interval spacing, thus far, for Indirect Spaced Repetition Software (ISRS) was counter-intuitive, as some efficiency is sacrificed in longer-term recall as Schuetze and Weimer-Stuckmann (2010, 2011)
have shown. However, expanded spacing is more practical in handling multiple study items/flashcards (e.g., 100 at a time and more) as review cards can be pushed further back more aggressively in later intervals (→ monthly, etc.). This aids to alleviate some of the review burden, which permits new cards to be introduced into the study mix more easily.

It is often overlooked that there are actually three types of “Expanding” spacing: +, ×, and a$^b$ (see Figure 2). What could be perhaps more interesting would be to compare these three types of expanded spacing among themselves in future research.

Beyond scheduling, another important area to address is the profundity of vocabulary study, a concept known as vocabulary breadth and depth.

On one hand, some SRSs place more importance on breadth (maximizing the number of words encountered) and less on study depth (aspects/levels of word knowledge). A typical example would be the unilateral use of a quick receptive knowledge recall exercise (side A = L2 word, and side B = corresponding L1 word).

On the other hand, some SRSs place more importance on depth, and thus less on breadth. In addition to L1/L2 words, other elements such as L1/L2 definitions, L1/L2 example sentences, audio, images/video may be included. These additional elements can enable a wider range of study, such as focus on forms, four language skills, both receptive and productive knowledge, and different task flows (L1 to L2, L2 to L1, and L2 to L2). One of the drawbacks of creating multiple tasks as additional items/flashcards is the creation of indirect overlap, where some of the prompts or affiliated sound/visuals of a flashcard may comprise the answer/recall elements of a different flashcard. This indeed upsets the desired cool-down period as it constitutes unscheduled review in between set intervals.

ISRS avoids such overlap by not adding extra tasks as additional items/flashcards but cycling between tasks according to the reached interval. This is similar to the concept of interleaving (the consecutive practice of multiple skills or concepts); however, in this case, the interleaved tasks are spaced across increasingly expanding intervals, interval-spaced interleaving (see Figures 3 and 4). ISRS’s interleaved tasks are based on three tiers of word knowledge: meaning, form, and function/use (Nation, 2001). ISRS’s sequencing is based on the work of Schmitt.
### Figure 3. Spaced-Interleaved Task Examples and Interval Breakdown for Indirect Spaced Repetition Software/System (ISRS).

<table>
<thead>
<tr>
<th>Tier Q/Task Level</th>
<th>Flow</th>
<th>Task Type</th>
<th>Task/Interval Route</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Meaning</strong> Q1/word or phrase</td>
<td>L2 audio to L1 recall</td>
<td>Recall Check</td>
<td>Session 1 (initial)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Session 2 (4h)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Session 3 (8h)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Session 4 (16h)</td>
</tr>
<tr>
<td><strong>Meaning</strong> Q2/word or phrase</td>
<td>L1 word/phrase to L2 recall</td>
<td>Recall Check</td>
<td>Session 5 (1d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Session 6 (2d)</td>
</tr>
<tr>
<td><strong>Form</strong> Q1/word or phrase</td>
<td>L2 audio to L2 word/phrase</td>
<td>Spelling</td>
<td>Session 7 (3d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Session 8 (6d)</td>
</tr>
<tr>
<td><strong>Form</strong> Q2/sentence</td>
<td>L2 (blank) to L2 sentence</td>
<td>Fill-in-the-blank</td>
<td>Session 9 (9d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Session 10 (18d)</td>
</tr>
<tr>
<td><strong>Use</strong> Q1/sentence</td>
<td>L2 sentence to L1 sentence</td>
<td>Writing</td>
<td>Session 11 (27d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Session 12 (54d)</td>
</tr>
<tr>
<td><strong>Use</strong> Q2/sentence</td>
<td>L1 sentence to L2 sentence</td>
<td>Writing</td>
<td>Session 13 (514d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Session 14 (162d)</td>
</tr>
</tbody>
</table>

*For increased effectiveness, ISRS can be combined with a fluency-building (voiced) reading task. For example, studying a set of word items, which compose a text could trigger such an activity.*

### Figure 4. Indirect Spaced Repetition System/Software (ISRS).

CALCULATING THE NEXT INTERVAL FOR IND.REPETITION

1. If the user answers the card/question successfully, the interval is increased by a factor of 1.5 or 2.
2. If the user answers the card/question successfully under the required time threshold, the interval is increased by a factor of 3. (optional)
3. Specific interval times may also be customized on-the-fly automatically through the collection of various user data.
4. Some developers may choose to set countdown timers, customize interval times, and amend task/interval progression routes for ISRS.
(2008) who implied that focusing on the meaning–form link at first and later enhancing context(use) may prove to be effective.

The following questions guided the ISRS testing:

1. Do the participants find the software useful for JHS (junior high school) students? (Usefulness assessment)
2. Do the participants find the software easy to use for JHS students? (Usability assessment)
3. Do the participants find the software enjoyable for JHS students? (Enjoyment assessment)
4. Do the participants think JHS teachers would like to integrate the software in their classes or in parallel to their classes (homework)? (Consideration assessment)

2 Methodology

The participants in this study were five undergraduate students of English education, one undergraduate student of elementary school education and one graduate student of psychology. Two of the participants were female, and the other five were male. Most had teaching experience through working at cram schools and most had undergone 4 weeks of teacher training in a public junior high school. They all completed pre-project and post-project surveys that collected quantitative data (Likert-scale, 1 = Strongly Disagree ~ 5 = Strongly Agree) and qualitative data (comments) about the participants’ perspectives of e-learning and the ISRS prototype software.

3 Results and Discussion

The pre-project survey revealed that four participants had prior experience using SRS software. Although the majority agreed that e-learning can make study more effective \( (M = 3.71, \text{standard deviation} [SD] = 0.76) \), the reality was that they did not implement e-learning in their own study of English to the level of their beliefs \( (M = 2.86, SD = 1.21) \) (Lafleur, 2015). Neither apathy nor enthusiasm regarding e-learning or SRS could be ascertained.

Regarding the interval scheduling and task interleaving of ISRS, participants were quite content with the prototype software and a good number of participants

<table>
<thead>
<tr>
<th>Perceptions (5 point Likert-scale, 5 = strongly agree) and Comments (P# = Participant #)</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Consideration) I think that teachers would like to incorporate such software in their class.</td>
<td>3.86</td>
<td>1.21</td>
</tr>
</tbody>
</table>

“*The all-surrounding Internet equipped environment required by the software is probably the biggest hurdle facing current classroom integration.*” (P3)

“One remaining problem is whether students have computers or not to enable review at home.” (P5)

“I think that with the cooperation of teachers this tool/software’s effectiveness can be even greater for students. I want to incorporate such a system as part of the routine of my future classes.” (P6)

Figure 5. Post-project (consideration) survey results and comments.
noted that they would consider using such a system as teachers \( (M = 3.86; \text{Figure 5}) \). However, it should be noted that most post-project survey comments/evaluations were not related to the scheduling and task interleaving of ISRS but the integration or lack of other external features (see Figures 6 and 7).

In terms of usability (see Figure 8), there were some concerns about the digital fluency level of potential users, especially younger students and older teachers. Also, the potential lack of in-class/in-school digital facilities and resources were also a point of contention (e.g., lack of [tablet] computers and/or lack of Wi-Fi connectivity). Finally, although few, there were also some apprehensions about assigning online review as homework because of differences in students' access to reliable internet.

In terms of enjoyment (see Figure 7), two important areas for improvement were identified: first, the need for more detailed progress boards (individual

<table>
<thead>
<tr>
<th>Perceptions (5 point Likert-scale, 5 = strongly agree) and Comments (P# = Participant #)</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Usefulness) I think that this software is an effective tool for JHS students learning English.</td>
<td>3.71</td>
<td>0.76</td>
</tr>
</tbody>
</table>

"It's good for students because the words they must memorize or recall are suggested automatically. So, I think it's effective." (P1)

"At the JHS level, I believe that students who have difficulty with English also have a general difficulty with memorization. With the use of such a tool, students can become more familiar with the concept of memory and increase it. Moreover, as the software asks users to type the words frequently, memorization becomes easy." (P2)

"The loaded content (questions/tasks) might be too easy for some students, so if the users could choose the difficulty level it would make the software better." (P4)

"I think that this software is a very effective tool for learning English because it utilizes a quiz style approach. Also when answers are checked we can listen to the pronunciation." (P5)

"I think this software would be better if it provided the users with more detailed records of their progress and a daily performance report; this could be used as a reminder to login. Also, some sentences took some time to be loaded (lag)." (P6)

Figure 6. Post-project (usefulness) survey results and comments.

<table>
<thead>
<tr>
<th>Perceptions (5 point Likert-scale, 5 = strongly agree) and Comments (P# = Participant #)</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Enjoyment) I think that JHS students would enjoy using such software as part of their study.</td>
<td>3.57</td>
<td>0.98</td>
</tr>
</tbody>
</table>

"As users progress through the software, I think that they need to have a better sense of their progress. For example, just as in a game, the screen could change and show their level going up." (P2)

"Students are used to writing on paper or listening to CDs, so using software with their PC or phone would be fun and interesting to them." (P4)

"Incorporating native English voices is a great point and they can feel the passion (of the characters/voice actors). The change in colors of buttons when running out of time was a nice touch too." (P6)

"The software could be more convincing with illustrations/animations of words, as just writing “good” or “wrong” in plain text seemed a bit regrettable." (P6)

Figure 7. Post-project (enjoyment) survey results and comments.
progress bar and leaderboards) and, second, the need for further gamification, such as an in-game reward system linked to achievement, but perhaps more importantly to consistency in participation. For increased efficiency, such progress and rewards should be linked to students’ overall participation scores.

4 Conclusion

The main goal of this research was to systemize, build, and test prototype software to demonstrate ISRS as a viable concept. This was achieved; however, the conducted survey revealed some areas that could be further improved as discussed above. To use the analogy of high cuisine where plating or presentation is as important as taste, in today’s context of e-learning, gamification features are just as important as the core structure of learning itself.

This is a testament to the importance of gamification for users, which has become a research area of interest for the author. A second area of importance is further research in spacing algorithms, especially comparing expanded types of algorithms (+, ×, and a\(^b\)) among themselves. A third area of importance is building various word lists appropriate for different levels of learners.

Some of the author’s current projects on a similar line of investigation include research on teachers’ perspectives on word counting units, creating common core elementary/junior high level word lists, and testing and supplementing the New Academic Word List (NAWL; Browne, Culligan, & Phillips, 2013) with example sentences and translations (Kanazawa & Lafleur, 2019).

Acknowledgments

The author is grateful to his former professors at Okayama University who supported this research in its early phases, to the staff at MageMontreal for their help in web application programing, and, finally, to the reviewer and journal editor for their advice and assistance.

Author’s contribution section

The empirical participant data used in this article are taken from the author’s 2015 master thesis study. Otherwise, other sections including changes made to the publicly available ISRS concept are unique to this article.

---

**Table 8. Post-project (usability) survey results and comments.**

<table>
<thead>
<tr>
<th>Perceptions (5 point Likert-scale, 5 = strongly agree) and Comments (P# = Participant #)</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Usefulness} I think that this software is an effective tool for JHS students learning English.</td>
<td>3.43</td>
<td>0.98</td>
</tr>
</tbody>
</table>

“*The font on the cellphone version is hard to read, even more so for JHS students. Moreover, the software doesn’t include enough explanations in terms of how-to use it.*” (P3)

“For people who are not accustomed to using computers or smartphones, using this software may prove a little difficult. But most students will be able to easily use it without any trouble.” (P4)
References


Vocabulary Learning and Instruction, 9 (2), 9–16.
On Creating a Large-scale Corpus-based Academic Multi-word Unit Resource

James Rogers
Meijo University

Abstract
This study outlines the steps taken to create an academic multi-word unit list derived from corpus data. It gives details on the procedure used and the rationale behind why certain approaches were utilised. It also compares existing resources and makes some suggestions for practical use of the resulting resource.

Keywords: English for specific purposes, academic English, collocation, formulaic language, multi-word units, corpora

1 Introduction
This article outlines a project involving the construction of a corpus-based academic English resource list which focuses on academic multi-word units (MWUs). In recent years, there has been increased awareness of the importance of collocational and MWU fluency for second-language learners, and there is an agreement on the value of such knowledge (Nation, 2013; Siyanova-Chanturia & Pellicer-Sanchez, 2019; Webb & Kagimoto, 2011). Lewis (2000) stated that mastering collocational knowledge should be “a top priority in every language course” (p. 8). Other researchers echoed this sentiment, remarking that much of language is made up of prefabricated chunks, and thus learning them is essential for obtaining language fluency (Hill, Lewis & Lewis, 2000; Hoey, 2005).

1.1 Defining Word Co-Occurrence
MWU is one of the many terms used by researchers to refer to word co-occurrence is referred to by researchers. Some narrowly define the linguistic phenomenon of two words occurring together in high frequency with the term collocation (Hoey, 1991). Others define collocations by considering syntactic structures (Gitsaki, 1996), with others utilizing a combination of multiple criteria (Lesniewska & Witalisz, 2007). When the term MWU is used, it can include collocations, but there is a lot of variation of what it specifically refers to. For instance, research has broken down such language into literals, figuratives, and core idioms (Grant & Nation, 2006). Others consider only two-word phrases as collocations and all others as idioms or lexical bundles (Biber, Johansson, Conrad, Leech, & Finegan, 1999). All of these definitions are appropriate, and depend on the type of research, but for this current study, the linguistic phenomenon of word co-occurrence will be defined as two or more words that co-occur in high frequency, and these will be referred to as MWUs.
1.2 Currently Available MWU Resources for Non-Native Learners of Academic English

Creating a usable MWU list for language learners is a complex process, and currently available resources are highly limited. The first issue is that they are of small scale. Liu’s (2012) study only identified 228 MWUs and Simpson-Vlach and Ellis’s (2010) study identified only 207 core items. Chon and Shin (2013) identified slightly more (934 written collocations) but their study was actually small in scope in that their list was derived from only 20 high-frequency academic node words. In comparison, this study begins with 500 node words and ended up with approximately 5,000 MWUs.

1.3 Concgramming as a Method of Identifying High-Frequency MWUs

A concgram “constitutes all the permutations of constituency and positional variation generated by the association of two or more words” (Cheng, Greaves, & Warren, 2006, p. 411). When this method is utilised, it allows for consolidation of overlapping items. This makes for more efficient learning in that very similar MWUs are not listed twice but at separate points in a list (e.g., this study found and these studies found). Concgramming also enables more accurate frequency ratings because consolidation of MWUs that only differ slightly by grammatical inflection are counted as one.

With concgramming, all lemmas of two co-occurring tokens are counted. They are counted while accounting for constituency variation. Therefore, when a corpus search is conducted for the lemma *make* and *money*, MWUs such as *make money* and *make some money* are counted. Concgramming also accounts for positional variation, so a search for the lemma *advice* and *give* will count instances of *advice you give* and *give you advice*.

Previous research has instead utilised the more simplistic n-gram corpus analysis to identify high-frequency MWUs that occur in a linear sequence. However, this method does not have the advantages as that of the concgramming method discussed above. The existence of non-consolidated partial duplicates and/or inaccurate frequency rankings in n-gram method-based studies creates limitations in the resources those studies result in, thus constituting a major gap in the research, since no large-scale studies on identifying academic MWUs have been done using the concgramming method.

1.4 The Need to Extend a Core MWU Beyond the Initial Identified Unit

When the concgramming method is utilised, concordance software will provide the most common MWUs in the corpus data, ranked by frequency. It is important to realise that the most frequent MWUs identified could be improved upon if they are extended beyond the core items. Often, extending MWUs in this way provides valuable information for learners on their extended contextual usage. A clear example of this is how a concgram search for the lemma *numerous* and *study* identifies *numerous studies*.
as the most frequent, but that when numerous studies occurred, the majority of the time the next word was have, and when numerous studies have occurred, the majority of the time the next word was numerous studies have shown (see Table 1). In Rogers (2017), experienced English as a foreign language (EFL) practitioners manually examined such data and decided whether or not to extend a MWU in this way with learners in mind. In that study, 53% of the items identified were extended, and the fact that the majority of the items were extended indicates that this may be an important criterion to consider.

2 Procedure

2.1 Search

This study began by using the most frequent 500 lemmas in Gardner and Davies’ (2013) high-frequency academic vocabulary list as pivot words to search for lemmatised collocates in the academic section of the Corpus of Contemporary American English COCA (Davies, 2008), a corpus consisting of over one billion tokens of American English sourced from material from 1990-2017, which is evenly divided into the following five genres: spoken, fiction, magazines, newspapers, and academic journals. For each collocate found, a file was created with 500 concordance lines from the academic section of the COCA containing both the pivot and the collocate. These files were analysed using the custom concordance software AntWordPairs (Anthony, 2013) to identify the most frequent MWUs that both the pivot and collocate occurred in.

