Vocabulary Learning and Instruction

Volume 1, Number 1,

August 2012

doi: http://dx.doi.org/10.7820/vli.v01.1.2187-2759
VLI Editorial Team

Editor: Raymond Stubbe

Editorial Board: Jeffrey Stewart, Luke Fryer, Charles J. Anderson, Aaron Gibson, Peter Carter

Copy Editors: Alex Cameron, Tim Pritchard, Andrew Thompson, Alonzo Williams

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

Copyright © 2012 Vocabulary Learning and Instruction, ISSN: Online 2187-2759; Print 2187-2767. All articles are copyrighted by their respective authors.
# Table of Contents

<table>
<thead>
<tr>
<th>Article</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreword</td>
<td>i</td>
</tr>
<tr>
<td><em>Tess Fitzpatrick &amp; Paul Meara</em></td>
<td></td>
</tr>
<tr>
<td>Letter from the Editor</td>
<td>ii</td>
</tr>
<tr>
<td><em>Raymond Stubbe</em></td>
<td></td>
</tr>
<tr>
<td><strong>First Annual JALT Vocabulary SIG Symposium</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Morning Session</strong></td>
<td></td>
</tr>
<tr>
<td>New Directions in L2 Lexical Development</td>
<td>1</td>
</tr>
<tr>
<td><em>Shigenori Tanaka</em></td>
<td></td>
</tr>
<tr>
<td>Lexical Development and Learners’ Practices in a Content-based Learning Course</td>
<td>10</td>
</tr>
<tr>
<td><em>Andy Barfield</em></td>
<td></td>
</tr>
<tr>
<td>The Frequency Model of Vocabulary Learning and Japanese Learners</td>
<td>20</td>
</tr>
<tr>
<td><em>Dale Brown</em></td>
<td></td>
</tr>
<tr>
<td>A Study of Learners’ Intuitions Behind the Use of Utterance Verbs in English</td>
<td>29</td>
</tr>
<tr>
<td><em>Yoshiaki Sato &amp; Aaron Batty</em></td>
<td></td>
</tr>
<tr>
<td>Utilizing Student-Generated Pictures for Formative Vocabulary Instruction</td>
<td>37</td>
</tr>
<tr>
<td><em>Charles J. Anderson</em></td>
<td></td>
</tr>
<tr>
<td><strong>First Annual JALT Vocabulary SIG Symposium</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Afternoon Session</strong></td>
<td></td>
</tr>
<tr>
<td>Four Empirical Vocabulary Test Studies in the Three Dimension Framework</td>
<td>44</td>
</tr>
<tr>
<td><em>Masamichi Mochizuki</em></td>
<td></td>
</tr>
<tr>
<td>A Multiple-Choice Test of Active Vocabulary Knowledge</td>
<td>53</td>
</tr>
<tr>
<td><em>Jeffrey Stewart</em></td>
<td></td>
</tr>
<tr>
<td>Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens?</td>
<td>60</td>
</tr>
<tr>
<td><em>Rie Koizumi</em></td>
<td></td>
</tr>
<tr>
<td>Identifying Dimensions of Vocabulary Knowledge in the Word Associates Test</td>
<td>70</td>
</tr>
<tr>
<td><em>Aaron Batty</em></td>
<td></td>
</tr>
<tr>
<td>Examining the Validity of the Lexical Access Time Test (LEXATT2)</td>
<td>78</td>
</tr>
<tr>
<td><em>Tatsuo Iso</em></td>
<td></td>
</tr>
</tbody>
</table>

Copyright © 2012 *Vocabulary Learning and Instruction*, ISSN: Online 2187-2759; Print 2187-2767. Articles are copyrighted by their respective authors.
Foreword: Vocabulary Learning and Instruction

It is with great pleasure that we introduce this inaugural issue of Vocabulary Learning and Instruction. The journal follows the agenda set this year at the first JALT Vocabulary Symposium: to harness the salient aspects of current knowledge in the field of lexical research and to identify future research questions, approaches and directions. Through all the contributions here runs a thread of relevance, context and application, and this commonality of research purpose lends the journal a collaborative and collegial sense. This is further enhanced by the inclusion of commentary chapters, which skilfully extract headline findings and priority areas for further study. In this way the journal reflects the belief that original and rigorous individual research activity realises its true potential when integrated into collaborative contexts, with multi-faceted approaches to shared research questions.

