Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

Vocabulary in a Second Language: Selection, Acquisition, and Testing: A Commentary on Four Studies for JALT Vocabulary SIG

Batia Laufer
University of Haifa
doi: http://dx.doi.org/10.7820/vli.v03.2.laufer

Abstract

Four papers by Charles Browne, Rachael Ruegg & Cherie Brown, Makoto Yoshii and Junko Yamashita were presented in the morning session of the Third Annual JALT Vocabulary SIG Vocabulary Symposium in Fukuoka, Japan on June 14, 2014. As discussant, it is my pleasure to comment upon each manuscript. These lexical researchers originate from all over Japan: Tokyo, Akita, Kumamoto and Nagoya. Their lexical topics are related to three themes that are central to vocabulary research: selection, acquisition and testing. The papers are concerned with the types of words that should be selected for teaching, with the optimal conditions for vocabulary acquisition and with instruments that measure lexical proficiency, or are used in lexical research. After commenting on each paper in turn, I shall present a few suggestions for their future research.

1 Introduction

Ten years ago, the late Paul Bogaards and I edited a book *Vocabulary in a Second Language* (Bogaards & Laufer, 2004). The papers in the book covered three main themes: selection, acquisition and testing. The authors of the papers were concerned with the types of words that should be selected for teaching, with the optimal conditions for vocabulary acquisition and with instruments that measure lexical proficiency, or are used in lexical research. It seems that these three themes are still central to vocabulary research in light of more recent publications in general (e.g. Nation and Webb, 2011; Schmitt, 2010) and this symposium in particular. The four papers I will comment on address precisely the above three important themes.

Vocabulary selection is of vital importance since L2 learners do not have the time and the opportunity to learn words in the same way native speakers do, via language input. Therefore, attempts have been made by L2 learning researchers and educationists to select a manageable quantity of the most important vocabulary for learners' needs. One principle behind vocabulary selection is the cost–benefit principle, which states that learners should get the best return for their learning effort, i.e. learn words that they will meet and use often. The first two papers (Browne and Ruegg & Brown) address the question of selecting the appropriate words for instruction and practice for specific learners. Because of the paucity of input that learners have in a non-L2 environment, an important source

Please cite this article as: Laufer, B. (2014). Vocabulary in a second language: selection, acquisition, and testing: a commentary on four studies for JALT vocabulary SIG. *Vocabulary Learning and Instruction*, *3* (2), 38–46. doi: 10.7820/vli.v03.2.laufer

of classroom vocabulary acquisition is word focused instruction. Such instruction combines drawing attention to specific words, in authentic, communicative activities, as well as practicing decontextualized vocabulary. Both types of activities can occur in incidental or intentional learning conditions. The third paper (Yoshii) deals with acquisition of new words through specific contextualized and decontextualized word-focused activities. Valid tests of lexical knowledge are essential when we want to assess learners' achievement and proficiency, or evaluate research results. Though numerous measurement instruments are available for testing single words, fewer tests are available to test multiword units. The fourth paper (Yamashita) investigates a question related to testing collocations.

2 The Four Studies

2.1 A New General Service List: The Better Mousetrap We've Been Looking for? By Charles Browne

As stated above, the cost–benefit principle in vocabulary learning states that learners should get the best return for their learning effort, i.e. learn words that they will meet and use often. However, what is frequent in a language may change from time to time following changes that occur in the language. Therefore, word frequency lists need to be updated. This is what Browne has done in his study. He produced an updated General Service List that is based on millions of modern text words from the Cambridge English Corpus (CEC), excluding academic and newspaper language, but including learner language. The reason for the exclusion is that the academic and newspaper subcorpora are larger than the other corpora and might therefore have biased the general CEC word frequency towards academic and newspaper words.

I wonder whether the academic subcorpus should have been excluded. Many academic words are found among the most frequent 2000 words. For example, 33 words from the first sublist of the Academic Word List (AWL; the 60 most frequent AWL words) and 25 from the second sublist of 60 AWL words are from the first 2000 words. Maybe the final corpus from which the new GSL (NGSL) was drawn could have included a sample of the academic subcorpus that is similar in size to the other subcorpora. In this way, the size of the inclusion of the academic subcorpus would not have skewed the frequency results, but academic vocabulary would have been duly represented. My suggestion is to compare the NGSL and the AWL lists to see how many 2000 academic words are in the NGSL. It would also be interesting to run the NGSL through the academic subcorpus of CEC and see how well it covers it. The authors may find that there are enough academic words in the list to provide good coverage without including the academic subcorpus in the list.

