Vocabulary Learning and Instruction

Volume 8, Number 1,

March 2019

doi: http://dx.doi.org/10.7820/vli.v08.1.2187-2759
Table of Contents

Articles
Letter from the Editor iv
Raymond Stubbe

Seventh Annual JALT Vocabulary SIG Symposium Morning Session: Vocabulary and Corpus Linguistics
Validating the Construct of Readability in EFL Contexts: A Proposal for Criteria Geoffrey G. Pinchbeck
Toward Written Error Correction with a Japanese-English Parallel Corpus: Data-Driven Learning in the Japanese EFL Classroom Michael McGuire
Evaluating Corpora with Word Lists and Word Difficulty Brent A. Culligan
Commentary on Four Studies for JALT Vocabulary SIG Yukio Tono

Seventh Annual JALT Vocabulary SIG Symposium Afternoon Session: Vocabulary Learning and Assessment
An Examination of the New General Service List Tim Stoeckel
Evaluating the Efficacy of Yes–No Checklist Tests to Assess Knowledge of Multi-Word Lexical Units Raymond Stubbe and Yumiko Cochrane
Comparing the Effectiveness of Word Cards and List Learning with Japanese Learners of English Magda L. Kitano and Katsuhiro Chiba
Foreign Language Vocabulary Learning and the Type of Processing-Resource Allocation Model Shusaku Kida
Vocabulary Instruction and Learning: A Commentary on Four Studies Jenifer Larson-Hall

Page
iv
1
8
17
29
39
53
62
70
76
83
Letter from the Editor

Dear Readers,

It is my great pleasure to offer you the March 2019 issue of *Vocabulary Learning and Instruction (VLI)*, featuring papers delivered at the 7th Annual JALT Vocabulary SIG Symposium at Meiji Gakuin University, Tokyo on September 15, 2018. In these pages you will find a variety of articles focusing on two topics: *Vocabulary and Corpus Linguistics* as well as *Vocabulary Learning and Assessment*. You will also find two expert commentaries by Yukio Tono and Jenifer Lawson-Hall.

As a reminder, VLI is an open-access international journal that provides a peer reviewed forum for original research related to vocabulary acquisition, instruction, and assessment. Submissions are encouraged from researchers and practitioners in both EFL and ESL contexts.

Please enjoy this issue,
Raymond Stubbe,
Editor, *VLI*
A Reconsideration of the Construct of "A Vocabulary for Japanese Learners of English": A Critical Comparison of the JACET Wordlists and New General Service Lists

Shin'ichiro Ishikawa
Kobe University

Abstract

By comparing the vocabularies included in the Japan Association of College English Teachers (JACET) wordlists (1993, 2003, and 2016 editions) and recently released New General Service Lists (Brezina & Gablasova, 2013; Browne, 2013), we tried to identify the construct of "the vocabulary for Japanese learners" with which JACET researchers have been concerned. Our quantitative analysis has shown that it includes the vocabulary concerning (a) family and people, (b) houses and daily life, (c) foods and cooking, (d) clothes and fashions, (e) sports, (f) social meetings, (g) transportation, and (h) emotions and mental states, as well as (i) spoken English vocabulary and (j) non-basic forms. This finding will shed new light on the discussion of what kind of vocabulary should be included in pedagogical wordlists.

Keywords: vocabulary selection; corpus-based approach; pedagogical adjustment

Introduction

It is true that recent developments of mega-corpora and the increasing sophistication of corpus query techniques have made the compilation of wordlists much easier than before, but this may not apply to the creation of pedagogical wordlists, for there exists a certain amount of pedagogically indispensable vocabulary that cannot be extracted easily from corpora (Ishikawa, 2007, 2015). Unlike L1 children, L2 learners, especially in English as a Foreign Language (EFL) situations, usually learn the vocabulary not from natural input but from institutionalized education at schools, which is undoubtedly influenced by their L1, nationality, and culture. This suggests the need to consider learners’ backgrounds in making a pedagogical wordlist.

The Japan Association of College English Teachers (JACET), which is one of the largest expert associations in the field of applied linguistics in Japan, has been engaged in the compilation of “a pedagogical vocabulary list for Japanese learners of English” (JACET 2003, p. 2). The first JACET list (J1) was published in 1981, and the most recent edition (J5) was published in 2016 (Fig. 1; Table 1). This means that the JACET has continued the challenging task of wordlist compilation for more than three decades.
Although the JACET researchers have regarded word frequencies obtained indirectly from established wordlists (J1, J2, J3) or directly from corpora (J4, J5) as an important base of judgment in vocabulary selection, they have also been interested in rational adjustments so as to make purely frequency-based wordlists more suitable for Japanese learners of English.

However, how the JACET researchers operationalized the construct of “a vocabulary for Japanese learners of English,” what kinds of adjustments they really made, and whether their adjustments were consistent and truly justifiable still remain uncertain. Therefore, this article aims to probe the construct of “a vocabulary for Japanese learners of English” with which the JACET researchers have long been concerned.

Identification of the Vocabulary for Japanese Learners of English

The history of JACET wordlists illustrates the developers’ lasting belief that there should exist some special set of vocabulary for Japanese learners of English that presumably cannot be extracted directly from corpora. In other words, they believe that words for learners should be chosen not only based on their frequency and dispersion (range) in corpora but also based on some pedagogical considerations.

What kind of vocabulary then have the JACET researchers regarded as “the vocabulary for Japanese learners of English” in the long history of their wordlist developments? To what extent is it different from the ordinary high-frequency words obtained from corpora? We discuss three research questions here.
RQ1. What kind of vocabulary items are regarded as the core of the JACET wordlists and frequency-based wordlists for general purposes?

RQ2. What kind of vocabulary is missing from the JACET core vocabulary?

RQ3. What kind of vocabulary is included only in the JACET core vocabulary?

Data
As Japanese learners currently learn 3000 words at secondary schools, we pay attention to the roughly 3000 words placed at the top of three JACET wordlists: J3_3K (2864 words in Levels 1–4), J4_3K (2750 words and the “plus 250” words), and J5_3K (3000 words), as well as two recently compiled corpus-based wordlists for general purposes (Table 2): Browne (2013) (2801 words and 52 supplementary words; Browne hereafter) and Brezina and Gablasova (2013) (2494 entries; B & G hereafter). These two corpus-based lists are intended as New General Service Lists (NGSL), namely, modernized alternatives to the General Service List (GSL) (West, 1953).

Method
Principles for dealing with tokenization differ greatly among the five lists to be compared. Some include derivative forms (e.g., ing-nouns and ly-adverbs), proper nouns (e.g., England and Japan), numerals (e.g., one and fourteen), and phrases (e.g., a lot of and according to) but others do not. Therefore, we re-lemmatized all five wordlists with the lemma table used for the development of J5 and identified the number of word forms to be compared in the current analysis, where we do not distinguish different meanings and functions of a single word form.

After re-lemmatization, the number of word forms included in each of the five lists was found to be between 2201 and 3016 (Table 3). We can say that the

Table 2. Outline of the Two Corpus-based Wordlists for General Purposes

<table>
<thead>
<tr>
<th>Wordlists</th>
<th>Source corpora</th>
<th>Entries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browne</td>
<td>Subsection from the Cambridge English Corpus (273 million words)</td>
<td>2801 words + 52 supplementary words (e.g., numerals, and names of months and days)</td>
</tr>
<tr>
<td>Brezina and Gablasova (B &amp; G)</td>
<td>The Lancaster–Oslo–Bergen Corpus (1 million words), BNC (100 million words), BE06 Corpus of British English (1 million words), and EnTenTen12 (11 billion words)</td>
<td>2494 entries including phrases. The word-forms of different POS are treated separately.</td>
</tr>
</tbody>
</table>

Table 3. The Number of Lemmas to be Compared in the Current Analysis

<table>
<thead>
<tr>
<th>Lemmas</th>
<th>J3_3K</th>
<th>J4_3K</th>
<th>J5_3K</th>
<th>Browne</th>
<th>B &amp; G</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2740</td>
</tr>
<tr>
<td>Merged</td>
<td>2740</td>
<td>2865</td>
<td>3016</td>
<td>2849</td>
<td>2201</td>
<td>3896</td>
</tr>
</tbody>
</table>
sizes of these lists are roughly equivalent to the size of the vocabulary needed by Japanese secondary students.

Concerning RQ1, we chose the words included in all three editions of the JACET wordlist, which will be called the JACET core vocabulary. Next, we chose the words included in both Brown’s and B & G’s lists, which will be called the NGSL core vocabulary. We then compared the JACET and NGSL core vocabularies and identified the words included only in the NGSL core (RQ2) and those included only in the JACET core (RQ3). The latter could be regarded as an embodiment of what JACET wordlists editors have regarded as the appropriate vocabulary for Japanese learners of English.

Results and Discussions

RQ1: JACET and NGSL Core Vocabulary

The number of words included in all of the J3_3K, J4_3K, and J5_3K was 2161. Table 4 shows the first 20 words in alphabetical order. We can say that this is core vocabulary of the JACET wordlists.

Meanwhile, the number of words included in both of the NGSL wordlists was 2100. Table 5 shows the top 20 words in alphabetical order. This can constitute the core vocabulary of corpus-based wordlists for general purposes.

It is of note that the sizes of the JACET core vocabulary and the NGSL core vocabulary were almost identical, which might suggest that the top 2000 words are relatively more stable than lower-ranked words.

RQ2: What Is Missing from the JACET Core Vocabulary

By comparing the JACET and NGSL core vocabularies, we identified 1729 words that were commonly included in both lists. Meanwhile, it was found that 373 words were included only in the NGSL and absent from the JACET core vocabulary. Table 6 shows the first 100 words in alphabetical order.

This suggests that JACET has regarded words primarily concerning business (agenda, asset, budget, buyer, colleague, consultant, consume, corporate, delivery) and higher-order mental activities (accuse, acknowledge, advise, analyze, anticipate, appreciate, approval, assess, awareness, complaint, comprise, confirm, consideration) as relatively less important for Japanese learners. This list seems to include many morphologically complex suffixed words.

Table 4. A Part of the JACET Core Vocabulary

| a, abandon, ability, able, about, above, abroad, absolutely, absorb, accept, accident, accompany, accord, account, accurate, achieve, acid, acquire, across, act |

Table 5. A Part of the NGSL Core Vocabulary

| a, abandon, ability, able, about, above, absence, absolute, absolutely, abuse, academic, accept, acceptable, access, accident, accommodation, accompany, accord, account, accurate |

Vocabulary Learning and Instruction, 8 (1), 1–7.
Ishikawa: Comparison of the JACET Wordlists and New General Service Lists

Table 6. Words Missing from the JACET Core Vocabulary

| absence, absolute, abuse, academic, acceptable, access, accommodation, accuse, achievement, acknowledge, ad, addition, additional, adjustment, administration, advise, agenda, album, alcohol, alongside, alternative, analysis, analyze, angle, anticipate, anxiety, apparent, appoint, appreciate, approval, approximately, arrival, assess, assessment, asset, assist, assistance, awareness, barely, barrier, beer, bind, boost, brand, breast, brilliant, budget, buyer, cancer, category, chapter, charity, chart, chat, childhood, chip, chocolate, cite, clinical, colleague, commit, commitment, complaint, component, comprehensive, comprise, concentration, confident, confirm, consequence, consequently, considerable, consideration, consistent, constantly, construction, consultant, consume, context, conventional, core, corporate, county, coverage, creation, creative, criminal, criterion, crucial, cultural, currently, decline, defense, definitely, definition, delivery, dependent, depression, derive, deserve |

Table 7. The Words Included only in the JACET Core Vocabulary

| abroad, absorb, acid, admire, advertise, ah, airport, alarm, altogether, anger, angry, anxious, anybody, apple, arrest, aside, asleep, aspect, assembly, assignment, assistant, aunt, authority, autumn, awake, badly, banana, bare, bark, baseball, basket, bat, bath, bay, bean, beginning, behave, being, bell, belt, bench, bend, beneath, beside, besides, best, better, bicycle, birthday, bite, blanket, blind, bomb, born, bowl, brick, broadcast, brush, building, burst, bury, bush, butter, calm, cap, captain, carpet, castle, cattle, ceiling, ceremony, champion, chase, cheer, cheese, classroom, clay, clerk, clever, clue, coat, coin, command, communicate, compete, complicate, concert, conference, confuse, conscious, continent, cook, cooking, corn, cotton, counter, courage, cousin, cow, crack |

(e.g., -able, -tion, -ment, -ate, and -ness), and the average length (number of letters) of these 373 words is 7.90 (SD = 2.38).

**RQ3: What Is Included Only in the JACET Core Vocabulary**

Our analysis found 435 words that were included only in the JACET core vocabulary, which we can conclude is the essence of the vocabulary for Japanese learners of English that JACET researchers have consistently prioritized. Table 7 presents the first 100 words in alphabetical order.

These words may be roughly classified into vocabulary concerning (a) family and people (*aunt, cousin, assistant, authority, captain, clerk, champion*), (b) houses and daily life (*bath, bench blanket, brick, brush, carpet, aisle, ceiling, counter*), (c) foods and cooking (*apple, banana, bean, bowl, butter, cheese, cook, cooking, corn*), (d) clothes and fashions (*belt, cap, coat, cotton*), (e) sports (*baseball, basket [ball], bat*), (f) social meetings (*assembly, concert, conference*), (g) transportation (*abroad, airport, bicycle*), (h) emotions and mental states (*anger, angry, anxious, asleep, awake, calm, cheer, clever, conscious*), and (i) spoken English (*ah, badly*), as well as (j) non-basic forms (*beginning, being, building, ceiling, cooking, best, better, born*).

This list includes many relatively simpler words in terms of not only semantics but also morphology. The average length (number of letters) of the 435 words is 5.83 (SD = 1.96), which is significantly shorter than the average length of the words missing from the JACET core vocabulary (Welch's two-sample *t*-test yielded: $t = 13.3643$; $df = 720.6922$; $p < 0.001$; Cohen’s $d = 0.959$).

*Vocabulary Learning and Instruction, 8 (1), 1–7.*
Summary
This article tried to identify the construct of “the vocabulary for Japanese learners” with which JACET wordlists editors have been concerned. Our data-based analysis has shown that it consists of the vocabulary concerning (a) family and people, (b) houses and daily life, (c) foods and cooking, (d) clothes and fashions, (e) sports, (f) social meetings, (g) transportation, and (h) emotions and mental states, as well as (i) spoken English vocabulary and (j) non-basic forms.

Although it is not necessarily clear how the JACET wordlists will change in the future, one thing that could be done is to incorporate the results of the analysis of Japanese learners’ L2 outputs into their vocabulary selection. The recent development of learner corpora has made it possible for researchers to identify the problems that learners with a particular L1 typically have in the use of L2 English vocabulary (Ishikawa, 2013). If JACET continues to try to choose words exclusively for “Japanese” learners of English, it would be indispensable for them to examine how they actually use English vocabulary in their speech and writing.

References


Validating the Construct of Readability in EFL Contexts: A Proposal for Criteria

Geoffrey G. Pinchbeck
Carleton University

Abstract
This article examines how English as a foreign language learners might be better matched to reading texts using automatic readability analysis. Specifically, I examine how the lexical decoding component of readability might be validated. In Japan, readability has been mostly determined by publishers or by professional reading organizations who only occasionally publish their lists of readability ratings for specific texts. Without transparent readability methods, candidate texts cannot be independently evaluated by practitioners. Moreover, the reliance on centralized organizations to curate from commercially available texts precludes the evaluation of the multitudes of free texts that are increasingly available on the Internet. Previous studies that have attempted to develop automatic readability formulas for Japanese learners have used surface textual features of texts, such as word and/or sentence length, and/or they have used word-frequency lists derived from large multi-register corpora. In this article, I draw upon the findings of a study that examines how such word-lists might be validated for use in matching Japanese learners to texts (Pinchbeck, manuscript in preparation). Finally, I propose a list of general criteria that might be used to evaluate the components of readability formulas in general.

Keywords: Readability, vocabulary, matching texts, Japan, cloze test

1 Background

A key dilemma in English as a foreign language (EFL) program design is how to choose course texts that are meaningful and interesting, and that also provide a source of useful new language input. This dilemma lies in the fact that texts – written or spoken – that are over-loaded with new (i.e., unknown) language forms will not be meaningful or interesting. Matching texts based on their content, interest, and language level to specific groups of learners and/or individuals is, therefore, a prerequisite for effective language pedagogy. Extensive reading has been proposed as a key learning strategy by which learners can make progress toward advanced proficiency in general and/or academic English (e.g., Nation, 2014). However, the dilemma inherent in matching learners to texts becomes more apparent when research that has examined the lexical difficulty thresholds is considered. These studies have converged toward the recommendation that at least 95% of the words in instructional texts be already known to the
Pinchbeck: Validating the construct of readability in EFL contexts

1.1 Readability formulas

Automatic readability formulas have been used for decades by publishers and practitioners in K-12 education. Theoretically, they allow candidate texts to be evaluated for the difficulty they will likely present to learners at different levels. Although commonly used formulas such as the Flesch–Kincaid grade level (Flesch, 1948; Kincaid et al., 1975) – which are packaged with MS-Word – have been shown to correlate with cloze test scores, the effect size of this relationship in EFL contexts varies dramatically depending on the study ($r = 0.48$ in Brown, 1998, and $r = 0.85$ in Greenfield, 1999). Even in K-12 contexts where they were originally conceived, they have been criticized for their lack of construct validity (see below), as they are based on statistical curve fitting, rather than on a theory of reading and/or language development. This problem becomes more obvious when the accepted difficulty of certain types of texts is not accurately predicted by this formula; for example, Shakespeare’s *Hamlet* is rated by Flesch–Kincaid analysis as a text appropriate for children in Grade 2. Because these formulas measure only the surface features of texts, rather than the underlying factors that determine text difficulty, they also fail in providing writers any accurate guidance for rewriting texts to be more accessible to readers at lower levels of language proficiency. These types of criticisms of readability formulas have been common in the K-12 reading research literature (Bruce, Rubin, & Starr, 1981; Connatser, 1999; Davison & Kantor, 1982; Duffy, 1985; Lange, 1982; Maxwell, 1978; Selzer, 1981), and in the L2 literature, where several authors have called for the development of more accurate readability formulas that have been optimized for specific L2 contexts (e.g., Carrell, 1987; Greenfield, 1999).

1.2 Readability in Japan

There are a range of approaches to determine text-level appropriateness to EFL practitioners in Japan. Some studies that have examined readability have used surveys that ask students or teachers to rate the difficulty of texts (e.g., Holster, Lake, & Pellowe, 2017). There are also professional practitioner organizations that provide reading list suggestions for learners at different grade levels and/or proficiencies. For example, the *Yomiyasusa Levels* (YL) are published biennially as a list of readability scores for English graded and leveled readers in Japan (Furukawa, 2007). This index depends on the interpretation of factors such as, “illustrations, the backgrounds of the books, the size of the fonts, and different text styles.” The details of how such interpretations are operationalized have not been made transparent. Similarly, a second readability system, *The Kyoto Scale* (MReader, 2018), uses a combination of the “headword” numbers that are provided by publishers and
information in the YL levels. The practice of using publisher’s headword levels is problematic because there are inconsistencies in how headword levels are derived by different publishers (Claridge, 2012). To my knowledge, such readability systems have not been validated directly against the reading difficulty of texts. Furthermore, none of the above approaches provides an individual learner, teacher, or librarian with an ability to choose language-level-appropriate reading materials that have not already been vetted by outside organizations, which precludes the evaluation of the growing multitude of texts that are available for free on the Internet.

1.2.1 Direct measurement of text difficulty

The use of traditional reading comprehension tests to measure text difficulty was recognized as problematic very early on (Lorge, 1939) because the difficulty of the test questions and/or distractors confound the measurement of the difficulty of the text per se (for a recent review of this issue, see Cunningham, Hiebert, & Mesmer, 2018). For this reason, cloze tests have been used as the criterion variable for many readability formula studies in academia and in the publishing industry since the 1960s. Greenfield (1999) tested whether readability formulas that had been created for English-as-a-first-language children in the USA were also valid for adult EFL learners in Japan. He administered cloze tests to Japanese university students and then used the mean test scores as a criterion variable for evaluating different readability formulas. He found that the classic formulas did not correlate as well with Japanese readability scores (ranges of $r = 0.691$ to 0.861) as they did in U.S. K-12 contexts (Bormuth, 1971). He concluded that readability in Japanese EFL contexts should be treated as an overlapping, but unique construct.