Following the parameters set by Rogers’ (2017) study, which accomplished the same task set in this study but for general English, one occurrence per million tokens was frequency cut-off. Previous research has also implemented a parameter using the statistical measure of mutual information (MI), and Stubbs (1995) and Hunston (2002) recommend an MI cut-off score of 3 or higher for collocates, and therefore this study also only considered collocates with such a score.
2.2 Manual Removal of Noise

High-frequency and an M.I. score of 3 or higher does not always result in useful collocations being identified. The results of Rogers’ (2017) study indicated that manual checking of data was essential, and therefore, six experienced EFL practitioners’ intuition was used to identify and remove items that were not useful for learners. Such items included proper nouns, noise in the data, and also MWUs which only occurred in a particular genre of academic English, and therefore, had little value for learners of general academic English. Each practitioner had their own list to review. After they reviewed it, a second practitioner also went through the entire list. Any items flagged as not being useful for learners were reviewed by a different practitioner, and if that practitioner agreed, the item was removed from the list. The practitioners were given a protocol to follow, which listed potential reasons why an item would be flagged. The protocol was to flag any items if they were:

- **Proper Nouns**: Items such as organisation titles, journal titles, descriptions of particular ethnicities, etc. Some examples that were already flagged include *Center for School Counseling Outcome Research, Census of Population and Housing, Native American population* and *Muslim population.*

- **Grammatical Formulations**: Items that are devoid of meaning as a whole but occur frequently due to grammar. An example that was already flagged is *positive negative.* When further data were analysed, this item was revealed to be part of a list, with the two words separated by a comma, such as in *positive, negative and neutral.*

- **Too Specific to Particular Academic Fields**: Items which only tend to occur in particular kinds of academic fields and therefore do not have general value for learners of academic English. Some examples that were already flagged include *the primary tumour, God’s presence and reading fluency.*

- **Too Technical**: Items which occur frequently due to the particular types of journal in the corpus, but which do not general value for learners of academic English. An example that was already flagged is *stance phase.*

- **Not Particular to Academic English**: Items which occur more often in general English and are not particular to academic English. An example that was already flagged is *obstacle course.*

- **Unnatural Formulations due to Formatting**: Items that are the result of the way a journal formats their papers with titles which are not actually MWUs. An example that was already flagged is *Results Preliminary Analysis.*

- **Others**: Any items deemed to have little or no value for learners who aim to improve their academic English fluency based on your teaching experience and native-like speaker intuition.

3 Results

3.1 Item Analysis

Initial results identified a total of 10,190 collocations. After these were analysed by the EFL practitioners, 5,057 of them (49.6%) were judged to be useful
for learners who aim to improve their general academic English fluency. The vast majority of the items judged to have little value only occurred frequently in particular genres, such as medical academic English. Table 2 lists every 500th item to highlight the type of MWUs that were deemed useful.

### 3.2 On Extending Core MWUs Beyond the Initial Identified Unit

A majority of the 5,057 items (53.9%) were extended beyond their initial identified unit. This percentage was similar to that found by Rogers (2017), in which intuition was utilised in comparison to this study’s usage of frequency data.

### 3.3 Accessing the Results

The resulting resource can be accessed via a custom webpage at https://www.smartsmart.org/academic-english. This website enables users to view the entire list ranked by frequency, along with a custom example sentence written with EFL learners for each multi-word unit. The site also allows users to search for any MWUs that contain or start with a particular word or words.

### 4 Discussion

#### 4.1 Findings

This study filled gaps in the literature of scope and methodology that previous studies that attempted to identify academic MWUs had. The experienced EFL practitioners who participated considered the approximately 5,000 MWUs identified to all be useful for learners intending to improve their general academic English fluency. This study also confirmed a previous study’s finding that manual analysis of data is essential for creating useable learning materials in that approximately half of the items initially identified were manually removed because they

Table 2. Every 500th Academic MWU Identified in This Project (with Core MWUs Identified in Bold)

<table>
<thead>
<tr>
<th>Rank in list</th>
<th>Frequency</th>
<th>MWU identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36,167</td>
<td>the results of this study suggest that</td>
</tr>
<tr>
<td>500</td>
<td>779</td>
<td>gain skills and knowledge</td>
</tr>
<tr>
<td>1000</td>
<td>477</td>
<td>described in detail</td>
</tr>
<tr>
<td>1500</td>
<td>330</td>
<td>highly effective</td>
</tr>
<tr>
<td>2000</td>
<td>253</td>
<td>samples were analysed</td>
</tr>
<tr>
<td>2500</td>
<td>209</td>
<td>situational factors</td>
</tr>
<tr>
<td>3000</td>
<td>176</td>
<td>consistent with the literature</td>
</tr>
<tr>
<td>3500</td>
<td>151</td>
<td>design features</td>
</tr>
<tr>
<td>4000</td>
<td>130</td>
<td>potential implications for</td>
</tr>
<tr>
<td>4500</td>
<td>113</td>
<td>requires detailed</td>
</tr>
<tr>
<td>5000</td>
<td>100</td>
<td>widely disseminated</td>
</tr>
</tbody>
</table>

Rogers: Corpus-based Academic Multi-word Unit Resource
were deemed to be not useful for learners despite falling within the parameters set. It should be noted that such manual checking of data is extremely time-consuming, and therefore, future software development should consider any possible ways to automatise this task and/or corpus compilers should consider ways in which this issue could be avoided when developing and organizing a corpus.

5 Conclusion

Much more research is still needed with regard to methodology of identifying MWUs. However, this study is a first step towards confirming the previous study’s findings. It also provides teachers, students and researchers with a large-scale resource that can be immediately used, which previously constituted a major gap in the research. Future research is also needed to discover how such a resource could be best studied or utilised in language courses which aim to improve learners’ academic English fluency.

References


Exploring the Effectiveness of Deliberate Computer-Assisted Language Learning

Andrew Obermeier
Kyoto University of Education

Abstract
This article presents a work-in-progress focused on developing an experiment to investigate the effectiveness of different types of deliberate paired-associate computer-assisted language learning (CALL). First, the rationale for Japanese EFL learners’ current need for doubling their efforts with this technique is explained. Next, an overview of research regarding the interface in second language acquisition is presented. This is followed by an explanation of results from a recent experiment. Questions and issues raised in that experiment are then discussed with regard to a proposal for a subsequent experiment that will be conducted during the semester starting in April 2020. In this proposed study, different conditions within Internet-based flashcard study will be the major experimental learning component. Psycholinguistic response time measures will be the main dependent variable, aimed at gauging gains in nondeclarative, or tacit L2 knowledge. In addition, online declarative measures and traditional offline measures of declarative knowledge will be used.

1 Background
For English learners in Japan, vocabulary size and vocabulary knowledge quality are two areas with deficiencies that must be addressed by institutions, researchers, teachers, and learners. Research concerned with vocabulary size emphasizes the knowledge gap of most L2 learners and the need for them to learn and remember large numbers of words. In Schmitt’s (2008) review of instructed second language vocabulary acquisition, he asserts that English as a Foreign Language (EFL) learners are faced with a daunting learning task if they truly wish to achieve adequate proficiency; in order to read common native speaker-level texts, 8,000 to 9,000 word families must be known, and for oral discourse estimates range from 5,000 to 7,000 word families. Considering these benchmarks, it can be seen that Japanese university graduates know far fewer words than the number that they need to know to function in English in meaningful, unhindered ways. In a cross-sectional design in which 3,449 Japanese university students took the Vocabulary Size Test (Nation & Beglar, 2007), McLean, Hogg, and Kramer (2014) estimated that the average university learner of English has a receptive vocabulary size of only 3,715 word families. To address the issue of deficient vocabulary size, Japanese schools ranging from elementary to university must continue to implement great increases in vocabulary size goals specified in curricula, and specify methods to help students learn the words.
Another important issue for teachers and researchers to address is with regard to the quality of vocabulary knowledge that is being developed and measured. In Japan, the overwhelming amount of test-focused English language learning required to pass entrance examinations has been well documented (Berwick & Ross, 1989; Gorsuch, 1995; Miyazaki, 1981). Such a situation has resulted in highly unbalanced high school curriculums, with a disproportional emphasis on deliberately teaching declarative knowledge that will be useful when taking tests. Schmitt (2008) stresses the need for principled and balanced learning, teaching, and material development. Vocabulary learning programs need to include both intentional and incidental learning opportunities. An example curriculum that balances intentional and incidental learning activities in a university-wide EFL reading program is explained by Hunt and Beglar (2005). For teachers and program designers to integrate intentional and incidental vocabulary learning, the four strands of meaning-focused input, meaning-focused output, language-focused learning, and fluency development, suggested by Nation, provide a useful structure (Nation, 2007; Nation & Yamamoto, 2012). In this way, previous studies concur that the average Japanese EFL learner needs to learn more than twice as many words as they currently learn by the time they graduate from university, and they need to learn these words in a way that balances intentional and incidental learning methods. That is to say, they need to learn many more words and develop higher quality knowledge of them.

This article is organized as follows: first, previous research on the interface in second language acquisition is reviewed. Next, in an experiment that operationalizes interface constructs in terms of online processing measures, the effects of flashcard-based learning will be discussed from a recently completed study by the author (under review). In this study, the popular Internet-based software Quizlet was used extensively for the deliberate paired-associate learning condition. Results focused mainly on measuring response times to investigate nondeclarative knowledge development and online processing abilities, and a more traditional declarative knowledge measure was also used. The final part of the article will discuss the next step that will be taken to build on findings from this recent research. The next experiment will similarly focus on deliberate learning activities and response time experiments to measure effects of this learning in terms of declarative, automatized declarative, and nondeclarative knowledge gains.

2 Deliberate Paired Associate Flashcard Learning and the Interface

The above rationale for the necessity to balance between intentional and incidental learning is rooted in the central theme of second language acquisition research on the interface. The overriding takeaway from this research is this need for balance between implicit and explicit learning and knowledge types. However, the greater effectiveness of explicit learning demonstrated in a multitude of studies confirms that deliberate learning should be positioned in the lead to drive acquisition (e.g., Doughty & Williams, 1998; Ellis, 2005; Hulstijn, 2003; Norris & Ortega, 2000). To close the problematic vocabulary size gap mentioned in the previous section, learners need to know that they can intentionally move forward on a clear path,
and systematically learn large numbers of words. Flashcard learning is an essential strategy that enables learners to do this and should be included within any well-balanced language learning program (Nation, 2007, 2013; Nation & Webb, 2011; Nation & Yamamoto, 2012). Thanks to modern technology, in addition to paper-based flashcards, learners can add digital flashcard learning to their language learning strategy repertoire to greatly increase the variety, convenience, systematization, excitement, enjoyability, and functionality of their practice. Digital flashcard learning is a strategy that has a vast potential to empower effective learning at all ages and levels, and there are abundant online applications available. Nakata (2011) conducted a thorough evaluation of nine popular, free, and available flashcard-based applications, offering valuable insights regarding which applications are best for different learning and teaching purposes. It is an essential strategy that is widely available on the Internet and must be incorporated into any serious L2 learner’s toolbox.2

Interface research has been central to the study of second language acquisition research for over 30 years. The first phase of the discussion was framed with regard to the relationship between explicit and implicit learning and knowledge, when Krashen (1982) famously began the discussion by controversially asserting his extreme position distinguishing between explicit learning and implicit knowledge acquisition. His no interface hypothesis stated that explicitly learned language can never be acquired as implicit knowledge. DeKeyser (1997, 1998) countered this position with the strong interface hypothesis, proposing that when explicit knowledge is practiced to the point of becoming automatized, it interfaces strongly with implicit knowledge development. This first stage of interface research culminated in the 2005 thematic issue of Studies in Second Language Acquisition. This seminal issue began with a brief introduction in which Hulstijn (2005) outlined much needed definitional distinctions between the constructs of implicit and explicit learning, implicit and explicit knowledge, and implicit and explicit memory. The issue consisted of five experiments concerned with the interface and was highlighted by a psychometric study by Ellis (2005b) that focused on measuring interface constructs. This issue culminated with a broad review by Ellis (2005a) that drew widely from the fields of applied linguistics, psycholinguistics, neurology, and cognitive science. His review took the middle ground between the no interface and strong interface positions, arguing that they are both partially correct. He argued that although implicit and explicit knowledge are completely separate neurological entities, based in distinct parts of the brain, they nevertheless interact dynamically in implicit and explicit learning processes to foster second language acquisition. Of these three positions, the no interface hypothesis has been for the most part rejected, and most researchers see the interface as ranging somewhere between the strong and the weak.

In 2015, a second special issue of Studies in Second Language Acquisition revisited the interface theme, and in the introduction, Andringa and Rebuschat (2015) focused on three central issues: the role of awareness; methodological concerns for processes of measuring awareness; and the implicit-explicit interface. In this phase of interface research, the discussion was reframed and operationalization became a central concern addressed in a variety of experiments (e.g., Andringa & Curcic, 2015; Bowles, 2011; Cintrón-Valentin & Ellis, 2015; Kim & Nam, 2017; Paciorek &

__Vocabulary Learning and Instruction, 9 (2), 24–38.__
Williams, 2015). Perhaps a precursor to this second special issue was the comprehensive book by Ellis et al. (2009). The first two chapters of this volume are written by Ellis, defining and then operationalizing the constructs of explicit and implicit learning, testing, and teaching. Key criteria proposed for operationalizing the constructs of explicit and implicit learning and knowledge were as follows: degree of awareness, time available, focus of attention, systematicity, certainty, metalanguage, and learnability. However, as can be seen by a brief perusal of these criteria, the discussion has primarily been concerned with the acquisition of grammar. Although degree of awareness, time available, focus of attention, and perhaps learnability can all be applied to the learning and knowledge of vocabulary, clearly systematicity, certainty, and metalanguage have more to do with grammar learning. A review of almost 30 years of interface revealed that interface research has mostly been about grammar acquisition.

There is a clear justification for considering the interface with a focus on lexical learning. Pinker (1998, 2000) explains that “the vast expressive power of language is made possible by two principles: the arbitrary sound–meaning pairing underlying words, and the discrete combinatorial system underlying grammar” (p. 219). If one accepts that the combinatorial system of grammar and the sound–meaning pairings of words are two discrete components of language, then it would make sense to argue for distinguishing the interface for vocabulary acquisition as separate from the grammar acquisition interface. Some may argue that words and rules are such intertwined language components, that proposing separate interfaces for grammar and vocabulary learning is problematic. Nevertheless, it makes sense to say that operationalizing implicit vocabulary knowledge gains is more straightforward than operationalization is for grammar. This is because words are more simple and cohesive units to investigate than grammar rules, which can be riddled with complexity, exceptions, and irregularities. In the next section, interface research with a lexical bias will be discussed.

3 Recently Completed Research Regarding the Vocabulary Interface

In pivotal breakthrough research, Elgort (2007, 2011) investigated the effects of deliberate learning on individual words, combining traditional explicit measures such as a meaning generation test with online response time measures to gauge nondeclarative knowledge gains. She found that deliberate flashcard learning resulted in effects that could be regarded as evidence of the acquisition of implicit knowledge. In a study on incidental learning during reading, Elgort and Warren (2014) explored the effects of contextual learning. Again, traditional declarative knowledge measures and nondeclarative tacit measures were used in the experimental design. Results revealed that in contextual L2 learning the number of encounters with words is the key predictor, but this learning is very complex and influenced by learner factors such as age, L2 lexical proficiency, first language, gender, learning strategies, and intrinsic versus extrinsic motivation. Further research on contextual learning investigated different aspects of word learning by monitoring participants’ eye movements in real-time while reading (Elgort, Brysbaert, Stevens, & Van Assche, 2018). Participants in this research
were advanced English learners with vocabulary size averages estimated at 8,855 word families. The eye-tracking procedures focused on the contextual learning of low-frequency English words such as *mulct* that the experimenters tested by inserting them to replace mid-frequency counterparts (for this example, the word *fine* was replaced. Both *mulct* and *fine* mean “monetary penalty,” or “compulsory payment”). Results from their investigation confirmed previous research findings that contextual word learning is a slow and incremental process. Furthermore, they explain that when cognitive processing resources are spent on understanding word meanings, reading conceptually sophisticated texts becomes far more problematic. They conclude the article explaining that their findings confirm that L2 reading materials should be carefully planned to meet the appropriate frequency levels required to provide conditions that make incidental learning possible. For teachers in Japan who teach learners with vocabulary sizes far less than half that of the participants found in Elgort et al. (2018), careful planning is necessary to assure that conditions favorable for incidental learning are present within course and curricular frameworks. Furthermore, a strong emphasis on deliberate learning activities is well justified.

So far in this article, the interface has been discussed with regard to grammar learning and individual word learning. However, the focus of the research conducted by the author aimed to expand from that in a different direction by investigating the declarative and nondeclarative knowledge development with regard to multiword figurative expressions. These are one of the types of expressions that Siyanova-Chanturia and Martinez (2015) refer to in their review of research in this area, which they umbrella under the term *multi-word expressions* (MWEs). To encompass all the different types, they defined MWEs loosely as fixed or semi-fixed, recurrent phrases, such as collocations (*strong tea*), binomials (*black and white*), multi-word verbs (*put up with*), idioms (*spill the beans*), proverbs (*better late than never*), speech formulae (*What’s up?*), and lexical bundles (*in the middle of*). Within these different types of MWEs, the objects of study would be called idioms.

However, the important distinction between *literal expressions*, *figurative expressions*, and *core idioms* must be mentioned. Figurative expressions such as *dog eat dog* (meaning “highly competitive”), *once in a blue moon* (meaning “rarely”), and *walking on air* (meaning “happy”) can be interpreted, so once a person learns their meaning, the metaphorical connection can be made. Originally Grant and Bauer (2004) and subsequently Nation and Webb (2011) explained that figurative expressions are different from core idioms, for which the metaphorical meaning cannot be readily discerned. Examples of core idioms are *kick the bucket* (meaning “die”), *red herring* (meaning “distraction”), and *spill the beans* (meaning “expose a secret”). In idioms such as these, there is a higher learning burden for L2 learners than the learning burden for figurative and literal expressions.