Much of the most influential recent research into vocabulary learning and assessment has been conducted by scholars based in Japan. The articles in this issue are representative of this region’s rich strand of scholarship, and are testament to the journal’s potential to tap into this energetic and purposeful research culture. On visits to Japan-based conferences and symposia we have always been impressed with the quantity and the consistently high quality of vocabulary-focussed research. Research hubs such as the JACET Vocabulary and Reading group, the CALS-Asia Swansea outreach group and the Temple University Japan Applied Linguistics group are central to this, and it is significant that the authors contributing to this journal are representative of these and other research groups. This makes for a research network which shares a clear aim: to further the understanding of lexical processing and the application of lexical assessment and acquisition tools. In contrast to this tight focus, though, studies are informed by literature which is empirically and theoretically wide-ranging. This volume includes studies informed by research into autonomy, fluency, syntax, learning strategies and multi-modality, as well as by approaches pivotal to lexical investigation such as word frequency, lexical richness and word knowledge frameworks. The individual studies are methodologically robust, and are presented in considered, proportionate terms; they offer a reliable foundation from which the commentary papers can identify common strands and mutually informative conclusions.

It is important to note that the relevance and applicability of the studies here extends far beyond the Japanese context. Vocabulary Learning and Instruction appeals strongly to all of us working at the interface of linguistic research, and teaching and assessment practice. We welcome the launch of the journal and are honoured to have worked with, and to continue working with, many of its contributors and editors.

Tess Fitzpatrick
Paul Meara
Dear readers,

Welcome to the inaugural issue of *Vocabulary Learning and Instruction*. *VLI* is an open-access international journal that provides a peer-reviewed forum for original research related to vocabulary acquisition, instruction and assessment. This journal will be published online in the growing tradition of open-access peer-reviewed research.

In the pages to follow you will find articles and commentaries written by ten participants (two discussants and eight presenters) from another inaugural event, the 2012 JALT Vocabulary SIG Symposium, held March 3 of this year in Fukuoka, Japan. This volume features commentaries by the two discussants: the renowned lexical researchers Dr. Shigenori Tanaka and Dr. Masamichi Mochizuki. The remaining eight articles cover a wide scope of second language vocabulary research interests ranging from word frequency levels through learner autonomy in lexical development, to classroom teaching techniques.

We welcome contributions for future issues of *VLI*. Submissions are encouraged from researchers and practitioners in both first language as well as foreign and second language contexts on any topics related to vocabulary acquisition, instruction and/or assessment. Readers who deem that any of the articles warrant an academic response are encouraged to submit *commentaries* for publication in an upcoming issue. Submissions as well as all comments, queries and/or concerns regarding VLI, including this launch, are welcome at editor@vli-journal.org.

Sincerely,

Raymond Stubbe,
Editor *VLI*
New Directions in L2 Lexical Development

Shigenori Tanaka
Keio University
doi: http://dx.doi.org/10.7820/vli.v01.1.tanaka

Abstract

Lexical competence can be defined as one's ability to use words differentially and fully. In this paper I shall first present my views concerning lexical development in order to comment on each of the following four articles: Lexical Development and Learners’ Practices in a Content-based Learning Course by Andy Barfield; The Frequency Model of Vocabulary Learning and Japanese Learners by Dale Brown; A Study of Learners’ Intuitions Behind the Use of Utterance Verbs in English by Yoshiaki Sato and Aaron Batty; and, Utilizing Student-Generated Pictures for Formative Vocabulary Instruction by Charles Anderson.

Keywords: lexical competence; lexical development; second language acquisition; vocabulary breadth; vocabulary depth; discussant; vocabulary symposium.