1.2.2 The lexical component of text difficulty

Among the readability formulas tested in the Greenfield’s (1999) study, only the New Dale–Chall readability formula (Chall & Dale, 1995) included a lexical component, which is calculated as the percentage of words not found on a list of 3,000 words known by 80% of U.S. grade four students. When Greenfield developed his own readability formula, the Miyazaki EFL Readability Index, he included the surface-level features of letters-per-word and words-per-sentence. While word and sentence length can be considered as proxies for lexical decoding and syntactic parsing, newer corpus-based and natural-language computational technologies allow the measurement of other factors that might represent such cognitive processes more closely (Baayen, Milin, & Ramscar, 2016). In a more recent study, Crossley, Greenfield, and McNamara (2008) used Greenfield’s original cloze score data to develop models of readability using the psycholinguistic indexes available in the Coh-Metrix package (Graesser, McNamara, & Kulikowich, 2011). The Coh-Metrix analysis package has several word frequency indexes, all based on the Centre for Lexical Information (CELEX) corpus (Baayen, Piepenbrock, & Gulikers, 1995). In their final model, the CELEX written word frequency index was shown to correlate $r = 0.61$ with Greenfield’s cloze test scores, and the combined readability model in this study, which included two additional text-indexes, was able to explain 84% of the variance (adjusted $R^2$).
2 The problem

The readability formulas that do exist for Japanese learners are lacking in several ways. Firstly, readability frameworks currently used in Japan depend on the ratings provided by publishers or by reading organizations. The methods by which publishers determine text reading levels have not been standardized or made available for scientific scrutiny. Similarly, other readability lists, such as the Yomiyasusa, are only published every 2 years and are based on an unpublished evaluation procedure, the reliability of which has not been determined. Secondly, to my knowledge, the few studies that have included a lexical component of readability for Japanese learners have not examined word-lists derived from different source corpora, registers and modalities for the purposes of optimization. Thirdly, commercially available readability frameworks used for the K-12 in the United States (e.g., the Lexile Framework; Stenner, Smith, & Burdick, 1983) are expensive, and they may not be valid for EFL learners. Finally, none of the available methods provides any diagnostic information that might be useful for teachers, learners, and/or material developers.

2.1 Proposal for readability criteria

I propose a set of criteria that the next generation of readability tools should meet. They should (1) be transparent in how they calculate a readability index for a text, (2) be available for use by teachers and learners at minimal or no cost, (3) be easy to use, (4) provide diagnostic information that is of pedagogical benefit to learners, teachers, librarians, and other stakeholders that would allow texts to be reliably rewritten to different readability levels, and (5) be empirically validated on the target population of learners’ reading abilities.

One way that readability formulas could be developed toward satisfying these criteria would be to create a scale of linguistic forms (lexis, syntax, morphology, etc.) that could be used not only for text evaluation, but which could also be used to develop diagnostic language tests. In this way, learners could be matched to texts using a common scale of linguistic difficulty. I will now summarize one study that has examined how the lexical component of such a scale might be optimized for Japanese learners.

2.2 Call for a re-examination of the lexical component of readability

In an exploratory study (Pinchbeck, manuscript in preparation), I operationalized readability as the average cloze test scores of Japanese learners (n = 200) using previously published data (Bormuth, 1971; Crossley, Greenfield, & McNamara, 2008; Greenfield, 1999). Using Spearman correlations to relate the readability predicted by 26 different lexical indexes with the actual text difficulty as seen by the Japanese students, it was observed that word frequency/distribution rank-lists derived from narrative-texts, conversational-texts, and from TV/movie-captions were the best predictors of text readability (e.g., ρ = 0.81) as compared to written informational texts (ρ < 0.5). Using a Hotelling-Williams test for correlation differences (Steiger, 1980), the word list ranks derived from conversational and/or narrative
registers of English were all significantly better predictors of readability than were published word lists based on general corpora of English, all of which have been promoted as general service lists of English (Brezina & Gablasova, 2013; Browne, Culligan, & Philips, 2013; Nation, 2012). These results are also consistent with a previous study that used vocabulary test-item difficulty as the criterion variable to compare word frequency lists (Pinchbeck, McLean, Brown, & Kramer, 2016).

An additional finding of this study was that the size of the corpus from which word frequency/distribution lists were derived was not found to be an important factor in predicting readability. An analogous finding was also reported by Brysbaert and New (2009), who examined word lists for their ability to predict lexical decision reaction times. They found that increasing the size of corpus samples of the British National Corpus (BNC) beyond 16 million tokens did not result in higher correlations with reaction times.

3 Discussion and Conclusions

The primary goal of this article was to illustrate how pedagogical instruments, such as readability formulas, might be developed, optimized, and validated for EFL learner populations. Many previous studies in readability have promoted cloze tests as a method that can be used to better estimate directly the difficulty of a candidate texts for a given learner population (e.g., Cunningham, Hiebert, & Mesmer, 2018).

The BNC, the Corpus of Contemporary American English (COCA), Ententen12, or the Cambridge English Corpus (CEC) are all large, general corpora of English that have been promoted as being representative of the English language. For these reasons, they have all been used by their proponents in the development of general-purpose word lists (Brezina & Gablasova, 2013; Browne, Culligan, & Philips, 2013; Nation, 2012), each of which have then been used as the basis for the creation of a number of other pedagogical tools such as specialized word lists or vocabulary tests. While the general principles that have driven these developments are sound, empirical validation of the pedagogical tools has been lagging, particularly with respect to how well these tools might work with different learner populations.

The results of the unpublished readability study cited here might be better interpreted in the light of Biber’s work (e.g., 1993), who first challenged the notion that a single corpus that has been deliberately “balanced” can truly be used to represent all registers of language. His work and that of his students have detailed the multitude of linguistic differences that exist between different registers and modalities of English. Through this lens, it seems arbitrary that when attempting to measure readability in Japanese EFL contexts, we should choose to use word lists derived from source corpora that includes U.S. or U.K. newspapers and/or academic research articles, which are unlikely sources of language input for Japanese learners. While it is beyond the scope of this article to speculate on the possible psycholinguistic reasons why corpora of fiction or of captions of TV/movies for children might better represent the lexical knowledge of Japanese learners, it nevertheless raises the question of whether a single index of lexical difficulty should be used to fit all pedagogical purposes, for all learner populations, and in all teaching contexts.
The necessary inclusion of a lexical component in any automatic readability system has several benefits, as follows. Firstly, word lists “provide a rational basis for making sure that learners get the best return for their vocabulary learning effort” (Nation & Waring, 1997, p.17). Word lists, based on corpus frequency, or some other psycholinguistic index, can be used as a proxy for a scale of lexical difficulty, which means that word lists can be used as strong predictors of text difficulty. Alternatively, such word lists can also inform how lexical items might be chosen from existing course materials for targeted instruction because they assist developers to avoid teaching words that learners already know. Additionally, word lists also allow materials developers to simplify texts using a principled approach that is likely to produce texts that can be used for a wider range of learner-proficiencies. These types of text manipulations are not possible when only surface textual features are modified. For example, text length (i.e., tokens) was reported to be the major factor of readability in a recent study of Japanese students (Holster, Lake, & Pellowe, 2017). However, we cannot purposefully make the text of a book easier to read by cutting a book in half; as the first half of a typical book is just as difficult to read as the second half. Finally, word lists also represent how useful individual lexical items are likely to be to learners, inside or outside the language classroom. When word lists are included as terms in readability formulas, it allows language programs and/or materials developers to better match learners to texts because the same lexical scales of difficulty that are used for all of the above purposes can also be used to create diagnostic vocabulary tests that measure learner knowledge.

Automatic readability programs are limited because they examine texts separately from their contexts of use and from a complex range of factors related to individual learners, such as learner interest, values, and motivation. Therefore, the experience and intuitions of developers, librarians, teachers, and learners will continue to be necessary in text selection decisions. The next generation of automatic readability technologies will show their value by allowing teachers and materials developers to more efficiently access and evaluate the wealth of materials increasingly available on the Internet. A range of classroom- or learner-suitable texts that match both the thematic content of learning modules and the lexical knowledge of learners might be identified much more efficiently than that which is currently possible. The readability frameworks that are currently in use in Japan would make this a very slow process indeed. The Yomiyasusa readability lists are published only on a biennial basis and rely on the judgment of a committee. Similarly, The Kyoto Scale readability depends on publishers to level their own products and to make that information public. Neither of these approaches allows users to quickly determine the appropriateness of new candidate texts for their own learners.

Using computers, it is not currently possible for any practitioners to easily and accurately analyze the deeper structural and semantic features of texts (i.e., dialect; rhetorical structure; topic background, specificity, and complexity; cohesive devices; allusion; metaphor; the level of inferencing required of the reader; and/or formulaic language; etc.). In order for these textual features to be included in automatic readability estimation, we must await further advances in natural language processing technologies. In the meantime, however, I propose
that a re-examination of the lexical component of readability is warranted, particularly in Japan and in other EFL teaching contexts. It is my hope that this article might generate some interest toward the development of large-scale validity studies that would develop new pedagogical tools for more valid and efficient matching of Japanese K-12 and/or adult EFL learners to English texts.

References


Abstract

The concept of data-driven learning (DDL) – direct student use of corpora – has been gaining attention among researchers and teachers. DDL gives learners the chance to take an inductive approach to learning by recognizing patterns in corpus data rather than following abstract rules. However, very few studies focus on presenting clear practices that other teachers can readily use in their classes, particularly related to written error correction. This paper begins by reviewing research pertaining to DDL in student writing as well as using Japanese-English parallel corpora. It then outlines a recent study by Jenifer Larson-Hall (2015), notes weaknesses in the study, and outlines a current project (McGuire & Larson-Hall, under preparation).

Key words: Data-driven learning; written error correction; parallel corpora; corpus linguistics.

1. Introduction and Background

Linguists and language educators have been using corpora for many years to gain deeper insights into the usage of language and to create more authentic learning materials. The ability to measure word and collocation frequency in massive collections of authentic text has led to more objective perceptions of language in use and the ability to reevaluate language instruction.

The concept of data-driven learning (DDL) – direct student use of corpora – has been gaining attention among researchers and teachers. DDL posits “that the language-learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data” (Johns, 1991, p. 2). The corpus becomes an environment for the learner to explore independently through their own inquisitiveness. A corpus provides students with extensive context, which can be used for discovery-based learning. Rather than just looking up the meaning of a word in a dictionary, students can use a corpus to see how common a word is, how it is used in sentences, what words frequently appear alongside it, whether it is used more often in spoken or written language, and a wealth of other useful information that cannot be provided by a rule-based learning approach. DDL gives learners the chance to take an inductive approach to learning by recognizing patterns in language rather than following abstract rules that have been handed to them.
One of the central elements of DDL that students will engage with is the Key Words in Context (KWIC) display format. This formatting aligns numerous in-line occurrences of a key word down the center of the page with sentence context on either side. This allows users to easily see the words that come immediately before and after the key word in many examples on the same page. Figure 1 shows KWIC results for the word *budge* from the Corpus of Contemporary American English (COCA). This particular search is focusing on the words that come before the key word, and it quickly reveals that the word *to* appears multiple times immediately before *budge*. Looking further to the left shows different forms of *refuse*, suggesting that *refuse to budge* is a recurrent collocation. DDL methods require students to find patterns in language on their own, and the KWIC format is particularly helpful at making such patterns more visible. This requires students to spend time analyzing search results and looking at example sentences, but that process is one of the focuses of DDL. “DDL may help learner development if it leads, for example, to increased language sensitivity, noticing, induction, and ability to work with authentic data” (Boulton & Cobb, 2017, p. 2).

One useful feature of KWIC format is that it works well as a printout for paper-based DDL. Several studies have shown that tasks involving teacher-prepared KWIC printouts help acquaint students with analyzing corpus search results (Allan, 2006; Boulton, 2008b, 2009b; Chambers, 2007; Tian, 2005). A few paper-based DDL activities could help prepare students for the real thing. In fact, strictly paper-based DDL activities have been shown to be just as effective as computer-based DDL or a combination of the two in TOEIC vocabulary acquisition (Chujo, Anthony, Oghigian, & Uchibori, 2012), so teachers who do not have access to computers in the classroom or do not wish to use them constantly can still employ DDL. However, there are some crucial differences between paper-based DDL and computer-based DDL. First, the preparation of paper-based DDL activities has been known to take a great deal of time (Boulton, 2008b; Johns, 1991), which may prevent many teachers from adopting it in their classes. Searching for, analyzing, and selecting appropriate KWIC concordance

Figure 1. KWIC results for the word “budge” from the Corpus of Contemporary American English.

Vocabulary Learning and Instruction, 8 (1), 17–28.
lines that clearly illustrate a specific pattern is no easy task. Instead, giving students the opportunity to freely examine the data themselves can be beneficial. Paper-based DDL deprives the learner of the explorative component of direct corpus use. Many times finding patterns requires multiple searches (Gaskell & Cobb, 2004), requiring students to test and refine search terms on their own. Through this practice, students can better build the metalinguistic awareness needed for effective DDL (Yang, Wong, & Yeh, 2013).

Both computer-based and paper-based DDL have been shown to be effective and efficient by two empirical meta-studies. These two meta-analyses used Cohen’s $d$ to measure effect size as the difference between either control and experimental groups or pre/posttest scores from the collected studies. Then, a mean Cohen’s $d$ was calculated for all of the studies combined. Boulton and Cobb’s extensive 2017 meta-analysis looked at 88 effect sizes from different DDL studies from 1989 to 2014 and found significant overall effects ($d = 0.95$ for control/experimental studies and $d = 1.50$ for pre/posttest studies). Mizumoto and Chujo’s meta-analysis of 32 effect sizes from DDL studies in Japanese EFL classrooms (2015) also returned positive results ($d = 0.99$). Both of these meta-studies indicated that DDL shows the greatest effects in learning lexicogrammar, or “how the target word fits into its surroundings” (Boulton & Cobb, 2017, p. 33), usually through collocation and colligation.

Data-driven learning can be applied in many different ways, but the majority of studies deal with inductive noticing from either paper-based or computer-based KWIC results. While most studies do not concentrate on practical classroom application, Chujo and Oghigian (2008, p. 5) lay out a series of clear DDL exercises which illustrate how students can use a corpus to learn lexicogrammar. In their study, students explored the KWIC results of several instructor-provided search terms in order to create hypotheses on particular patterns or rules. One such search term was the single word quality. Students were asked to note the words that appear before and after it in order to induct basic noun phrase structure. The search reveals a common pattern: an article before quality and a prepositional phrase after, as in the quality of education. Other searches included: enjoyed to find what often follows (gerunds rather than infinitives), discuss to find wh-clauses (discuss when), furniture* and passenger* to distinguish countable and uncountable nouns, and several more. Once students had been given time to explore and come up with their own hypotheses, they were given explicit explanations of the rules in order to verify their ideas. After this, students were given follow-up worksheets where they needed to apply the rules in language production. In this manner, Chujo and Oghigian applied a combination of inductive and deductive approaches which proved effective.

Other studies have examined an inductive approach to many different language features. Smart’s (2012) students used corpus data to study a variety of grammar features: subject–verb agreement, phrasal verbs, active and passive voice, relative clauses, and finite compliment clauses. Other studies have looked at learning collocations (Alshaar & AbuSeileek, 2013; Ashouri, Arjmandi, & Rahimi, 2014; Chan & Liou, 2005; Chen, 2011; Sun & Wang, 2003), phrasal verbs (Boulton, 2008b), linking adverbials (Boulton, 2009b; Cotos, 2010, 2014), and prepositions (Hadi & Alibakhshi, 2012; Someya, 2000). Gaskell & Cobb (2004) found that providing teacher-selected KWIC examples as feedback for particular kinds of written errors helped students to correct them in revisions.
Johns (1991) initially imagined DDL as a tool for advanced learners, but a number of studies have shown it to be beneficial to lower level learners as well (Boulton, 2008a, 2009b; Chujo, Anthony, Oghigian, & Yokota, 2013; Chujo & Oghigian, 2012; Cobb, 1997, 1999a, 1999b). There is some debate as to how much training, if any, is necessary before students can make use of concordance data. Some studies assert that extensive training is essential before using a corpus (Breyer, 2006; Gaskell & Cobb, 2004; Turnbull & Burston, 1998; Yang et al., 2013), while others have found evidence that students can benefit from DDL with little training (Boulton & Wilhelm, 2006) or with no training at all (Boulton, 2009b). The key challenges that are revealed from the discussion of training are (a) if students can switch from a deductive to inductive approach and (b) if students can make sophisticated use of a corpus autonomously. The hope is that with enough hands-on practice, students can do both, so emphasis could be placed on independent practice with feedback rather than guided training.

Despite all of this research on DDL, very few studies focus on presenting clear practices that other teachers can readily use in their classes. The majority of the studies mentioned above used offline corpora with concordance software and the article discussions concentrated on the results of pre/posttests rather than the instruction and activities involved. So, while the results of studies are often encouraging, the procedures cannot be easily replicated by other teachers. It is clear that for DDL to be more widely accepted as a viable methodology, it needs to move from theory to more accessible practice.

### 1.1 Data-Driven Learning with Parallel Corpora

Data-driven learning is a promising approach to language learning, but it is easy to see some of the challenges that might be encountered. Learning is based on student autonomy and motivation. Students are put “at the center of the process, taking increased responsibility for their own learning rather than being taught rules in a more passive mode” (Boulton, 2009a, p. 82). Under DDL, the roles of both the teacher and student shift, which can be problematic. Students must take a more active role, and teachers must step back from controlling the learning process. On the technical side, students have reported that they find concordance software difficult to use (Farr, 2008) and the monolingual concordance results difficult to understand (Aston, 2001; Tian & Liu, 2004).

A parallel corpus (a bilingual corpus built from translations between two languages that displays both simultaneously) may help to alleviate some of the challenges of using DDL in the EFL classroom. When a student searches a parallel corpus, the results are given in both languages, giving students the opportunity to analyze word usage in English and compare it to their native language. A number of studies have looked at using Japanese-English parallel corpora in the Japanese EFL classroom. Students responded positively and found a parallel corpus useful (Chujo, Utiyama, & Miura, 2006), and students in a DDL treatment group made greater improvements in learning TOEIC vocabulary and grammar than a non-DDL control group (Chujo, Anthony, & Oghigian, 2009).
While these studies are promising, challenges remain. McGuire (2018) argues that students who are not familiar with a parallel corpus may initially treat it like a bilingual dictionary. They use it to look for a quick one-to-one word replacement rather than looking at the usage and context. Many students expect a quick “answer” from their electronic dictionaries (and more recently machine translation websites and apps) and thus may treat a corpus in much the same way. Some students simply use the first result they see, which can create bizarre word-choice problems. Another impediment is that strict grammar-translation teaching is still prevalent in Japan, even at universities. Many students are required to remember exact sentence translations and reproduce them accurately. This trains them to treat language translation as mechanical and prevents them from “developing the mind-set for the deeper syntactic analysis of language needed to use a parallel corpus effectively” (McGuire, 2018, p. 257).

There has up to now been a paucity of research regarding the use of a parallel corpus for error correction in student writing. Chujo, Utiyama and Miura’s 2006 study asked participants at one point to attempt to translate Japanese phrases into English using a parallel corpus. This study was only looking at student reactions to using the corpus in this manner and did not try to quantifiably measure if students could successfully make the translations. Yang, Wong, and Yeh (2013) had participants look at a Chinese sentence and then rearrange scrambled English words into the “correct” translation as taken from a Chinese-English parallel corpus. This did not involve any original writing on the part of the students, and the focus of the study was on developing students’ metalinguistic awareness rather than correcting written errors.

In her 2015 study, Larson-Hall asked if Japanese university EFL students could identify and correct written mistakes by using a Japanese-English parallel corpus. In this study, the parallel corpus was treated as a reference for students to use in place of a bilingual dictionary. As a pretest, students in academic writing classes were asked to correct a flawed English translation of a Japanese paragraph written by an anonymous Japanese student. The translation contained many grammatical and lexical mistakes, and the students were asked to correct any problems they found. None of the mistakes were marked or identified, but the students were allowed to use dictionaries or other resources. Following this, the DDL focus group was given two 90-min training sessions to use the parallel corpus to correct written mistakes. This training involved making corrections to teacher-led worksheets of erroneous English translations of Japanese sentences. The control group was given no training or in-class access to the parallel corpus. Over 12 weeks, both groups continued their regular studies in their academic writing classes. The DDL focus group was encouraged to use the parallel corpus over the course of the semester to help with their writing. At the end of the 12 weeks, all students were given a posttest identical to the pretest. They were again asked to correct any mistakes that they found, and all students were allowed to see what they wrote on their pretests. Pretests and posttests were analyzed to see if the students were able to improve in error identification and correction over the course of the 12 weeks. The results were mixed and inconclusive. Some students in the treatment group were able to improve, while others were not.
While the questions asked in Larson-Hall’s 2015 research are very important, a few issues call for a new revised study:

First, error identification was left to the students and played a very central part. It seems obvious that foreign language students have difficulty identifying mistakes in their target language, especially in their own writing. Many of the students in both the focus and control groups made corrections where none were needed or made corrections in the wrong place. Students reported that they could sense that something was wrong with a sentence, but they could not specifically identify the error. A new study should not make error identification such a key component, but rather test if students can sufficiently correct instructor-identified errors.