In a recent research by the author, figurative expressions were the object of learning in the experimental treatment, and comparisons were made between deliberate paired-associate learning and incidental contextual learning on 73 participants. Figurative expressions that were learned through deliberate flashcard learning resulted in significantly faster online processing of semantic probes than did those that were learned contextually in the self-paced reading task. This was evidence that deliberate practice resulted in higher quality nondeclarative
knowledge gains, an important finding that builds on understanding of individual word knowledge development gained from Elgort (2007, 2011). Also, deliberate learning resulted in significantly superior results on the meaning generation test of declarative knowledge, which makes sense and confirms abundant previous interface research in the relationship between deliberate learning and declarative knowledge. Finally, compared to controls, contextual learning also resulted in slightly faster processing of the semantic probes. This was evidence that contextual figurative expression learning also results in tacit knowledge development, but the developmental process may be slower and more uncertain. In the long run, however, phrases that are learned contextually may become more stable fluent knowledge, as this is the way they are naturally learned in the L1.

To summarize this discussion of the interface and the experiment conducted by the author, four main conclusions can be drawn:

1. Acquisition of nondeclarative tacit knowledge is a by-product of repetition and retrieval, regardless of whether that repetition and retrieval is incidental or deliberate. Tacit knowledge develops below the surface, underlying declarative knowledge. With regard to the learning of figurative expressions, there is some interfacing between the two learning and knowledge types.
2. Tacit learning is more likely to result in tacit, nondeclarative knowledge development and deliberate learning is more likely to result in declarative knowledge.
3. The simpler and more regular the L2 feature, the more quickly its tacit knowledge elements will develop. Thus, tacit individual word knowledge development is typically faster than tacit knowledge development for grammar.
4. Deliberate and contextual learning and declarative and nondeclarative knowledge development go hand in hand. All are essential for SLA development, and more effective when done together with their interface counterpart.

4 Future Experiment Proposal

This section explains a study proposed to be conducted in the spring of 2020, within two sections of an identical required English class for second year Japanese university EFL students who are non-English majors ($n = 80$). The class title is “Intensive Reading,” which means that a large proportion of class activities should be focused on different types of deliberate learning in connection to text that are slightly above the learners’ ability level. There is no need to balance between deliberate and incidental learning activities within this class because students also attend another class called “Extensive Reading” that is focused on incidental learning activities and promoting fluency development through easy reading. However, within this framework, different types of deliberate learning are conducted to balance along a scale between granular rote memory and broader, more contextual types of deliberate learning. The class is taught by the author.
4.1 Research Question Options

With the above conclusions in mind, many further questions come to mind that should be addressed in future research following this agenda. In a future experiment, one of the following research questions will be investigated:

- **What kind of paired-associate lexical learning is most effective?**
  - English target word / Japanese meaning (paired associates)
  - English target word / English meaning (paired associates)
  - English target word / CLOZE use sentence example + meaning (paired associates)
- **Are there other learning pairs that will extend the capacity of Quizlet to accommodate for more contextual types of deliberate learning?**
  - Sentence stems / Sentence completions containing target word (paired associates)
  - Sequential sentence stems / Sequential sentence completions (paired associates)
  - Text comprehension questions / answers (paired associates)
- **What kind of learning target is most effective?**
  - English target word
  - English target word embedded in a short phrase to show grammatical collocations
- **Actual tasks conducted throughout the class: For longer texts, are two-sided paper-based contextual paired-associate cards effective?**
  - Ten sentence sequence completion tasks (see Appendix A)
  - 200-word CLOZE memorization task (see Appendix B)

4.2 Materials

The experimental conditions will be presented within the Quizlet sets themselves. Counterbalancing of learning conditions will be managed with Google Classroom. That is, at the beginning of the semester, students will be assigned to experimental treatment groups, and counterbalanced Quizlet sets will be assigned to each group accordingly within Google Classroom. Therefore, students will not be aware of the different learning groups that they are assigned into and will be doing nearly identical learning tasks side by side in the same classroom.

4.3 Experimental Design Issues

The experiment can be conducted over 6 to 13 weekly 20-minute sessions. Time on task will be managed within the framework as all experimental conditions and testing instruments will be conducted during the class time under the auspices of a graded strand of the class called “Learning Strategy Training.” To control for word learning burden and prior learning, pseudowords can be used.
4.4 Instrumentation
There will be a balance of declarative, online-declarative, and online nondeclarative knowledge measures. Declarative knowledge tests could be a paper-based meaning generation test or online multiple-choice test. Frequent online multiple-choice tests will be administered using Google Classroom and scored using Flubaroo. Online declarative knowledge measures will be administered using E-Prime to test receptive retrieval, most likely in a multiple-choice format measuring response times and accuracy. Online nondeclarative knowledge will be measured using E-Prime in a lexical decision task format.

5 Conclusion
In this article, the acute need for Japanese students of English to learn many more words than they currently do was addressed. Furthermore, the need for learners, teachers, materials developers, and curriculum designers to balance between intentional and incidental learning opportunities was stressed. The review of interface research supported this necessity for balancing these two complementary learning types. This broad overview of over 30 years of interface research in second language acquisition research was explained in three main stages. The initial stage attempted to establish how implicit and explicit knowledge types interface. This was followed by a stage aimed at operationalizing interface constructs. The third and current stage of interface research focuses on awareness, methodological concerns, and measurement. Next, it was argued that the interface should be considered differently with regard to lexical learning versus grammar learning. Following this assertion, recent lexical (as opposed to grammatical) interface research was presented. Finally, the article concluded with a proposal for a future lexical interface experiment.

Note
1. In Schmitt’s (2008) review, he refers to this as depth of knowledge, following a long tradition within the field. However, throughout this article the term knowledge quality will be used instead. This is done mainly because the concept of quality more accurately describes the useful, durable, fluent knowledge representations that should be developed. Furthermore, the construct of knowledge quality is more closely connected to speed and efficient retrieval than the traditional construct of depth, which is more about the complexity and elaborateness of the representations.
2. However, before getting too enthusiastic about this method, it must be noted that the knowledge gains it brings are greatly weakened if done in isolation. Although this article argues strongly in favor of the effectiveness of flashcard learning, this does not mean that it should result in neglecting incidental types of language learning. For maximum benefits, this activity must be balanced with other activities that provide abundant input and ample opportunities for incidental learning. Incidental and intentional learning are both equally essential, yet distinct parts of the language learning process.
3. Elgort prefers the terms declarative and non-declarative or tacit over explicit and implicit because they more clearly define the nature of L2 knowledge here being discussed. The terms declarative and non-declarative more accurately specify the key distinction between knowledge about the L2 versus more intuitive, subconscious, tacit knowledge that is applied in fluent use. For the remainder of this article, these terms will be used exclusively.

References


Appendix A
Story Retelling Sentence Completion Activity (Prompts Side of Paper)

Chapter 3 Retelling

1. Chapter 3 is . . .

2. Mr. Mimura regards time management . . .

3. Mr. Mimura is irritated by . . .

4. Mr. Mimura doesn't think Mr. Yagami should . . .

5. Mr. Yagami explained that he was happy to . . .

6. Mr. Uehara says that Mr. Mimura . . .

7. Mr. Mimura realized that he should . . .

8. The key is not to . . .
Appendix A (continued)
Story Retelling Sentence Completion Activity (Correct Responses Side of Paper)

Chapter 3 Retelling

1. Chapter 3 is about putting first things first.

2. Mr. Mimura regards time management as the most important thing at work.

3. Mr. Mimura is irritated by meetings that are longer than planned.

4. Mr. Mimura doesn't think Mr. Yagami should wait for customers who are late for their reservation.

5. Mr. Yagami explained that he was happy to wait because they are special for his restaurant.

6. Mr. Uehara says that Mr. Mimura should see life more broadly.

7. Mr. Mimura realized that he should take more time to discuss things with his boss.

8. The key is not to manage time but to manage ourselves.
Appendix B
Intensive Reading CLOZE Memorization Task (Prompts Side of Paper)

**Habit 2:**
*Begin with the End in Mind*

(1) Everything made by people is created. First, things are imagined, which means they are created in the mind. Second, things are created in real life. For example, when an architect is going to build a house, first they imagine a good design and draw a clear image that shows their plan. Second, the plan is used to build the house. In this way, the image of the house is created first in the mind, and the actual house is created second. Building your life is like building a house. You need a clear plan and plan of the life you want to make. With such a plan in mind, you can know exactly what needs to be done at any time. Beginning with the end in mind means carefully and thoughtfully about what you want to have and be in the future. The diagram on page 68 in the textbook shows the importance of having clear goals. When you have a clear goal in mind, it's always very simple to decide what is best to do. Having this clear image makes it much easier to know what you should do every moment of every day.

(2) A very effective way to clarify your life goals is to write a mission statement. A mission statement focuses on what you want to be (character), what you want to do (contributions and achievements), and what values you want to have. It's a clear written statement that you read often, reflect on, and revise regularly. Mission statements are unique, and different for everyone, but they reflect each person's individual core. On page 70, there's an explanation of mission statements. They reflect who you are, what you value, and what you want to become.

(3) For teachers, having a clear mission is essential. When teachers have a clear mission, their professional life reflects their clear purpose, and students are inspired by their presence. Likewise, beginning with the end in mind is also essential for teachers. They should have a clear plan for class every day, and also for the whole school year. With a clear plan and a clear mission, a teacher can easily decide what to do in the various situations that arise every day. When interacting with students, they praise good behavior, scold bad behavior, and give grades according to the goals they want their students to achieve.

(4) There's a song about this strong, purposeful feeling called *I Can See Clearly Now*. It's about the feeling of moving toward clear goals. Life has many obstacles, which are difficult things that we must do something about. But as the song says, when we see our goals, we can overcome them. Having such a clear sight is like seeing a rainbow after a storm. The dark clouds of rain, doubt, fear, and uncertainty are no longer in our way. We look straight ahead on these goals, and see nothing but blue skies. When we begin with the end in mind, we can see clearly both what we need to do now and what we want to become.
Appendix B (continued)

Intensive Reading CLOZE Memorization Task (Correct Responses Side of Paper)

Habit 2:
Begin with the End in Mind

(1) Everything made by people is created twice. First, things are imagined, which means they are created in the **mind**. Second, things are created in real **life**. For example, when an architect is going to build a house, first they imagine a good design and draw a clear **image** that shows their plan. Second, the plan is used to build the house. In this way, the image or **plan** is created first in the mind, and the actual **house** is created second. Building your life is like building a house. You need a clear **image** and plan of the life you want to make. With such a plan in mind, you can know exactly what needs to be done at any time. Beginning with the end in mind means thinking carefully and creatively about what you want to have and be in the future. The diagram on page 68 in the textbook shows the importance of having clear **goals**. When you have a clear goal in mind, it's always very simple to decide what is best to do. Having this clear image makes it much easier to know what you should do every moment of every day.

(2) A very effective way to clarify your life goals is to write a mission statement. A mission statement focusses on what you want to be (character), what you want to do (contributions and achievements), and what **values** you want to have. It's a clear written statement that you read often, reflect on, and revise regularly. Mission statements are unique, and different for everyone, but they reflect each person's individual **changeless** core. On page 70, there's an explanation of mission statements. They reflect who you are, what you value, and what you want to become.

(3) For teachers, having a clear mission is **essential**. When teachers have a clear mission, their professional life reflects their clear purpose, and students are inspired by their presence. Likewise, beginning with the end in mind is also essential for teachers. They should have a clear **plan** for class every day, and also for the whole school year. With a clear plan and a clear mission, a teacher can easily decide what to do in the various situations that arise every day. When interacting with students, they praise good behavior, scold bad behavior, and give grades according the **goals** they want their students to achieve.

(4) There's a song about this strong, **purposeful** feeling called *I Can See Clearly Now*. It's about the feeling of moving toward clear goals. Life has many **obstacles**, which are difficult things that we must do something about. But as the song says, when we see our goals clearly, we can overcome them. Having such a clear **vision** is like seeing a rainbow after a storm. The dark clouds of rain, doubt, fear, and **uncertainty** are no longer in our way. We look straight ahead, **focused** on these goals, and see nothing but blue skies. When we **begin** with the end in mind, we can see clearly both what we need to do now and what we want our life to become.
Vocabulary and Computer Technology: A Commentary on Four Studies for JALT Vocabulary SIG

Tatsuya Nakata
College of Intercultural Communication, Rikkyo University

Abstract
Four papers by Clint Denison and Imogen Custance, Louis Lafleur, James Rogers, and Andrew Obermeier will be presented at the Eighth Annual JALT Vocabulary SIG Symposium in Tokyo, Japan, on September 20, 2020. The topics covered in the four papers are vocabulary learning using online student-created vocabulary lists, development of a flashcard program that manipulates the review schedule and question format, creation of a list of multi-word units based on corpora, and examination of the acquisition of declarative and tacit vocabulary knowledge from deliberate computer-assisted learning. This commentary briefly summarizes each study and offers suggestions for future research. All of the four studies exhibit how computer technology can be used to facilitate vocabulary research, teaching, and learning.

1 Vocabulary and Computer Technology
According to Nation (2013), Tom Cobb once observed that “the relationship between computer technology and vocabulary is a marriage made in heaven” (p. 32). As Tom Cobb nicely puts it, computer technology can help vocabulary research, teaching, and learning in many ways. For instance, thanks to computer technology, researchers today can analyze large quantities of text with a single click of the mouse. Suits of computer programs, such as those developed by Tom Cobb and Laurence Anthony, enable researchers to create word lists, identify keywords, analyze vocabulary load of texts, or extract common collocations (Nakata, 2013). Computers have also advanced how researchers measure vocabulary knowledge through computer-adaptive testing (Laufer & Goldstein, 2004) or priming experiments (Elgort, 2011). Computer-based resources such as online dictionaries, vocabulary learning software, lexical profilers, and concordancers have also helped us teach and learn vocabulary effectively and efficiently. Given the benefits offered by computer technology, it is apt that the 2020 Japan Association for Language Teaching (JALT) Vocabulary SIG (Special Interest Group) Symposium organizes a session that exclusively focuses on computer technology and vocabulary. In what follows, I will briefly summarize and discuss four papers presented at the symposium.
2 The Four Studies

2.1 Vocabulary Learning Using Student-created Class Vocabulary Lists by Clint Denison and Imogen Custance

Clint Denison and Imogen Custance reported a study that examined the effects of collaborative vocabulary notebooks (class vocabulary lists) implemented with Google Spreadsheets. The entries in the class vocabulary lists were compared against frequency-based lists (British National Corpus/Corpus of Contemporary American English [BNC/COCA], New General Service List, Test of English for International Communication [TOEIC] Service List), as well as an information technology keyword list (for informatics students only). The analysis suggested that most items entered in the class vocabulary lists were words that were useful and relevant for the learners.

Many researchers and teachers would agree that ready-made vocabulary lists are useful for vocabulary teaching, learning, and materials development. At the same time, teachers sometimes may need or want to prioritize vocabulary items that are relevant for their students and immediate context, rather than those taken from ready-made lists. Class vocabulary lists may be especially beneficial for this purpose. One significant aspect of this study is that the authors not only had the students create the class vocabulary lists but also incorporated them into their classes. The authors created quizzes and activities based on the class lists, and some of the items were also tested in the mid-term and final exams. Denison and Custance did not describe that in their paper, but I believe that designing classes around student-created lists required an extensive amount of effort and time, and I would like to congratulate them on being dedicated teachers as well as good researchers.

The authors have used Google Spreadsheets for class vocabulary lists, but they could have also used Group Lex on Compleat Lexical Tutor developed by Tom Cobb (https://www.lextutor.ca/group_lex/demo_lab_mod/). Group Lex allows learners to keep a log of unfamiliar words, together with their definitions and example sentences, in a database. The online nature of the software also enables learners to share the entries with their peers (Horst, Cobb, & Nicolae, 2005). If the authors could clarify why they chose to use Google Spreadsheets over Group Lex, it would strengthen their paper. In addition to being a part of class activities and tests, if items in the class vocabulary lists were studied with flashcard software or practiced in online exercises, it might facilitate independent learning outside of the classroom and increase learning even further (Group Lex has a function to automatically generate fill-in-the-blank exercises based on words entered by users). Class vocabulary lists in this paper included pronunciation codes (number of syllables in the word and position of the stressed syllable). In addition to the pronunciation codes, the International Phonetic Alphabet may also be included in the class lists. If learners could hear audio recordings of the pronunciation of the words, it would also be helpful. This can be done by creating links from class vocabulary lists to entries in online dictionaries, most of which provide audio recordings of the pronunciation (e.g., https://www.dictionary.com/). Websites such as youglish.com (http://youglish.com/) also allow learners to hear how a given English word or expression is used in authentic speech. Creating links to online
databases such as Just the Word (http://www.just-the-word.com/) from class vocabulary lists would also help learners acquire common collocations and understand how the word is actually used in authentic texts. The authors wrote that they analyzed the entries with Compleat Web Vocabulary Profiler and manually highlighted words in New General Service List and TOEIC Service List. This can be done automatically by preparing the word lists as separate sheets and using a vlookup function in Google Spreadsheet.

2.2 The Indirect Spaced Repetition Concept by Louis Lafleur

Louis Lafleur developed a prototype flashcard program for vocabulary learning called Indirect Spaced Repetition Software and described its features in his paper. One significant aspect of his software is that it keeps a record of the learner's performance on individual items and controls not only the review schedule but also the practice format (e.g., receptive recognition, productive recognition, and productive recall). Although research suggests that learning vocabulary form flashcards is a useful strategy (e.g., Elgort, 2011), one potential weakness of paper-based flashcards is that it is not very effective for learning multiple aspects of word knowledge such as pronunciation, collocations, or grammatical functions. Lafleur's software is significant because it has the potential to offer the benefits of flashcard learning, while at the same time, providing opportunities for learning multiple aspects of word knowledge that are usually hard to learn with paper flashcards.