1 Introduction

The study of lexicon is too broad and complex and many-sided to fit easily into the standard academic divisions of “disciplines.” Relevant fields include lexicography, etymology, corpus linguistics, cognitive linguistics, developmental psychology, cultural studies, language acquisition, and language teaching, just to name several representative domains. This vocabulary symposium is, however, primarily concerned with second language acquisition and second language teaching. The key questions to be addressed are as follows: How does a second language learner develop lexical competence in the target language? How does a language teacher facilitate the process of acquiring that competence?

In recent years, there have been an increasing number of both theoretical and empirical studies concerning how lexical competence develops in a second language learner, and under what conditions (for recent studies see Boggards & Laufer, 2004; Daller, Milton, & Treffers-Daller, 2007; David, 2008; Fitzpatrick, Al-Qarni, & Meara, 2008; Hilton, 2008; Jiang, 2000; Laufer & Hulstijn, 2001; Meara & Fitzpatrick, 2000; Milton, 2008; Nation, 2007; Vermeer, 2001). However, it seems that researchers have yet to reach a consensus on the definition of lexical competence (i.e. the objective of learning vocabulary). Defining what we are after operationally is the crucial first step. Once operationally defined in a valid way, the concept of lexical competence becomes a guiding principle when designing empirical research on lexical acquisition, devising measures of vocabulary testing, and developing curriculum for teaching vocabulary. Without knowing what the target construct is, we cannot be on the right track when pursuing our research projects.

Regarding the question of what word knowledge is, Nation (2001) made a three-way distinction: knowledge of form (phonological and graphic forms), knowledge of meaning (form and meaning, concepts, referents, and associations),
and knowledge of usage (grammatical functions, collocations, and constraints on use). This is nothing more than a list of elements constituting word knowledge, and yet the list conveys what to consider when dealing with this matter. Daller et al. (2007) characterize the complexity of lexical knowledge in terms of “breadth,” “depth,” and “fluency.” Vocabulary breadth roughly means the number of words a learner knows, while vocabulary depth refers to the learner’s knowledge about their usage. Fluency is a notion which explains how easily the learner can recognize and use those words.

It appears that useful concepts are on the table, and yet they have to be combined in an intuitively more appealing way to produce a useful definition. In my opinion, lexical competence can be defined as one’s ability to use words differentially and fully. This definition suggests that lexical competence is a composite of inter-lexical competence (i.e. the ability to use words differentially) and intra-lexical competence (i.e. the ability to use a given word fully).

In addition, I would make an additional distinction between basic lexical competence and extended lexical competence. Basic lexical competence refers to one’s ability to select the most appropriate basic or common word in context (inter-lexical competence), and to use a given basic word through its full range of usages (intra-lexical competence). For example, inter-lexical competence makes it possible for a learner to use “put” and “set” differentially. With intra-lexical competence, the learner is able to use “put” freely in different contexts such as “put some eye drops in one’s eyes,” “put a ladder against the wall,” “put a stamp on the envelope,” “put a 100 dollar bill in the envelope”. In reality, we intuitively know that the learner’s use of “put” is limited to certain exemplars, thus exhibiting the case of semantic under-extension.

Basic lexical competence should be a common foundation of the learner’s lexical competence. In other words, basic lexical competence refers to the common lexis required to allow individuals to communicate effectively in the target language in common situations. Once established, the learner needs to develop his or her extended lexical competence in order to more deeply explore his/her particular interests and concerns. The learner expands his or her mental lexicon in terms of not only “thematic range,” but also “vocabulary size.” Thus we can expect that learners’ extended lexical competence should differ according to each individual’s hobbies and occupations.

Generally, lexical competence is one’s ability to use words differentially and fully. More specifically, we have basic lexical competence and extended lexical competence. The meaning potential of basic words is normally much greater than that of extended words. Thus, the general definition above applies directly to basic lexical competence. Extended lexical competence, on the other hand, involves the number of both productive and receptive words. So you can say that in addition to the basic words, the learner knows 6000 words or 10,000 other words. However, vocabulary size alone does not suffice as a lexical indicator; thus, we add the concept of thematic range. Extended lexical competence should be naturally skewed towards to the learner’s interests and needs; it is topic-oriented lexical competence. Essentially the important questions are: How many words does the learner have in his or her mental lexicon? and What is the topical coverage? or simply How many words in what topics?