Additionally, in order to have a standardized pre/posttest, students were required to correct the translation mistakes of another student. The idea was that the students would be able to read and understand the author’s original ideas in Japanese and then find mistakes in the English translation. However, several students who participated in the study reported that they didn’t know what the author wanted to say. Since the writing was not their own, it is understandable that they felt removed from the ideas. This issue makes using a standardized pre/posttest difficult. A new study would need to look directly at the mistakes in each participant’s compositions over the course of the study period. By doing so, changes in error rate and improvements in error correction can be measured for each student.

Finally, the DDL focus group classes were not conducted in a computer lab after the corpus training was complete. Because of this, students could not use the parallel corpus while they were writing in class or receive direct feedback on their search skills. Students were asked to mark areas in their compositions where they consulted the corpus, but there was no way to judge how much time each student actually spent doing hands-on DDL. The DDL focus group in a new study should have constant access to the parallel corpus and be encouraged to use it in all stages of composition.

1.2 Planned Methodology
To address the above-mentioned issues with Larson-Hall (2015), a new study will need to be conducted.

To fix the problem of student error identification, an indirect error tagging system will be introduced. Rather than the common practice of explicitly correcting mistakes with a red pen, instructors in the study will use highlighters to mark the words involved in an error. Different colored highlighters will be used to differentiate grammatical mistakes from problems with word choice. Once the instructor has tagged mistakes, the students will then be given time to use the parallel corpus to try to correct them. After attempting correction, the instructor can judge if the student has adequately fixed the mistake or if more work is needed. This process could be repeated on multiple drafts as necessary.

This system could easily be used by any teacher and will not be as time-consuming as explicit correction. Additionally, this error tagging system will give students more focused and personal practice using the parallel corpus, which
should then lead to more effective utilization of it. As students improve at using the parallel corpus, they will likely use it more effectively for their subsequent writing assignments as well. The new study will measure whether or not this system leads to improved error correction and ultimately to a decrease in the number of mistakes made in student writing submissions.

1.3 Eijiro on the WEB Parallel Corpus
The parallel corpus utilized in Larson-Hall’s 2015 study deserves more attention, as it does not seem to appear in any other studies thus far. The parallel corpus is publishing company ALC’s online implementation of the Eijiro (英辞郎) corpus, from here on referred to as Eijiro on the WEB. It is developed and constantly updated by the Electronic Dictionary Project (EDP) and currently contains over 2 million lemma entries from over 700,000 bilingual newspaper texts. The PRO version which will be discussed here contains over 1 million aligned parallel sentences. It is one of the most accessible Japanese-English parallel corpora for the following reasons:

First, Eijiro on the WEB is accessible online and does not require additional concordance software. It can be accessed by computer or smartphone, making it usable in a regular classroom without a computer for each student. As long as students have smartphones they can access the corpus.

Second, it is a general purpose corpus compiled from translated newspaper articles and is not overly technical or field specific. Other larger English-Japanese parallel corpora exist, but they have a limited range of sources such as the Asian Scientific Paper Excerpt Corpus (ASPEC) which is comprised of parallel scientific abstracts, the Japanese-English Subtitle Corpus (JESC) which is a collection of parallel film subtitles, or the Kyoto Free Translation Task (KFTT) created from parallel Wikipedia articles about Kyoto. Chujo et al. (2006, 2009, 2012) used the Japanese-English News Article Alignment Data (Utiyama & Isahara, 2003), which is larger than the Eijiro corpus. However, all of these corpora are just raw data without any web interface or search engine and require separate concordance software to use. Because the accessible web interface is easy for students to use, McGuire and Larson-Hall have always chosen to work with the Eijiro corpus.

Third, Eijiro on the WEB’s search engine is familiar, straightforward, and easy to use. It is a simple search bar; and users can search in English, Japanese, or even both at the same time without adjusting settings. Granted, it is not as powerful as a full feature corpus interface like Mark Davies Corpus of Contemporary American English (2008-), but it is much more approachable for people who are unfamiliar with corpus linguistics. ALC’s search engine offers some more advanced syntax search abilities and includes bilingual KWIC search results (see Figure 2) and collocation frequency charts (see Figure 3).

Finally, it is not particularly expensive to use. The PRO version costs 324 yen per month or 3564 yen for a yearly subscription. However, the first month is free, which makes it viable for short-term use with large numbers of students. The basic version that only contains word and phrase translations is always free to use, but it does not include all of the sentence data. The basic version is more of a bilingual phrase dictionary than a corpus, but it can also be useful for students.
Figure 2. KWIC search results for “disruption” in ALC’s Eijiro on the WEB PRO. Entries are expandable to show Japanese translation and source.

Figure 3. Collocation frequency chart for “align” in Eijiro on the WEB PRO.

Vocabulary Learning and Instruction, 8 (1), 17–28.
The combination of a parallel corpus and bilingual dictionary on the same site makes it more versatile.

Ultimately, wider adoption of DDL practices will require more accessible tools, and more studies will be needed to empirically show the benefits of such tools. As mentioned before, most DDL studies up until now use raw corpus data in concordance software, which greatly restricts where and when students can engage with the data. *Eijiro on the WEB* is a viable option for EFL teachers in Japan who want to bring DDL to their classrooms and for students who want to directly explore concordance data.

References


**Websites**

Asian Scientific Paper Excerpt Corpus (ASPEC) - http://orchid.kuee.kyoto-u.ac.jp/ASPEC/

Eijiro on the WEB PRO - https://eowp.alc.co.jp/

Eijiro on the WEB - https://www.alc.co.jp/

Japanese-English News Article Alignment Data (JENAAD) - http://www2.nict.go.jp/astrec-att/member/mutiyama/jea/index.html

Japanese-English Subtitle Corpus (JESC) - https://nlp.stanford.edu/projects/jesc/

Kyoto Free Translation Task (KFTT) - http://www.phontron.com/kftt/
Evaluating Corpora with Word Lists and Word Difficulty

Brent A. Culligan
Aoyama Gakuin Women's Junior College
https://doi.org/10.7820/vli.v08.1.Culligan

Abstract
This study examines the application of an IRT analysis of words on lists including the General Service List (GSL), New General Service List (NGSL), Academic Word List (AWL), New Academic Word List (NAWL), and TOEIC Service List (TSL). By comparing line graphs, density distribution graphs, and boxplots for the average difficulty of each word list to related lists, we can get a visualization of the data’s distribution. Japanese EFL students responded to one or more of 84 Yes/No test forms compiled from 5,880 unique real words and 2,520 nonwords. The real words were analyzed using Winsteps (Linacre, 2005) resulting in IRT estimates for each word. By summing the difficulties of each word, we can calculate the average difficulty of each word list which can then be used to rank the lists. In effect, the process supports the concurrent validity of the lists. The analysis indicates the word family approach results in more difficult word lists. The mean difficulties of the GSL and the BNC COCA appear to be more divergent and more difficult particularly over the first 4000 words, possibly due to the use of Bauer and Nation’s (1993) Affix Level 6 definition for their compilation. Finally, just as we should expect word lists for beginners to have higher frequency words than subsequent lists, we should also expect them to be easier with more words known to learners. This can be seen with the gradual but marked difference between the different word lists of the NGSL and its supplemental SPs.

Key words: IRT, word difficulty, corpus validity, measurement, vocabulary testing, Yes/No test

1. Measuring Word Difficulty Using IRT

At present, while there is a healthy debate on what constitutes a word for counting purposes, there are few ways to describe and evaluate the products of corpus analysis, whether their consistency or difficulty. In this study, I will look at one method to compare word lists derived from frequency analysis of corpora using word difficulty. Under this approach, word difficulty is measured using item response theory (IRT) applied to responses on Yes/No tests. IRT is a statistical model that estimates the probability of a person getting an item correct. More formally, IRT attempts to describe and predict the relationship between a random person and a random item, specifically how the person will respond to the item, given the ability of the person and the difficulty of the item. IRT posits that the
higher the ability of the person, the higher the probability the person will respond correctly to the item. Conversely, the more difficult the item, the less likely the person is to correctly respond. For example, we would not expect a beginning English as a Foreign Language (EFL) student to know the word residue, but we would be very surprised if an advanced learner did not know the word room. If we know both the difficulty of the item and the ability of the person, we can predict the probability of the person correctly responding to a given item. IRT values are measured in logits, which can range from $-\infty$ to $+\infty$ but are more generally constrained to $-7$ to $+7$ for vocabulary difficulty. For word difficulties, the higher the logit, the more difficult a word is. For example, room at $-6.43$ is far easier than elicit at $5.38$.

By summing the difficulties of each word, we are able to find the average difficulty of the word list. In this way, we can assess and rank word lists by difficulty. If we know the ability of a student, we will be able to ascertain how many of the words will be known to the student much more accurately than by testing a stratified random sample of the words in the list. For example, Nation’s Vocabulary Size Test samples one word from 100, so to estimate the number of words known for each thousand-word band, they test 10 items. However, by using IRT estimates for word difficulty, we can get a better estimate of how many words are known by summing the probabilities of each word being known on the word list.

The purpose of this study is the application of the findings from the IRT analysis of the large data set to the assessment of the difficulty of a word list. The research question is to investigate what the average IRT difficulty estimates for the words on the General Service List (GSL), New General Service List (NGSL), Academic Word List (AWL), New Academic Word List (NAWL), and TOEIC Service List (TSL) can tell us about the use and validity of these lists?

2. Method
2.1. Participants

In this study, over 1,200 students from various schools responded to one or more of 84 different Yes/No test forms. Participants were primarily first and second year students enrolled in either a 2-year women’s junior college or 4-year coed universities in Japan.

2.2. Materials
2.2.1. Yes/No tests

Real words. The real words from this study were compiled from word lists that are commonly found in the ESL/EFL literature. First to be considered were those lists that can be seen as useful for general English education. These lists were derived from corpora that tried to incorporate a diversity of genres and modalities. The first important list was West’s (1953) A General Service List of English Words. This list of approximately 2,000 words has seen many revisions. As West claimed that “no attempt has been made to be rigidly consistent in the method
used for displaying the words: each word has been treated as a separate problem, and the sole aim has been clearness” (p. viii), his definition of what to count as a word was not explicitly stated. However, from the examples given within the text, he includes the frequency counts of the word used in different parts of speech and its inflections. Thus, we have nouns, verbs, and verbals (which include participles used as nouns and adjectives) contributing to the final frequency counts. For example, in West’s entry for feel (p. 178–179), the total count for feel includes the verb (it feels soft) and the noun (the feel of silk), as well as the gerund feeling and its plural feelings (you’ve hurt my feelings). The version of the GSL used here was Bauman and Culligan’s (1995) adaptation using Bauer and Nation’s (1993) Word Family Level 3 affix definitions and ranked according to frequency counts from the Brown Corpus (Frances & Kučera, 1982). This resulted in a list of 2,284 words. Next examined was one of the updates to the GSL, the 2,801-word NGSL (Browne, Culligan, & Phillips, 2013a). The NGSL uses an extended lemma based on West’s GSL for its definition of what is a word. Three other lists were derived from the Brown Corpus, the British National Corpus (BNC), and the BNC_COCA. The BNC_COCA (Nation, 2012) was designed by Paul Nation and colleagues and is an amalgamation of his BNC list (Nation, 2006) and data from the Corpus of Contemporary American English (COCA). It comes with Nation’s Range software and is compiled into 1000-word bands using Bauer and Nation’s Affix Level 6 definition (1993). For the BNC (Leech, Rayson, & Wilson, 2001), all words with a frequency of ten or more were used.

Second to be considered are the special purpose lists. The first of these is the long-established AWL (Coxhead, 2000) and the newer 963-word NAWL (Browne, Culligan, & Phillips, 2013b), followed by an examination of two newer lists, the 1,754 words Business Service List (BSL) and the 1,259 words TSL (Browne & Culligan, 2016a, 2016b). The AWL uses Bauer and Nation’s (1993) Affix Level 6 definitions, while the NAWL, BSL, and TSL use the extended lemma. The list was then sorted, and all duplicates were eliminated, leaving 5,880 unique words.

Nonwords. The source of the 2,520 nonwords was the ARC database of nonwords (Rastle, Harrington, & Coltheart, 2002a, 2002b), which provided a list of nonwords with information on many characteristics including the number of neighbors. Nonwords that appeared to be inflected, as well as words that had too many neighbors with real words, were eliminated, thus providing an ideal source for nonwords.

2.2.2. Test forms

Each of the 84 test forms created for this study consisted of 70 real words and 30 nonwords. The 5,880 words and 2,520 nonwords were arranged in order of number of letters and randomly assigned to one of 120 tests. The words and nonwords on each form were then sorted by random number to determine their order on the test form. Each test form was assigned a unique test identification number (TestID) that was converted to a scanner-readable bar code. On the test, the 100 randomly ordered words and nonwords were arranged in four columns, with each word item directly followed by a bubble font version of a Y or an N.
2.3. Procedure

The procedure for administering the Yes/No tests was simple and efficient. The tests were given at the beginning of each class. The whole procedure for each test took less than 10 minutes. The number of test forms taken by each student ranged from one to five. The tests were staggered to ensure common items among the population.

The tests were scanned using the Remark Office OMR software (Remark Office OMR, 2000), resulting in a data file composed of a line for each test. Each line consisted of 102 comma-separated variables consisting of the student number, a TestID, and responses to 100 items, with Y for a Yes response, N for a No response, or an X if the item was omitted.

3. Analysis

To obtain estimates for difficulty based on IRT, the data were analyzed using Winsteps (Linacre, 2005). Before analysis, individual test forms were eliminated based on false alarms rates, as Meara (2010) suggests that proportions of 0.50 or above render a test unreliable. After individual test forms were eliminated, the resulting test data set was converted to where each line of data represented the students’ responses to all items on all test forms, and then analyzed by Winsteps. This resulted in IRT estimates for each of the tested words.

4. Results

Initial analysis of the data resulted in 5880-word difficulty estimates. A total of 554 words received perfect scores where all students who saw the word said they knew it, while 75 received zero scores where no students indicated knowledge. Winsteps uses a different algorithm that depends on the responses of the students who take similar items to compute the perfect and zero scores (Linacre, 2005).

The first research question looked for the average difficulty of various word lists. All data were reported in logits, the log of the odds unit. The lists used for general ESL and EFL programs were first compared. The GSL (West, 1953) had an overall mean of −1.86 logits (SD = 2.20), while the NGSL was slightly easier (M = −1.92, SD = 1.97) (see Figure 1). The violin plots in the figure show boxplots surrounded by the density distributions of the difficulties. The boxplots show the median, the 25 and 75 percent quantiles, and the minimum and maximum. The black dots represent extreme scores. The peak of the density distributions represents the mode, the most common score in the distribution. We can see that the distributions are relatively similar in the bottom half but differ toward the top. However, the graphic shows that the median score, the 25th percentile, and the 75th percentile of the GSL are very similar to the NGSL.

The first 1,000 words of the GSL had a mean word difficulty of −3.22 (SD = 1.56), while the second 1,000-word list had an average of −0.79 (SD = 2.03), a difference of 2.43 logits. The mean of the last 284 words was −0.83 (SD = 2.10), which is slightly easier than the previous thousand. For the same breakdown, the NGSL had means of −3.30, −1.59, and −0.62 (SD = 1.54, 1.68, and 1.71), respectively.
Graphically this is represented in Figure 2, where we can see the slight dip in the line graph for the GSL. Under the assumption that less frequent words tend to be more difficult, we would expect to see this line continue to incline upward given that these words are less frequent according to the frequency data from the BNC.
Next, we will look at the distribution of other lists derived from general corpora, specifically the Brown Corpus, the BNC, and the BNC_COCA. As can be seen in Figure 3, the BNC, Brown, and NGSL track fairly similarly, particularly over the first 4,000 words. However, the GSL and the BNC_COCA appear to be more divergent and more difficult. One possible explanation for the difference in overall difficulty is that the BNC_COCA uses Bauer and Nation’s (1993) Affix Level 6 definitions. By including derivations with high frequencies with their base form, they are effectively removed from the lists. This means that in each band, and compounded subsequently, more words with lower frequencies will be included. For example, government is not found on the BNC_COCA as it is part of the govern word family, but in the BNC, it is in the first thousand, while govern is found in the third thousand. This concentration of words has an effect on base words as well. The third thousand of the BNC and Brown include words such as rabbit and slight, which are found in the BNC_COCA first thousand. The BNC_COCA third thousand includes words like discriminate and affirm that appear on the BNC’s sixth and seventh thousand lists. For these reasons, we can see that how we define the lexical unit of counting has many ramifications for word lists extracted from corpus analysis.

Figure 3. Growth in Average Word Difficulty.
With the introduction of the University Word List (Xue & Nation, 1984) and its replacement with the AWL (Coxhead, 2000), we see the beginnings of the use of Special Purpose (SP) vocabulary lists to supplement the GSL. Many such lists proliferated, such as a Medical Academic Word List (Wang, Liang, & Ge, 2008), a nursing list (Yang, 2015), and an engineering list (Ward, 2009). The NAWL was designed to supplement the NGSL in the same way the AWL supplemented the GSL. Figure 4 shows the distribution of the original AWL \((M = -0.12, SD = 2.07)\) and more recent iteration, the NAWL \((M = 0.85, SD = 2.05)\). The mean score of the NAWL is almost one logit more difficult than the AWL. From the plots, we can also see from the length of the tails that there are more extreme scores in the NAWL, particularly for words over six logits in difficulty. We can also see that the median of the AWL is clearly below the median and mode of the NAWL, thus affirming that the majority of the words in the NAWL lie above the mean, median, and mode of the AWL.

To get a more complete picture, it is necessary to see how the special purpose list fits with the base word lists. In Figure 5, we can see density distributions of the word difficulties for the GSL and the AWL. For this graph, the second word list of the GSL contained 2,284 words. As can be seen from the violin plots, while there seems to be a good separation between the first and second thousand bands of the GSL, there is a considerable overlap with word difficulty between the GSL 2 \((M = -0.80, SD = 2.19)\) and the AWL \((M = -0.12, SD = 2.07)\), with the box plots showing similar medians and the density distributions displaying similar modes. Visually, they appear to be occupying similar space on the difficulty spectrum.

In Figure 6, the NGSL and its supplemental SP word lists show a gradual, but much more marked difference between the different lists when compared to

Figure 4. Violin Plots of Word Difficulty for Two Special Purpose Academic Word Lists.
Figure 5. We can see that both the medians and the mode for each distribution are below their neighbor in the majority of cases. There are fewer differences between the SPs, but this is to be expected as the SPs are not mutually exclusive.
However, we can see some differences between the related SPs, the TSL and the BSL, which both deal with business type discourse. As the TOEIC is a test of business communication, this corpus is more limited in scope than the Business Corpus that covers a wide range of situations, genres, and text types. This is reflected in the increased difficulty of the BSL compared to the TSL.

5. Conclusion

This study looked at the application of difficulty to word lists produced from corpora. By comparing line graphs, density distribution graphs, and boxplots for each word list with related lists, we can get a better picture of how the data are distributed. This allows us to assess the validity of our word lists as well as investigate how well they work in conjunction with each other. Just as we should expect word lists for beginning courses to have higher frequency words than subsequent lists, we should also expect them to be easier, with more words known to the learners. In effect, the process described in this article provides a way to support the concurrent validity of our word lists. Finally, this analysis seems to indicate that using a word family approach, particularly using word affixation beyond Level 2, will result in more difficult word lists as was evidenced by the increased difficulty at successive thousand-word bands of the BNC_COCA based on Bauer and Nation’s (1993) Affix Level 6 (see Figure 3). Similarly, the GSL based on Affix Level 4 showed the second thousand to be more difficult than those of the word lists using West’s extended lemma approach.

References


Commentary on Four Studies for JALT Vocabulary SIG

Yukio Tono
Advanced Research Unit, Graduate School of Global Studies,
Director of the World Language Center, Tokyo University of Foreign Studies
doi: https://doi.org/10.7820/vli.v08.1.Tono

Abstract
The four papers in this volume focused on mainly three areas: readability (Pinchbeck), wordlist evaluation (Ishikawa; Culligan), and data-driven learning (McGuire). The author would argue that whilst applied corpus linguistics and L2 vocabulary research are closely related to each other, there has not been much interaction between the two disciplines. The review of the four papers has also shown such lack of interaction between the two fields. By reviewing each study, some additional concepts and previous related studies in corpus linguistics will be presented in order to fill those gaps.

Key words: corpus linguistics/ wordlist/ data driven learning/ parallel corpus

Introduction
Corpus linguistics used to be a paradigm supported by a small number of empirical linguists in Europe who were against Chomsky’s mentalist approach (Sampson, 2005). Back in the 1960s, the sizes of corpora were tiny, around 1 million words at most, and only a large mainframe computer could process the data on magnetic tape using punch cards for inputting queries. Half a century has passed, and now the size of corpora available to general users can be twenty or thirty billion words, and one can handle a corpus of 100 million words in plain text files on a small laptop computer by using a user-friendly corpus query program. The Chomskyans’ negligence in relying on empirical data has given way to more usage-based views of language, which revived the use of a large body of linguistic data as the mirror of language input and output that shapes the organization of an individual’s or a language community’s linguistic competence and usage (Bybee, 2006; Gries & Divjak, 2012; Sampson, 2009).