As for the inclusion of learner language in the corpus, I wonder how different it is from samples of native speakers' language in terms of vocabulary use. Learners' selection of words is not necessarily based on word frequency. For example, speakers of Romance languages are known to overuse infrequent Latin vocabulary because of its similarity to their L1. Learners may avoid frequent words in their writing if they are uncertain about some features of their use (grammatical or semantic). It would be interesting to analyze word frequency in the learner

subcorpus and compare it to the larger corpus. If the profiles are suspiciously different, the learner subcorpus could be taken out.

The NGSL defines a word as a 'modified lexeme'. Different parts of speech are not separated if they are spelled in the same way. So 'to list' and 'list' are counted as the same modified lexeme. But different parts of speech that are spelled differently, i.e. with different morphemes, are not counted together (avoid/ avoidance). Moreover, according to this definition, if words have different meanings in different parts of speech, they are still considered to be the same modified lexeme, e.g. 'account' (v) in 'account for the differences in the results' and 'account' (n) in 'bank account'. However, the two meanings would not be in the list if one of them is not in the most frequent 2000 modified lemmas. The 'modified lemma' principle raises the question whether frequency should be the only principle for selecting when to teach which vocabulary. It seems to me that learners who learn 'avoid' would easily learn 'avoidable', 'unavoidable' and 'avoidance', but these derivatives would not be taught if they are of different frequencies and do not appear on the NGSL. The two meanings of 'account', on the other hand, are more difficult to learn even though they are one modified lemma since the lemma represents two completely different meanings. If the frequency lists are intended for learners, we may try to take into account the principle of ease and difficulty of learning as well.

The most important feature of any frequency list is its coverage and here is the main strength of the NGSL. The NGSL covers more of the CEC than the GSL (90.34% as opposed to 84.24%). One may argue that if the lists were constructed on the basis of the CEC, the most frequent words of the list would necessarily cover a large portion of the corpus. But the paper presents coverage figures of different texts as well of older novels, and more modern texts of Scientific American and Economist. For the modern non-fiction texts, NGSL provides better lemma coverage than the other lists, about 5% more than GSL and 3% more than ONGSL. (However, if we measure coverage in terms of word families, the advantage is less clear since there are about 300 more families than in the GSL.) Even if the differences in lemma coverage are small, small differences in coverage may lead to substantial results in comprehension (Laufer and Ravenhorst-Kalovski, 2010). We found that the difference in about 3% (from 87 to 90) leads to an increase in 8% of comprehension.

The paper can be related to several general questions about the construction of frequency lists, degrees of coverage, vocabulary learning and testing.

- How much overlap is there between lists of the most frequent vocabulary (BNC-based, COCA-based, GSL, NGSL, the Longman defining vocabulary, Longman Communication 3000, the JACET 8000 list)? My own experience shows that many words have different frequencies in British National Corpus (BNC) and Corpus of Contemporary American English (COCA). What is the implication of such differences for setting teaching priorities and constructing tests?
- What is the relationship between increase in coverage and increase in comprehension? An important goal of constructing the best possible lists is to improve comprehension via improving lexical coverage.
- What definition of a word (by lemmas or word families) is preferable in list construction in the context of vocabulary teaching? The list of families

includes more lemmas, but is there much more learning effort than with a lemma-based list if the additional lemmas include common derivatives?
- How can we construct vocabulary size tests with representative samples from each word frequency if the same words appear in different frequencies in different corpora and hence in different frequency lists?

2.2 Analyzing the Effectiveness of Textbooks for Vocabulary Retention by Rachael Ruegg and Cherie Brown

Vocabulary selection is also the central theme of the second paper. The authors examine a series of textbooks to see whether the vocabulary that is supposed to be important and necessary for learners at a certain proficiency level is indeed selected for practice by book writers.

The paper starts with examples showing that the widely used labels for proficiency levels – beginner, intermediate and advanced – are unclear and can mean different things in terms of vocabulary knowledge. Particularly important is the observation about the Common European Framework of Reference for Languages (CEFR) levels. As vocabulary size is not specified in the CEFR guidelines, the CEFR proficiency levels may be interpreted in different ways, as Table 1 shows (see Ruegg and Brown, p. 13, of this issue).