As a discussant at this symposium, my role is to review and comment on the following four papers: Lexical Development and Learners’ Practices in a
Content-based Learning Course by Andy Barfield; The Frequency Model of Vocabulary Learning and Japanese Learners by Dale Brown; A Study of Learners’ Intuitions Behind the Use of Utterance Verbs in English by Yoshiaki Sato and Aaron Batty; and, Utilizing Student-Generated Pictures for Formative Vocabulary Instruction by Charles Anderson.

2 On Barfield’s Paper

Barfield is concerned with the relationship between lexical development and learner autonomy. This relationship is one of our most important pedagogical objectives; helping learners become independent, responsible, and autonomous in their attempts at learning a second language. “Learner autonomy” is a concept that captures this objective. When applied to lexical learning in English, as Barfield suggests, learners take on the task of learning vocabulary and try to develop their lexical competence on their own, and by trial and error finding their personal ways of learning. The assumption behind this idea is that there is a tremendous amount of individual variation with respect to how individuals learn a second language in general, and L2 vocabulary specifically. This is not limited to inter-individual variation, but it also applies to intra-individual variation. A learner’s learning style is subject to developmental change as the learner grows in the target language.

Barfield provides an enlightening account of how learner autonomy fosters sound lexical acquisition. The adjective “sound” here can be replaced by “useful” or “communicative.” It is possible that learners’ lexicon remains simply knowledge, failing to become part of the learner’s communicative competence. This brings up a situation where the learner knows a lot about English words, but cannot use them for communicative purposes. Barfield explains this gap in terms of “school knowledge” and “action knowledge,” a distinction originally made by Barnes (2008). For instance, learners may learn English vocabulary in the same way as they learn Japanese history. Knowing does not guarantee the ability of usage.

Then, how does the learner develop action knowledge readily available for language use? Barfield agrees with Barnes’ constructivist view of learning, and cites Little (2007) on “learner autonomy” to introduce the concepts of learner involvement, learner reflection, and appropriate target use. “Learner involvement” is a prerequisite of developing learner autonomy. It is a prerequisite of good learning in general. Obviously, one can acquire practical competence in language only by using it in an appropriate way. Thus, “appropriate target use” is also an essential condition for learner autonomy. The most difficult condition to satisfy is “learner reflection.” Conceptually it is easy to understand what this is: the learner makes his or her own analysis and judgment about how he or she is doing. Involvement and language use are straightforward even in practice, but reflection is troublesome because a set of criteria is required for distinguishing “good reflection” from “poor reflection.” It is not a matter of quantity, but a matter of quality.

Learner autonomy is a general concept in education. When it is applied to lexical development in particular, we need to show how learner autonomy in L2 lexical development is fostered. In this respect, Barfield seems to favor content-based lexical development and points out that learner-generated materials, which
are based on the learners’ own interests and priorities, are more personal and thus meaningful in the sense of topic-based network building.

In the mental lexicon, words do not exist in isolation of other words; rather, they form lexical networks which can be easily triggered when a related topic is given. This is connected to the concept of extended lexical competence, discussed above.

Lexical restructuring is an interesting concept. Learner reflection should be a driving force for restructuring. Yet the nature of reflection still remains to be explored. Barfield (2012, p. 11) asks an interesting question, as shown below:

The pressure to cram vocabulary for university entrance exams in Japan is one example of how an overbearing emphasis on school knowledge forces learners into specific vocabulary practices that distance them from using the language meaningfully for their own purposes. How does this impact their later vocabulary practices and goals when they continue learning English beyond (high) school?