One suggestion I have for the paper is that perhaps the author could have discussed the strength of his software in comparison to existing vocabulary learning programs such as Quizlet, Memrise, or Anki. It would have also been useful to evaluate the effectiveness of his software using frameworks such as Technique Feature Analysis (Nation & Webb, 2011) or criteria for evaluating flashcard software (Nakata, 2011). Also, I am not quite sure whether the name of his software (Indirect Spaced Repetition) accurately captures its significant features. The most innovative aspect of the software seems to be its ability to promote elaborate processing and systematic review of vocabulary simultaneously, both of which are considered important in vocabulary learning (Hulstijn, 2001). I wonder if the software could be given a name that highlights these two important design features.

Regarding the review schedule, Lafleur wrote that there are two ways to manipulate spacing: expanding and equal (uniform). In expanding spacing, the intervals between encounters of a given item are gradually increased (e.g., 1 week, 2 weeks, and 3 weeks). In equal spacing, the intervals between encounters of a given item are held constant (e.g., 2 weeks, 2 weeks, and 2 weeks). Perhaps it is useful here to point out that some studies examined the effects of contracting spacing, where the intervals between encounters are gradually decreased (e.g., 3 weeks, 2 weeks, and 1 week). Lafleur stated that many studies have failed to find any significant difference between the effects of equal and expanding spacing. Most of the studies cited in his paper, however, examined the learning of materials other than second language (L2) vocabulary such as name pairs (Landauer & Bjork, 1978, Experiment 1) or face-name pairs (Carpenter & DeLosh, 2005; Landauer &
Bjork, 1978, Experiment 2). The author could have also pointed out a number of L2 vocabulary studies have compared equal and expanding spacing, failing to find the significant advantage of expanding spacing over equal spacing (Kang, Lindsey, Mozer, & Pashler, 2014; Karpicke & Bauernschmidt, 2011; Nakata, 2015; Pyc & Rawson, 2007). At the same time, although results of empirical studies are not necessarily in favor of expanding spacing, I understand why Lafleur decided to employ expanding spacing for his software as expanding spacing is perhaps easier to implement than equal spacing. Expanding spacing may also have a positive effect on learners’ motivation because it may lead to higher retrieval success than equal spacing during the learning phase.

Regarding the question format, the author argued that the questions were designed to cover three tiers of word knowledge proposed by Nation (2001): form, meaning, and use. However, according to Nation, each of the three tiers consists of three aspects, resulting in a total of nine aspects. Form consists of the written form, spoken form, and word parts. Meaning consists of form and meaning, concept and referents, and associations. Use consists of grammatical functions, collocations, and constrains on use. It is not clear how some of these aspects (e.g., word parts, associations, or constraints on use) are addressed in the software. I also wondered about the sequencing of the different question formats. According to the retrieval effort hypothesis (Pyc & Rawson, 2009), difficult but successful retrievals facilitate retention more than successful but easy retrievals. This hypothesis suggests that when sequencing different types of questions, they should be given in the order of increasing difficulty. In other words, easy formats should be given before difficult formats because this may help us to gradually increase difficulty while ensuring that retrieval will be successful. In Indirect Spaced Repetition Software, L1 sentence to L2 sentence translation (Level 5) is given before L2 sentence to L1 sentence translation (Level 6; Table 4). However, considering that translating from L2 to L1 is usually easier than translating from L1 to L2, I wonder if the order of these two question formats should be reversed. The author also discussed the concept of interleaving. Effects of interleaving have been studied extensively in the field of cognitive psychology (e.g., Kang & Pashler, 2012; Rohrer & Taylor, 2007; Taylor & Rohrer, 2010), as well as L2 learning (Carpenter & Mueller, 2013; Nakata & Suzuki, 2019a). In these studies, interleaving refers to a schedule where multiple concepts or skills are practiced simultaneously, as opposed to blocking (where only one concept or skill is practiced at a time). Lafleur, however, used the term to refer to multiple exercises for a given target word, so it is not clear whether what he calls interleaving in his software has similar effects on learning as suggested by earlier research.

The paper also presents results of a 2-week trial with seven university students. Questionnaire results suggested that participants generally expressed positive attitudes toward the software. One possible limitation of the study is that the questionnaire was designed to explore four constructs (usefulness, usability, enjoyment, and consideration), but there was only one question for each category. Including multiple questions for each construct would have increased the validity and reliability of the questionnaire. Nonetheless, the software described in the paper has a number of features that may facilitate learning, and I would like to congratulate the author on developing it.
2.3 On Creating a Large-scale Corpus-based Academic Multi-word Unit Resource by James Rogers

James Rogers presented his attempts to create a list of academic multi-word units in English based on corpora. When Michael West created General Service List (West, 1953), frequency counts had to be done manually. Even with help from an army of research assistants, such endeavor must have required a tremendous amount of effort and time. Nowadays, thanks to corpus linguistics software such as AntConc (Anthony, 2019), researchers can analyze large quantities of text, create frequency-based vocabulary lists, or extract common collocations with just a single click of the mouse. This paper is a great example of computer technology helping advances in vocabulary research. The paper by Rogers is also significant because his vocabulary list concerns multi-word units rather than single words. Despite the increasing recognition that the knowledge of multi-word units is important for language learning and processing, most vocabulary lists to date have been on single words, although several lists of multi-word units have been published in recent years, such as the Academic Formulas List (Simpson-Vlach & Ellis, 2010), PHRASEList (Martinez & Schmitt, 2012), Academic Collocation List (Ackermann & Chen, 2013), and PHaVE List (Phrasal Verb Pedagogical List; Garnier & Schmitt, 2015). The word list described in this paper will be a very useful addition for researchers, teachers, learners, and materials developers.

The multi-word unit list in this paper has been carefully created with a number of considerations in mind, and the justification for adopting certain approaches is clearly explained in the paper. The information provided in this paper will be useful for researchers who attempt to create a list of multi-word units in the future. In addition to objective data such as frequency, the author also turned to survey data involving English as a Foreign Language (EFL) practitioners when creating the list. Providing more details about the survey would have been useful. For instance, how many EFL practitioners were surveyed? How were they selected? Were they native speakers or non-native speakers? Were there large variations among respondents, or were they mostly consistent (issue of inter-rater reliability)? Is there any data suggesting that EFL practitioners’ intuition is accurate and useful? The author wrote that about half of the collocations surveyed “were judged to be useful for learners” by the EFL practitioners. But what was the threshold for collocations to be considered useful? Although it is not the main purpose of the paper, I was also interested in how the author thinks the multi-word units identified in the paper should be taught and studied. Although we can expect multi-word units to be learned mostly in the same way as single words, it is possible that effective principles for learning single words may not necessarily apply to the learning of multi-word units. For instance, although most studies suggest that increasing temporal spacing facilitates the learning of single words (e.g., Karpicke & Bauernschmidt, 2011; Nakata, 2015; Nakata & Suzuki, 2019b; Nakata & Webb, 2016), one recent study suggests that spacing may not necessarily increase the learning of multi-word units (collocations), especially when they are learned incidentally (Macis, Sonbul, & Alharbi, 2019). The findings suggest that more research examining how to teach and learn multi-word units is warranted.
2.4 Exploring the Effectiveness of Deliberate Computer-assisted Language Learning by Andrew Obermeier

Traditionally, in most empirical studies, vocabulary learning has been measured by an offline, paper-based test such as a translation or multiple-choice test. In other words, vocabulary acquisition has essentially been synonymous with the acquisition of declarative knowledge. In recent years, thanks to advances in technologies such as eye tracking, priming, or event-related potential (ERP), an increasing number of studies have attempted to measure tacit (or non-declarative) knowledge. This line of research was pioneered by Irina Elgort (2011), who will be a discussant at the JALT Pan SIG conference in 2020. Obermeier, who worked closely with Elgort, described a research design that examined the effects of deliberate, computer-assisted learning of vocabulary. The proposed study included measures of tacit knowledge, as well as online and offline measures of declarative knowledge.

In order to communicate effectively in L2, declarative knowledge alone is not sufficient. Learners should have fluent and automatic access to the formal and semantic representations of L2 words. However, since most existing studies have measured only declarative knowledge, our understanding of how tacit knowledge can be effectively taught and learned is still limited. A small number of studies examining the acquisition of declarative and tacit knowledge have suggested that effective principles for developing declarative knowledge may not necessarily apply to the acquisition of tacit knowledge. For instance, Nakata and Elgort (2019) suggested that although spacing facilitates the acquisition of declarative knowledge, it may not necessarily be effective for the development of tacit knowledge. Due to the importance of tacit knowledge for vocabulary development and paucity of research examining its acquisition, findings from the proposed research by Obermeier will be valuable. The paper lists a number of research question options. It would be useful if the rationale for investigating these questions, along with hypotheses, were also given. The paper also mentions the possible use of pseudowords in the experiment. As the author argued, pseudowords are very useful when controlling for prior knowledge. Lexical variables that might affect learning and processing can also be controlled more easily with pseudowords, which is especially important when measuring response times on an online lexical decision task as described in the paper. Ethical issues, however, need to be addressed, especially when data are collected during regular class hours.

3 Conclusion

Although the four papers differ in their focus, all of them are great examples of how computer technology can be used to facilitate vocabulary research, teaching, and learning. The paper by Denison and Custance demonstrates how online student-created vocabulary databases may facilitate vocabulary teaching and learning. The paper by Lafleur shows how computer technology can be used to maximize vocabulary learning from flashcards. Data gained from the flashcard program may also provide valuable data for researchers. For instance, data might allow researchers to identify what factors (e.g., parts of speech, word length, abstractness, and polysemy) make some words harder to learn than others.
Researchers may also be able to track learners’ vocabulary retention or attrition over time based on data collected from the software. The paper by Rogers demonstrates how computer technology facilitates the creation of a list of multi-word units and is a great example of computers helping vocabulary researchers. The resulting list, at the same time, will be useful not only for researchers but also for teachers and students. The paper by Obermeier is primarily concerned with how computers enable researchers to measure the development of both declarative and tacit vocabulary knowledge. His proposed research may also demonstrate the potential for computers to increase deliberate vocabulary learning, just like the paper by Lafleur clearly shows.

One suggestion for future research is to examine actual vocabulary learning gains from computer-assisted language learning. The main purpose of the paper by Denison and Custance was to analyze the usefulness of the words entered in the class vocabulary lists, and the authors did not examine how much vocabulary was actually learned. Because the class vocabulary lists were created by learners throughout the semester, I understand that it was not feasible to measure vocabulary gains using a standard pretest-treatment-posttest design. At the same time, considering that some items from the class vocabulary lists were tested in the mid-term and final exams, it may be useful to analyze learners’ performance on these items. The paper by Lafleur involved trialing the flashcard program with students for 2 weeks. Although a questionnaire was administered at the end of the trial, no vocabulary gains were measured using a posttest. A study comparing the effects of Indirect Spaced Repetition Software and other learning methods may also be useful because it may allow researchers to tease apart which factors of the software are especially beneficial for learning. Investigating vocabulary learning gains may also increase the chance of research being published. For instance, submission guidelines of *Language Learning & Technology*, one of the top journals in the field of computer-assisted language learning, state that “Articles containing only descriptions of software, pedagogical procedures, or those presenting results of surveys without providing systematic empirical data and analysis on language learning outcomes or processes will not be considered” (https://www.lltjournal.org/submission-guidelines/).

The four papers reviewed here demonstrate many ways in which computer technology helps us research, teach, and learn vocabulary in L2. I hope that this marriage made in heaven will last a lifetime and look forward to seeing more research being published, further demonstrating a strong bond between computer technology and vocabulary.

**References**


Teaching Ideas for Improving Oral Performance through Formulaic Language Instruction

Chie Ogawa
Kyoto Sangyo University

Abstract
This article suggests three teaching ideas to help L2 learners improve speaking performances through form-focused instruction using formulaic language. Formulaic language is considered an effective way to foster speaking fluency because prefabricated chunks are faster to retrieve than constructing sentences word by word (Wray, 2002). In spite of the benefits of learning formulaic language in L2 learning theory, few empirical studies were found which examined the effects of formulaic language instruction in intact classrooms, in particular in the EFL (English as a Foreign Language) context. By introducing some effective classroom tasks to foster L2 learners’ speaking fluency focusing on formulaic language in this article, the author emphasizes the need for empirical research involving EFL learners.

Key words: Speaking, Formulaic Language, Automatization, Proceduralization, Focus on Form, Fluency, CALF, Instructed SLA

1 Introduction
According to Wray (2008), formulaic language refers to large units of processing, in other words, lexical units that are more than one word long. She explains formulaic language as sequences of words that are in some regard not entirely predictable, a function that is only achieved with the whole expression or features of structure such as morphology or word order that are non-canonical by acknowledging that it is not easy to have consensus across domains (Wray, 2013, p. 317). The words in a formulaic sequence are glued together and stored as a single big word (Ellis, 1996).

The study of formulaic language is related to automaticity, and it also plays an important role in second language acquisition. Learners’ automatic access to prefabricated chunks, which are stored in memory, can lead to fluency development (Boers & Lindstromberg, 2012; Segalowitz, 2003; Wood, 2009, 2015) through the repeated use of formulaic language. Linguistic chunks can become part of production rules and can be retrieved directly from declarative memory without the need for computations in working memory (Wood, 2010, p. 3). If learners process formulaic language automatically, they can use more attentional resources for other areas.

According to Boers, Eyckmans, Kappel, Stengers, and Demecheleer (2006), there are three reasons why acquiring formulaic language is believed to be
beneficial to L2 learners. First, the mastery of the idiomatic aspects of natural language can help learners sound more native-like. Second, the mastery of formulaic language can help learners to speak more fluently because prefabricated sequences or ready-made chunks can be retrieved faster than sentences generated word by word under real-time conditions. Third, formulaic language can help learners produce more accurate language provided that the prefabricated chunks are stored correctly in memory.

Wray (2000) acknowledges the difficulty of mastering formulaic language for L2 learners because of the following reasons. The first reason is that L2 learners may not be exposed enough to the formulaic language. Formulaic language is often omitted in the interaction with L2 speakers although formulaic language is common for native speakers. The second reason is that formulaic language may not be taught very well. It seems difficult to match to the real-world experience of language in the classroom.

Gatbonton and Segalowitz's (1988) article has a significant contribution to practical classroom L2 teaching of automatization of formulaic language. They claimed that traditional repetition of the target form in a monotonous drill is not sufficient to acquire the target language because it lacks communicative context. Their suggestions demonstrate ways of incorporating repetition and rehearsals of useful formulaic language into a genuinely communicative task. They call this type of process creative automatization because students themselves generate and create appropriate utterances based on their understanding of the communicative situation. Automatizing utterances requires students to repeat utterances that would occur naturally in a normal communicative situation. Communication tasks can promote automatization of the target phrases within a communicative framework in the classroom. Gatbonton and Segalowiz suggest that the tasks should be designed for L2 learners to elicit short and memorizable utterances. The formulaic language should be multisituational so that it can be usable in many situations with little or no modification. By doing so, the short and memorizable formulaic language can help automatization take place within a short period of time. Rehearsal promotes fluency and familiarity with formulaic expression.

According to Boers and Lindstromberg (2012), ways to foster the use of greater breadth of formulaic sequences have been examined in many intervention studies, but few researchers have examined the proceduralization of formulaic sequences. In the following section, the author suggests form-focused instruction and demonstrates some types of tasks which can promote proceduralization of target formulaic language.

2 Focus on Form

2.1 Text Enhancement and Pre-Task Planning

Ellis (2016) explained four ways to incorporate focus on form into task-based language teaching: text enhancement, corrective feedback (CF), pre-task planning, and task repetition.
In text enhancement, linguistic features are underlined, bolded, or italicized so that learners could notice the target linguistic feature through typographical enhancement. There was a high probability that the participants noticed the target linguistic forms, as their salience was increased by the underlining (see Doughty, 1991; Lee & Huang, 2008; Sharwood-Smith, 1993 for studies concerning the effectiveness of typographical cues such as underlining, bolding, and italicization).

Lee and Huang (2008) examined 20 text-enhancement studies in their meta-analysis and concluded that text enhancement has an overall positive effect, but the effect size is quite small. Lower proficiency learners may struggle to engage in processing both comprehending the meaning of the text and consciously attending to linguistic forms (Ogawa, 2019). Even if learners notice the target form, they may not acquire it and it might fail to be passed into long-term memory; therefore, Ellis (2016) recognized the importance of combining text enhancement with other instructional techniques that encourage intentional learning.

2.2 Peer Feedback

As one of the form-focused instruction methods, CF is considered effective in promoting noticing and is thus conducive to L2 learning (Lyster & Saito, 2010; Mackey & Goo, 2007; Russell & Spada, 2006; Sato, 2017). Compared to teacher CF, peer corrective feedback (PCF) might be considered as less effective because many L2 learners might not feel confident enough to provide CF or they may trust teacher CF but not PCF. However, research has shown that PCF has positive effects on L2 learners’ language development (Kim, 2013; Sippel & Jackson, 2015). The main difference between teacher CF and PCF is that L2 learners can act as a feedback receiver and provider. In other words, PCF serves two beneficial functions (Sato, 2017). From a feedback provider’s point of view, a learner first needs to detect an error in the input which her peer produces. In order to do so, a provider must notice the gap between the error and the target-like production. This gives an opportunity to compare the peer’s error and her own interlanguage and notice that she might make the same error and correct it internally. This cognitive process may contribute to restructuring of the feedback provider’s L2 knowledge (Sato, 2017). On the other hand, from a receiver’s point of view, PCF given by her peer may trigger noticing and push the speaker to modify the original utterance.