A sample of 20 textbooks was analyzed in terms of text length and "target words", i.e. words that appeared in the text and at least in one exercise. One can wonder whether one text sample is representative of the entire book. The authors may consider taking several (three or four) random samples per book in several (three or four) books of each type (reading and integrated skills), and show that there is no difference between the texts within each book in the number of target words and their frequencies. This may increase the validity of the sample.

I will not relate to text length results because I do not consider text length very important, unless it can be shown that the target words appear more times in the longer texts. The paper assumes this is indeed so. (This can easily be checked in a follow up study.) Since the textbook writers claim that the books are written for an intermediate level or above, the assumption is that at least the 2000 most frequent words are familiar (according to Table 1), and therefore the books should be focusing on the "beyond 2000" vocabulary. I will comment on two interesting results. The first shows how well, or badly, the books focus on the beyond 2000 words. The second result is the number of target words the books have selected to practice in a text.

The reading textbooks include 77% of beyond 2000 words among the target words and the integrated skills textbooks include 52%. If you want to make a statistical comparison between the two types of book, I would not compare them at each frequency level, but at the vocabulary level the books are supposed to teach: the 'beyond 2000' vocabulary, that is all the vocabulary that is not included within the first 2000 most frequent words. If you think that second 1000 vocabulary should be taught to the intermediate learners, then you can compare the 'beyond 1000' vocabulary. You will probably find that that reading books have a larger number of

these words. To get back to the percentage of the 'beyond 2000' words, the average statistics do not look too bad. Even if 23% of target words in the textbooks are from the first 2000 words that are familiar to the learner, the practice of such words may provide information on additional word features thus deepening the already existing word knowledge. Furthermore, focus on known words acts against forgetting them, and forgetting is a common and troublesome phenomenon in vocabulary learning. In other words, just as practicing new vocabulary is necessary for learning, practicing familiar words is beneficial for vocabulary maintenance. But the mean percentages of the 'beyond 2000' target words obscure the results of the individual books. Some of the books seem inadequate for the intended learners in terms of their target vocabulary as there seem to be too many high-frequency words.

The importance of the analysis carried out by the authors is that it demonstrates the lack of principled selection of target vocabulary in widely used textbooks and the confusion that the terms beginners, intermediate and advanced introduce. (It is unclear whether there is any other guiding principle for vocabulary selection in these textbooks.) One implication of this analysis is that each textbook should be examined in terms of the target vocabulary for suitability for particular learners before it is used in the classroom. Another implication, or rather suggestion, is not to rely on the above proficiency labels in reporting and interpreting research. In my own research, whenever I describe my participants, I state their proficiency in terms of vocabulary size on the Levels Test.

The second result, the number of target words in each text, is not discussed in the paper and I would suggest looking at it closely in a follow-up study. The authors have kindly provided me with the numbers of texts in each of the books they analyzed. The average number of texts in a reading book is 20; the average number of texts in an integrated skill book is 14. If an integrated skills book offers seven words for practice per text and only half of them are new to the learners, then about $3.5 \times 14 = 49$ new words are expected to be learnt from the entire book. A reading book includes 20 texts, each text has 13 target words and 10 of them are new to the learners (see percentages of beyond 2000 words in Table 3, Ruegg and Brown, p. 15, of this issue). In this case, 200 new words will be introduced in the book. If in a course of 15 weeks students complete one book (personal communication with the authors), the question is whether additional 49–200 new words (assuming all the target words have been learnt) can be regarded as sufficient progress in the course.

Textbook analysis of the kind the authors have performed could be an essential component of course planning and should be linked to setting vocabulary teaching goals. Since learners at different learning stages are supposed to reach specific vocabulary sizes, suitable textbooks should be chosen with these targets in mind.

2.3 Effects of Glosses and Reviewing of Glossed Words on L2 Vocabulary Learning Through Reading by Makoto Yoshii

This paper is concerned with acquisition of new words through specific word-focused activities during online reading. The first activity is looking up word

meaning in an online gloss, the second activity is revision, which is supposed to provide an additional focus on some of the looked up words that are left to the learner's choice. The results confirmed what we know from previous studies, that the use of glosses is beneficial for learning. Out of 10 words that were available for learning, 8.2 were learnt and 6.7 retained on a delayed test. The new feature of the paper is the revision activity. The author makes use of log files to examine individual patterns of looking up words and reviewing some of them.