Now most researchers in theoretical and applied linguistics believe that corpora are a useful source of data that, if used correctly, shed light on various aspects of languages. L2 vocabulary acquisition/learning research is no exception. Due to the strong influence of Chomskyian linguistics, many second language acquisition (SLA) researchers worked primarily on syntax until the 1980s, when a gradual shift of attention from syntax to lexicon began. It was interesting that such pioneers in L2 vocabulary acquisition/learning research as Paul Meara, Paul
Nation, and Batia Laufer first paid little attention to corpora, partly because it was still too early then to realize its power. It was not until the growing influence of a balanced, representative corpus such as the British National Corpus (BNC) at the beginning of the 21st century that many applied linguists started to appreciate the power of mega-corpora. Another reason might be that many L2 vocabulary researchers were interested in vocabulary teaching and learning techniques or strategies, which seemed to be somewhat removed from what corpora can tell us about languages (Schmitt & McCarthy, 1997). It was Nobert Schmitt that started to use corpora intensively for L2 vocabulary research (Schmitt, 2010). His experience with Ronald Carter and Michael McCarthy in the Nottingham school helped him a great deal to appreciate the power of corpora for investigating L2 vocabulary research.

Among the corpus linguistics community, applications of corpora in language learning/teaching research started in the beginning of the 1990s, when the participants of the ICAME Conference, which was one of the first conferences for those who developed the Brown Corpus and its family, discussed the possibility of holding a conference dedicated to teaching and language corpora (TALC). The first TALC conference was held at Lancaster University in 1994 (Wichmann et al., 1997). I attended the second TALC in 1996, where I met numerous fantastic scholars at Lancaster University, such as Geoffrey Leech, Tony McEnery, Nick Smith, and Paul Rayson. It was around that time that the BNC was released in 1994 and made available within European countries only. In 1995, as fruit of the first 100 million balanced corpora, two principal monolingual learner’s dictionaries, the Longman Dictionary of Contemporary English (LDOCE, the third edition) and the Oxford Advanced Learner’s Dictionary (OALD, the fifth edition), were revised based on the analyses of the BNC. Together with COBUILD, the impact of corpora was such that it changed the views of language researchers and teachers toward the importance of ‘big data’ in language pedagogy. That was also about the time when I started to compile a corpus of learner English. The founder of the International Corpus of Learner English (ICLE), Sylviane Granger at Catholic University of Louvain, also launched the project ICLE at the beginning of the 1990s (Granger, 1998). Learner corpus research has been multiplying since then and has contributed to our understanding of lexico-grammatical features of learner language. This is a field of studies closely linked to L2 vocabulary research, but L2 vocabulary and learner corpus researchers did not talk to each other very often until recently.

The same kind of lack of interaction between the two fields can be seen in the present collections of papers as well. I notice that in corpus linguistics, some of the questions addressed by L2 vocabulary researchers have been investigated in more detail and sometimes provide more concrete ideas and heuristics concerning methodologies. This kind of gap can be and needs to be filled. Therefore, it would indeed be an excellent opportunity for the two fields to meet and discuss various issues at this JALT Vocabulary SIG.

The four papers in this volume focused on mainly three areas: readability (Pinchbeck), wordlist evaluation (Ishikawa and Culligan), and data-driven learning (DDL) (McGuire). In the following sections, I will discuss the four studies regarding the issues of wordlist evaluation, readability, and DDL, respectively.
The four studies


Researchers have taken different approaches toward wordlist evaluation. There are two main types. One is an internal validation approach, which is to evaluate the wordlist internally by examining the validity of sampling and representativeness of the source data based on which the wordlist was created. Ishikawa (in this volume), for instance, examined two New General Service Lists (NGSLs) (Brezina & Gablasova, 2015; Browne, Culligan, & Phillips, 2013) and their source data in his review. The sources for the two wordlists happened to come from native speaker (NS) corpora only (see Table 2 in Ishikawa), and Ishikawa argued that there might be a question of whether a wordlist based solely on NSs’ usage is valid as a pedagogical list for language learners.

The other approach is an external validation approach, in which wordlists are compared against each other regarding the coverage of a particular text. For this, one could evaluate the quality of the wordlist by comparing the coverage of texts of different genres or text types against the wordlist (Chujo 2012; Chujo & Oghigian, 2009). One could also cross-compare different wordlists. A good example is Nation (2004), in which a comparison between the General Service List (GSL) and the BNC frequency list revealed the characteristics of each list (e.g. educational and dated for the GSL and the British nature of the BNC) due to their different text sources. Ishikawa used this method and compared the new JACET 8000 against two NGSLs. Culligan (in this volume) is unique in this sense because he uses new techniques (i.e., responses to Yes/No tests scaled using Item Response Theory (IRT)) to assess the contents of the wordlists.

Ishikawa described the history of the development of JACET 8000, arguing for the value of blending frequency information from corpora and intuition-based lists. Also, he questioned the validity of wordlists based only on NS corpora. In his case study, he compared JACET 8000 against two NGSLs and found that while there is much core vocabulary covered in the three lists, NGSLs tend to include words denoting “business and higher-order mental activities,” which are missing in JACET 8000. On the other hand, JACET 8000 contains words related to personal and daily life situations and morphologically much shorter words.

The results seem to show a difference in what source texts were used for wordlist creation: NS corpora only versus other sources. The two NGSLs mainly use NS corpora (cf. Table 2 in Ishikawa) while JACET 8000 was based on various materials relevant to English learners in Japan. The fundamental concept of corpus compilation has to be revisited here, where such concepts as the target population, balance, representativeness, and sampling are to be discussed carefully in order to fulfill the aims of target wordlist construction. The NGSLs aim to serve English learners in general, which is the main reason why NS corpora were used for word selection. However, it could be argued that the wordlist merely reflects the characteristics of vocabulary regarding words used most frequently by NSs, which may not be the same as the learning order of L2 vocabulary. Ishikawa argues that an educational wordlist should have a clear target user and use in mind,
to which I could not agree more. The GSL sounds good, but what is meant by “general service” has to be critically evaluated. In this sense, I hasten to add, Ishikawa’s argumentation about the definition of “vocabulary for Japanese learners of English” may need further scrutiny, regarding proficiency levels, situations of use, and target learners’ age among other things.

Related to Ishikawa’s paper, it should be noted that recently more dynamic models of wordlists have been proposed. For example, the English Vocabulary Profile (EVP; Capel, 2015), created based on analysis of the Cambridge Learner Corpus, shows the difficulty level information based on the Common European Framework of Reference for Languages (CEFR; A1 to C2) not only for each headword but also for each word sense, colloquial expressions, and phrasal verbs. While how to determine the CEFR levels of word meanings is still an empirical question, this wordlist sheds new light on the direction of pedagogical wordlist creation.

Another wordlist is the CEFR-J Wordlist (Tono, 2013). It contains 7500 words covering CEFR levels A1 to B2. This wordlist was based on the analysis of English textbooks used in China, Korea, and Taiwan, together with the information from the CEFR-related resources. The motivation behind this project is to compile a wordlist reflecting current best practice in Asian countries/regions, where English has been taught at elementary school, as well as some pedagogical implications from the CEFR-based lexical resources such as the EVP, Threshold Level (van Ek & Trim, 1990), and the Core Inventory for General English (North, Ortega, & Sheehan, 2010). The CEFR-J Wordlist provides the information about nouns for their general versus specific notions indicated in Threshold Level as well as thematic categories from the Core Inventory for General English. This information is vital as a word is used in conjunction with certain constructions. Together they serve to perform a specific function in given contexts or situations. This specification of language points is realized, in the CEFR terms, by a set of illustrative functional descriptors that will produce certain situations and tasks where grammar and vocabulary are united to produce utterances. Vocabulary should be seen as a realization of functions in context. Thus, this kind of wordlist does not stand on its own. It should work most effectively as part of a broader framework and other linguistic inventories such as a grammar profile or text profile (Tono, 2017).

**Evaluating Corpora with Word Lists and Word Difficulty by Brent A. Culligan**

Culligan proposed a unique method of evaluating the quality of different wordlists regarding the average IRT difficulty estimates for the words on the wordlist. The title of the paper contains the phrase “evaluating corpora,” but actually what he did was another type of external validation of the wordlists. Although I am not entirely convinced whether a sufficient number of data points were obtained for each lemma from Yes/No tests in order to produce reliable theta scores, the results still show some interesting observations.

Firstly, the results show that there is considerable overlap in word difficulty between the GSL 2000 and the Academic Word List (AWL), which suggests that the difference between the GSL 2000 and the AWL was not as clear as was intended. It is possible that many words added in the AWL could have been in the...
GSL 2000 if more texts had been added to the original data examined by Michael West. The purpose of the AWL is to supplement the GSL with general-purpose academic words found in various academic texts, which assumes that a group of words in the AWL should not be too domain-specific. In this sense, this similarity between the GSL 2000 and the AWL might indicate that the control of word difficulty levels for the GSL 2000 (−0.79 logits, SD = 2.03) may not be so successful, compared to that of the NGSL (−1.59, SD = 1.68).

Secondly, Culligan concludes that “this analysis provides evidence that seems to indicate researchers that use a word family approach, particularly using word affixation beyond Level 2 will result in more difficult word lists.” I could not agree with this claim more, because in English corpus linguistics at least, we have never used word family as a solid counting unit for a word. Also, as one scrutinizes the GSL entries, there are many cases where lemmas and word families were not treated consistently. Some of the decisions made by Michael West seem to be arbitrary and need fine-grained reanalysis. In this sense, the approach adopted by the NGSL team, both in the UK and Japan, seems to be reasonable, since it is not a good idea to follow the GSL blindly.

Methodologically, Culligan’s study suggests an interesting direction of evaluating the wordlists in light of learners’ perception of word familiarity. While what is known in Yes/No tests is still very partial, it would certainly provide useful information about how the wordlists behave against such measures. This reminds me of the recent development of Pearson’s Global Scale of English (de Jong et al., 2016) and its accompanying word and grammar resources. Their attempt at estimating word difficulty levels on the GSE 10 to 90 scales is also an excellent example of applying logit scores converted to the GSE scores. In their survey, however, the word difficulty level was assessed by the teacher questionnaire survey, in which teachers were asked whether a particular word is necessary to learn at a given CEFR level. It would be interesting to compare the results of Culligan’s study to the word difficulty level proposed by the GSE.

Another interesting issue as a follow-up study would be to investigate whether the results of Yes/No tests will be affected by Japanese learners’ prior knowledge of katakana loan words. Many entries in the GSE and other wordlists are already part of Japanese katakana vocabulary, and once students can recognize the correspondence between English spellings and their katakana equivalents, they can tell the meaning in many cases. Thus, the item difficulty scores might be affected by such knowledge as recognizing katakana words as cognates in the test items. Also, this judgment of cognates is heavily affected by L2 learners’ meta-cognitive knowledge, and thus, the target population of Japanese learners of English might yield a different result.

 VALIDATING THE CONSTRUCT OF READABILITY IN EFL CONTEXTS: A PROPOSAL FOR CRITERIA BY GEOFFREY G. PINCHBECK

Pinchbeck’s paper focused on the lexical component of readability and proposed a list of criteria to evaluate candidate components of readability formulas for Japanese learners, including (a) transparency in calculating a readability index, (b) availability, (c) non-commercial, (d) accessible technology, (e) diagnostic, and (f) empirically valid measures. They all look reasonable, and some of the
readability measures available in Japan (e.g., Yomiyasusa Level and Kyoto Scale) need further refinement in these respects.

Pinchbeck relies on cloze tests as the basis of readability scales, but I am not entirely convinced whether what is measured by cloze tests is equivalent to text difficulty. It is not clear that one can define readability by the results of cloze tests. Cloze tests measure various aspects of linguistic competence: grammatical, lexical, collocational, and discoursal knowledge. While these different types of knowledge are closely related to learners’ linguistic competence, learners do not usually require them as they read. Most test items asked in cloze tests tap on L2 learner’s productive knowledge. Readability, however, is more closely linked to the difficulty level of receptive knowledge. I would argue that assessing readability levels using cloze tests provides only an indirect link to text difficulty as a construct and that a more valid and clear operational definition of readability is needed.

Most readability formulas developed so far are based on simple surface measures such as the total text length, the average word or sentence length, and the word difficulty level. In corpus linguistics, we can find other useful substitutes for existing readability measures. For example, Tom Cobb’s Compleat Lexical Tutor (Cobb, 2018) provides various measures for text characteristics based on different wordlists such as the GSL, the BNC, or the Corpus of Contemporary American English (COCA). Many concordancers such as WordSmith (Scott, 2018), AntConc (Anthony, 2018b), and CasualConc (Imao, 2018) have functions for measuring text statistics. Text Inspector developed by the English Profile team can produce various readability measures and their original CEFR level estimation.

Our project called the CEFR-J (Tono, 2017) aims to construct a CEFR-based original framework together with various profile information including the CEFR-J Wordlist, Grammar Profile, and Text Profile. There is an online CEFR level assignment tool called the CEFR-based Vocabulary Level Analyzer, developed by Satoru Uchida (see Figure 1). When you select a text whose CEFR level you want to know, the tool will produce an estimated CEFR level based on the regression model of different CEFR-J word proportions, existing readability measures, the average number of verbs per sentence, and the ratio of B-level words to A-level words.

These new metrics are all produced as a part of reference level descriptions for the CEFR. This will inform L2 vocabulary researchers of new types of lexical and grammatical profile.

**Toward Written Error Correction with a Japanese–English Parallel Corpus: DDL in the Japanese EFL Classroom by Michael McGuire**

Data-driven learning is becoming popular as an inductive approach to learning a language. McGuire rightly commented that there are very few studies on actual DDL activities, primarily related to written error correction. He went on to suggest that the use of Japanese–English parallel corpora would make a DDL approach more accessible and useful. Regarding the effects of DDL on error correction in L2 writing, Tono, Satake and Miura (2013), and Satake (2018) tackle the very issues of how DDL helps L2 learners identify and correct errors. Satake (2018) found that prior identification of errors by teachers would facilitate
CVLA: CEFR-based Vocabulary Level Analyzer (ver. 1.1)

[Legend]
A:example, B:example, B1:example, B2:example, C1:example, C2:example,
NA content word, other:example

Small numbers indicate the ranking in COCA (added only to nouns, verbs, adjectives, and adverbs).

[Input]
Using computers, it is not currently possible to easily and accurately analyse the deeper structural and semantic features of texts (i.e., dialect, rhetorical structure, topic, background, and complexity) of cohesive devices, allusion, metaphor, the level of inference required, and/or formulaic language, etc. For this we must await further advances in natural language processing technologies. In the meantime, I propose the examination of the lexical component of readability, particularly, in Japan and in other EFL teaching contexts. It is my hope that this editorial might generate some interest towards the development of large-scale validity studies that would develop new pedagogical tools for matching Japanese K-12 and adult EFL learners to English texts.

Figure 1. The CEFR-based Vocabulary Level Analyzer (Uchida, 2018).
learners’ correction of those errors using corpora. However, there is an interaction between the ease of error corrections and the types of errors; omission and addition errors were more easily corrected than misformation errors, due to the saliency in error status on concordance lines.

McGuire also encouraged the use of parallel corpora and Eijiro on the Web as a kind of parallel corpus. This issue is closely related to the three challenges for using DDL in the classroom that Laurence Anthony mentioned in his plenary speech given at Asia Pacific Corpus Linguistics Conference (APCLC) 2018 (Anthony, 2018a). The first challenge is to find a suitable corpus. In order to make a useful DDL material, it is essential to adjust the difficulty level of the texts on concordance lines to the level of target learners. McGuire recommends the use of parallel corpora to bridge the gap of learners’ lack of understanding of difficult examples. A similar approach has been already taken in previous projects such as Chujo and her colleagues (Chujo et al., 2006).

DDL researchers usually use various corpus building tools to compile so-called DIY corpora. For example, AntFileConverter (Anthony, 2017) can convert PDF or Word files to plain texts, so if supplementary teaching materials are available in PDF or Word format, it can add those texts to the DIY corpus for specific teaching purposes. People also use tools such as BootCat, Sketch Engine, and WordSmith to crawl the Internet and collect web materials automatically.

Secondly, Anthony (2018a) brought up the issue of interacting with a corpus, which involves how to interpret patterns of use found on KWIC lines of a particular search word, in the information about word frequency and distribution, and in a list of collocations and related association measures. This issue is complicated, because the purposes of DDL are diverse, as it could aim for convergent learning, in which learners are expected to come to the same generalization through the interpretation of corpus evidence, while it could also aim for more divergent learning, where each learner will make their own discoveries through data exploration in a serendipitous manner (Leech, 1997).

The final challenge mentioned by Anthony (2018a) was using corpus tools. In order to have an active DDL environment, it would be necessary for learners in the classroom to have hands-on experience with corpus data. Printed versions of corpus data have some positive effects, but many corpus query tools need a certain degree of mastery for flexible data extraction and learners need to overcome computer literacy problems if they are really serious about exploring the data in a corpus. The development of user-friendly corpus tools such as SKELL (Baisa, 2014) and SCoRE (Chujo, Oghigian, & Akasegawa, 2015) is quite useful in this sense.

McGuire recommends the Eijiro on the Web because it has a function for showing lists of example sentences with Japanese translations in Key Word In Context (KWIC) format. Many electronic dictionary interfaces such as Eijiro or CASIO’s pocket dictionaries now have this function, which looks similar to corpus data. However, it should be noted that the number of examples in the Eijiro or other electronic dictionaries may not be large enough to show significant patterns of use of a particular search word in context. Thus, learners need to be warned about how to interpret the example list on such systems.
Overall, the effects of DDL have to be evaluated in light of the role it plays in the lesson and to answer such questions as whether DDL is used for presenting a new structure or vocabulary, practicing vocabulary or grammar, or for more task-based production activities. Many DDL studies did not seem to take into account the overall syllabus contents and how DDL works within that syllabus. The inductive approach sounds fine, but it may take too much time to use DDL for every stage of a lesson. Then, one should seriously ask what aspect of learning (e.g., comprehension vs. production; lexis and grammar; and general vs. ESP) must be most effectively delivered using DDL.

Some Suggestions for Future Research: The Interface between Grammar and Lexis

Concerning McGuire’s suggestion to use parallel corpora for DDL, I will finish my commentaries by making a brief report on the application of parallel learner corpora for finding Japanese EFL learners’ overuse and underuse of grammatical structures. In L2 vocabulary acquisition research, there are many studies on wordlists, but not many studies have been done to summarize the overall usage patterns of grammar constructions. Since there is no longer a clear boundary between lexis and grammar, more research needs to be conducted to profile L2 learners’ use of grammar items.

The Japanese EFL Learner Corpus (JEFLL Corpus; Tono, 2006) contains English essays written by 10,038 junior and senior high school students in Japan (the total number of running words: 669,304 tokens). All the essays were proofread by NSs, thus making the original version comparable against its corrected version. A set of pattern-matching queries for extracting grammar items was prepared and automatically extracted frequencies of grammar items from both the original and the corrected versions of the JEFLL Corpus (see Table 1).

By extracting all the frequencies of each grammar item from the original and corrected versions of the JEFLL Corpus, we can contrast the two versions regarding the overuse versus underuse of each grammar item, shown in Figure 2. The JEFLL Corpus is classified by the CEFR levels, which makes it possible to observe the transition of overuse and underuse phenomena of a particular grammar item throughout the CEFR levels.

### Table 1. Extraction rules for CEFR-J grammar profile

<table>
<thead>
<tr>
<th>ID</th>
<th>Item</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>INDEFINITE PRONOUN: none</td>
<td>\none_NN\none\b</td>
</tr>
<tr>
<td>49</td>
<td>COMPARATIVE and COMPARATIVE (the same adjective)</td>
<td>\b(\S+_(\JJR</td>
</tr>
<tr>
<td>66</td>
<td>TENSE/ASPECT: PAST PROGRESSIVE (AFFIRMATIVE DECLARATIVE)</td>
<td>(was</td>
</tr>
<tr>
<td>145</td>
<td>AUX+PERFECT (AFFIRMATIVE DECLARATIVE)</td>
<td>(?!cannot(b))S+_MD_S+ have_VH_have \S+_V.N_S+</td>
</tr>
</tbody>
</table>
**Figure 2.** The overuse versus underuse of grammar items in the JEFLL parallel corpus
Tables 2 and 3 show the 10 most underused and overused grammar items, respectively. The number in each cell shows the ratio of the number of grammar items in the original essay to that in the corrected essay. If the number is below 1, it means the item is underused. If it is above 1, it is overused. Our analysis shows that Japanese junior and senior high school students tend to underuse grammar items rather than overuse them. Underused items are typically those constructions which involve more than one grammar item, for example, ‘being + past participle [PROGRESSIVE+PASSIVE]’.