In other words, learners come to our university classrooms with crammed-at-school knowledge about English vocabulary. Their knowledge may consist of a list of 6000 words, most likely in the form of a one-to-one correspondence between the English word and its Japanese counterpart. However, most words are not personalized. They are not semantically connected to each other to form lexical networks. They are not readily available as language resources for spontaneous use in communication. However, Japanese learners arrive in class with something; a list of English words which could be restructured into something personal, meaningful, and authentic.

Barfield is correct to suggest that content-based research can help bridge the gap between school knowledge and action knowledge, fostering lexical development linked with content knowledge construction.

3 On Brown’s Paper

Brown is concerned with what he calls “the frequency model of vocabulary learning.” The general claim is that students learn words in the order of their frequency of occurrence. Along with Ellis (2002), Brown claims that “frequency is recognized as a key driver of all aspects of language learning, and indeed of human learning in general.”

This has an intuitive appeal. Few would argue against the idea that some words occur more often than others. The logic of the frequency model is this: the more one encounters certain words, the better one tends to learn them. In other words, exposure to words is a strong condition for lexical acquisition.

To use the frequency model empirically or in curriculum design, however, we need to have frequency information beforehand, so that we can rank words in the order of frequency. Frequency information is generally drawn from corpus data, and so suffers inevitably from sampling bias. A corpus of any magnitude cannot be the general model of the English language, which is a theoretical sum of possible English uses.
Even the well-known British National Corpus (BNC) is, as Brown points out (along with Nation, 2004), “a corpus of British, adult, formal, informative language.” Moreover, it is no more than a sample drawn from an ever-increasing population. English has become the “global language,” which is spoken virtually all over the world for various practical purposes. The degree of diversity will easily intensify as the spread of English-speaking people widens.

In theory, no corpus can be representative of the ever-diversifying population. If the source corpus is biased, then the frequency of occurrence attached to each word loses its empirical validity. With this in mind, Brown suggests that Japanese learners of English can benefit more, if we use a corpus representative of the kinds of English they experience. He suggests that the JACET 8000 list can be a better source of information when applying the frequency model to Japanese learners.

Brown argues that frequency is an important measure in second language learning and teaching. It seems reasonable to assume that the most frequent words are also generally the most useful because those words enjoy wide coverage in any text. However, what matters is how we obtain reliable and valid frequency data and how we use this data in the practice of teaching English as a second language.

In order to use frequency data effectively, a distinction is needed between “domain-general lexicon” and “domain-specific lexicon.” Domain-general lexicon refers to a set of common words irrespective of semantic fields or topics. The aforementioned basic words are at the center of this set. To a certain extent any corpus may be useful in that the high frequency words remain the same regardless of the corpus selected. In other words, there should be a stable common core of basic words irrespective of linguistic diversity.

In contrast, however, domain-specific words are subject to thematic ranges. Thus, English has a lexicon for cooking, finance, government, education, ecology, information technology, and the like. Thematic ranges differ according to one’s interests and concerns. For example, some learners may be more interested in learning vocabulary related to climate and weather than cooking and education. In that case, they would be willing to learn words and phrases such as tropical zone, rain forest climate, monsoon climate, savanna climate, approaching disturbance, El Niño, La Niña, and so on.

Given this division of lexical domain, it may be possible to employ a general corpus for common words to obtain their frequency. However, for domain-specific words, a domain-specific corpora will be required. For that purpose, we need to have a theory which enables the handling of the huge domain of human activities, and then develop domain-specific databases or corpora. This is a formidable task, but surely one worth the challenge.

4 On Sato and Batty’s Paper

Sato and Batty conducted an interesting study concerning Japanese learner’s lexical intuitions about the use of four utterance verbs: “speak,” “talk,” “say,” and “tell.” A considerable number of studies have been reported with respect to the acquisition of intra-lexical competence. The most notable case is “break” as in the study by Kellerman (1977). However, few studies have been conducted in the area of
the acquisition of inter-lexical competence. So I consider Sato and Batty’s paper a pioneering study.

As Sato and Batty correctly claim, the aforementioned verbs are generally called “basic verbs” in English, and learners are likely to encounter them in the earliest stages of language learning. In fact, those four verbs are ranked among the top 100 in terms of the frequency of occurrence in the BNC. We naturally expect that learners all know these verbs and they tend to perceive them to be easy words to learn, as compared to, for example, “inform,” “declare,” “utter,” and “proclaim”.