Although that the effects of PCF are theoretically and empirically acknowledged, L2 learners still might find it difficult to correct their classmates’ errors due to the social and psychological nature of PCF. Philp, Walter, and Basturkmen (2010) explained that L2 learners hesitate to provide PCF because they feel less confident of their proficiency (e.g., readiness to correct as a learner) and social relationship (e.g., face saving). Related to this issue, Sato (2017) also suggests that it is essential for L2 learners to be trained how to interact with each other and how to provide CF to each other because providing CF to their peers is influenced by learners’ social dynamics between peers. Therefore, teachers’ role is important to create learners’ mindset which is positive toward PCF so that they can avoid underscoring peer feedback and help L2 development.
Yet, few studies, to my knowledge, have been carried out to explore the role of peer feedback on a speaking task, especially on the usage of the formulaic language. It is uncertain if and how the L2 learners are able to contribute to peer feedback and on their usage of the target linguistic features. Experimental studies of pedagogical interventions are much needed.

2.3 Task Repetition

Another important way to help L2 learners proceduralize and automatize the target formulaic language is through repetitive practices. However, merely repeating the same rules has been criticized because it does not provide a meaningful context in which students genuinely need to communicate. Automaticity is best achieved by the repeated use of language rules in a context of authentic communication (DeKeyser, 2003; De Ridder, Vangehuchten, & Gómez, 2007; Segalowitz, 2003).

Task repetition is helpful for developing fluency because repetition allows learners to activate concepts and linguistic forms so that they are more easily and quickly accessed. Based on the limited attentional model of speech production (Skehan, 1998), low-proficient L2 learners face a number of challenges in the speaking process from conceptualization to articulation in Levelt’s speech model (see Levelt, 1989) because demands of thinking of preverbal message and how to formulate the message efficiently are not yet efficiently processed. Repetition might reduce the attentional demands on learners to conceptualize, encode, and monitor their messages simultaneously.

One effective activity is the 4/3/2 task (e.g., Boers, 2014; Nation, 1989; Nation & Newton, 2009; Thai & Boers, 2016). In the 4/3/2 task, students talk about the same topic for 4 minutes, then 3 minutes, and finally 2 minutes. When the students talk about the same topic three times with increasing time pressure to perform more quickly, they must speak faster. Therefore, this task can be used in language classrooms to foster speaking fluency.

According to Nation (1989), the 4/3/2 task has three important features: repetition, reducing time, and a change of audience. These features directly affect fluency by encouraging L2 speakers to focus on the meaning under a time constraint. Nation’s suggestions and rational were empirically proven by previous researchers (e.g., De Jong & Perfetti, 2011; Thai & Boers, 2016). Repeating the same topic helped L2 learners improve fluency more than the group who did not repeat the same topic (De Jong & Perfetti, 2011). Shrinking time condition (4-3-2) was able to promote more fluency than having the constant time condition (3-3-3) (Boers, 2014; Thai & Boers, 2016).

3. Suggestions for Future Studies

Considering many researchers have suggested form-focused instructions, few studies have examined the effects of form-focused instruction with formulaic language. For example, Wood (2009) conducted a case study of the classroom teaching of formulaic language and fluency development with a female
Japanese learner of English. This study took place in an intensive study abroad class but only the female participant’s speaking progress was analyzed. The fluency workshop consisted of 9 hours of instruction over 6 weeks. The sessions included an (1) input stage, (2) automatization stage, (3) practice and production stage, and (4) free talk stage. In the input stage, the student listened to native English speakers’ personal stories. In the automatization stage, the learners shadowed the recording with a focus on the formulaic language and also did a dictogloss, which includes texts rich in formulaic language. In the practice and production stage, the student did the 4/3/2 task, in which she told personal narratives. In the free talk stage, students in small groups took turns listening to individuals speaking spontaneously about the topics they had been assigned.

The female participant did a monologue narrative recording before she started the formulaic language instruction and again after the 6-week training session. The results showed that she made a 13.8% gain (123.2 to 140.2) in speech rate with syllable per minute and a 26.3% (5.1 to 6.4) gain in mean length of runs with syllables per a run between the pretest and posttest. The workshop also improved her complexity, as she used a greater variety of formulaic language. Before the fluency workshop, she produced 18 tokens of formulaic language. After the workshop, she used 52 tokens of formulaic language. This result suggests that the fluency workshop provided the participant with samples of formulaic language, which she added to her repertoire; thus, her utterances became more fluent. Although Wood’s study shows an interesting finding, it is quite difficult to generalize the result due to sample size.

To summarize, formulaic language is an important factor to help L2 speakers sound more fluent because automatic retrieval of prefabricated chunks is faster than retrieving word by word. The effects of teaching formulaic language showed that an L2 speaker improved speaking (Wood, 2009). Yet, it is difficult for L2 learners to master formulaic language partly because L2 learners have little exposure with formulaic language although they are common among native speakers’ speech (Wray, 2000). Indeed, few studies have examined the effects of pedagogical intervention on specific target formulaic language on the L2 learners’ speaking development as an experimental study. More studies are needed to focus on the formulaic language and speaking development.

References


Kim, Y. (2013). Effects of pretask modeling on attention to form and question development. *TESOL Quarterly, 47*(1), 8–35. doi:10.1002/tesq.52


Assessing Low-level Cognitive Processes of Word Recognition

Michael Holsworth
Kyoto Sangyo University

Abstract
A fundamental skill required for vocabulary development is word recognition ability. According to Perfetti (1985), word recognition ability relies on low-level cognitive processing skill to be automatic and efficient in order for cognitive resources to be allocated to high-level processes such as inferencing and schemata activation needed for reading comprehension. The low-level processes include orthographic knowledge, semantic knowledge, and phonological awareness. These low-level processes must be efficient, fluent, and automatic in second language readers in order for them to achieve the ultimate goal of reading comprehension. This article briefly describes the concept of word recognition, its relation to vocabulary, and three tests that were designed to measure the three components of word recognition (orthographic, semantic, and phonological knowledge) in a longitudinal study that investigated the effects of word recognition training on reading comprehension.

Key words: Reading, Word Recognition, Orthographic Knowledge, Semantic Knowledge, Phonological Awareness, Testing

1 Introduction
Reading is a fundamental aspect of communication and is a critical skill needed for successful L2 language learning. This skill, however, is difficult to develop for many learners when one considers the variety of different orthographic systems used around the world. This difference is easily seen between two orthographically different languages such as English and Japanese. It is vital to have a strong subset of reading skills that are automatic, such as word recognition, in order to develop a strong vocabulary for speaking and reading.

In order for an L2 learner to build his or her vocabulary and reading proficiency, the fundamental skill of word recognition is required. Richards and Schmidt (2002, p. 557) defined word recognition, a straightforward definition, as “higher level information is knowledge of permissible words as well as actual words of a language, while the lower level information is the actual phonetic input (or orthographic input in the case of written word recognition)” . This definition is rather narrow and focuses on the orthographic element of word recognition. Koda (1988, p. 135) defined word recognition as “the process of deriving both a phonological code (i.e., sound used in speech) and the meaning from graphemic representation.” Koda (2004, p. 29) added to this definition of word recognition as
“the processes of extracting lexical information from graphic displays of words” and relates only to a word’s sounds and meaning. She further stated that decoding pertains to the extraction of phonological information.

In order for reading to be successful, several complex cognitive processes must occur rapidly, accurately, and automatically. These cognitive processes include low-level processes such as word recognition, semantic parsing, and semantic proposition encoding. In fact, according to Perfetti’s Verbal Efficiency Theory (1985), reading comprehension is limited by a reader’s low-level processing efficiency. When lower-level processes are automatic and efficient, a reader should be able to reallocate some of his or her cognitive resources to higher-level processes such as schemata activation, global knowledge application, and inferencing (Grabe, 2009; Perfetti, 1985). It is these high-level processes that lead to reading comprehension. It has been hypothesized that these higher-level processes rely upon lower-level processes to be efficient and automatic. In other words, high-level processing used for reading comprehension relies on automatic low-level processes such as word recognition.

Perfetti and Hart (2001) defined word recognition as the interaction between three components: orthographic, phonological, and semantic processes. According to Cunningham, Perry, and Stanovich (2001), the visual recognition of word forms or orthographic knowledge is the first component and is one of the vital processing skills of word recognition. Orthographic processing includes identifying not only letters and letter groups but also word roots and morphological affixes. In addition, an understanding of how words are assembled using graphic representations is crucial to word recognition skill. The second component of word recognition is phonological processing and can be thought of as the awareness of the sound structures of a language (Stanovich, 1988). It is necessary to understand what sounds orthographic symbols of a language represent. Phonological awareness is known to be a key predictor of reading ability for both L1 and L2 English reading (Huang, Lin, & Su, 2004; Johnson & Tweedie, 2010; Jongejan, Verhoeven, & Siegel, 2007; Swank & Catts, 1994). This means that in order for L1 and L2 readers to process reading efficiently, they need to understand the graphophonic mapping that occurs when decoding letters and words from a given language. It is seen as the precursor to semantic activation and is triggered by the onset of orthographic processing. For example, when a reader processes a word, the orthographic and phonological processes begin by processing the visual and sound cues. This leads to the third component of word recognition, semantic knowledge. A reader activates his or her semantic knowledge from his or her lexicon until the combination of orthographic, phonological, and semantic information finds the best match to the word being read if it is a word known to the reader, in other words, successful word recognition (Grabe, 2009). For L2 reading teachers, the logical goal should then be to foster the development of automatic word recognition skills by providing large amounts of comprehensible texts and training that targets word recognition skill development. A study by Burrows and Holsworth (2016) investigated this aspect of word recognition training and found that when students received training in all three components of word recognition, reading comprehension scores increased. This type of training seems to help learners
move on from processing texts on a word-by-word basis. Teaching reading in the L1 and L2 requires knowledge of word recognition components in order to make sound pedagogical decisions to help learners improve their reading skill.

2 Measurement of Word Recognition

In order to have sufficient vocabulary in a second language to achieve reading comprehension, a learner must first develop and automatize word recognition skills. According to the Verbal Efficiency Theory (Perfetti, 1985), these skills are considered low-level cognitive processes and are easily automatizable. They include orthographic knowledge, semantic knowledge, and phonological awareness. These three components of word recognition are the key constructs to measure in order to accurately measure word recognition. To investigate this, a lexical decision test was used to measure orthographic knowledge, an antonym pairs test was selected to measure semantic knowledge, and a pronunciation speaking test was used to assess phonological awareness. The following is a description of each of these tests as they were used in a longitudinal study of the effects of word recognition training on reading comprehension.

2.1 Lexical Decision Test

The lexical decision test was designed to measure the orthographic knowledge of a reader. Participants are presented with a series of real English words and non-words. The original test was created by Matsuo (2017) and consists of 144 items (72 real English words and 72 non-words). Because reaction tests such as these can produce a cognitive test bias toward either positive or negative responses, and equal number of real-word items and non-word items were used (Jiang, 2012). By doing so, this reduces the possibility that a test taker would preemptively respond to an item based on his or her previous response. The real-word items were chosen from the British National Corpus with 18 items per level selected from the 1,000, 2,000, 3,000, and 4,000 level of frequency. These 18 items per frequency level were then equally divided into three subgroups, based on word length of four-, five-, and six-letter words. Finally, each word length subgroup contained two words representing a part of speech with two nouns, two verbs, and two adjectives. The non-words (or pseudowords) were created using the ARC Non-word Database (Rastle, Harrington, & Coltheart, 2002). These non-words were based on English language spelling conventions in order to resemble real words and consist of legal bigrams and orthographically existing onsets. The non-words were also created to match the real-word items in word length.

This test was administered using the reaction time software called SuperLab (Version 5.0, 2014). Participants took the test individually using a standard laptop computer. Instructions appeared in the test taker’s L1 (Japanese) on the screen followed by 14 practice items to acclimate the test taker to the test format. The 144 items were presented in random order for each participant. Prior to each item, visual primer image (++++) appeared on the screen for 1,000 ms so test takers were aware of the word location. After the test item appears, the
test taker selected either the V-key for a real word or the N-key for a non-word. The responses were then recorded in millisecond reaction times. In L1 research, responses with <200 ms are considered to be a preemptive guess, and a response of >2,000 ms is considered to be a distracted response, and thus, they were considered outliers (Balota & Chumbley, 1984). However, according to Jiang (2012), L2 reaction time boundaries need to be reconsidered due to the additional processing time required by participants. For this reason, responses with 200 ms or less were still considered preemptive guesses and were removed, and there was no maximum limit for response times.

The original test created by Matsuo consisted of one version only. However, the longitudinal study consisted of three test times, so two additional tests were created in order to have multiple measures and to avoid test bias. The original version had 72 target words, so with the addition of 12 anchor items present in the first test, the second and third tests contained 84 target words with corresponding non-word items. The anchor items allow for the measurement of the same items on all three tests and thus alleviate any potential issue with differences among the three tests. Table 1 shows the 12 anchor items present in the lexical decision test for all three test times.

### 2.2 Antonym Pairs Test

The antonym pairs test addresses the component of semantic knowledge and is categorized as a semantic classification task (Jiang, 2012). It has been used extensively as a semantic knowledge measure in word recognition research (Nation & Snowling, 1998; Perfetti & Zhang, 1995; Segalowitz & de Almeida, 2002). The main goal of this test for test takers is to identify pairs of items as either antonyms or non-antonyms. The original version of this test was created by Matsuo (2014), and it consists of 72 antonym pairs and 72 non-antonym pairs. Examples of antonym pair items include life-death and come-go. Examples of non-antonym pairs

<table>
<thead>
<tr>
<th>Anchor item</th>
<th>Item counterpart</th>
<th>Part of speech</th>
<th>Frequency level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>Meve</td>
<td>Noun</td>
<td>1,000</td>
</tr>
<tr>
<td>Laugh</td>
<td>Cromb</td>
<td>Verb</td>
<td>1,000</td>
</tr>
<tr>
<td>Late</td>
<td>Pson</td>
<td>Adjective</td>
<td>1,000</td>
</tr>
<tr>
<td>Aunt</td>
<td>Kest</td>
<td>Noun</td>
<td>2,000</td>
</tr>
<tr>
<td>Smile</td>
<td>Phows</td>
<td>Verb</td>
<td>2,000</td>
</tr>
<tr>
<td>Empty</td>
<td>Skump</td>
<td>Adjective</td>
<td>2,000</td>
</tr>
<tr>
<td>Guest</td>
<td>Muilt</td>
<td>Noun</td>
<td>3,000</td>
</tr>
<tr>
<td>Shout</td>
<td>Telve</td>
<td>Verb</td>
<td>3,000</td>
</tr>
<tr>
<td>Angry</td>
<td>Orked</td>
<td>Adjective</td>
<td>3,000</td>
</tr>
<tr>
<td>Kilometer</td>
<td>Queens</td>
<td>Noun</td>
<td>4,000</td>
</tr>
<tr>
<td>Kiss</td>
<td>Hosh</td>
<td>Verb</td>
<td>4,000</td>
</tr>
<tr>
<td>Ugly</td>
<td>Yoaf</td>
<td>Adjective</td>
<td>4,000</td>
</tr>
</tbody>
</table>
include *plant-brother* and *fat-gray*. In alignment with the lexical decision test, test items were selected to represent the first four 1,000 word frequency levels of the British National Corpus.

As with the original lexical decision test, the original antonym pairs test by Matsuo had only one version. Therefore, two new versions were created following the same criteria as the original test. The two new tests each contained 72 antonym pairs and 72 non-antonym pairs, with 18 pairs at each of the first four 1,000 word frequency levels. In following with the original test, the two new tests were controlled to the extent possible for part of speech and word length for the target words. The pair and non-pair items were also controlled to ensure that the primer word (first word shown) and the target word (second word shown) were at the same frequency level, but in some cases due to the limited choices for accurate antonym pairs, the difference between several items was no more than one frequency level. Once the three versions were created, they were combined and items were redistributed using random stratified assignment to ensure that each new version of test one, two, and three all had equal item representation from the original and new tests. Once again, the antonym pairs tests each contain the same 12 anchor items used in the lexical decision tests. Table 2 shows the anchor items used and their corresponding antonym or non-antonym pair.

Once again, the SuperLab software was used to measure participant reaction times in milliseconds. After reading the instructions, the next screen informed them they had 14 practice items before the test began. A visual primer appeared on the screen (++++) for 1,000 ms to show the test taker where the words would appear. The primer (++++) appears for 1,000 ms followed by the primer word for 1,000 ms, which was then followed by the target word. Participants determined if the pair of words was an antonym pair or not. Participants pressed either the V-key to show they were antonym pairs or the N-key to show that it was a non-antonym pair. The reaction time was measured from the time the target word was shown until the student pressed either the V-key or N-key.