I was impressed by the seriousness of the participants. In the look up stage, they looked up more words than were unfamiliar to them, thus verifying what they had already known. In the reviewing session, fewer words were focused on, altogether 364 words. Three hundred and twenty-six of them were unfamiliar on the pretest. Even though students were not told what the purpose of the reviewing activity was, they were serious enough to follow the teacher's instructions and review some of the words. Two hundred and thirty-one of all unfamiliar words were reviewed, i.e. 71%. Hence the degree of reviewing was not bad. My guess, which could be tested in a follow-up study, is that in other cultures, the look up and reviewing behavior might be different. (See, for example, Laufer and Yano, 2001 for a comparison of Japanese, Chinese and Israeli learners.) The results are presented not only in tables, but in grids (Figures 4–6, see Yoshii, pp. 24–25, of this issue), which is an interesting visual display I was not familiar with.

And yet the reviewing activity did not affect the acquisition scores. The relatively easy test (recognition of meaning) and the possible ceiling effect (94% were learnt regardless of reviewing) are, in my opinion, secondary reasons. The main reason could be the lack of goal specification of the reviewing activity. The learners were not asked to review the words for any specific reason, such as a task completion, or an upcoming test. An obvious reason for a learner to review a list of words is a test in which these words may be tested. In my own work (Laufer, 2006), I found an impressive increase of scores in the same learners between two stages of learning: an incidental stage and a following stage of intentional learning that involved reviewing the target words for a test. This leads us to a thorny issue of defining the differences between incidental and intentional learning. Was there an intentional learning condition in the study presented in the paper? The author intended to introduce such a condition by the reviewing task. However, according to Hulstijn (2001), intentional learning means deliberate memorization of words for a specific purpose. The most obvious one is an upcoming test, even though one can decide to commit words to memory for other reasons as well. If there is no intention to memorize vocabulary, then learning is incidental. If, however, we examine the literature on incidental and intentional learning, we will often find conflicting operationalizations of the two types of learning. In many papers, "intentional" is taken to be word focused, or practiced in non-communicative activities, e.g. writing sentences with the target words. This is probably the definition that was adopted in the paper. However, this definition is fraught with problems which are beyond the scope of this commentary. Suffice it to say that if the participants in the study had been told there was a test following the reading activity, they would probably have reviewed the words more thoroughly than they did.

I think the paper is a good preliminary study that can be followed by another study with some changes: more target words to avoid the ceiling effect, a more

challenging test or even two tests, e.g. active recall and passive recall that will yield a better spread of scores, and test announcement in order to introduce an intentional learning component. If a within-subject design is chosen, then the test announcement should come after the incidental stage. If a between-subject design is preferred, then one group could be assigned to each condition. If a mixed design is chosen (and this would be my preference), then one group would be assigned to the intentional condition only and the other to an incidental condition followed by an intentional one. This would show word gains of each condition separately and the increase that the addition of deliberate memorization would add to the incidental learning.

2.4 Effects of Instruction on Yes-No Responses to L2 Collocations by Junko Yamashita

The paper addresses an issue in research methodology – whether the form of a prompt, the precise question about phrasal expressions, can affect the answer. Participants were asked to judge phrasal expressions as 'acceptable', 'commonly used' or 'natural' in English. The slight differences in the accuracy results did not overshadow the congruency effect that was observed in other studies. Collocations that were incongruent with L1 yielded lower number of correct responses. The author concludes, therefore, that all three types of instruction can be used in studying interlingual influences of phrasal language.

I would prefer to call them 'prompt types' rather than 'types of instruction' since the phrases 'effect of instruction' and 'types of instruction' are often associated with 'instruction' in the sense of 'teaching'. Even though the three prompts are claimed to be equally legitimate, learners provided more accurate answers to 'E-only' (correct in English, but not in its Japanese word for word translation) items when the prompt was 'acceptable' rather than 'natural' or 'commonly used', and native speakers were more accurate on J-only items (correct in Japanese but not in its English word for word translation) with the 'acceptable' prompt. The author says that the reason for this difference in the effect of the prompt is unclear. My hunch is that the term 'acceptable' is a more neutral or less demanding term than the other two, as it addresses lexical knowledge only. 'Acceptable' seems to be similar to 'correct', both for learners and native speakers. For L2 learners, the terms 'commonly used' or 'natural', on the other hand, could be associated with pragmatic knowledge as well. Since lexical knowledge of a word or expression is often acquired without its pragmatic nuances by L2 learners, it is easier to answer correctly to the 'acceptable' prompt. For native speakers, 'commonly used' or 'natural' could be associated with common usage, which is not necessarily identical with correctness. Therefore, when studying L1–L2 congruency effect, the three prompts would probably yield similar results, as was the case in the paper. However, if the purpose is to study some features of the depth of knowledge that L2 learners developed, such as sensitivity to frequency or naturalness of a phrase, or to study L2 influence on L1 that native speakers start experiencing after living in L2 environment, my preference would be for 'natural' or 'commonly used' prompts.