As pointed out in relation to Barfield’s study, knowing something does not guarantee the ability of using it. Sato and Batty, thus, asked Japanese university students of different English proficiency levels to participate in their empirical study. Sato and Batty had participants choose the most appropriate utterance verb, while completing a cloze (fill in the blank) sentence activity.

It turned out that Japanese students found it troublesome to use the target words fully and differentially. The results were basically consistent among the participants, irrespective of their levels of proficiency. Even returnees, who had spent several years in English-speaking countries, failed when using their intuitions about certain usages of “speak” and “talk.” By contrast, native speakers of English gave uniform responses to almost all test items.

This study shows that the native speaker’s intuitions are hard to develop by only having exposure to the target language. Returnees had been in an input-rich environment, and yet, they failed to match the native speakers’ accuracy in the use of these semantically related basic verbs. This means that incidental learning has its limitations; the quantity and quality of language exposure and language use alone do not guarantee full-fledged linguistic competence in the target language.

“Speak” and “talk” sometimes behave alike. You can say either “I want to talk to you” or “I want to speak to you” to mean basically the same thing. However, you can talk without speaking as you can hear without listening. A gorilla can talk, but it can’t speak. What does this differential use of “speak” and “talk” suggest? I think it suggests that the Bolingerian principles of “one form for one meaning” and “one meaning for one form” hold true when it comes to lexical semantics (Bolinger, 1977). In other words, “speak” and “talk” are different in form, and then, in principle, they should differ in meaning.

For example, we may ask: Why is it that “a nice speaking voice” as in “She has a nice speaking voice” is much more common than “a nice talking voice”? The explanation should rest on the semantics of “speak” and “talk.” As Sato and Batty explain, the verb “speak” highlights the aspect of emitting language sounds, carrying a sense of “one-way” direction. Thus, we have expressions such as: “She has a nice speaking voice,” “This painting really speaks a lot,” “My parrot speaks in a loud voice,” and “Actions speak louder than words.” By contrast, “talk” is a verb that emphasizes two-way interaction. Thus, we have expressions such as: “We have a talk show tonight,” “We talked a lot about the issue, but didn’t come up with a good solution,” “My baby began to talk at the age of 7 months,” and “Money talks.” Talking is the act of two-way interaction, which, in turn, implies “the negotiation of meaning” in the broadest sense of the term. Thus, we can say, “They are talking without speaking,” when they are interacting with the use of sign language. The saying “Money talks” sounds right because “talk” implies some kind of negotiation...
between the parties and the expression implies that money has a negotiating power. The spontaneous expression “Money speaks” may sound bizarre in that money does not emit a linguistic sound.

As Sato and Batty correctly argue, syntactically marked expressions such as “talk business” and “talk religion” equally imply verbal interaction of some kind, and the use of the verb “talk” is motivated here. It must be noted, however, that “talk business” is not the paradigm case of the transitive use of “talk” – at least, the degree of transitivity is very low with the verb “talk” here. Likewise, business and religion are not topics per se, but rather, they probably indicate types of talk (i.e. business talk and religion talk). When business is topicalized in the form of “our business” or “a new business,” then we need to use a preposition like “about” or “on” as in “Let’s talk about our business.” The main point here is that verbal interaction is associated with the verb “talk.”

If mere exposure to the target language does not suffice to develop lexical intuitions, then Sato and Batty are correct when they claim that carefully designed pedagogical devices should be developed to help learners improve their lexical competence.

5 On Anderson’s Paper

In Anderson’s study, the key issue is whether the use of a picture-drawing technique helps students learn English vocabulary and in what way. Given words such as “true,” “quick,” and “surprised,” a Japanese student easily comes up with the translation equivalent for each target word. However, Anderson argues that when we ask students to draw pictures which may illustrate their interpreted meanings, we are giving them opportunities to personalize the word meaning, which, in turn, leads to deeper lexical understanding.