<table>
<thead>
<tr>
<th>Anchor item</th>
<th>Antonym/non-antonym pair</th>
<th>Part of speech</th>
<th>Frequency level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>Right</td>
<td>Noun</td>
<td>1,000</td>
</tr>
<tr>
<td>Laugh</td>
<td>Drive</td>
<td>Verb</td>
<td>1,000</td>
</tr>
<tr>
<td>Late</td>
<td>Early</td>
<td>Adjective</td>
<td>1,000</td>
</tr>
<tr>
<td>Aunt</td>
<td>Uncle</td>
<td>Noun</td>
<td>2,000</td>
</tr>
<tr>
<td>Smile</td>
<td>Cry</td>
<td>Verb</td>
<td>2,000</td>
</tr>
<tr>
<td>Empty</td>
<td>Occupied</td>
<td>Adjective</td>
<td>2,000</td>
</tr>
<tr>
<td>Guest</td>
<td>Manager</td>
<td>Noun</td>
<td>3,000</td>
</tr>
<tr>
<td>Shout</td>
<td>Leap</td>
<td>Verb</td>
<td>3,000</td>
</tr>
<tr>
<td>Angry</td>
<td>Final</td>
<td>Adjective</td>
<td>3,000</td>
</tr>
<tr>
<td>Kilometer</td>
<td>Earth</td>
<td>Noun</td>
<td>4,000</td>
</tr>
<tr>
<td>Kiss</td>
<td>Arrest</td>
<td>Verb</td>
<td>4,000</td>
</tr>
<tr>
<td>Ugly</td>
<td>Beautiful</td>
<td>Adjective</td>
<td>4,000</td>
</tr>
</tbody>
</table>
2.3 Pronunciation Test

The third test was a pronunciation test designed to measure participants’ phonological awareness. The foundation for this test comes from the speaking section of the Wide Range Achievement Test (WRAT-4) by PAR (Wilkinson & Robertson, 2006). The original WRAT-4 speaking section contains two tests with 55 items per test. Each test contains 55 items selected from the Education Development Laboratories (EDL) Core Vocabularies in Reading, Mathematics, Science, and Social Science (Taylor, Frackenpohl, & White, 1989); and A Writing Vocabulary of Elementary Children (Hillerich, 1978). Each test begins with 15 warming up items (upper case letters of the English alphabet) to acquaint the test taker with the test format. The remaining items then move from common high-frequency words to less common low-frequency words. Test takers simply read the test items aloud, and they are judged to be either correct or incorrect. The original test was designed to be taken in a 1-to-1 format with a test taker and judge.

In order to have three equal versions of the pronunciation test for the longitudinal study, an additional new version was created. The original WRAT-4 test manual did not contain a description of the criteria for item selection, but contains only the source for the items. Therefore, the original two versions were analyzed for word length, phoneme count, syllable count, and frequency level on the British National Corpus. Once an additional set of items were selected from the same source as the original two tests, the three tests were then combined, and items were redistributed using randomly stratified allocation to create three new and equal pronunciation tests. Pilot data showed that each new test was not statistically significantly different from the other.

Test takers completed the pronunciation tests using a standard laptop computer and a headset. Test takers viewed the items using Microsoft PowerPoint and were instructed to say each item to the best of their ability and not be concerned about the word meaning. Time was not recorded for this test because accuracy was the main measure, not semantic or orthographic access speed (Jiang, 2012). The test results were then analyzed using five North American English speakers each with a high Japanese L2 speaking ability and over 10 years of post-secondary teaching experience in Japan. Each rater was coached to ensure they clearly understood the test and how to assess each speaker. Each interrater listened to their assigned audio files and determined if the speakers were correct or incorrect. Replies from test time one, two, and three were also randomly assigned so as to avoid rater bias on improvement. Individual accents were not considered a factor in the assessment; instead, each word was assessed as incorrect if it was misread by using grapheme–morpheme rules of English. For example, if a student added an additional voiced vowel to any part of a word, such as great pronounced as gu-reat-o, it was marked incorrect. Each interrater also used the Merriam-Webster Online Dictionary to access the syllable counts and audio samples to calibrate and clarify any item pronunciations (www.merriam-webster.com). The main research evaluated all audio files, and each interrater evaluated approximately 50 items, with an overlap of 10 items per interrater. This allowed for accurate Rasch assessment of interrater reliability of scoring.
3 Summary

As an important part of reading, vocabulary knowledge relies upon strong word recognition skill. Understanding this foundational skill and its components provides greater insight into vocabulary development and reading skill growth. Through longitudinal research, measurement, and analysis of orthographic, semantic, and phonological knowledge, researchers add to the knowledge of vocabulary assessment and second language reading. In addition, L2 teachers will also be able to make better pedagogical choices to help their students achieve their learning and reading goals, for example, by including training tasks designed to move readers from a word-by-word low-level cognitive processing to a more high-level cognitive processing of texts that better facilitate reading comprehension.

References


The Challenges of Measuring Multi-Word Expression Use in Conversation

Haidee Thomson

Hokusei Gakuen University Junior College Department
and Victoria University of Wellington

Abstract

This article introduces three important challenges and possible solutions when using spoken dialogue to measure the use of specific multi-word expressions. The first challenge is deciding whether to count precise and accurate use of target expressions only or whether to extend the count to include variation. The second challenge requires addressing the indirect nature of dialogue as a testing method. The third challenge is organizing data and preparing ways to clearly identify speakers within the dialogue. These challenges are illustrated with examples and potential solutions from my recent research investigating spoken use of multi-word expressions.

1 Background

Learners of English in Japan are known to struggle with speaking fluency (Herder & Sholdt, 2014; Nishino & Watanabe, 2008). Research has suggested that speakers can boost their speaking fluency by using common multi-word expressions (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Wood, 2009). Therefore, replicable classroom interventions that are proven to increase knowledge and use of multi-word expressions would be very useful.

In response to the need for proven teaching interventions, I conducted a classroom-based study to measure participants’ ability to use target multi-word expressions (such as I think I will and how do I get). This article does not report results but rather discusses the challenges and solutions found to measuring multi-word expression use. Multi-word expressions were sourced from classroom materials and checked whether they were in current general use by checking the occurrence of the expressions in the spoken sub-corpus of the Corpus of Contemporary American English (COCA) (Davies, 2013). There were 30 selected expressions, each made up of four words. Participants were from an engineering university in northern Japan, with minimal exposure to English outside of a 90-minute English class once a week (Thomson, 2017). This research was conducted as an action research project (Burns, 2010), wherein after each study was completed, results and feedback were used to inform adjustments for follow-up studies and replication. The experimental group was introduced to the target expressions through travel themed units where they were encouraged to use the expressions in activities such as shadowing, role-play, dictogloss, and more. The control group was not exposed to the target expressions, as they studied engineering topics in English using a
linked skills format. In order to assess the efficacy of the experimental teaching intervention, learners were randomly partnered together and given a scenario to role-play and audio record before and after the 6-week intervention. Participants then uploaded the audio recordings to a class Moodle page, where I could measure the use of the target expressions within the conversations.

As I endeavored to assess the impact of the classroom intervention through participant dialogue recordings, I faced three challenges. The first challenge was deciding whether to adhere strictly to counting the exact target expressions or whether to include variation when measuring multi-word expression use. This decision had the potential to limit or broaden the scope of conclusions about learning. The second challenge was the use of role-play, where one cannot be sure that unused expressions are actually unknown. Role-play allows speakers to choose their expressions, so that conclusions from such an indirect test are necessarily partial and methods of triangulation need to be considered. The third and final challenge regarded technical and logistical decisions about how to organize data collection and how to identify speakers. These decisions affected sample size and in turn how generalizable the results would be.

1.1 Challenge One: Decisions About How to Measure Multi-Word Expressions

When measuring multi-word expressions, it is important to be clear about what is to be counted and how it is to be counted, so that fair comparison or replication is possible (Porte, 2012). Running analysis on a small subset so that various counting options can be compared is likely to assist decision-making on the most appropriate counting method for a particular purpose and context. In order to identify whether the participants used the target expressions in my research, I first transcribed participant conversations (pre- and post-intervention) to Notepad UTF-8 files. I was then able to use AntConc (Anthony, 2014) to search for sequential words that made up the target expressions using the concordance word search function. Use of complete four-word expressions was rare, so in order to discover how much learning had actually taken place, a more sensitive measure was required. I decided to count partial use from two-word target combinations upward. There were also instances where participants used two or more words from a target expression with alternative words to complete the expression: for example, one target expression was “how do I get (to),” and a participant said, “how can I go (to).” The first and third words were from the target expression, whereas the second and fourth words were variants of the target. If the variant words are counted, then “how can I go” could be counted as four words. However, if only the target words are counted, then “how can I go” could only be counted as two words. I chose to count both ways and found that including such instances of variation in the measurement did not change the overall comparison of results; it simply lifted the numbers. For instance, in my first study (of three), I investigated the length of multi-word expressions and found the overall comparison post-intervention between the control group ($n = 8$, $Mdn = 2.35$) and experimental group ($n = 15$, $Mdn = 2.63$) was not statistically different whether variation was counted $U = 86.5$, asymp $p = 0.085$, $z = 1.72$, $r = 0.359$ or not: control
group $Mdn = 2$, experimental group $Mdn = 2$, $U = 83.0$, asymp $p = 0.081$, $z = 1.75$, $r = 0.364$. There was a medium-sized effect between the experimental group and the control group either way. Therefore, for the sake of clarity, I decided to keep the measurement simple by only counting the target words used in target expressions in follow-up studies (not the variant words used in target expressions).

1.2 Challenge Two: Using Role-Play as an Indirect Test of Multi-Word Expression Knowledge

As my participants were a convenience sample from conversation-based classes, I wanted to use conversation in my tests. Measuring knowledge of specific multi-word expressions through use in natural-like conversation is an inherently indirect measurement method because speakers are under no compulsion to use the specific expressions being measured in conversation. Previous research that has used dialogue to measure language production includes Tavakoli (2016) who used a discussion task between two learners to compare fluency with a monologic task in an English as a Second Language (ESL) environment, while Taguchi (2007) used dialogue between learner and researcher to investigate the use of memorized chunks in Japanese as a foreign language. However, the use of dialogue between learners to investigate use of multi-word expressions seems to be rather unique. The beauty with scenario role-play is that learners are free to choose the words or multi-word expressions that they wish to use. If a speaker chooses to use an expression under the natural time pressure of conversation, it can be assumed to some degree that they know it. The flip-side of conversational freedom is that speakers can choose alternative expressions to complete the role-play. Participants may be able to retrieve and use the expressions being measured, but in the absence of task-essentialness they can complete the role-play without using them (Thomson, Boers, & Coxhead, 2019). There are a multitude of reasons why learners may choose not to use known vocabulary or expressions in their output, including nonnecessity or lack of confidence (see Coxhead, 2018).

While spoken and written modes are not the same, the use of target expressions in either speech or writing reflects some knowledge of the expression. My solution to the indirect nature of role-play was to directly test target expression knowledge through a cloze test after the role-play. In this way, knowledge of meaning and written form of the target expressions (as tested in the cloze test) could be triangulated with the ability to produce and use orally (as measured through the role-play). It was informative to be able to cross-check (triangulate) evidence of knowledge of expressions from the cloze test with use in the role-play. The combination of these two complementary measures revealed more direct evidence as to what extent the expressions were known.

1.3 Challenge Three: Technical and Logistical Challenges

When setting up a dialogue assessment situation, decisions need to be made about the allocation of partners, identification of individuals, and roles. It is preferable to keep the speaker partnerships the same for all tests. Every speaker has their own style of speaking in their first language (e.g., long pausing patterns or
quick-fire speaking), which may influence their second language speaking style (de Jong, Groenhout, Schoonen, & Hulstijn, 2015; Derwing, Munro, Thomson, & Rossiter, 2009). And, every combination of speakers is likely to have their own dynamic, as speaking patterns converge (de Jong, 2018; Pardo, 2013; Wilson & Wilson, 2005). I had a pre- and post-experiment model; so, in order to keep the conditions as similar as possible, I strived to have participants record their pre- and post-dialogue with the same partner, playing the same role, in the same scenario role-play. Absenteeism was difficult to control, and as a result there were different partnerships in some cases, which I had to exclude from my analysis. In class situations, absenteeism will impact continuity of conditions, so checking for schedule clashes and perhaps providing incentives may help to reduce sample shrinkage.

Identification of speakers was an important challenge to consider when designing the study. I chose to record conversations using audio only (rather than using video). I made this choice for two reasons: the availability of audio recording software that participants could operate simply, and also data capacity, as storing video data for many conversations would be problematic. In my first study, participant pairs labeled the audio file with their names but were not instructed to say their names or roles at the beginning of the recording. Unfortunately, when I listened to the recordings, I found that it was extremely difficult to match the pre-intervention voice with the post-intervention voice. Most of my participants were male, and many had similar voice tones. Therefore, I ran my speech analysis by dyad rather than individual, which halved my sample size (from 46 to 23). Naturally, halving the sample size reduced the power of the data set (Thomson, 2017). Learning from this experience, the next time I collected conversational data. I instructed the participants to say their names, identification numbers, and role at the beginning of the interaction, which helped me to associate their voice and name on each recording. For those considering dialogue analysis or assessment, the use of video recording might be worthwhile, as it would add visual evidence, which would help when identifying who is speaking.

In a before and after intervention measurement of conversation, ideally the speakers would play the same roles in both conversations. In a separate follow-up study, I instructed participants to play the same role as they had in the pre-test. However, in spite of this instruction, close to half of the group ended up playing opposite roles to what they had played in the pre-test. I excluded these dialogues, which reduced my sample size in the follow-up study from 52 to 26. A fairly simple initiative to ensure that participants play the same roles in pre- and post-dialogues would be to make a list of names with roles to be shared with participants prior to doing the final recording. Such planning is simple but can be easily overlooked.

2 Implications for Future Research

The three challenges described above for assessing knowledge and the ability to use multi-word expressions through dialogue can be mitigated through advanced consideration and planning. Firstly, decisions regarding which words to count and not to count need to be thoughtfully considered. Second, measuring target expression use from conversation can be fraught with difficulties because a speaker

Vocabulary Learning and Instruction, 9 (2), 63–68.
may have been able to use an expression, but the opportunity to use it did not arise in the conversation or perhaps they chose to use an alternative word combination. Therefore, combining role-play with a more direct measure (such as a cloze test) can provide triangulation to help show the expressions that are known even if they are not used in the role-play. Finally, technical and logistical planning can maximize continuity between pre- and post-intervention conditions and dialogue partnerships. Anticipating and planning for these challenges will take some of the stress out of data analysis and/or assessment. I hope that these insights from my own data collection and analysis will help readers plan and implement their own spoken language assessments and research involving multi-word expressions.

References


Deliberate Vocabulary Learning from Word Cards

Darrell Wilkinson
Tampere University, Tampere, Finland

Abstract
While word cards are a widely supported method of deliberately studying foreign language vocabulary, there is a surprising lack of research-based evidence supporting them. This paper first reviews some of the key literature on word cards and then briefly describes two experiments concerning word card methodology. The first experiment described in this paper examined the learning outcomes of making word cards while the second experiment compared the use of self-made word cards with premade cards. The results of the first study indicated that the simple process of making word cards results in significant short-term vocabulary learning, but this new knowledge is sensitive to attrition if no further study is carried out soon after making the cards. The results of the second experiment indicated that while both methods are effective in the short and long-term, learners may be better studying from premade cards. Taken together, the results offer support for the use of word cards for foreign language vocabulary learning.

1 Introduction
Vocabulary knowledge is a central component of language acquisition and communication (Nation, 2013; Schmitt, 2008). In order for learners to master the thousands of words needed to become competent in a foreign language, most researchers now recommend a combination of both deliberate and incidental techniques (Folse, 2004, 2011; Laufer, 2003; Nation, 2013; Schmitt, 2008). While incidental vocabulary learning is very useful for consolidating and broadening the knowledge of previously studied vocabulary, deliberate learning techniques are a much more time-efficient method for learning the meaning of new words or phrases. The literature and experiments discussed in this paper are focused on deliberate vocabulary study, something which involves learners engaging in activities with the specific purpose of acquiring new vocabulary knowledge (Nation, 2013; Schmitt, 2008). Specifically, I investigate the use of word cards as a deliberate study method.

The reason for this focus came from (1) a belief in the necessity of deliberate vocabulary study in general, (2) the theoretical support regarding the strength of word cards described by Mondria and Mondria-de Vries (1994) and Nation (2013), and (3) the important and surprising gaps in the literature regarding word cards, especially those pointed out by Nation and Webb (2011).
2 Word Card Design

In their most basic form, word cards involve having the target language word written on one side, with the native language meaning of the word written on the other side. However, depending on learner or teacher preferences, other information such as part of speech, collocations, pronunciation information, or example sentences can also be included on the cards (Folse, 2004; Hulstijn, 2001; Nation, 2013; Nation & Webb, 2011). The most common type of word card is the traditional paper version, but there are a great number of digital word card programs and smartphone applications available these days, such as Quizlet. Based on my analysis and qualitative data collected from students, when using paper word cards, students can easily add whatever information they like, but some of the digital versions offer less freedom or ease of input.

3 Theoretical Support for Word Cards

Many vocabulary experts support the use of word cards (Laufer, 2003; Mondria & Mondria-de Vries, 1994; Nation, 2013; Waring, 2004); however, Paul Nation is probably the most well-known and ardent supporter of word cards for repeated deliberate vocabulary learning. Under the four-strand approach, this method falls under the language-focused learning strand, something often also referred to as deliberate vocabulary learning (Nation, 2013). Generally, while other deliberate study techniques such as learning from lists or notebooks are generally seen as effective for building vocabulary knowledge in a relatively short period of time (Nation, 2013; Nation & Waring, 1997; Schmitt & Schmitt, 1995; Walters & Bozkurt, 2009), studying from word cards has been suggested to be a more effective method (Mondria & Mondria-de Vries, 1994; Nakata, 2008; Nation, 2013; Waring, 2004).

Overall, while any deliberate learning approach will produce positive learning outcomes, word cards are commonly seen as a more effective option due to the unique affordances they provide, as discussed below.