The detailed corpus analysis of such overuse/underuse grammar patterns at each CEFR level will help materials developers focus on those grammar items that need special attention at different CEFR levels and change the organization of grammar textbooks drastically. Many of these grammar items involve the use of essential verbs and function words, which will shed new light on the treatment of core vocabulary more systematically.
Conclusions

This review convinces me even more that there is a very close relationship between applied corpus linguistics and L2 vocabulary research. The four studies reviewed in this volume clearly show some directions toward more communication between the two fields. As was mentioned in the introduction, corpus linguists working on learner language or pedagogical materials sometimes do not realize the background of SLA, which prevents them from interpreting the corpus analysis results in a way that is meaningful to SLA researchers. Likewise, SLA researchers are not fully aware of the wealth of resources and tools available in corpus linguistics and natural language processing, which will greatly facilitate data analysis. This has been said many times, but I have the impression that the more corpus and Natural Language Processing (NLP) technologies advance, the wider the gap will be for those who try to fill it. I would recommend that those colleagues who are active in an interdisciplinary Special Interest Group (SIG) like JALT Vocabulary (JALT VOC) should promote the exchange of ideas and people even more to help bridge the gap.

Note

1. This grammar pattern extraction was mainly done by Yasutake Ishii. See Ishii and Tono (2018) for more detail.

References


An Examination of the New General Service List

Tim Stoeckel
University of Niigata Prefecture

Abstract
The New General Service List (NGSL; Browne, Culligan, & Phillips, 2013b) was published in 2013 as a modern replacement for West’s (1953) original General Service List (GSL). This study compared GSL and NGSL coverage of a 6-year, 114-million word section of the Corpus of Contemporary American English (COCA), and used COCA word frequencies as a secondary data source to identify candidates for addition to the NGSL. The NGSL was found to provide 4.32% better coverage of the COCA than the GSL. Moreover, several candidates were identified for inclusion to the NGSL: three are current members of the NGSL’s companion list, the New Academic Word List (Browne, Culligan, & Phillips, 2013a); five are words whose usage has increased in recent years; and five are individual types that appear to have been miscategorized during the original development of the NGSL. Because NGSL word selection was based on not only empirical but also subjective criteria, the article calls for the addition of annotations to the NGSL to explain decisions regarding low-frequency NGSL constituents and high-frequency non-constituents.

Key Words: New General Service List

Introduction
The New General Service List (NGSL; Browne, Culligan, & Phillips, 2013b) was published in 2013 as a modern replacement for West’s (1953) original General Service List (GSL) for the purposes of both research and pedagogy for second-language learners of English (Browne, 2014). As the NGSL is currently considered an interim list open to examination and debate by interested researchers (Browne, 2014), this article investigates NGSL coverage and word frequencies in a large, modern corpus in hopes of promoting discussion and possible refinement of the list.

The General Service List
The original GSL (West, 1953) was constructed based upon frequencies in a 5-million word corpus (for some words, 2.5 million) as well as subjective criteria on the usefulness of words to learners of English. Generally, word forms with shared roots were grouped together, but there was no single guiding principle for what to include in each group. Under breadth, for example, are only breadth and the compound broadcast. The GSL was extensively revised by Paul Nation in the early 1990s so that many inflected and derived forms were added (e.g., broader and broadly under broad) and some compound forms (e.g., broadcast) were removed. Numbers, months, days of the week, and letters of the alphabet

Please cite this article as: Stoeckel, T. (2019). An Examination of the New General Service List. Vocabulary Learning and Instruction, 8 (1), 53–61. https://doi.org/10.7820/vli.v08.1.stoeckel
were also added to the list. This is the most widely used version of the GSL today, as it is utilized in Lextutor VocabProfiler (http://www.lextutor.ca/vp/) and accompanies the Range (Heatley, Nation, & Coxhead, 1994) and AntWordProfiler (Anthony, 2013) programs.

Although the GSL has been valuable in teaching and research, the corpus from which the list was derived is now dated and considered small by modern standards. Additionally, while the inclusion of derivational and compound forms may be useful in raising awareness of related word forms, it is arguably not the most efficient way to help students learn the most important words for second language acquisition. For many learners, derivational knowledge cannot be assumed from knowledge of a headword (McLean, 2017; Mochizuki & Aizawa, 2000; Ward & Chuenjundaeng, 2009), which implies that the learning burden for the GSL entails learning not only headwords but also many separate constituents.

The New General Service List

The NGSL was developed in part to address these limitations. It was derived from the analysis of a 273-million word section of the more modern Cambridge English Corpus (CEC). As an organizing principle, words in the NGSL are grouped into modified lemmas. A regular lemma consists of a headword plus inflected forms that are of the same part of speech as the headword (e.g., the nominal headword *show* plus the plural *shows*). A modified lemma is comprised of a headword in all of its various parts of speech together with the inflected forms for each part of speech. Thus, the modified lemma for *show* includes the noun forms above as well as the verbs *shows*, *showed*, *showing*, and *shown*. This use of the modified lemma is consistent with McLean (2017), who found that most L2 learners with knowledge of a headword also have receptive knowledge of its inflected forms. This word-grouping principle determined what was to be included in each NGSL modified lemma regardless of whether individual word types actually occurred in the CEC (Browne, 2014). Exceptions included cases in which a word form was both the head of one modified lemma and a constituent of another, in which case the word form in question was sometimes considered only as the canonical form of its own modified lemma (B. Culligan, personal communication, March 28, 2018). Thus, *rose* is not found under *rise*, but it is considered the head of the modified lemma comprised of *rose* and *roses*. Other exceptions were the inclusion of pluralized gerunds (e.g., *teachings* under *teach*) for approximately 10% of the gerunds in the list. The NGSL consists of 2801 modified lemmas, and there is a supplementary list of 52 entries comprised of numbers, months, and days of the week.

Word selection for the NGSL was based on both empirical evidence and subjective considerations. To empirically rank words, Carroll's *U* (Carroll, 1971) was used; this statistic expresses an adjusted words-per-million (wpm) that accounts for differences in both dispersion across and size of corpus sections. As shown in Table 1, the most common 1000 words in the CEC occur with an adjusted frequency of at least 93.4 wpm, and the top 2000 words are of at least 36.5 wpm. Because the NGSL has 2801 entries, the modified lemma *utility*, with a *U* value of 21.2 and rank of 2801, can be used as a benchmark for assessing whether a word belongs in the NGSL based on empirical evidence alone (CEC *U* values available...
at http://www.newgeneralservicelist.org/). Regarding subjective criteria, there was ongoing input from Paul Nation as well as comparisons of the NGSL to other existing lists to “make sure important words were included or excluded as necessary” (Browne, 2014, p. 40).

We get a sense of the balance between these quantitative and subjective considerations from Table 2, which shows the number of entries in the NGSL from different adjusted-frequency levels of the CEC. These follow an expected pattern, with most 1–2K level words present in the NGSL or its supplementary list. Notably however, seven 1K and eight 2K words are absent; with $U$ values between 37.4 and 137.5, the adjusted frequencies for these words in the CEC far exceed many NGSL constituents. Because the list is intended to give maximal coverage of English texts with the fewest words, and as explanations regarding the subjective criteria for individual word selections are unavailable, an examination of non-empirically grounded decisions would seem worthwhile. It may also be advisable to further investigate coverage provided by the NGSL. Browne (2014) has reported that the NGSL (90.34%) offers approximately 6% better coverage than the GSL (84.24%) of general English discourse, but this comparison may have favored the NGSL because it used the section of the CEC from which the list itself was derived. Moreover, although the precise composition of this CEC section is unavailable, the complete CEC is comprised mostly of British English sources (65.4%; D. Moser, Cambridge University Press, personal communication, April 16, 2018), suggesting the NGSL may reflect British usage tendencies. There would therefore be merit in cross-checking NGSL membership with word frequencies in corpora composed primarily of North American or other world English sources.

Table 1. Carroll’s $U$ Values for Words at Selected Frequency Ranks in the CEC

<table>
<thead>
<tr>
<th>Modified Lemma Head</th>
<th>Rank</th>
<th>$U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1</td>
<td>60,909.9</td>
</tr>
<tr>
<td>somebody</td>
<td>1000</td>
<td>93.4</td>
</tr>
<tr>
<td>revolution</td>
<td>2000</td>
<td>36.5</td>
</tr>
<tr>
<td>mortgage</td>
<td>2500</td>
<td>26.0</td>
</tr>
<tr>
<td>utility</td>
<td>2801</td>
<td>21.2</td>
</tr>
<tr>
<td>quit</td>
<td>3000</td>
<td>19.0</td>
</tr>
<tr>
<td>explicit</td>
<td>3500</td>
<td>14.7</td>
</tr>
</tbody>
</table>

*Note: Carroll’s (1971) $U$ values downloaded from http://www.newgeneralservicelist.org/

Table 2. NGSL Entries from Various Adjusted-Frequency Levels of the CEC

<table>
<thead>
<tr>
<th>List</th>
<th>Adjusted-Frequency Band</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1K</td>
</tr>
<tr>
<td>NGSL</td>
<td>981</td>
</tr>
<tr>
<td>NGSL Supplementary</td>
<td>12</td>
</tr>
</tbody>
</table>

*Note: Adjusted frequencies based on Carroll’s (1971) $U$ values downloaded from http://www.newgeneralservicelist.org/
Purpose

This study investigated NGSL coverage and word frequencies in the Corpus of Contemporary American English (COCA; Davies, 2008). The research questions (RQ) were:

1. How does the NGSL compare to the GSL in terms of coverage of the COCA and each of its individual sections?
2. Does analysis of the COCA as a secondary data source reveal candidates worthy of consideration for addition to the NGSL?

Methods

Materials

The study corpus was comprised of word/lemma/part-of-speech files for a 6-year section (2010–2015) of the COCA. Regarding GSL and NGSL word lists, to make comparisons using the most readily available variants today, the version of the GSL that accompanies Range (downloaded from https://www.victoria.ac.nz/lals/about/staff/paul-nation) and version 1.01 of the NGSL (downloaded from http://www.newgeneralservicelist.org/) were used. Because of the additions Paul Nation made to the GSL (described above), the NGSL supplementary list, with letters of the alphabet added by the researcher, was included in order to make fair comparisons between the NGSL and GSL.

Corpus Cleaning

Part-of-speech tags in the complete set of COCA files were used to identify proper nouns, foreign words, marginal content, non-words, and all other discourse. To assess the accuracy of these classifications, a 2000-token sample was also categorized manually, with a 98.3% agreement rate. Tokens categorized as non-words were then removed from the COCA files, and the size of each corpus section was recorded (Table 3).

Analyses

For RQ1, AntWordProfiler (Anthony, 2013) was used to obtain the coverage percentage provided by the GSL and NGSL of the COCA and each of its sections.

Table 3. Number of Tokens Before and After Removal of Non-Words from the 2010–2015 Section of the COCA

<table>
<thead>
<tr>
<th>Section</th>
<th>Before Cleaning</th>
<th>After Cleaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction</td>
<td>29,985,804</td>
<td>23,705,513</td>
</tr>
<tr>
<td>Spoken</td>
<td>29,621,441</td>
<td>23,348,630</td>
</tr>
<tr>
<td>Academic</td>
<td>27,208,629</td>
<td>20,913,397</td>
</tr>
<tr>
<td>Magazine</td>
<td>29,122,476</td>
<td>23,317,601</td>
</tr>
<tr>
<td>Newspaper</td>
<td>28,965,412</td>
<td>22,825,277</td>
</tr>
<tr>
<td>Total</td>
<td>144,903,762</td>
<td>114,110,418</td>
</tr>
</tbody>
</table>
For RQ2, AntWordProfiler was first used to obtain frequencies for NGSL constituents and for off-list types. A file of modified lemmas was then made for the 1000 most frequent off-list types, and this was used with AntWordProfiler to obtain frequency counts for off-list modified lemmas in each COCA section. Carroll's $U$ (Carroll, 1971) was then calculated for all NGSL, NGSL supplementary, and off-list modified lemmas. These $U$ values and those derived from the CEC in the original development of the NGSL (downloaded from: http://www.newgeneralservicelist.org/) were examined to identify candidates for addition to the NGSL.

Results and Discussion

For the 3000 modified lemmas with the highest adjusted frequencies in the CEC, the Pearson’s correlation for $U$-values in the CEC and the COCA was 0.991, indicating a very high degree of similarity in the occurrence of high-frequency vocabulary in the two corpora. Regarding coverage, though somewhat lower than Browne’s (2014) original comparison, the NGSL plus its supplemental list performed well, providing 4.32% better overall coverage than the GSL (83.66% vs. 79.34%, Table 4). Notably, even though the academic and newspaper sections of the CEC were excluded from analysis during the development of the NGSL (Browne, 2014), the advantage in coverage provided by the NGSL was greatest in these two COCA sections, +9.07% and +4.87%, respectively.

RQ2 asked whether analysis of the study corpus would reveal candidates for inclusion to the NGSL. The empirical benchmark for NGSL membership in the COCA was $U = 22.7$ for the modified lemma con (ranked 2801), which is similar to that of the CEC reported above. There were 282 off-list modified lemmas above this threshold in the COCA. Although this is partially due to differences that exist between any two corpora (see Nation, 2016 p. 99), three categories of words above this threshold will be presented as NGSL candidates.

First are three high-frequency words that currently belong to the NGSL companion list, the New Academic Word List (NAWL; Browne, Culligan, & Phillips, 2013a). Thirty NAWL members have $U$ values greater than the empirical threshold for NGSL membership in both the COCA ($U \geq 22.7$) and the CEC ($U \geq 21.2$). However, many of these could be considered specialized academic vocabulary on the basis of higher adjusted frequencies in the academic corpus used.

Table 4. A Comparison of GSL and NGSL Coverage of the 2010–2015 Section of the COCA

<table>
<thead>
<tr>
<th>Section</th>
<th>Proper Nouns</th>
<th>Foreign</th>
<th>Marginal</th>
<th>NGSL</th>
<th>Supp</th>
<th>Total</th>
<th>GSL</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fiction</td>
<td>3.62</td>
<td>0.04</td>
<td>0.25</td>
<td>85.38</td>
<td>0.66</td>
<td>86.03</td>
<td>85.11</td>
<td>+.92</td>
</tr>
<tr>
<td>Spoken</td>
<td>4.65</td>
<td>0.01</td>
<td>0.65</td>
<td>88.08</td>
<td>0.70</td>
<td>88.78</td>
<td>85.59</td>
<td>+3.18</td>
</tr>
<tr>
<td>Academic</td>
<td>6.58</td>
<td>0.10</td>
<td>0.04</td>
<td>79.57</td>
<td>0.84</td>
<td>80.41</td>
<td>71.34</td>
<td>+9.07</td>
</tr>
<tr>
<td>Magazine</td>
<td>5.29</td>
<td>0.03</td>
<td>0.07</td>
<td>81.06</td>
<td>0.75</td>
<td>81.81</td>
<td>77.70</td>
<td>+4.11</td>
</tr>
<tr>
<td>Newspaper</td>
<td>8.42</td>
<td>0.06</td>
<td>0.05</td>
<td>79.57</td>
<td>1.25</td>
<td>80.82</td>
<td>75.95</td>
<td>+4.87</td>
</tr>
<tr>
<td>Total</td>
<td>5.67</td>
<td>0.05</td>
<td>0.22</td>
<td>82.82</td>
<td>0.84</td>
<td>83.66</td>
<td>79.34</td>
<td>+4.32</td>
</tr>
</tbody>
</table>

Note: Coverage totals for each list are in boldface.
to create the NAWL (e.g., impact and authority; data available at http://www.newgeneralservicelist.org) or because in the COCA they occur most frequently in the academic section (e.g., aspect and distribution). Setting aside such cases, candidate, conference, and click remain as candidates for inclusion to the NGSL (Table 5).

The second category includes five words whose usage has recently increased (Table 6), probably due to changes in technology (website, blog, click), world events (immigration), or both (solar). Changes over time were documented with data from the full online version of the COCA (Table 7).

The final category consists of five types (best, better, rose, born, and criteria) that could have been listed as constituents under current NGSL headwords but were instead classified as heads of their own modified lemmas. Let us consider best as an example. Although best is the superlative form of the NGSL headword good,

<table>
<thead>
<tr>
<th>Modified Lemma</th>
<th>U Values</th>
<th>Occurrences per Million in the 2010–2015 Section of the COCA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COCA</td>
<td>CEC</td>
</tr>
<tr>
<td>candidate</td>
<td>117.40</td>
<td>94.29</td>
</tr>
<tr>
<td>conference</td>
<td>90.81</td>
<td>117.04</td>
</tr>
<tr>
<td>click</td>
<td>42.09</td>
<td>24.61</td>
</tr>
</tbody>
</table>

Note: CEC and NAWL Corpus U values downloaded from http://www.newgeneralservicelist.org/

<table>
<thead>
<tr>
<th>Modified Lemma</th>
<th>U Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>website</td>
<td>76.07</td>
</tr>
<tr>
<td>blog</td>
<td>31.40</td>
</tr>
<tr>
<td>immigration</td>
<td>51.69</td>
</tr>
<tr>
<td>solar</td>
<td>43.66</td>
</tr>
<tr>
<td>click</td>
<td>42.09</td>
</tr>
</tbody>
</table>

Note: CEC U values downloaded from http://www.newgeneralservicelist.org/

<table>
<thead>
<tr>
<th>COCA Section</th>
<th>Occurrences per Million</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>website</td>
</tr>
<tr>
<td>1990–1994</td>
<td>0</td>
</tr>
<tr>
<td>1995–1999</td>
<td>3.34</td>
</tr>
<tr>
<td>2000–2004</td>
<td>13.39</td>
</tr>
<tr>
<td>2005–2009</td>
<td>27.90</td>
</tr>
<tr>
<td>2010–2014</td>
<td>67.87</td>
</tr>
</tbody>
</table>

Note: Data from the online version of the COCA at https://corpus.byu.edu/coca.
Table 8. Constituents of Two Modified Lemmas

<table>
<thead>
<tr>
<th>Modified Lemma</th>
<th>COCA</th>
<th>CEC</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>best</td>
<td>445.48</td>
<td>1.47</td>
<td>223.48</td>
</tr>
<tr>
<td>better</td>
<td>433.92</td>
<td>3.13</td>
<td>218.53</td>
</tr>
<tr>
<td>rose</td>
<td>78.05</td>
<td>12.01</td>
<td>45.03</td>
</tr>
<tr>
<td>born</td>
<td>102.46</td>
<td>0.25</td>
<td>51.35</td>
</tr>
<tr>
<td>criteria</td>
<td>18.83</td>
<td>(unlisted)</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: CEC U values downloaded from http://www.newgeneralservicelist.org/

Table 9. Occurrences of Modified Lemma Constituents for Good and Best in the COCA

<table>
<thead>
<tr>
<th>Type</th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
<th>Adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>good</td>
<td>329</td>
<td>2</td>
<td>500,960</td>
<td>6,688</td>
</tr>
<tr>
<td>goods</td>
<td>19,797</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>best</td>
<td>86</td>
<td>1795</td>
<td>179,702</td>
<td>55,846</td>
</tr>
<tr>
<td>bests</td>
<td>2</td>
<td>131</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>bested</td>
<td>0</td>
<td>361</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>besting</td>
<td>0</td>
<td>127</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: Data from the online version of the entire COCA at https://corpus.byu.edu/coca/.

It is not listed in the NGSL because it was treated as the headword for the modified lemma consisting of best, bests, besting, and bested, and the U value for this modified lemma was too low for NGSL membership. In the COCA, best is tagged as a noun, verb, adjective, or adverb depending on the context, and its 445.48 COCA U value shown in Table 8 is derived from all of these. It is unclear how the CEC U value for best was derived, but the very low value in Table 8 (U = 1.47) suggests it might only have included occurrences for constituents of the verbal lemma rather than across parts of speech. If so, this would explain the exclusion of this very common word from the NGSL.

For three of the words shown in Table 8, an argument could be made for inspecting frequencies at each part of speech to determine which headword to place it under. Using this approach, best would belong under good because both best and good occur predominantly as adjectives (Table 9). Similarly, better and rose would become constituents of good and rise, respectively (data omitted due to space limitations). Born would be listed under bear as it was in the original GSL, but because many modern dictionaries list born as an adjective rather than as the past participle of bear, and because the connection between bear and born is probably no longer felt (Online Etymology Dictionary, www.etymonline.com), perhaps born could be listed separately. Whichever approach is taken, if the frequency with which born occurs is more accurately reflected in the COCA data, it ought to be included in a general service vocabulary list. Finally, the case of the plural form criteria appears to be straightforward: it is currently a stand-alone headword in the NAWL, but it belongs under the NGSL headword criterion.