One of Anderson’s points is that personalization has a pedagogical effect on vocabulary learning when the student’s interpretation of the word meaning comes into play. In order to draw a picture of the word meaning, the student needs to interpret the meaning in his or her own way. Drawing a picture of the word meaning is a very personal experience. Additionally, personalization is the condition for meaningful learning. Once grasped in a personal way, the meaning is not simply public knowledge, but it becomes the learner’s own personal knowledge.

Thus, drawing pictures becomes an effective learning strategy. Anderson further claims that it also becomes an effective pedagogical or teaching strategy when it is linked with feedback. Feedback can be a one-way direction from teacher to student. It can also be a two-way interaction between the teacher and the student, or even between students themselves. Intrinsically, feedback is both cognitive and affective; it is also either positive or negative in nature. Anderson uses the term “formative feedback,” which is defined as feedback that drives the student to move forward in the pedagogical or real task he is engaged in. Formative feedback should be necessarily interactive and as it is a form of communication, the act of giving feedback necessitates giving unspoken messages; cognitive or affective and positive or negative. Anderson correctly points out that the student’s pictures can become good foci for formative feedback. Some pictures are just apt and semantically transparent, and some are otherwise. Semantically opaque pictures motivate verbal
interaction between the teacher and the student, through which the student modifies his or her ideas behind the picture in question. This is how the formative feedback session proceeds.

Overall, Anderson’s paper is suggestive and instructive in terms of how the use of pictures or image schemas can be both an effective learning strategy and a teaching strategy. Here one comment should be added. Most common words are polysemous, having multiple senses. If such is the case, then a simple one-to-one correspondence between the word sense and the sense picture does not suffice. What is needed here is a theory of lexical semantics. For that purpose, we may be tempted to turn to the field of linguistics and choose what looks like the best-fitting theory among many. However, this temptation has to be resisted because linguistics does not take second language learning and teaching into consideration when they do their lexical studies. What we really need is a pedagogically sound theory of lexical semantics. On the basis of the findings of lexical acquisition research, we need to construct a theory of L2 lexical semantics with a cross-linguistic perspective.

6 Conclusion

In summary, imagery or pictures, frequency, learner autonomy, and lexical intuition are the key concepts explored in these four papers. Imagery is associated with effective lexical teaching when backed up by formative feedback. Frequency is a powerful indicator in determining what words should be introduced in which developmental stages. Learner autonomy requires a shift in teaching: the subordination of teaching to the promotion of learning. Lexical intuition should be the target of lexical teaching, with the pedagogical question being how teachers help learners develop their lexical intuition.

All in all, these papers address the question of who learns what in what ways. On the basis of the claims made by the authors, we could envision the following scenario: an autonomous learner responsibly takes on the task of developing lexical intuitions: a set of criteria for freely using high frequency basic words as well as topic-related words of personal relevance in a multimodal and network-building way. It may be true that each construct needs further elaboration, but the four papers discussed have enabled us to see the light at the end of the tunnel of lexical learning and teaching in a second language.

References


Lexical Development and Learners’ Practices in a Content-based Learning Course

Andy Barfield
Chuo University
doi: http://dx.doi.org/10.7820/vli.v01.1.barfield

Abstract
Developing a better understanding of the beliefs and practices that students hold concerning their English vocabulary development should assist lexical instruction and learning. To explore different connections between lexical development and learner autonomy, changes and developments over time in the vocabulary practices and goals of several students engaged in content-based learning projects through English were tracked. After reconstructing their vocabulary histories, participants kept dedicated vocabulary notes and reflections. This study examines the developments that one individual goes through over time in his vocabulary practices and goals, in order to better understand processes of lexical restructuring and network building. From this analysis, connections to a Vygotskian view of conceptual development are drawn, particularly with regard to what such a model may indicate about learners’ situated lexical development.

Keywords: learner autonomy; lexical development; vocabulary histories; vocabulary practices; content-based learning; conceptual development; lexical restructuring.