The success of word cards can be attributed to a number of factors described by Nakata (2008), Nation (2013), and Mondria and Mondria-de Vries (1994). First, by having the L1 and L2 forms of a word written on different sides, learners are given opportunities for retrieval, that is, retrieving the meaning of the word from memory. Word lists or notebooks provide both the L1 and L2 information side-by-side, meaning that learners must cover up one of the forms in order to retrieve the other language meaning from memory, something that many students might not do. However, because the L2 and L1 forms are presented on different sides, when studying from word cards, learners are forced to try to recall the word before checking if they are correct, and this process is deemed more efficient. Learners are essentially giving themselves a mini test each time they study with their word cards, and as a result, positive test effects are likely to take place. Word cards also afford a time-efficient method for repeated rehearsal because students can quickly study the words again and again without the use of any other study aids. Furthermore, expanded spaced rehearsal can be implemented more easily than when using notebooks or word pair lists. Learners can separate their word card packs in order to focus on more unknown or new words, spending less time on easier or better-known...
words that would be presented together on a list. Also, as words are presented separately and can be regularly shuffled, there is little chance of the list effect or serial learning. Finally, as Mondria and Mondria-de Vries (1994) point out, the word card system is flexible in terms of content; a variety of information such as translations, synonyms, language functions, and pronunciation can be added to the cards.

4 Research-Based Support for Word Cards

4.1 Previous Research

In addition to the theoretical and indirect research-based support provided by the literature discussed above, there have been some empirical studies concerned specifically with the efficacy of word card study. Given the space constraints in this article, rather than describing the studies in detail, the main results of the papers are presented in Table 1.

4.2 My Research on Word Cards

In addition to the above research findings, the results of my own recent experiments have offered further support for word cards as a deliberate vocabulary study method.

In the first experiment, I aimed to determine the amount of short-term and long-term vocabulary learning that takes place from simply making words cards. This experiment was carried out as no research showing how much learning takes place from simply making word cards could be found. Nation and Webb (2011) state that “We know from experience that after making the cards, learners can recall from 30 to 70 percent of the meanings on the first run though the cards” (p. 42); however, experimental data backing up this claim have not been reported, as far as I am aware. Therefore, my experiment aimed to address this gap in the literature.

Table 1. Results of Studies on Word Cards

<table>
<thead>
<tr>
<th>Findings</th>
<th>Researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word card study results in significant vocabulary knowledge gains</td>
<td>Alliner (2011); Dodigovic (2013); Emrah and Omur (2010); Komachali and Khodareza (2012); Kuo and Ho (2012); Nakata (2008); Nikoopour and Kazemi (2014); Waring (1997, 2004)</td>
</tr>
<tr>
<td>Word cards are a time-efficient method of deliberate vocabulary study</td>
<td>Komachali and Khodareza (2012); Kuo and Ho (2012); Mondria and Mondria-de Vries (1994); Nakata (2008); Waring (1997),</td>
</tr>
<tr>
<td>Word cards help students learn both receptive and productive vocabulary knowledge</td>
<td>Komachali and Khodareza (2012); Kuo and Ho (2012); Waring (1997)</td>
</tr>
<tr>
<td>Word card learning is superior to list learning</td>
<td>Kuo and Ho (2012); Mondria and Mondria-de Vries (1994); Nakata (2008)</td>
</tr>
<tr>
<td>Learners show positive perceptions toward word card use</td>
<td>Altiner (2011); Dodigovic (2013); Emrah and Omur (2010); Kuo and Ho (2012); Nakata (2008); Nikoopour and Kazemi (2014)</td>
</tr>
</tbody>
</table>
The participants were first- and second-year students recruited from the two private universities in Tokyo. Three groups of learners were taken from three intact classes at one woman's university and one group from another private university. There were 72 participants (62 female and 10 male students) aged 18 to 21 years old. This was a convenience sample as the participants were chosen mainly due to ease of access. TOEIC (Test of English for International Communication) scores ranged from approximately 250 to 600, and all participants were enrolled in courses working towards improving their general English proficiency, but some wished to improve their scores on the TOEFL (Test of English as a Foreign Language), IELTS (International English Language Testing System), or TOEIC tests.

Based on the analysis of pretest, immediate posttest, and delayed posttest data from three word knowledge measures (passive recall, active recall, and written cloze), results suggest that the simple act of making word cards, without any deliberate study, produces statistically significant ($p < 0.001$) and pedagogically impressive learning outcomes. Immediately after making the cards, the mean gain score on the active recall knowledge measure was 60%, that of the passive recall measure was 70%, and the gain was 52% for the cloze measure. However, up to 50% of this new knowledge was lost by the delayed posttest stage. The results show that the process of making word cards results in considerable short-term memory gains and some long-term retention of target items. The results also indicate that the process of making word cards leads to better passive recall knowledge than active recall and knowledge of how to use the words in a given context. Overall, however, given that it took approximately 1 minute to make each card (15 minutes for 15 cards), and given the simple nature of the activity, these results seem quite impressive. However, significant forgetting also occurred during the 2-week period between the posttest and delayed posttest, proving that repeated study is needed after initially making the cards.

In another experiment, the learning outcomes attributable to (1) studying from premade word cards and (2) making and studying from word cards were examined. While we know that word cards are an effective deliberate study technique, I have been unable to find any carefully designed studies that compare the learning outcomes of studying from premade and self-made word cards. Therefore, it was unknown if it was better to (1) have students first make their own word cards before repeatedly studying from them or (2) spend the same total amount of time studying from premade word cards.

The participants were first- and second-year students recruited from the same two private universities mentioned above and taking similar classes. There were 85 participants aged 18 to 21 years old, whose TOEIC scores ranged from approximately 270 to 620. The results of both the cross-sectional and longitudinal studies showed that both methods are effective for developing various types of word knowledge. Both word card methods produced significant short- and long-term vocabulary knowledge; however, in the short term, premade cards were superior in terms of all three types of vocabulary knowledge. These results suggest that in the short term, the extra number of repetitions afforded by the premade card method is more important than the extra involvement load induced by making word cards. However, in the long term, there was no significant difference between
any of the measures of word knowledge. This showed that after 8 weeks of repeatedly studying the cards, the method of word card creation was not so important. Overall, the results provide further support for the use of word cards, whether they are self-made or premade, via a teacher’s resource folder or a study app such as Quizlet. Furthermore, the results highlight the need for repeated study if learners are to transfer knowledge into long-term memory.

5 Conclusion

In summary, given the theoretical and empirical evidence provided by the current word card literature, and the results of the experiments briefly described above, there is ample evidence to support the use of word cards as a deliberate study method. Word cards offer a practical and portable method of deliberately studying vocabulary while being able to switch between active and passive recalls. In addition, students can implement expanded retrieval strategies by separating cards into packs based on how well they know the words and then reviewing more well-known words less frequently than lesser known words. In addition, even a small amount of word card study can result in considerable short-term knowledge gains, and with repeated study this knowledge can be successfully transferred into long-term memory. Finally, it appears that studying from premade word cards may be just as effective, if not more so, as studying from self-made cards, meaning that many word card applications could be a suitable option for students who prefer this new platform.

References


Vocabulary Learning and Assessment: A Commentary on Four Studies for JALT Vocabulary SIG

Irina Elgort
Victoria University of Wellington

Abstract
Four papers by Chie Ogawa, Haidee Thomson, Michael Holsworth and Darrell Wilkinson were presented in the Vocabulary Learning and Assessment session at the Eighth Annual Japan Association for Language Teaching (JALT) Vocabulary Special Interest Group (SIG) Symposium, at the University of Niigata, Japan, in May 2020. The papers raised methodological questions, proposed approaches to assessing spoken and written word knowledge and fluency, and presented some experimental findings. It is my pleasure to discuss these papers in terms of the ideas proposed by the four researchers, their implementation and potential future directions.

1 Introduction: The Golden Age of L2 Vocabulary Research
Forty years ago, Paul Meara (1980) referred to vocabulary as a neglected area of language learning research. The situation has changed dramatically since then and L2 vocabulary research is now thriving. A new lexical studies conference, Vocab@, first held at the Victoria University of Wellington, New Zealand, in 2013, with two subsequent conferences, Vocab@Tokyo in 2016 in Japan and the most recent gathering, Vocab@Leuven in 2019 in Belgium, shows an impressive array of research topics, high quality of papers from prominent vocabulary researchers and, most encouragingly, many new, talented vocabulary researchers entering the field. Key themes represented at Vocab@Leuven included corpus studies, deliberate and incidental learning of words and multiword expressions, and tests of vocabulary knowledge. However, there are areas of vocabulary research that do not appear to enjoy the same level of attention as the research topics above; in particular, I want to draw our attention to two such topics: fluency development that, according to Nation (2007), is a key component of a balanced language learning curriculum, and the development of non-declarative lexical knowledge, including implicit, procedural and automatised knowledge (Godfroid, 2020). Recently, I have identified this gap in a methodological review of L2 studies in technology-mediated vocabulary development (Elgort, 2018) but, I believe, the situation is similar in L2 vocabulary research, more generally.

It is really exciting, therefore, to see that fluency development is considered in the majority of the papers presented in the Vocabulary Learning and Assessment session at this year’s JALT Vocabulary SIG Symposium. In particular, Ogawa and Thomson consider questions related to speaking fluency and Holsworth considers...
fluency of lexical processing in reading. Another encouraging trend in this year’s papers is attention to multi-word expressions. A relative newcomer to L2 lexical studies, research into processing, using, learning and teaching of multi-word expressions has been rapidly gaining momentum in the last 10–15 years, no doubt, at least partially due to the vastly improved access to large-size language and learner corpora and a simultaneous associated boost in the number of accessible and user-friendly corpus and text analysis tools (such as Lawrence Anthony’s AntLab www.laurenceanthony.net; Mark Davies’ English-Corpora.org; and #LancsBox – the Lancaster University corpus toolbox corpora.lancs.ac.uk/lancsbox spearheaded by Vaclav Brezina, and Tom Cobb’s Lextutor tools www.lextutor.ca, among others).

In the symposium, Ogawa’s paper looks at approaches to teaching formulaic language, while Thomson brings to our attention challenges associated with assessing the knowledge of multi-word expressions in the classroom setting. An innovative take on vocabulary assessment is also considered in the paper by Holsworth, who proposes a set of measures for evaluating three components of lexical quality outlined by Perfetti and colleagues (e.g., Perfetti, 2007; Perfetti & Hart, 2001): recognition of orthographic form, access to semantic knowledge and accuracy of the spoken form production. Finally, Darrell Wilkinson considers the use of word cards – one of tried-and-true methods of deliberate word learning that is experiencing a resurgence as a result of the recent advent of online and mobile flashcard applications. Let us now consider each of the four papers. I will start with the two studies by Ogawa and Thomson that look at instructional approaches to facilitating productive use of multi-word expressions as a bridge to attaining production fluency. I will then turn to Holsworth’s paper that considers fluency of visual word processing. After that, I will discuss Wilkinson’s experiments that investigate the effectiveness of self-made word cards.

2 The Four Studies

2.1 Teaching Ideas of Improving Oral Performance through Formulaic Language Instruction by Chie Ogawa, Kyoto Sangyo University

Chie Ogawa’s paper suggests ideas on how to facilitate the learning of formulaic language by English as a Foreign Language (EFL) learners, as a way to improve their spoken fluency. As mentioned in the “Introduction”, research into *formulaic language* – a cover term for different types of multi-word expressions (Schmitt, 2010) – is gaining momentum among L2 vocabulary researchers. However, so far, the bulk of this research has been focused on whether formulaic sequences are processed faster than individual words and whether non-native speakers are able to attain native-like knowledge of such sequences. There is less published research that looks at pedagogical interventions that facilitate the learning of multi-word expressions (but see the work of Frank Boers and colleagues on the topic). Therefore, Ogawa’s paper is timely.

We know that the acquisition of formulaic language is challenging and even more proficient L2 speakers may have trouble producing multi-word expressions.
One of the reasons for this learning difficulty is that formulaic sequences are not salient in the input; in reading, for example, they are not visually different from individual words and thus are less likely to be explicitly noticed as multiword units, unless they have a completely different meaning from that of their component words, as may be the case for idiomatic expressions, such as *kick the bucket* or *another cattle of fish*.

Ogawa proposes three form-focused pedagogical treatments that could promote proceduralisation of the knowledge of formulaic language: text enhancement, peer feedback and repetition in language learning activities and tasks. The first approach, text enhancement (Sharwood Smith, 1991, 1993; Szudarski & Carter, 2016), aims to draw learners’ attention to multi-word expressions, making them more salient in the input through externally induced focus. Ogawa refers to a meta-analytic synthesis by Lee and Huang (2008), suggesting that the effect of typographic enhancement on learning grammar from input is likely to be positive but small. In addition, only very few studies have examined how typographic enhancement affects proceduralisation and the development of implicit knowledge. In the first (to my knowledge) study of its kind, Sonbul and Schmitt (2013) did not see any effect of typographic enhancement on the development of implicit knowledge of lexical (medical) collocation, while learners’ explicit knowledge was enhanced by this treatment. In a conceptual replication and an extension of this study, Toomer and Elgort (2019) confirmed the positive effect of orthographic enhancement on the development of explicit knowledge of L2 collocations and no effect on implicit knowledge. Nevertheless, Toomer and Elgort found that enriched input (nine contextual occurrences over 2 days) may facilitate the development of implicit knowledge of lexical (medical) collocation, while learners’ explicit knowledge was enhanced by this treatment. In a conceptual replication and an extension of this study, Toomer and Elgort (2019) confirmed the positive effect of orthographic enhancement on the development of explicit knowledge of L2 collocations and no effect on implicit knowledge. Nevertheless, Toomer and Elgort found that enriched input (nine contextual occurrences over 2 days) may facilitate the development of implicit knowledge of lexical collocations for advanced English as a Second Language (ESL) learners (measured using a collocation priming task) when the target collocations are not typographically enhanced in the text. Furthermore, Ogawa rightly points out that typographic enhancement may change learners’ engagement with the rest of the text (Choi, 2017) and that learning from input alone may be challenging for lower proficiency learners.

Ogawa, thus, proposes the second treatment, peer feedback. The advantage of corrective peer feedback compared with teacher feedback is that learners are both feedback-givers and feedback-receivers, which doubles their opportunities for noticing the gap in their knowledge. With this in mind, Ogawa argues that it could be an effective way of increasing the noticing of multi-word expressions in the output, while warning that the provision of useful corrective peer-feedback is not assured by simply requesting it. Participants need to be willing to point out errors and inaccuracies in their peers’ output, which could be a face-threatening act. For this reason, it is critical for the learners to be trained in the provision of peer-feedback in a constructive and polite manner, and for them to understand the benefits of giving and receiving peer-feedback. What’s missing from this section, in my opinion, is concrete suggestions on how such peer feedback may be elicited, that is, what kind of activities and tasks could be designed to ensure that peer feedback is focused on multiword expressions and is useful in promoting their acquisition. Using peer-feedback to learn formulaic language may also be somewhat problematic: even if advanced L2 learners are having trouble with acquiring
formulaic language, would low and intermediate proficiency learners be able to provide useful corrective feedback to their peers? Another question is: what kind of teacher support would be most beneficial in helping students understand how to engage in peer-feedback activities in an effective way?

The third approach to improving the knowledge of formulaic language proposed by Ogawa and, in my opinion, the most promising one, is developing and using learning activities and tasks (such as 4/3/2, Nation, 1989) that create conditions for the proceduralisation of the established declarative knowledge and its eventual automatisation, as argued by Segalowitz (2010) and DeKeyser (2003, 2015). However, as shown by Solovyeva and DeKeyser (2018) and Hui (n.d.), it may take a very long time to develop automaticity of vocabulary processing through input only. Moving beyond Nation’s 4/3/2, it would have been helpful to see concrete examples of fluency development activities that are suitable for formulaic language.

Here, I want to reiterate a key requirement for fluency development activities, that is, the target multi-word expressions used in such activities need to be familiar to their participants. This helps increase the rate of production without compromising accuracy. Prior knowledge of the target multi-word expressions will also increase the likelihood of their use in retelling and other free production tasks. This brings us to one of the challenges outlined in Thomson’s paper, that is, how to ensure that the target language is used by the learners in free production.

There is one well-known caveat in teaching and learning formulaic language, that is, not all multi-word expressions are created equal (see Carrol & Conklin, 2019; Martinez & Schmitt, 2012; Siyanova-Chanturia & Martinez, 2015). Instructional approaches chosen to facilitate the learning of idioms, binomials, collocations and lexical bundles may need to be different. For example, in an eye-tracking study, Carrol and Conklin (2019) found that, beyond frequency, decomposability matters most for idioms, predictability and semantic association for binomials, and mutual information for collocations. This is why it is helpful to clarify what type or types of multiword expressions are targeted in the proposed instructional and learning treatments.

Overall, I enjoyed reading this paper because it took a careful and balanced approach in proposing research informed pedagogical treatments that could facilitate acquisition of multiword expressions. The author suggests that the combination of the three teaching ideas could promote better knowledge of L2 multi-word expressions and, overall, I agree. However, if fluency is, indeed, a key goal, I would prioritise multiple opportunities for processing and re-use over peer corrective feedback. Moreover, different stages of acquisition would require different instructional approaches: deliberate learning and explicit noticing through input enhancement and peer feedback may be particularly effective at the point of initial familiarisation and early learning, while increasing frequency of encounters (e.g., via input enrichment) and opportunities for re-use under time pressure may be more beneficial at later learning stages. These conjectures need to be tested and verified in future L2 learning studies with multi-word expressions.
2.2 The Challenges of Measuring or Assessing Multi-Word Expression Use in Conversation by Haidee Thomson, Hokusei Gakuen University and Victoria University of Wellington

Haidee Thomson set out to investigate whether classroom interventions that require students to use and re-use four-word multi-word expressions, such as I think I will or how do I get, in structured learning activities and tasks promote their transfer into students’ spontaneous interactions, such as those taking place in a role-play. In this study, she faced three key challenges: (1) how to score the accuracy of use, for example, whether or not to count incomplete or altered renditions of the original expressions as correct; (2) how to interpret the non-use of the target expression; and (3) technical and logistical challenges of assessing recorded spoken interactions in classroom research.