Tim Stoeckel: An Examination of the New General Service List 59
Conclusions

This study corroborates the findings of Browne (2014) regarding the substantially better coverage provided by the NGSL in comparison to the GSL. Using the COCA as a secondary data source, it also found a very high level of agreement between NGSL members and words occurring with high frequency in the COCA. Twelve candidates were tentatively identified for inclusion in the NGSL. Three currently belong to the NAWL but appear commonly across discourse types; several others have seen increased use in recent years; and five others, each of which is both the headword of one modified lemma and a constituent of another, may have been inadvertently excluded from the NGSL due to the use of only partial CEC frequency values.

It is hoped that this study will promote further discussion and debate of the NGSL. Although quantitative evidence was found to support the addition of a small number of words to the NGSL, it would be helpful to know the original rationale behind the decisions to exclude these words. More broadly, similar to the commentary that supplements some entries in West’s original GSL, perhaps a next useful step in the development of the NGSL could be to add brief explanations to low-frequency NGSL constituents and high-frequency non-constituents so that interested parties would better understand non-empirically grounded membership decisions.

Note

1 This GSL variant is not organized according to Bauer and Nation’s (1993) word family levels, as it was made prior to that publication (P. Nation, personal communication, September 27, 2017).

Acknowledgements

I would like to thank Brent Culligan and Charlie Browne for helping me understand technical aspects of the NGSL. In addition, this article is part of a larger collaborative study with Phil Bennett, whose many valuable insights are reflected here. This research was funded in part by a 2018 JALT Vocabulary SIG grant.

References


Evaluating the Efficacy of Yes–No Checklist Tests to Assess Knowledge of Multi-Word Lexical Units

Raymond Stubbe\textsuperscript{a} and Yumiko Cochrane\textsuperscript{b}
\textsuperscript{a}Nagasaki University; \textsuperscript{b}Kyushu Sangyo University
https://doi.org/10.7820/vli.v08.1.stubbe.cochrane

Abstract

One of the many challenges facing Japanese university students studying English is the multi-word phrase. The English language contains a large number of such multiple-word items, which act as single words with a single meaning. This study is concerned with evaluating the efficacy of yes/no checklist tests to assess knowledge of multi-word units. Participants (n = 206) took a yes–no test of 30 real words and 15 pseudowords. The 30 real words were selected from the students’ textbook, based on the teacher’s intuition of the words and multi-words posing the greatest learning burden for the students. Twenty-one of the selected words were single-word items. The remaining nine were multi-words, such as “get up” and “take turns”. Forty-five minutes following completion of the yes–no test, an English to Japanese translation test of the same 30 real words was taken by the same participants to evaluate the efficacy of yes/no test. Results suggest that the yes–no vocabulary test format may be able to measure student knowledge of multi-word lexical units as (or more) effectively than single-word units.

Key words: Multi word lexical units; yes no vocabulary tests; translation tests; overestimation; Japanese EFL learners.

1 Introduction

One of the many challenges facing Japanese university students studying English is the multi-word phrase. This study is concerned with evaluating the efficacy of Yes/No checklists to assess student knowledge of multi-word units. As this enquiry involves comparing multi-word lexical units with single-word items using a yes–no checklist test and an English to Japanese (L2 to L1) translation test, these three entities will be briefly reviewed.

1.1 Yes–No Checklist Tests

Recognized as the easiest vocabulary test to administer to a group of learners, yes–no checklist tests may be an efficient means of enquiring about student knowledge of multi-word units. According to Read (2007, pp. 112–113), “Despite its simplicity, the Yes/No format has proved to be an informative and cost-effective means of assessing the state of learners’ vocabulary knowledge, particularly for
placement and diagnostic purposes.” Yes–no tests present learners with a list of decontextualized words and have them signify their knowledge of each item by either checking (or circling) that word, or by selecting either “yes” or “no”. As yes–no tests depend on self-reporting, the actual lexical knowledge of test-takers cannot be verified. One concern with this format is whether yes–no results accurately reflect the test-takers’ knowledge of the tested items, or do they overestimate the number of words truly known (Read, 1993, 2000). To compensate for the potential of students claiming knowledge of words they do not actually know the meaning of (labeled overestimation), pseudowords were introduced to the yes–no format by Anderson and Freebody (1983). In such yes–no tests, claiming knowledge of a real word is known as a “hit”, while claiming knowledge of a nonword is called a “false alarm” (FA). Not claiming knowledge of a real word is labeled a “miss” and not claiming knowledge of a nonword is a “correct rejection”. A number of scoring formulae have been devised to adjust yes–no test scores using FA and real-word hit data. The simplest formula h-f, subtracts the FA rate from the hit rate. Along with overestimation, yes–no tests are also liable to underestimation, where students do not signal knowledge of items they actually do know if tested with a translation test or interview. Unlike pseudowords for overestimation, nothing has been developed to indicate possible underestimation in yes–no tests.

A number of prior studies have investigated yes–no test results by comparing them with a criterion measure. Mochida and Harrington (2006) used a multiple-choice test, the Vocabulary Levels Test (VLT) (Nation, 1990; Schmitt, Schmitt & Clapham, 2001). Pellicer-Sánchez and Schmidt (2012) utilized personal interviews as the criterion measure to ascertain participants’ actual vocabulary knowledge in order to determine the amount of overestimation of vocabulary knowledge on their yes–no test. Eyckmans (2004, p. 77) selected L2 to L1 translation tests as her criterion measure because “asking participants to provide mother-tongue equivalents of the target language words was the most univocal way of verifying recognition” on the yes–no tests. As multiple-choice test results are liable to guessing effects (Stewart & White, 2011) and personal interviews are very time-consuming, the L2 to L1 translation test was selected as the criterion measure.

1.2 L2 to L1 Translation Tests

Similar to yes–no tests, the L2 to L1 translation test to be used in this study presents learners with a list of decontextualized words and has them translate the English words into Japanese. L2 to L1 translation tests measure passive recall ability (Laufer & Goldstein, 2004), which is the ability to recall the meaning of an L2 word in the first language, or the ability to recall the meaning of an L3 word in the L2. Echoing Eyckmans (2004), other researchers agree that translation ability is a strong indicator of which words students can actually understand while reading (Waring & Takaki, 2003) and listening (Pellicer-Sánchez & Schmitt, 2012).

1.3 Multi-word Units

As the name suggests, multi-word units are lexical items made up of two or more words, which combine to have a single meaning that differs (however slightly)

English has a large number of these multiple-word-item lexemes that behave as a single word with a single meaning ... There are a number of different kinds of multiword units, including compound words (playpen), phrasal verbs (give up), fixed phrases (ladies and gentlemen), idioms (put your nose to the grindstone), and proverbs (A stitch in time saves nine). (p. 747)

2 Methodology

Participants in this study were first-year university students enrolled in one of five mandatory English classes \( n = 206 \) at a public university in southern Japan. Although English levels were not available for each student, the five participating departments tend to have students at the high beginner level.

All tested items \( k = 30 \) were selected from the students’ textbook, based on the teacher’s intuition of the words and multi-words posing the greatest learning burden for the students. Twenty-one of the selected items were single-word items. The remaining nine were the following multi-words: brush (your) teeth; easy to; fish bowl; get dressed; get ready; get up; put on makeup; take turns; and, wake up.

For the yes–no test, the 30 items were combined with 15 pseudowords, and all 45 items were randomly ordered and then numbered 1 to 45. Each numbered item was followed by a bubbled “y” (for yes I know this item) and a bubbled “n” (for no I do not know this item). For the translation test, the 30 items were also randomly ordered, then numbered 1 to 30. Following each numbered item, two spaces appeared ( ___________ ___________) and students were encouraged to provide two answers if possible. This was done to avoid the possibility of students writing only one answer (as naturally occurs when only one space is provided) that happens to be incorrect while still knowing a different, correct answer.

The yes–no test was given at the beginning of the second class of the term, prior to commencing the textbook, and took under 10 min. The textbook commenced in the third class. The translation test was given toward the end of that second class, to maximize test-pairings and avoid any possible acquisition of some of the tested items between testing. This second test also was completed in about 10 min. The yes–no test was scored using an optical scanner. The translation test was hand-marked by a native Japanese English university teacher with high-level English ability. Ten percent of the translation tests (21) were copied prior to marking and also marked by a second native Japanese university teacher with high-level English ability. Co-rater agreement was good at 92%.

3 Results

One participant signaled knowledge of 11 of the total 15 pseudowords. As an extreme outlier, his results have been deleted from the data. The overall mean on the yes–no test was 21.97 of the 30 items (73%; Table 1). Compared to other studies
involving Japanese learners (Mochida & Harington, 2006; Stubbe, 2012; 2015), the FA rate was high, likely resulting from the composition of the pseudowords used in this study. Five of the 15 created by this author were based on Laufer’s (1998) concept of synforms, and they proved unexpectedly attractive. Removing these 5 synform pseudowords would result in a FA rate of 4.4%, which is in line with the rates mentioned in the other aforementioned studies. For this reason, only the 10 pseudowords will be used hereafter. The overall mean for the translation test was 18.52 of the 30 words (62%). The correlation (Pearson Product Moment) between the yes–no test and the translation test was moderate at 0.539.

Table 2 presents a comparison of the single and multi-word results for both tests. As there were different numbers of single words (21) and multi-words (9), the mean percentage figures are the most useful. On the yes–no test, students signaled knowledge of 77.2% of the 21 single words and 64.1% of the nine multi-words. The students signaled that they knew 13.1% more single words. However, on the translation test, the students correctly translated 63.7% of the single words (a decrease of 13.5%) and 57.2% of the multi-words (a decrease of 6.9%).

Comparing the responses of the yes–no test directly with the answers on the translation test on a student-by-student item-by-item basis yields four possible outcomes:

1. *yes* on the yes–no test matched with a correct answer on the translation test, labeled “known”.
2. *no* on the yes–no test matched with an incorrect answer on the translation test, labeled “unknown”.
Table 3 presents these possible matches between the two tests. The second and third columns present outcomes $a$ and $b$, in which the yes–no test response is verified by the translation test (known and unknown). The single-word items were more accurate than the multi-word items by 3.9% (79.4% – 75.5%). The last two columns show the amount and type of inaccurate yes–no test responses. With the single-word items, the participants overestimated their word knowledge by 17% but underestimated it by only 3.6%. However, with the multi-word items, less overestimation occurred (15.7%), while much more underestimation occurred (8.8%). This greater amount of underestimation for the multi-word items offsets their overestimation amount, which accounts for the smaller decrease in mean scores between the yes–no and translation tests (6.9%), compared to the single-word decrease (13.5%; see Table 2).

In a further analysis, the scoring formula $h-f$ was used to determine the usefulness of the 10 pseudowords for adjusting yes–no results to better reflect demonstrable knowledge of the tested items (Table 4). For each group (the 21 single-word items, and the 9 multi-word items) applying the $h-f$ scoring formula resulted in adjusted yes–no scores that were closer to the translation test scores than the original yes–no test scores, but considerably closer for the multi-word items. However, the correlations between the translation test scores and the yes–no results were weaker for the multi-words, especially when adjusted by $h-f$.

An item analysis was also undertaken. As can be seen in Table 5, the mean difference between the yes–no scores and the translation test scores for the single words was 27.52, representing 13.4% of the 205 students. Noteworthy items included helmet, the only underestimated single word (-11), as well as monitor and schedule, which had equal scores. The word mechanic was the most overestimated single word on the yes–no test.

Table 3. Direct Comparison of the Yes–No and Translation Tests – Person-by-Person, Item-by-Item

<table>
<thead>
<tr>
<th>Item</th>
<th>Known</th>
<th>Unknown</th>
<th>Overestimation</th>
<th>Underestimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single words</td>
<td>60.1%</td>
<td>19.3%</td>
<td>17.0%</td>
<td>3.6%</td>
</tr>
<tr>
<td>Multi-words</td>
<td>48.3%</td>
<td>27.1%</td>
<td>15.7%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Differences</td>
<td>11.7%</td>
<td>-7.80%</td>
<td>1.30%</td>
<td>-5.20%</td>
</tr>
<tr>
<td>Single–Multi</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Applying h-f Correction for Guessing Formula to Yes–No Test Results

<table>
<thead>
<tr>
<th>Test/items</th>
<th>Mean%</th>
<th>Diff ( __ - Tr)</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YN single-word items (21)</td>
<td>77.1%</td>
<td>13.5%</td>
<td>0.534</td>
</tr>
<tr>
<td>Single-word items $h-f$</td>
<td>68.8%</td>
<td>5.2%</td>
<td>0.387</td>
</tr>
<tr>
<td>Tr single-word items</td>
<td>63.6%</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>YN multi-word items (9)</td>
<td>64.1%</td>
<td>6.9%</td>
<td>0.383</td>
</tr>
<tr>
<td>Multi-word items $h-f$</td>
<td>59.7%</td>
<td>2.5%</td>
<td>0.313</td>
</tr>
<tr>
<td>Tr multi-word items</td>
<td>57.2%</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Note: As the number in each group differs (i.e., 21, 9), means are given as percentages; diff denotes difference; $r$ denotes correlation with Tr scores.
Table 5. Item Analysis: Single words ($n = 205$)

<table>
<thead>
<tr>
<th>Word item (21)</th>
<th>YN single</th>
<th>Tr single</th>
<th>Diff YN-Tr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boring</td>
<td>193</td>
<td>152</td>
<td>41</td>
</tr>
<tr>
<td>Deliver</td>
<td>190</td>
<td>161</td>
<td>29</td>
</tr>
<tr>
<td>Dorm (dormitory)</td>
<td>55</td>
<td>37</td>
<td>18</td>
</tr>
<tr>
<td>During</td>
<td>203</td>
<td>192</td>
<td>11</td>
</tr>
<tr>
<td>Helmet</td>
<td>184</td>
<td>194</td>
<td>−10</td>
</tr>
<tr>
<td>License</td>
<td>178</td>
<td>164</td>
<td>14</td>
</tr>
<tr>
<td>Mechanic</td>
<td>196</td>
<td>43</td>
<td>153</td>
</tr>
<tr>
<td>Monitor</td>
<td>187</td>
<td>187</td>
<td>0</td>
</tr>
<tr>
<td>Qualification</td>
<td>59</td>
<td>33</td>
<td>26</td>
</tr>
<tr>
<td>Saw (noun)</td>
<td>108</td>
<td>29</td>
<td>79</td>
</tr>
<tr>
<td>Schedule</td>
<td>197</td>
<td>197</td>
<td>0</td>
</tr>
<tr>
<td>Shave</td>
<td>144</td>
<td>122</td>
<td>22</td>
</tr>
<tr>
<td>Shelf</td>
<td>142</td>
<td>114</td>
<td>28</td>
</tr>
<tr>
<td>Sink</td>
<td>157</td>
<td>141</td>
<td>16</td>
</tr>
<tr>
<td>Skill</td>
<td>205</td>
<td>197</td>
<td>8</td>
</tr>
<tr>
<td>Strategy</td>
<td>191</td>
<td>166</td>
<td>25</td>
</tr>
<tr>
<td>Textbook</td>
<td>202</td>
<td>201</td>
<td>1</td>
</tr>
<tr>
<td>Tool</td>
<td>199</td>
<td>177</td>
<td>22</td>
</tr>
<tr>
<td>Tough</td>
<td>196</td>
<td>156</td>
<td>40</td>
</tr>
<tr>
<td>Tutor</td>
<td>32</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>Vase</td>
<td>100</td>
<td>64</td>
<td>36</td>
</tr>
<tr>
<td><strong>Mean ($n = 205$)</strong></td>
<td>158</td>
<td>130.48</td>
<td>27.52</td>
</tr>
</tbody>
</table>

Table 6. Item Analysis: Multi-Words ($n = 205$)

<table>
<thead>
<tr>
<th>Word item</th>
<th>YN Multi</th>
<th>Tr Multi</th>
<th>Diff YN-Tr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brush (your) teeth</td>
<td>196</td>
<td>196</td>
<td>0</td>
</tr>
<tr>
<td>Easy to ~</td>
<td>158</td>
<td>142</td>
<td>16</td>
</tr>
<tr>
<td>Fish bowl</td>
<td>57</td>
<td>19</td>
<td>38</td>
</tr>
<tr>
<td>Get dressed</td>
<td>104</td>
<td>99</td>
<td>5</td>
</tr>
<tr>
<td>Get ready</td>
<td>162</td>
<td>150</td>
<td>12</td>
</tr>
<tr>
<td>Get up</td>
<td>197</td>
<td>173</td>
<td>24</td>
</tr>
<tr>
<td>Put on makeup</td>
<td>26</td>
<td>53</td>
<td>−27</td>
</tr>
<tr>
<td>Take turns</td>
<td>84</td>
<td>62</td>
<td>22</td>
</tr>
<tr>
<td>Wake up</td>
<td>198</td>
<td>161</td>
<td>37</td>
</tr>
<tr>
<td><strong>Mean ($n = 205$)</strong></td>
<td>131.33</td>
<td>117.22</td>
<td>14.11</td>
</tr>
</tbody>
</table>

Table 6 presents an item analysis for the multi-words. The mean scores of 131.33 and 117.22 for multi-words on the yes–no and translation tests were substantially lower than for the single-word items (158.00 and 130.48). However, the difference between the yes–no scores and the translation test scores for the multi-words was 14.11, only 6.8% of the 205 participants. In addition, the correlation between the number of correct responses for each target word in the two
test formats was stronger for the multi-word items than for the single-word items ($r = 0.951$ and $0.852$, respectively). Finally, the item *put on makeup* accounted for 21.5% of the underestimation noted in Table 3.

### 4 Conclusion

This has been an investigation into the efficacy of yes–no checklist tests for assessing student knowledge of multi-word units. A yes–no test comprises 21 single-word items, and 9 multi-word items were taken by 205 university students in Japan. Students checked more single-word items than multi-word items on this yes–no test (77.1% and 64.1%, respectively; see Table 2). The yes–no test was then followed by a translation test of the same items. Again participants correctly translated more single-word than multi-word items (63.6% and 57.2%, respectively). However, a greater decrease in scores between the two tests was observed with the single-word items compared to the multi-word items (13.4% vs. 6.9%, respectively; see Table 2). While the single-word items primarily suffered from overestimation on the yes–no test (17.0%) with only a small amount of underestimation (3.6%), the multi-word items suffered from both overestimation (15.7%) and underestimation (8.8%). This underestimation helped to reduce the effect of the overestimation for the multi-words on the yes–no test, resulting in closer yes–no and translation test scores. The application of the $h$-$f$ correction formula to the yes–no test scores resulted in adjusted scores that were much closer to the translation test result for the multi-words than for the single words. While correlations presented in Table 4 favored the single words (0.534 vs. 0.383), the item analysis correlations between the two test formats favored the 9 multi-words (0.951) over the 21 single words (0.852).

The above results suggest that the yes–no vocabulary test format may be able to measure student knowledge of multi-word lexical units as (or more) effectively than single-word units. One weakness of the present study is the small number of items tested, especially the nine multi-word items. Future studies should increase the number of multi-word items to match the number of single-word items.

### References


Comparing the Effectiveness of Word Cards and List Learning with Japanese Learners of English

Magda L. Kitano and Katsuhiro Chiba
Bunkyo University
https://doi.org/10.7820/vli.v08.1.kitano.chiba

Abstract
This study investigated the recall of words learned through two deliberate learning techniques, word cards and list learning. While the literature points to word cards as being more effective, Japanese learners of English are seen to prefer list learning, which may indicate unique learning styles stemming from a non-alphabetic L1. To test the efficiency of the two techniques for Japanese learners, 25 university students of varying English proficiency were divided into four groups. Following the within-subject design, all groups were subjected to both treatments. Twenty low-frequency English words were learned within a 20-min period using one method, and then 20 more words were similarly learned with the other method. Subjects were tested immediately after the treatments, after a 20-min distraction period, and after an interval of 2 weeks. Results from all three testing stages indicated that list learning was more effective than word cards for these students.

Key Words: deliberate learning; word cards; list learning; Japanese learners

Deliberate learning is an important component of L2 vocabulary learning, and the usage of word cards has been shown to be a time-efficient strategy for learning large amounts of words, while also resulting in long-term retention (Nation, 2008, 2013). However, in EFL (English as a Foreign Language) classes in Japan, students are more likely to be seen using the list learning strategy or the repetitive writing of new words in columns, sometimes even after instruction in word card usage. Having developed their L1 reading and writing skills with a phonetic writing system along with logographic Chinese characters (kanji), it could be possible that different strategies may better suit their learning styles than those found to be efficient with learners with alphabetic L1s. This study investigates the effectiveness of learning English vocabulary using the word card strategy compared with list learning for Japanese learners of EFL.