Even though the purpose of the paper was to examine the effect of different prompts on phrasal lexical decisions, the other important finding is the congruency effect. The importance of interlingual influence in L2 learning is not as widely acknowledged and incorporated into teaching as it should be. Psycholinguistic experiments of the kind we witness in the paper corroborate the findings of error analysis, elicitation studies and experimental studies that demonstrate the pervasive role that familiar languages, particularly L1 have on learning an additional language.

3 Some Suggestions for Future Research

The four papers in the symposium are related to more general research issues in selection, acquisition and testing that can be explored in further studies.

In the domain of selecting vocabulary for language syllabi and tests, we should ask ourselves whether word frequency is indeed the only criterion for selection. From the learner's point of view, a crucial factor in L2 acquisition regardless of word frequency is word "learnability", i.e. the ease or difficulty with which a particular word can be acquired. For example, an infrequent word may be a cognate in the learner's L1, or a loan word (if the languages are genetically unrelated). This makes such words easy to learn. Similarly, some derivatives of a frequent word may be infrequent, but they too are not difficult to learn. On the other hand, a frequent word that has no semantic equivalent in L1 is difficult. Further research could explore the effect of lists based on frequency word learnability on expanding the lexical repertoire of the learners.

Learnability is an important theme of the acquisition domain. There is still work to be done on the effect of interlingual factors on vocabulary learning, by studying a particular L1 group of learners and by comparing different L1 groups, by exploring the facilitating effect of easy words and finding effective methods for teaching difficult words or chunks of words. For example, cognates and loan words are easy to learn and introducing them in large number can enrich learners' functioning in a foreign language and consequently positively affect their confidence and motivation. (If these words changed their original meaning and became 'false friends', learners should be made aware of this.) I am not aware of any empirical work that measures the effect of teaching numerous cognates/loan words to low-level learners, possibly because many EFL classes have learners with different L1s. But such a study can be carried out in monolingual classes in Japan. Intentional learning, as defined by Hulstijn (2001), has not been sufficiently explored. Most studies I know of investigate the intentional learning of *new* words. But intentional learning can also be examined as a method of consolidating the knowledge of words that have been practiced in incidental learning tasks, particularly difficult words or multi-word units like collocations. Research on collocations has received a lot of attention recently. Serious research depends on good and original tests of collocations. Unlike the other existing tests, the phrasal lexical decision can measure fluency. Though there are studies comparing the knowledge of single words and collocations, I am not familiar with research comparing the fluency of the two. Lexical Decision Task (LDT) and Phrasal Decision Task (PDT) tests can provide us with the necessary tools for this kind of research. I hope that the JALT symposium, the papers and the discussions

will contribute to the already growing body of vocabulary studies conducted by scholars who work in Japan with Japanese learners.

References

Bogaards, P., & Laufer, B. (Eds.). (2004). *Vocabulary in a second language: Selection, acquisition and testing*. Amsterdam, the Netherlands: Benjamins.

Hulstijn, J. H. (2001). Intentional and incidental second-language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge, UK: Cambridge University Press.

Laufer, B. (2006). Comparing focus on form and focus on FormS in second language vocabulary learning. *Canadian Modern Language Review*, *63*, 149–166. doi:10.3138/cmlr.63.1.149

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15–30. Retrieved from: http://nflrc.hawaii.edu/rfl/April2010/articles/laufer.pdf

Laufer, B., & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand. *Reading in a Foreign Language*, *13*, 549–566. Retrieved from: http://nflrc.hawaii.edu/rfl/PastIssues/rfl132laufer.pdf

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle Cengage Learning.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York, NY: Palgrave Macmillan.