Challenges (1) and (2) are related. The solution to the challenge of scoring the accuracy of the target expressions in free production (e.g., How can I go [to], instead of the target, How do I get [to]) and non-use (when the target expression is not produced), proposed by Thomson, is to follow free production with a cloze test, thus providing a form of triangulation. I agree that it is helpful to triangulate different measures, particularly so when a quantitative measure is accompanied by a qualitative one, in order to create a richer picture of the phenomenon or construct under investigation. It is important to think how the measures being triangulated are related to the phenomenon of interest. In a study that investigates fluency, Thomson proposes to use measures of controlled off-line production of multi-word expressions to triangulate measures of their free online production. Free spontaneous production relies, at least to some extent, on non-declarative (procedural, implicit) lexical knowledge, while a cloze test relies on declarative (explicit knowledge) and the use of task and linguistic strategies to come up with the right answer. Although the two knowledge types can support similar lexical processes, fluent speech is not possible when only explicit knowledge is used. So, even if it turns out that the results of the cloze test are correlated with the fluency measures, there is no guarantee that the cloze test scores would be predictive of fluent production, in general, or even the fluent production of the target multi-word expressions.

No doubt, measuring effectiveness of instructional and learning treatments in free production is hideously difficult. This is where psycholinguistic measures and approaches may help. Using behavioural (e.g., response time) measures of production under time pressure (e.g., using a shadowing paradigm), online access to the knowledge of multiword expressions can be tested under the conditions that limit participants’ ability to use deliberate task and memory strategies, while the production of the target items is inevitable. Combing naturalistic free production with such behavioural production measures can get us closer to the desired triangulation that addresses the same kind of knowledge and access to this knowledge.

Thomson provides an excellent overview of the methodological issues and pitfalls that will be very useful for future studies that plan to audio-record students’ classroom interactions. This experience underscores the importance of working out the study procedures and labelling approaches during the piloting stage. Researchers may also seek technological solutions to this problem, for example,
by using qualitative data analysis tools that link media files with participants’ IDs, such as NVivo, or use multi-track recording software with separate input feeds for the two participants in dialogues (such as, Audacity).

Let us now return to the challenge of scoring partial responses. Thomson gives the following example: *How can I go (to)* produced instead of the target *How do I get (to).* Because there were too few instances of complete expressions with all four words, she decided to score their production accuracy word-by-word. This presented its own challenge – whether to assign such an instance a score of 2 (because two out of the four words are the same as in the original lexical bundle) or a score of 4 (if the substitutions are counted as acceptable variants). In my opinion, there is a third option – not counting such expressions as correct, thus acknowledging that the classroom interventions used in this study did not have an observable effect on the development of free productive knowledge of the target expressions. Conceivably, *How can I go (to)* represents a free combination of four words put together in order to communicate a message, using individual known words. Indeed, is there a reason to assume that *How can I go (to)* is a pre-fabricated multi-word expression? This situation raises an important question: where do we draw the line between formulaic language and free word combinations and, at the risk of sounding radical, how important is it to have such a clear line? When teaching adults, in particular, it is just as important (if not more so), in my opinion, to have high quality of semantic knowledge of individual words, for example, *get* signifies a completed action (akin to the perfective aspect in other languages) and *go* signifies an incomplete action (akin to the non-perfective aspect). Presumably, such semantic/grammatical awareness would help learners to select *get over go* in the example provided. This thinking is aligned with Csomay’s (2013) conclusion based on an analysis of the use of lexical bundles in university classroom discourse that, on the discourse level, there is a “strong relationship between grammar and lexis”.

Not to put in doubt the importance of having the knowledge of formulaic language, let me put it this way: if we know that formulaic language is resistant to explicit teaching and deliberate learning (especially if we are talking compositional, more transparent sequences, such as lexical bundles and grammatical collocations), why not simply encourage and facilitate learner exposure to comprehensible input (both spoken and written) that contains frequent lexical bundles and let statistical learning (Ellis, 2002) do its job? And, in the meanwhile, explicit teaching and deliberate learning time could be spent on the aspects of vocabulary knowledge that are more likely to benefit from this kind of learning, such as creating a robust form-meaning mapping for individual words (Elgort, 2011), idiomatic expressions (Obermeyer & Elgort, under review) and less compositional phrases, such as *not at all* (as a response to, *thank you*). This is where the teaching ideas proposed by Ogawa may also come handy; initially, input enhancement could promote noticing. Once learners start attempting these sequences in production, peer feedback could be helpful and, finally, as their accuracy of knowledge of these formulaic sequences develops, activities that facilitate fluency development become important. This would be in line with the *Four Strands* (Nation, 2007).

Now, let us briefly go back to the question of fluency. Both Ogawa and Thomson start their papers with the statement that fluency is an important goal.
in second language learning, and that acquisition of formulaic language is a way to facilitate fluency attainment. So far, so good. However, the actual papers then proceed to the discussion of how best to facilitate the learning of multi-word expressions and evaluate the resulting knowledge, stopping short of making an explicit connection to fluency. Furthermore, the construct of fluency is not at all straightforward. Is fluency about a speed-up of some cognitive and motor processes or does it involve a qualitative change, a restructuring of lexical processing that optimises lexical access in comprehension and production (Anderson, 1982; DeKeyser, 2015; Segalowitz, 2000)? Another important question is the following: whose perception of fluency is accepted as the norm and what linguistic and non-linguistic features affect this perception? For native-speaker raters, for example, Saito, Ilkan, Magne, Tran, and Suzuki (2018) found articulation rate and pausing to be the distinguishing factors in fluency ratings of native English speakers and Japanese speakers of English of different proficiency levels. L2-speakers’ perceptions of fluency, on the contrary, appear to be affected by a whole host of factors, in addition to the delivery speed and pausing (Magne et al., 2019), including lexical richness (the use of varied and sophisticated vocabulary) and the perception of speech as being native-like. Interestingly, an accurate online use of formulaic language could be seen as both a feature of lexical richness, sophistication (e.g., the use of idiomatic expressions) and of native-likeness (the use of lexical and grammatical collocations and lexical bundles).

2.3 Assessing Low-Level Cognitive Processes of Word Recognition by Michael Holsworth, Kyoto Sangyo University

Michael Holsworth’s paper considers three components of word recognition: orthographic, semantic and phonological knowledge, and proposes a battery of tests for measuring these components. Similar to Ogawa and Thomson, Holsworth is interested in the fluency and automaticity of access to lexical knowledge, but, this time, instead of spoken production, the focus is on visual word recognition and processing, such as that involved in L2 reading. Holsworth refers us to the early work by Charles Perfetti (Perfetti, 1985; Perfetti & Hart, 2001), suggesting that quality and efficiency of lower level word processing affects readers’ ability to allocate cognitive resources to higher level processes, such as the construction of the global and local meaning of the text, integration of the new and old information and making inferences, which are needed for reading with understanding.

The author does not appear to make a distinction between knowledge and access to knowledge, often referred to as declarative and procedural knowledge in the L2 literature. Instead, if I understand correctly, Holsworth sees processing and access to knowledge as inherent in the quality of the knowledge itself. Taken one step further, this means that word learning is a process of improvement of the quality of lexical representations (Perfetti, 2007), understood as the establishment of (1) precise, fully specified orthographic representations, (2) redundant, word-specific phonological representations and (3) generalised and flexible semantic representations, accompanied by an increasingly stronger binding of these three types of representations.
To help language teachers evaluate the quality of L2 word knowledge of their students, Holsworth adopted and extended three tests of the constituent components of lexical quality. To measure orthographic knowledge, he used a lexical decision (LD) test that required participants to make word–non-word decisions; to measure semantic knowledge, he used an antonym-pairs test that required participants to identify pairs of items as antonyms or non-antonyms. According to Holsworth, both these tests had been originally developed by Matsuo (2017) and extended by Holsworth to include two additional sets of parallel items for longitudinal testing purposes. The items used in the tests were selected from the most frequent 4,000 words of the British National Corpus (BNC) corpus. The scoring of these tests only concerned response times (with incorrect responses excluded), as far as I understand. The third, pronunciation test, was based on the speaking section of the Wide Range Achievement Test (WRAT) (Wilkinson & Robertson, 2006) and extended by Holsworth to have an additional set of items. In this test, the scoring concerned the accuracy but not the speed of responses.

Overall, I really like the idea of evaluating the three components of lexical quality to guide EFL teachers’ understanding of their students’ lexical knowledge and planning of instructional interventions to extend this knowledge. However, I also have a number of questions and concerns about the proposed tests related to their validity, item selection, scoring procedures and the interpretation of their results. Firstly, let us consider the validity of the measures used to evaluate each of the components. Personally, I am not convinced about the use of response times in the LD task as a measure of the precision of L2 orthographic representations. The speed of LDs is not a very good measure of orthographic knowledge; rather, LDs measure lexical access. It has been shown that making an LD will not only involve accessing formal–lexical representations, but will also engage lexical–semantic representations of the word stimuli (Joordens & Becker, 1997; Masson, 1995; Neely, 1991). This is precisely why studies interested in orthographic representations and orthographic processing usually use orthographic priming (e.g., Andrews & Hersch, 2010; Forster & Veres, 1998) rather than non-primed LDs. Furthermore, higher accuracy of LDs may point to more precise lexical representations, while faster responses are more likely to indicate higher ease/fluenzy of lexical (but not, specifically, orthographic) processing. So, if the LD test described in this paper was, indeed, “designed to measure the orthographic knowledge of a reader” using response times, then the construct validity of the test may be violated. In fact, orthographic knowledge may be better measured by a dictation or speeded spelling judgment/verification test.

In addition, using Reaction times (RTs) in LDs as a measure of lexical processing assumes that LDs can be made at a certain level of accuracy (as pointed out in the paper). This does not appear to be the case for lower proficiency EFL speakers, who are the target audience for this test (based on the fact that only the first 4000 frequency-level words are tested). Grainger and Jacobs (1996) explain that both local and global lexical activation processes are involved in making ‘yes’ decisions in the LD task. The local activation criterion is met when the activation level reaches a threshold for a particular word, while the global criterion is a function of the overall level of activation in the lexical processor. In L1 LDs, the
local activation criterion is likely to be reached before the global one (provided the overall level of activation in the lexical processor is set high by using word-like non-words); that is, the decisions are likely to be based on the activation of the lexical representations of the individual stimuli rather than on their word-likeness. However, with EFL participants who may only have very weak lexical representations for L2 (English) words, the global criterion of word-likeness may be reached prior to the local one. This means that the LD may not be indicative of the knowledge of the test items. One of the signs that L2 lexical representations are unstable for lower proficiency EFL participants is their high rate of false alarms – “yes” responses to word-like non-word distractors.

Secondly, I have noticed some seeming inconsistencies in the selection of items used in the test of semantic representations, that is, the antonym pairs test. Some of the pairs appear to be much easier to judge than others, based on the strength of the semantic relationships between the two items. For example, it is very easy to make a decision that right and left are antonyms and that laugh and drive are not, but the decisions to word pairs, guest – manager and kilometre – earth are not straightforward. This means that the measure of response time in this semantic classification task does not only reflect the participants’ knowledge of and access to the semantic features of the test items but also the presence – absence of similar semantic features in the pairs of items and the level of ambiguity in the critical semantic feature. For example, for late – early, the critical feature is located on the time continuum and is in common for the two words in the pair, that is, early means before some expected time or near the beginning of a time period, while late means after some expected time or near the end of a time period; in other words, the participants are comparing (−)n with (+)n, where n is the set feature in common. The same analysis would work for life – death, that is, (−) being alive versus (+) being alive. For laugh – drive, the decision is also relatively straightforward (provided the meanings of the two words are known) because there is no semantic feature in common between the two items with a +/− sign. However, this fast decision-making approach will not work for guest – manager and, therefore, some degree of deliberation may be involved in making the antonym/non-antonym decision. For example, in a hotel, guest is a visitor and manager is someone who receives visitors; therefore, these two words could be antonyms; on the contrary, if the hotel schema is not invoked as part of the decision-making process, the two items could be perceived as not having semantic features in common and a “no” decision could be made. Thus, response times in this task reflect more than semantic knowledge; they reflect conscious explicit decision-making processes that are task specific and unrelated to the quality of semantic representations of the items involved. For a speeded antonym pairs test to be a valid measure of semantic knowledge, therefore, pairs of items need to have a consistent, unambiguous semantic relationship that would allow test-takes to identify them as antonyms or non-antonyms, without the need to engage metacognitive strategies.

Thirdly, the author states that the proposed pronunciation test was “designed to measure participants’ phonological awareness”. To me, this framing of the test is confusing; this is because, to measure phonemic or phonological awareness, some form of recognition or segmentation test is usually envisaged; however, instead, the test proposed by Holsworth involves a word pronunciation (i.e., production) task.
I would also question the rationale for specifically selecting the WRAT-4 speaking section to measure phonological awareness because the word reading section of this test was designed to measure “word decoding through letter identification and word recognition”, rather than phonological awareness (online information available from www.pearsonclinical.com.au/products/view/567).

I also want to comment on the practical aspects of conducting and scoring these three tests by language teachers. Because the measure used in the lexical decision and antonym pairs tests is response times (rather than response accuracy), the test will need to be taken in a quiet (ideally, sound-proof) computerised environment because millisecond-level response time differences need to be recorded. In addition, because Holsworth proposes not to remove extreme outliers (very slow responses), the integrity of the data analysis and, consequently, test results would be threatened by the presence of such outliers. In relation to the pronunciation test, the scoring appears to be quite laborious and would require seeking and training inter-raters for the purposes of test scoring. In my opinion, using a recognition or segmentation test format to measure phonological awareness would eliminate these requirements and make the test much easier to score and administer.

Finally, similar to the two papers discussed above, clear conceptualisations and/or operationalisations of automaticity and fluency are missing in this paper. Holsworth correctly points out the need for the lower level processes, such as word recognition and processing during reading to be “automatic and efficient”. However, the three tests proposed in the paper may not assess automaticity of word processing as conceptualised by Segalowitz and colleagues (e.g., Segalowitz & Hulstijn, 2005; Segalowitz & Segalowitz, 1993). Since the development of fluency and automaticity of low-level lexical processing is, indeed, a very important goal of language learning, I would recommend to explicitly operationalise them, clearly state what criteria would be indicative of their attainment (or different levels of attainment) and show how the proposed tests speak to these criteria. An idea of developing a battery of tests that assess lexical quality (as defined by Perfetti, 2007) goes a long way to address this; these tests need to be technically appropriate to target well-defined constructs representing specific knowledge dimensions while, at the same time, remaining accessible to language teaching practitioners. This is a very challenging task and Holsworth’s paper facilitates our thinking on how we can get there.

2.4 Deliberate Vocabulary Learning from Word Cards by Darrell Wilkinson, Tampere University, Tampere, Finland

The paper by Darrell Wilkinson takes another look at using word cards as a deliberate word learning method. He goes over some key reasons why the use of word cards may be more effective than using word lists, highlighting the importance of retrieval of form and meaning, opportunities for implementing different spacing regimes, learner control over the selection of items and flexibility in terms of the time and place of practice. He then shares the findings from his two experimental studies that evaluated the benefits of learners making their own word cards. In the first study, Wilkinson confirmed that the sheer act of making
word cards results in knowledge gains, and that these gains are greater on passive than on active recall. Unfortunately, some details of the study are not provided in the paper, for example, whether the target words were provided to the learners or whether learners chose their own learning targets, what measures were used to evaluate passive and active recall, and whether the initial knowledge of the target items was measured prior to the word card intervention. The second experiment compared the learning from pre-made and self/learner-made word cards. Wilkinson found that, on all measures (passive and active recall and retrieval), the short-term learning from pre-made word cards was superior compared to that from the self-made cards. However, long-term, this advantage of the pre-made cards dissipated and, in the end, the learning outcomes from these two approaches were not different.

These results are reassuring and suggest that both approaches to making word cards can be used, as and when appropriate. For example, teachers can make electronic word cards (using Quizlet) for the words students need to learn as part of their foreign language curriculum, and students can create their own word cards, that are relevant to their interests and learning goals and add them to their personal card decks and practice routines. This line of research could be further extended to check whether the use of audio recordings made by the teacher or by students themselves would better support word learning.

3 Conclusions and Recommendations

The four papers presented at the Eighth Annual JALT Vocabulary SIG Symposium raise important issues and are representative of some key direction in vocabulary research, namely, teaching, learning and acquisition of multi-word expressions, the development of fluency and automaticity in processing and production, and approaches to measuring and testing aspects of word knowledge. The authors demonstrate keen interest in drawing on and applying interdisciplinary insights and methods from cognitive psychology, psycholinguistics and education/learning science to applied linguistics research, which is laudable. This is in line with recent calls for interdisciplinarity in SLA research, for example, Rebuschat, Meurers, and McEnery (2017) in Language Learning; Suzuki, Nakata, and DeKeyser (2019) in The Modern Language Journal and our upcoming Special Issue of Second Language Research on Issues in L2 lexical acquisition and processing: Setting an interdisciplinary research agenda (with Anna Siyanova). Importantly, these issues are considered with practical outcomes in mind, that is, how to optimise foreign language learning and teaching practices. It has been a real privilege to discuss the ideas put forward by the four papers and suggest how they may be fine-tuned and extended. Thank you for the opportunity.

References