1. Background
As a deliberate learning strategy to complement a vocabulary program, Paul Nation maintains that the usage of cards with the new foreign word on one side and the L1 meaning on the other is superior to list learning, where the learner memorizes new L2 words by looking at a list of those words paired with their
L1 meanings. One reason is that the order of the words cannot be changed on a paper list, so serial learning interferes with the ability to later recall words independently. In addition, there are effects such as the primacy and recency effects, where words at the beginning and end of the list are better learned. Second, word cards force retrieval, and long-term memory is dependent upon the number of times a word is seen and its meaning recalled (Nation, 2013, p. 437–478).

While there have been many studies that test the efficiency of the word card technique and also those that compare deliberate learning via the word card technique with incidental learning (as introduced in Nation, 2013), there have not been studies that directly compare the two deliberate techniques of word cards and list learning. Various studies have compared learner strategies by following learners through a semester or program as they record their vocabulary-related activities by method and times in a log (e.g., Fitzpatrick, Al-Quarni, & Meara, 2008; Sannaoui, 1995). They often look at a variety of aspects of learner behavior and time spent. In some cases, such log studies follow instruction in a particular method (e.g., Elgort, 2010; Pauwels, 2015). Studies that take place in a controlled environment tend to test aspects of vocabulary acquisition rather than strategy effectiveness (e.g., Pyc & Rawson, 2009; Tinkham, 1993).

Yet, while list learning has been dismissed by researchers in favor of word cards, the popularity of list learning among learners is apparent in a variety of studies. Some log studies identify a preference among learners for list learning (e.g., Pauwels, 2012). Surveys also show list learning to be widely used, and Schmitt (1997) found list learning to be the most popular among Japanese learners in terms of both actual usage and learners’ perception of helpfulness, although this preference was seen to drop off with age. In 2016, the authors carried out a survey of 71 undergraduates training to be English teachers. In a multiple response question, 37 (52%) responded that they utilize a list and hide the meanings when learning new English words, 23 (32%) responded that they use a list without hiding the meanings, and only 16 (23%) used word cards.

1.1. Research Question

This popularity of list learning indicates that there may be cause to take a closer look at the effectiveness of list learning compared to word cards. In the classroom, convincing Japanese students to utilize word cards seems to be an uphill battle. Instruction, hands-on trials in class, and presentation of the benefits of word cards over list learning result in only a very few students adopting this strategy. Is this resistance simply due to unfamiliarity with the method and lack of motivation to try something new? Or, are learners instinctively staying with the method that works best for them? This study aims to compare the effectiveness of list learning and word cards in learning new English words.

Research question:

Are word cards a more effective deliberate learning technique than list learning for Japanese EFL learners?
2. Methods

2.1. Participants

The participants in this study were 25 Japanese undergraduate students of a private university in Kanagawa, Japan. They assembled specifically to participate in this study in return for compensation. There were first through fourth year students, all majors in International Studies. CASEC scores ranged from 424 to 786, with a mean of 560.32 ($SD = 97.07$). CASEC (Computerized Assessment System for English Communication) is an online test of English communication skills, and a score of 550 is said to have equivalency to a TOEIC (Test of English for International Communication) 465 score (Data & Information, n.d.).

2.2. Materials

Word lists provided for participants to learn were sourced from the 10th 1000 word list of Paul Nation’s BNC/COCA Headwords of the first 10,000 words (Nation, 2017) in order to maximize the chances that students had not encountered the words previously. Only nouns were used, and culled from the list were words that were probably familiar to Japanese students, those that would be translated as loanwords or are used in Japanese as false friends, those that are inappropriate for a school setting or shocking (sexual, violent, death, illness), words whose Japanese translations may not be familiar to the students, and rare specific biological or scientific items (plants, animals, body organs, chemicals, and terms from geology and astronomy). After excluding words of less than five and more than eight letters in length, 209 words remained.

Four lists of 20 words each were made by choosing an equal number of words of each length (five through eight letters), and from a variety of starting letters of the alphabet. In addition, words of highly similar form and meaning were avoided within the same list (such as audacity and tenacity).

Japanese definitions provided to the students were taken from Advanced Favorite English-Japanese Dictionary (Asano, 2002). The most prominent noun definitions of the words were chosen. If there was a choice between a loanword and a Japanese word as a definition, the Japanese word was taken.

2.3. Procedures

Procedures followed the within-subject design, where all participants are subject to both treatments. A resulting order effect (Nation & Webb, 2011, p. 35) was controlled for by dividing the participants into two groups who then experienced the treatments in a different order. The following procedure was carried out on two different occasions with 12 and then 13 different participants, using four different lists of words to learn. On each occasion, the participants were separated into two groups, resulting in a total of four groups of six, six, six, and seven participants.

On the first meeting, each of the first two groups was given a list of 20 words to learn within 20 minutes. After vocal introduction of the words, pronunciation only, students in one group were told to utilize the list in learning the words. They were
given cardboard to cover up the words and paper to practice writing the words if they so wished. The other group was given prepared word cards, and the list was taken away. They were told to first go through the pack until they could recall the Japanese meanings of each word, then go on to the reverse while practicing the spelling from memory, as outlined in Nation (2013, p. 458). They were told to shuffle the cards after each round, and to utilize a drop-out schedule (put aside learned cards).

After the 20-minute study period, students were given a recall test where they wrote the English words in response to the Japanese meanings. The groups were then reversed to undergo a second study period and recall test, this time using the list and method of the other group. At this point, both groups had experienced the same methods paired with the same lists, yet in a different order.

The two groups were then brought together for a 20-minute distraction activity that did not involve the use of English. At the end of that activity, they were given an overall test of all 40 words they had learned that day, the two lists mixed together in random order. Two weeks later, the participants assembled and took a second mixed test. The meeting had been set in advance, but participants were not told that they would be tested.

This entire procedure was then repeated on a different date with 13 new members and two new lists of words.

3. Results

Results at all three stages indicated that list learning was more effective than word cards for these students (Table 1). Tests had a maximum of 40 points for 20 words. Two points were given for each correctly spelled word, and one for each word with one spelling mistake which did not change the meaning of the word. The means from words learned by list learning are higher for immediate, after 20-minute distraction, and after the 2-week interval, although for the latter very few words were recalled from either method of learning.

4. Discussion

Although these results are far from conclusive, preliminary results saw higher retention from list learning over word cards. The word card strategy utilized in this study differs from Nation’s guidelines (2013:446–468) in that students

Table 1. Descriptive Statistics of Posttests

<table>
<thead>
<tr>
<th>Retention interval</th>
<th>Method</th>
<th>Mean</th>
<th>SD</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immediate</td>
<td>List Learning</td>
<td>34.12</td>
<td>7.623</td>
<td>85.3</td>
</tr>
<tr>
<td></td>
<td>Word Cards</td>
<td>26.40</td>
<td>10.532</td>
<td>66.0</td>
</tr>
<tr>
<td>After 20-min distraction</td>
<td>List Learning</td>
<td>27.28</td>
<td>9.423</td>
<td>68.2</td>
</tr>
<tr>
<td></td>
<td>Word Cards</td>
<td>19.80</td>
<td>10.352</td>
<td>49.5</td>
</tr>
<tr>
<td>2-Week delay</td>
<td>List Learning</td>
<td>5.92</td>
<td>5.937</td>
<td>14.8</td>
</tr>
<tr>
<td></td>
<td>Word Cards</td>
<td>3.76</td>
<td>5.262</td>
<td>9.4</td>
</tr>
</tbody>
</table>

Note: Maximum score = 40, n = 25
do not choose the words themselves, do not make their own cards, and do not review the cards at increasingly spaced intervals, or in fact at all after the 20-minute study period. However, if word cards themselves are in fact superior to list learning, more ambivalent results would be expected, if not those in favor of word cards.

Of course word learning does not usually happen in a controlled environment, and one advantage of word cards is that the act of making them and their availability for quick review throughout a day or week can make vocabulary learning more enjoyable. This in itself may be a good reason for a teacher to provide instruction in word card learning. In the classroom, effectiveness is only one aspect of a study method that needs to be considered. If a task is enjoyable, learners will be more inclined to spend more time on it, increasing learning. If the making of word cards and their usage is favored by students, then whether they are spending more time per word learned may not be important.

However, Japanese learners may not have the same preferences. In a questionnaire taken at the conclusion of the first meeting of this experiment, 16 of the 25 students responded that they preferred list learning to word cards, while only 6 preferred the word cards after having just experienced both methods. If Japanese learners prefer list learning and if they are finding more success utilizing it, then there may be no reason for teachers to encourage reluctant students to change to the word card strategy.

There is a possibility that list learning is not as ineffective as has been assumed, although that could not be evaluated in this study. Another possibility is that Japanese learners may be particularly receptive to list learning due to their L1 background, where repetitive writing is mainly used to learn over 2000 kanji characters. Due to the small sample size and limitations in the methodology of this study, further experimentation is called for in clarifying this issue.

5. Limitations and Further Research

The fact that different word lists were used for the two methods prevented a statistical analysis of the differences in effectiveness of list learning and word cards. In the next stage of this study, we will test a larger number of participants utilizing common lists learned by different participants using different methods.

Further studies that can be envisioned from these results include looking into whether the results here are due to students' background in a non-alphabetic L1. To do so, similar studies could be held with learners of different L1 backgrounds. For example, one variable may be the methods participants had used for their L1 vocabulary learning.

In addition, as noted above, there were several aspects of word card learning that departed from Nation's guidelines due to the 20-minute limit for learning time. In order to re-examine the effectiveness of word cards in comparison to list learning, a longer-term study would be necessary, including aspects such as word choice and spaced repetitions. Further, the low overall recall in the third test prevented a comparison of long-term retention. Testing for recognition rather than
recall at that point, shortening the interval, or providing a practice period midway between the second and third tests may result in more usable data.

References


Foreign Language Vocabulary Learning and the Type of Processing-Resource Allocation Model

Shusaku Kida
Hiroshima University

Abstract
This paper argues the nature of adult foreign or second language (L2) vocabulary learning and points out the importance of word-form learning at the initial stage of vocabulary development. The Type of Processing-Resource Allocation (TOPRA) model is introduced as a prominent theoretical framework to capture it. Finally, some future studies are proposed to reinforce TOPRA-based studies with respect to (1) types of vocabulary processing tasks and (2) dimensions of vocabulary development.

Keywords: L2 vocabulary learning, form-meaning connection, the type of processing-resource allocation model, semantic and structural processing, acquisition of word knowledge and fluent processing

1. Introduction
It has been acknowledged that the quality and quantity of processing play crucial roles in foreign or second language (L2) vocabulary learning (Hulstijn, 2001). The quality of processing refers to how learners process unfamiliar words (e.g., semantic processing, structural processing), whereas the quantity of processing refers to how many times they process the same words (e.g., exposure frequency). Numerous studies have examined the effects of the two factors (e.g., Eckerth & Tavakoli, 2012; Joe, 2010; Laufer & Rozovski-Roitblat, 2011, 2015; Paribakht & Wesche, 1999; Schmitt, 2008). Yet, the discussion is ongoing, and there is no universally accepted resolution.

This paper focuses on a specific theoretical account of the quality of processing: the Type of Processing-Resource Allocation (TOPRA) model (Barcroft, 2002, 2003). By looking at previous studies, some possible future studies will be proposed to reinforce the TOPRA model and related studies.

2. The Nature of Foreign/Second Language Vocabulary Learning
Foreign language vocabulary learning entails at least the following three aspects: word form, meaning, and form–meaning connection (Barcroft, 2003). When it comes to adult foreign language learning, language learners usually already have the semantic representations of most L2 words through their L1 acquisition experience, which helps them understand the meanings of new L2 words
using the L1 translation of the word in question. Therefore, at the beginning stage of L2 vocabulary development, the primary task is not to learn meaning itself, as the meaning of the word could be understood by the L1 translation (Ellis, 2004; Jiang, 2000; Laufer, 1997; Nation, 2001), but rather to learn the target word form and form–meaning connection.

Another aspect that L2 learners need to develop is fluency. Language learning is not limited to simply memorizing word knowledge mentioned above (i.e., word form, meaning, and form–meaning connection); we also need to be able to use the language fluently. In the field of second language acquisition, Bialystok and Sharwood Smith (1985) once proposed the distinction between language knowledge and its control. The former was explained as “the way in which the language system is represented in the mind of the learner,” while the latter was referred to as “the processing system for controlling that knowledge during actual performance” (p. 104).

Similar distinctions have been proposed in L2 vocabulary learning research, such as the distinction between lexical knowledge and lexical competence (Jiang, 2000) and between explicit lexical knowledge and tacit lexical knowledge (Elgort & Warren, 2014). These proposals emphasize the importance of both the acquisition of word knowledge and the fluent processing of words. For example, according to Jiang, lexical knowledge refers to “the knowledge or information an L2 learner remembers about the form, meaning, grammatical usage, and sociolinguistic use of a word that is stored in a general memory system, rather than integrated into the lexical entry of a word” (p. 65), whereas lexical competence refers to “the semantic, syntactic, morphological, and formal knowledge about a word that has become an integral part of a lexical entry in the mental lexicon and can be retrieved automatically in natural communication” (pp. 65–66). Elgort and Warren (2014) also emphasized the distinction between explicit knowledge (which is measured in conventional tests such as multiple-choice tests or translation with some sort of controlled retrieval of words) and tacit knowledge (which is measured in word processing tasks such as lexical decision and underlies fluent processing of words). Therefore, it has been assumed that successful L2 vocabulary learning includes not only the acquisition of new word knowledge, such as word form and meaning, but also the acquisition of fluent and efficient processing of newly learned words.

As these distinctions show, it is common to assume that L2 vocabulary acquisition consists of multiple stages or aspects. Therefore, in L2 vocabulary learning research, it is crucial to examine vocabulary learning from various perspectives using different types of measurement.

3. The Type of Processing-Resource Allocation Model

The levels-of-processing (LOP) framework (Craik & Lockhart, 1972; Craik & Tulving, 1975) has been widely used to account for the role of the quality of processing in L2 vocabulary learning research (Laufer & Hulstijn, 2001). The LOP framework assumes that some levels or depths of processing exist, and that deeper information processing leads to better memory performance. According to the original LOP framework, semantic processing was assumed to be deeper than
orthographic or phonological processing, and therefore should facilitate memory of the target information.

On the other hand, the TOPRA model proposed by Barcroft (2002, 2003) assumes that semantic processing does not necessarily help new L2 word learning. According to this model, L2 learners’ processing-resource allocation to L2 word form or meaning corresponds to their learning performance. For example, when L2 learners allocate their processing-resources to unknown L2 word meaning, the likelihood of remembering the (L1-based) meaning increases. Simultaneously, learning of structural aspects, such as word form, is inhibited when overall processing demands are sufficiently high. On the other hand, when L2 learners allocate their processing-resources to the structural aspect (e.g., word form) of the word in question, the possibility that they learn it (e.g., the new L2 word form) increases. The TOPRA model predicts this kind of trade-off between different aspects of L2 vocabulary processing and learning. A series of studies by Barcroft (2002, 2003, 2004, 2009) demonstrated that the TOPRA model offers a better explanation than the LOP hypothesis for adult L2 vocabulary learning.

3.1. Vocabulary Processing Tasks in Type of Processing-Resource Allocation Studies

Although the prediction of the TOPRA model has been tested by previous studies, some criticism may be possible. One criticism may be of the ecological validity of vocabulary processing tasks when it comes to foreign/L2 pedagogy in L2 classrooms. For example, Barcroft (2002) presented new L2 Spanish words and their pictures to learners. They were instructed to learn these word–picture pairs in the following conditions: semantic processing, structural processing, and control. In the semantic and structural processing conditions, the participants were asked to remember new L2 words and their pictures while making pleasantness ratings about the meaning of each word as well as while counting the number of letters of each word. In the control condition, there was no such additional task; learners were simply instructed to remember the word–picture pairs. The semantic and structural processing tasks were originally based on LOP research (e.g., Hyde & Jenkins, 1969) in psychology and were not necessarily intended to be used in L2 pedagogy. Therefore, one may argue that the results of the previous studies do not have direct implications on L2 classroom practice, and that studies with more pedagogically applicable tasks would be necessary when the TOPRA model predictions are applied to foreign/L2 pedagogy.

3.2. Aspects of Vocabulary Development in Type of Processing-Resource Allocation Studies

Another possible critique is that the aspects of vocabulary acquisition examined in previous studies are limited. So far, the predictions of the TOPRA model have been tested using the following four measurements: L1 free recall, L2 free recall, L1-to-L2 cued recall, and L2-to-L1 cued recall. They assess whether the L2 learners remember target word forms, meanings, and form–meaning connections.
For example, after the L2 Spanish word learning session described above in the study by Barcroft (2002), the participants took these four measurements. In the two free recalls (L2 and L1 free recalls), the participants were asked to recall L2 target words or their pictures in L1, respectively. On the other hand, in the two cued recalls, they were asked to produce L2 target words for the presented L1 words in the case of L1-to-L2 cued recall, and vice versa in the case of L2-to-L1 cued recall. It was assumed that L1 and L2 free recalls are relatively pure measurements of the effects of semantic and structural processing, respectively. If the prediction of the TOPRA model is valid, the semantic processing condition should outperform the structural processing condition in the case of L1 free recall, whereas the opposite result should hold good in the case of L2 free recall. The results of Barcroft (2002) were consistent with this prediction. However, as argued above, language acquisition is not limited to simply acquiring new word knowledge such as word form or meaning; it also includes the acquisition of fluent and efficient processing of the newly learned words (Elgort & Warren, 2014; Jiang, 2000). The four types of recall used in the previous studies seem to measure only acquisition of new word knowledge. Therefore, it is not clear whether the TOPRA model prediction is applicable to the acquisition of fluent processing of learned words.

4. Some Possible Future Directions

Thus far, the present paper has described the nature of adult foreign/L2 vocabulary learning and points out that at the beginning stage of L2 vocabulary development, the learning of word form is more important than the learning of meaning. The TOPRA model captures this by proposing a trade-off between the different types of processing and learning as a result of processing-resource allocation. We can further reinforce the results of studies on the TOPRA model by (1) examining the effects of other types of vocabulary processing tasks and (2) assessing other aspects of vocabulary development. Figure 1 shows areas for possible future studies.
study of the TOPRA framework in the two dimensions of (1) types of vocabulary processing tasks (vertical) and (2) aspects of vocabulary development (horizontal).

According to this framework, as Figure 1 shows, previous studies regarding the TOPRA model only focused on one of four possible research areas. Future research is therefore possible in at least three remaining areas.

If one focuses on tasks directly applicable in the L2 classroom (the top two research areas in Figure 1), it is necessary to examine the effects of other types of vocabulary processing tasks which have not been introduced in previous studies, yet induce semantic and structural processing. One example of such a task is a word association task. For example, it is possible to ask L2 learners to choose a word that is semantically or structurally similar to the target L2 word. When they are instructed to produce/choose semantically similar words, they must allocate their processing-resources to the semantic aspect of the target words, whereas when they are instructed to produce/choose structurally similar words, their processing-resources are allocated to the structural aspect. After the learning session, the same vocabulary posttests (e.g., L2 and L1 free recall, L1-to-L2 and L2-to-L1 cued recall) as in previous studies could be used as a first step.

On the other hand, if one focuses on the acquisition of fluent processing of newly learned L2 words (the right two areas in Figure 1), it is necessary to use measurements that could separate the processing of word form from its meaning. One such example is the so-called lexical decision task (LDT). Since lexical decision does not require semantic processing of the target, it would be an interesting test of the TOPRA model to examine how different types of vocabulary processing affect the accuracy and reaction time of the lexical decision regarding new L2 words. In this case, the same vocabulary processing tasks (e.g., pleasantness rating, letter counting) as previous studies could be used as a first step. These potential studies would be conceptual replications of previous studies in the TOPRA framework and should demonstrate the range of applicability of the TOPRA model.

Kida (2018a, 2018b) addressed these two issues; the preliminary analyses demonstrated that (1) the results of Barcroft (2002) were replicated when the word association tasks mentioned above were used and (2) structural processing of new L2 words led to more accurate and faster reaction times than semantic processing in the LDT. These studies demonstrated the possibility that the notion of the TOPRA model may be applicable in the L2 classroom and suggested that the prediction of the TOPRA model may be valid not only for the acquisition of word knowledge (i.e., word form, meaning, or form–meaning connection) but also for the acquisition of fluent processing of newly learned L2 words.

References


Laufer, B. (1997). What’s in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In N. Schmitt & M. McCarthy (Eds.),
Vocabulary: Description, acquisition and pedagogy (pp. 140–155). Cambridge University Press.


Vocabulary Instruction and Learning: A Commentary on Four Studies

Jenifer Larson-Hall
University of Kitakyushu
https://doi.org/10.7820/vli.v08.1.Larson-Hall

Abstract

Four papers were presented by Shusaku Kida, Magda Kitano and Katsuhiro Chiba, Tim Stoeckel, and Raymond Stubbe. In this article, my thoughts about the issues raised in these thought-provoking papers about vocabulary resource processing allocation and the Type of Processing Resource Allocation (TOPRA) model, the use of flashcards in vocabulary learning, an empirical evaluation of the New General Service List (NGSL), and the use of yes–no checklists for multi-word units are given. The first paper provided some theoretical underpinnings for thinking about vocabulary acquisition, while the last three papers were valuable for their extremely timely and practical examination of issues that are highly important for the acquisition of vocabulary.

I have four children, all bilinguals. Raising my children (and I am not done yet) is the hardest but most rewarding work that I have done in my life. You might wonder, though, why I mention raising children in the context of discussing the four papers we have heard today. I mention it because to raise four children without going crazy means that you must be extremely organized and practical, and I like to think I am. In the same vein, I like practical research that has details about the study methodology, concrete numbers and results, and I like to ask practical questions that relate to how research can apply to real-life problems.

I am pleased and honored to be asked to comment on these studies today. All of them are very interesting. Three out of the four studies I will discuss today are empirical studies and can be evaluated on the strength of their conceptual design, methodology, and results. The fourth study, Shusaku Kida’s paper, “Foreign Language Vocabulary Learning and the Type of Processing-Resource Allocation Model,” is not an empirical study but rather a summary of a type of approach to studying vocabulary learning, namely, the TOPRA model. Kida’s paper gave me food for thought and got me the most engaged in thinking about a number of overall issues.

I actually hope that the issues I bring up here, inspired by all of the papers, will be very practical issues that we can think about.
1. Practical question 1: How can we judge that a word has been learned?

The biggest issue I came away with was how to judge that a word has been learned. The TOPRA model that Kida examines proposes that at least at the beginning stages of learning a language, “the learning of word form is more important than the learning of meaning.” Kida is arguing that since the meaning of the word in the L1 is already known, what needs to be learned is not the meaning of the word itself, but the mapping from the L2 word (and its form) to the L1 word. Kida invokes Barcroft’s TOPRA model that says there will be tradeoffs in processing ability so that trying to add another component to the memory exercise of forming those L1–L2 mapping links, such as a semantic evaluation component (like evaluating whether the word is pleasant or not) or a structural evaluation component (such as counting the number of letters in the word), is going to take away from, rather than add to, the ability to make that mapping link. Kida and Barcroft (2017) make the pedagogical recommendation to not add any extra processing demands such as these when asking language learners to learn new words. They also make the point that learning a word is a process that develops in stages.

Of course, the question of how we judge whether a word has been learned is not limited to Kida’s paper. Magda Kitano and Katsuhiro Chiba also explore the question of what kind of learning is the most effective way for Japanese learners to acquire vocabulary in their paper entitled “Comparing the Effectiveness of Word Cards and List Learning with Japanese Learners of English.” Of course, to answer the question we must also think about what it means to have acquired or learned vocabulary. The answer to this question will dictate what kinds of evidence we will find convincing to answer questions like whether “list learning” or “word card (flashcard)” learning is more effective.

Would we like to say that someone has learned a word if they can correctly give an L1 or L2 translation to a word immediately after studying the word? I think none of us would like to say that this is acquisition, because acquisition seems to mean it is something that we can use in the course of our daily life. We do not talk about “acquiring” an ice cream cone, because that is quickly bought and then eaten, and not available to us anymore. In the same way, we all know that with the decay of memory, vocabulary words in a second language that are studied at only one point in time and not rehearsed will fade away quickly too, melting in the heat of disuse like ice cream in the hot sun.

Kitano and Chiba did not ask us to accept that a one-time encounter with the words they tested meant language acquisition. They very responsibly went back and tested the students after a 2-week delay and found that the students had not forgotten everything, but they had forgotten most of what they had known. Since Kitano and Chiba deliberately picked less frequent words to test that they supposed their students would not know, and since that probably implies that the students did not encounter those words again, it is not surprising that they forgot most of what they had been able to produce immediately after learning the words and even after a 20-min distraction on the day of testing.
I am sure almost all of you have seen Ebbinghaus’ (1885) forgetting curve. If you recall, Ebbinghaus tested himself with nonsense strings of 13 randomly generated CVC syllables per list (so, e.g., one word might be “kat-bup-set-mak-tip-lum-kak-vur-zip-lur-kul-wam-lal”). He memorized these “words” until he could remember them perfectly, then tested himself at different time periods to see how much he forgot. The results showed a big drop-off in ability at the beginning of the time period, but then the forgetting rate flattened out to an asymptote at a low percentage; however, it did not dissolve to zero. I would like to show you a graph I made that shows the forgetting sequence of a study by Bierling (1990), who taught students 100 less frequent words during their first 20 hours of an intensive Spanish class but then never mentioned the words again. They were words like madrastra “stepmother.” Bierling then tested the students at different points to see how well they recalled these words that they did not use again (she ensured they did not by excluding words they said they had had contact with in the interim). This is the data for the productive test, and each line represents the data of one person. The y-axis ranges from 0% to 100%, and the x-axis shows the time in weeks [five data points were tested: 0 (after the initial 20 h), 3.5, 7.0, 49, and 90] (Figure 1).

Figure 1. A parallel coordinate plot of Bierling’s (1990) productive data.
As you can see, most people were not able to produce more than 50% soon after the end of the semester, but even about a year later, this information had stabilized so that many people retained a little bit. So, this is similar to Kitano and Chiba’s study, in that most of the students had forgotten these words that they had only spent a little time learning and had not subsequently used.

In fact, Kitano and Chiba did not include a figure in their study, but I have created a figure, similar to that of Bierling, which shows individual scores at the two testing periods for the group who learned with flashcards and the group who just looked at a list (Figure 2). By the way, here is the R code for the two graphs in Figure 2:

```r
library(ggplot2)
ggplot(KitanoList, aes(Time, List, group=Student)) + geom_line(aes(color=Student))+geom_text(aes(label=Student), size=2)+theme(legend.position="none") + ggtitle("List Learning")
ggplot(KitanoList, aes(Time, List, group=Student)) + geom_line(aes(color=Student))+geom_text(aes(label=Student), size=2) + +theme(legend.position="none")+ggtitle("List Learning")
```

Here we see that for the list learning group, most of the students are still retaining most of the information after 20 min, while for the flashcard learning group, many are starting to lose ground. Actually, the list learning group started out with more even at the 0 point. So there is the point that the list learning group did better than the flashcard group, which I will return to. But the first point I want to make is that these data look very similar to Bierling’s data. In other words, this study does not seem to be so much about language acquisition as about language forgetting. So maybe what is most important is not really the method you use to first learn a word, it is really the method that you use to
retain a word, because retaining a word in order to use it in your daily life is what we are really interested in.

Coming back to the point that the list group did better than the card group, I was actually surprised that Kitano and Chiba did not find any difference between these two methods. If the methods were just to have a list of words to learn that could be covered up and guessed at as the student moved down the list, and to have flashcards that would also be guessed at, there really should not be much difference between the methods. However, Kitano mentioned (personal communication) that the list learning students were able to use a piece of paper and a pencil to write down the words, and when I asked her about it, she told me that in fact, all of the students had used the paper to practice writing the words as well. My initial thought was that since Kitano and Chiba’s test was a production test, perhaps the list learning students were able to do better on the test than the flashcard group due to practice with production rather than superiority of list learning over flashcards. However, later Kitano mentioned that the flashcard students were able to use a paper and pencil to practice writing down the words as well, and again, all of them did, so that explanation does not appear to be valid. I am therefore left with no explanation for why list learning works better than flashcard learning in this study, and explaining why the list learning participants surpassed the flashcard participants could be an interesting idea for the authors to pursue. It would of course also be good to see the study replicated with different populations and circumstances.

So I guess my main thought about Kitano and Chiba’s study is that it cannot really tell us much about acquiring a vocabulary word. Of course, this study does not stand alone in using this methodology. It is the methodology of numerous research designs for vocabulary, and in that sense, I do not want to single it out for any special mention of being problematic. What I would like to do is to note that the entire field of vocabulary studies ought to examine this issue very thoughtfully. Unless we want to admit that we are not interested in how students actually acquire vocabulary words and subsequently incorporate them into their lexical system, a research methodology of conducting a short-term treatment and then doing an immediate test and even a slightly delayed posttest should not be the way we go about investigating this question. Of course, proposals have been made regarding this issue (Jiang, 2000 discusses some of them, but that is beyond the scope of this article).

It may be much more difficult to teach students vocabulary that they indeed acquire than we have been heretofore assuming. A very informal study I did in a 15-week English reading class comes to mind. The textbook I used asked the students to learn 20 words a week, and before the beginning of the semester, I gave the students a yes–no test to see how many of the words that were going to be studied they already knew. I then tested the students on their vocabulary words every week with a 10-item test, using a fill-in-the-blanks format. Students were encouraged to review words by having a midterm that tested all the words up to that point and a final vocabulary test where students were told they had to know all of the words from the entire semester. Then, at the end of the semester, I administered the same yes–no test. My finding was that most students gained about three or four words over the course of the semester! This finding was astounding to me.
in its scarcity of learning and made me highly skeptical of how much vocabulary could/would be learned by students in the long run, even if they performed well in the short run.

2. Practical Question 2: What vocabulary words should we study?

At the end of his paper, Kida brings up a number of possibilities about how the TOPRA model could be applicable in answering practical questions about vocabulary acquisition. Kida suggests that different kinds of language processing could affect both the effectiveness of ways of learning words and the fluency with which they can learn to use those words. This is a very intriguing line of thinking, and clearly, more research should be done in this area.

So in thinking about how we should learn words, we are first faced with the choice of what words to learn. In Tim Stoeckel’s paper, “An Examination of the NGSL,” we find here again another highly practical implied suggestion for language learners, which is to use the NGSL. I was actually surprised to learn that this list is based on a British corpus (the Cambridge English Corpus), so matching up a section of Corpus of Contemporary American English (COCA) (Davies, 2008) and the NGSL was a very good idea. Stoeckel finds that the NGSL provides 4.32% better coverage than the General Service List (GSL) (83.66% vs. 79.34%), meaning that knowing the words from the NGSL will buy you a higher percentage of words from the 6-year section of COCA that was examined than if you only knew the words in the GSL. This is even more impressive given that the number of lemmas in the GSL is reported (at the NGSL website: www.newgeneralservicelist.org) to be 3623, while the number of lemmas in the NGSL is 2818. Higher coverage with fewer lemmas is quite a feat! And of course, Stoeckel tried the comparison out on a new corpus that neither one of the lists had been taken from, making it seem to be a very fair comparison.

So in practical terms, the NGSL seems to be a wonderful tool for students to use and for teachers to recommend for learning vocabulary. At the moment, the NGSL is available in Quizlet stacks for flashcard learning. Quizlet provides Spaced Repetition System (SRS) reviews but only if you pay the yearly fee, which is about $20. I know a lot of students who learn Japanese like Anki for flashcards, and Anki has an SRS function as well, but it is free, and I see online that there are NGSL stacks for Anki as well.

Many of you are aware as well that there are analogues to the NGSL in Gardner and Davies’ Academic Vocabulary List (AVL) (no date) and Brown, Culligan, and Phillips’ (2013) New Academic Word List (NAWL). Both provide better coverage of academic corpora than the original Academic Word List created by Coxhead (Gardner & Davies, 2014). If your students are fairly advanced and/or would benefit more from studying academically oriented words, then this list would also be a wonderful place for them to start. There are many resources for both lists. Emily Crandell’s MA thesis (2017) details her process of creating Quizlet flashcards for the first 500 words of the AVL with translations into seven languages, including Japanese (also Mandarin, Korean, Arabic, Spanish,
Portuguese, and Russian). There are links to flashcards in Quizlet with English definitions and in Memrise with Japanese definitions on Browne’s NGSL site (www.newgeneralservicelist.org, then navigate to the NAWL).

I would like to suggest that any of you with free time might create a paid app for learning these words. I am personally using an app to learn kanji called WaniKani, and I find it much more compelling than my Quizlet cards. Why? I am not sure, but maybe because the program is all laid out for me and I just have to go in and spend 15–30 min every morning and night, and it feels like a game where I am progressing, rather than just a collection of decks of words. If you do not spend time on it, the work piles up so you feel you must do it every day to keep up. There are five levels and you move up or down the levels as you correctly or incorrectly input the words. With Quizlet if you use the SRS function, you do get some feedback about your overall progress, but my Quizlet suggests I learn only 10 words at a time. With WaniKani, it is like this train is moving on, you are going to have more words piling up every day if you do not keep up, and so it is more motivating to me. Also, with Quizlet there are a large number of stacks instead of having everything all in one program as I find in WaniKani. Perhaps the motivation of thinking you have to get your money’s worth for something you paid for also factors in.

And some people might think that students would not want to study really basic words, but even though I speak Japanese pretty well, I have not minded doing WaniKani from the beginning because each level has a number of words that I never learned, or only have half-learned, so I felt even from the beginning it was helpful for improving my Japanese (Figures 3 and 4).

Figure 3. A screenshot of the WaniKani program showing how many words I have learned in each of the five spaced repetition intervals (the longest being the “burn” level, which takes at least 6 months to reach).
3. Practical Question 3: Are yes–no vocabulary tests really as good as they seem?

Yes–no vocabulary tests have been around for a long time. There have been many tests and much work on verifying that this very simple format might be a good one to test vocabulary.

Raymond Stubbe’s article, “Evaluating the efficacy of yes–no checklists tests to assess knowledge of multi-word units,” is a very interesting look into this area. It seemed very concrete and practical to me, as I am currently working on a Japanese yes–no vocabulary test myself. Stubbe very astutely points out that there have been no checks on possible underestimation in yes–no tests, so I read very anxiously to find out the results of his test. He investigated the yes–no results by comparing them with an L2–L1 translation test, which seems to be a very direct way of testing whether the participants know the words. If I were him, I would justify this as the most reasonable and logical way of testing whether participants know the words, rather than saying it was the fastest measure. In fact, it seems funny to have to justify that translation tests are good indicators of which words students can understand receptively, since that is exactly what the translation test does, but I suppose justification must be made.

I think the biggest drawback to this research study is that it tries to look at too many issues but cannot really do all of the issues justice in the small amount of time that is available. I see in fact three very important issues raised in this paper:

1. How much of a problem is underestimation of knowledge?
2. Is a yes–no test a good tool for looking at knowledge of multi-word items?
3. To what extent can correction for pseudowords help adjust for the difference between reported knowledge and actual knowledge?
I think all of these are important questions and deserve a thorough inspection. However, in his introduction, Stubbe seems to be focusing only on question 2, which is concerned with the efficacy of using yes–no checklists for multi-word items.

Stubbe notes that the issue of underestimation in yes–no tests has not been investigated previously. Researchers have worried more that students will overinflate their knowledge and say they know words that they are really not able to define. Stubbe, on the other hand, worried that students may indicate that they do not know words that they actually can give a definition for. However, his results showed, at least in the aggregate for single words (which have been the only kind of data used previously), that students responded with an “yes” to 77.2% of the 21 single words, while they were only able to correctly give a translation equivalent for 63.7% of those words, meaning they overestimated their knowledge instead of underestimating it. However, in looking at individual cases, Stubbe did find some instances where students said “no” on the yes–no test but then did in fact give a correct translation – 3.6% of the total number of answers. I would assume that this level of mismatch is simply a careless mistake, something like what Chomsky called the difference between competence and performance for language. From this study, I would conclude that we do not really have to worry very much about underestimation, and perhaps further study is not really necessary, but I may be wrong about that. The study was small and further investigation may reveal more pervasive underestimation.

For the issue of multi-word units, my feeling is that this issue is not really answered clearly enough in this paper to make strong conclusions. The whole area of multi-word units is clearly gaining attention in the field but is still nebulous, and not a lot of research traditions exist. Investigating knowledge of multi-word units or formulaic sequences with a yes–no test (and a translation test) is a really interesting idea, but I am not sure why single words and multi-word units were put into the same study. Stubbe said he wanted to compare them, but I am not sure there is anything comparable about them. We have different numbers of items, which Stubbe noted as a problem, and fixing that would be good, but probably it is more important to ascertain whether the words are of the same frequency level. We can talk about frequency level for words, but I am not sure we have the tools yet to actually talk about frequency levels for phrases. Presumably, different processes may be acting on words versus phrases, and in fact, Stubbe makes different recommendations about the use of yes–no tests and adjustments for false alarms with them. Likewise, the translation test gives the nearest evidence we can get to the actual recognition ability of the words, and we see that the correlation between the yes–no test and the recognition test is, on the surface of it, somewhat higher for single words than for multi-word units (0.53 vs. 0.38 of Stubbe’s paper from Table 4, unadjusted scores), although this difference narrows when scores are adjusted for false alarms (0.39 vs. 0.31).

I guess I am also worried that yes–no tests even used with single words have a number of problems that we should not forget. I know they have been used successfully in many cases, but a little looking can uncover some unsettling research on yes–no tests. For example, Stubbe noted that Eyckmans (2004) also compared L2-to-L1 translation as a criterion measure to examine the
validity of yes–no tests, but I was wondering what the results of that study were. This paper was Eyckmans doctoral thesis, and she looked at the yes–no test for French learners of Dutch. Eyckmans actually did seven different experiments because she found a lot of unreliability in the yes–no test and low correlations with the translation test. Eyckmans states that “whenever a substantial false alarm rate is encountered in the data, this is evidence of the fact that the participants perform a different task than is expected of them hence the validity of the test becomes doubtful. In these cases, cases, the competence that is measured with the Yes/No Test does not correlate well with word recognition as measured by a translation task.” I am afraid the testing issues Eyckmans brings up are too complicated to discuss here, but we should keep in mind that although yes–no tests do seem to work well sometimes, they also seem to have some serious drawbacks as well. So the fact that Stubbe found some substantial false alarm rates also has me worried that the yes–no test may not have worked well enough in this administration of his test to measure even single words, much less multi-word units.

The last issue that came to the forefront here was the third one I mentioned, which is whether adjustments for false alarms can effectively reduce the difference between reported knowledge and actual knowledge, and I think that is a difficult question that deserves further consideration. Again, Eyckman’s (2004) thesis does explore this in depth and opines that more sophisticated techniques might need to be applied to the data in order to better adjust for false alarms. Clearly that was not done with these data but if recommendations are going to be made about adjustments with one type of data over another, I think it might be worth a look.

If we have a moment to indulge me, I would like to bring up just one more issue in regard to this study. That is the issue of outliers. Stubbe noted that one participant chose 11 out of the 15 pseudowords, so he was removed from the data set. Stubbe also did some analyses with only 10 of the 15 false alarm words included. I know it is very tempting to the data of (participants) whose results look very abnormal, but statisticians say that just removing what seem like outliers by using our own judgments will result in a loss of independence in the data. It is better to use a principled approach to removing outliers, which involves robust estimators. So I made up a scatterplot of Stubbe’s data, comparing the scores on all 30 of his items for the yes–no test versus the translation test. Here is the scatterplot and the R code I used (Figure 5):

```r
scatterplot(Tr.30.Score~YN.Scores.30, regLine=FALSE, smooth=FALSE, boxplots='xy',
jitter=list(x=1, y=1), xlab="Yes-No test score", ylab="Translation test score",
main="Stubbe data for all 30 items and all participants", xlim=range(c(0,30)), ylim=range(c(0,30)), cex=1.6,
data=Stubbe)
```

Now looking at this, we can see there is definitely a positive relationship between scores on the yes–no test and scores on the translation test, although
we do not have data along the whole length of our regression line since most students were near the upper end of the scale (that might be another issue in this test). I am not sure if the person who had 11 false alarms is in these data or not, but if you look at these data, you will probably think that you can identify some outliers.

Next I want to show you a scatterplot with what are considered the “robust” data, using a tolerance ellipse plot, where the ellipse is defined by those data points whose distance is equal to the square root of the 0.975 chi-square quantile with 2 degrees of freedom. Here is the R code and the scatterplot (Figure 6):

library(robustbase)
covPlot(cbind(Stubbe$YN.Scores.30, Stubbe$Tr.30.Score),which="tolEllipsePlot", classic=T, cor=T)

There may be more outliers here than what you had considered! What I want you to think about is that you may not be very good at identifying what is really an outlier and what is not, especially when you have multivariate data. So let us allow mathematical formulas to help us figure it out by using robust statistics.
4. Conclusion

The four papers we heard today have presented us with a variety of ideas to think about and challenged us to apply them to real-life problems. As a field, clearly we need to think about a number of important big-picture ideas as we seek to refine the practical techniques which are being used in our classrooms and taught to language teachers. Novices to the area of vocabulary studies almost invariably ask the question, “How can vocabulary be learned most effectively,” but today we have seen that that question involves many different components, starting with how we define “learning.” We have also seen ideas about how that learning can be tested and what words and what forms of words students should concentrate on learning. The types of processing that go on while students learn words should also be taken into account and explored. As our research on vocabulary studies goes on, we continue to make progress on how, what and when to learn vocabulary, but we also open up many new questions and avenues for continued research.

References