1 Background
Understanding the beliefs and practices that students bring to bear on their English vocabulary development in content-based courses has long been a pedagogic interest of mine. In these courses, where students do self-directed research projects through English on social, political, legal, and global issues that they choose to explore, appropriate vocabulary development figures as a major challenge. How do learners develop their vocabulary development preferences and practices? What effect does a test-oriented transmission mode of education have on learners’ vocabulary goals? When lexical development is situated within particular educational practices and individual histories, the assumption that individuals should find the single best way that works for them and then apply that one way to achieve uniform success from that point onwards becomes questionable. It may rather be that learners change, re-configure, and reconstruct their ways of learning and using vocabulary as they reflect on their vocabulary histories, develop a greater engagement with content, see more clearly their own purposes for using English, and become more critically aware of the constrained choices that they have available for them. This adjustment in focus takes us towards exploring connections between lexical development and learner autonomy in ways that bring the learner’s side of the story more fully into view. It also makes it necessary to give greater heed to how learners themselves see and organize their histories, changing practices, and goals within the shifting contexts in which they learn and use English. This study considers some partial answers to the kind of questions raised above by looking at...
how learners interpret for themselves their vocabulary learning histories, as well as developments in their vocabulary practices and goals. This article also addresses how this kind of research may help us understand processes of lexical restructuring and network building.

At the outset, a brief review of learner autonomy is warranted, starting with an important distinction that has been made between two modes of learning within formal education. These modes are known as ‘school knowledge’ and ‘action knowledge.’ School knowledge is associated with the knowledge of others (e.g. teachers, texts that learners read, books that learners study) and action knowledge with the constructivist capacity of the learner to re-interpret others’ knowledge, re-shape it to their own purposes, and make it part of their understanding of the world where “the pupil’s ability to re-interpret knowledge for himself is crucial to learning” (Barnes, 1976/1992, p. 142). David Little has theorized from Barnes’s work to propose three general pedagogic principles for the development of learner autonomy: learner involvement, learner reflection, and appropriate target use. By learner involvement Little means that learners “are brought to engage with their learning and take responsibility for key decisions.” The principle of learner reflection requires that “learners are taught to think critically about the process and content of their learning” (Little, 2007, p. 7). Finally, the principle of appropriate target language use entails learners “using the target language as the principal medium of language learning” (Little, 2006, p. 2). That is, learners should use the target language not only for their own communicative purposes, but also for the metacognitive functions of reflecting on and evaluating their performance and development in the target language. Interesting accounts of the changes from school knowledge to action knowledge that classrooms geared towards autonomous learning can lead to are available in Dam (1995) and Little (2009).

While Dam and Little’s approaches offer useful insights into connections between learner autonomy and lexical development, there are however some noticeable gaps in the lessons that we can draw, particularly with regard to what learners do after recording vocabulary that they choose as important, and why learners tend to list vocabulary in L2-L1 columns as they gain in language proficiency. A major question here concerns how intermediate level and above learners may be guided to keep developing their lexical practices in ways that work best for their own purposes. Such learners have necessarily had to reconcile their own purposes with the institutionalized learning that the education system has required of them. The pressure to cram vocabulary for university entrance exams in Japan is one example of how an overbearing emphasis on school knowledge forces learners into specific vocabulary practices that distance them from using the language meaningfully for their own purposes. How does this impact their later vocabulary practices and goals when they continue learning English beyond high school? What do different individuals do to re-configure their vocabulary practices towards their own communicative purposes when they have choice over content in self-directed research projects?

2 Method

To explore different connections between lexical development and learner autonomy, changes over time in the vocabulary practices and goals of several students engaged in content-based learning projects through English were tracked.
The students ranged from high intermediate to advanced in their English proficiency. The class met once a week and had a general focus on international issues, with the students doing four research projects over periods of 5–6 weeks during the second year. Some projects were pair-based, others individual, and while students made research notes and research diaries outside class each week, much of their in-class activity focused on explaining their research to each other and considering what problems they faced in different research and explanation processes. They also focused on how their knowledge of an issue was growing, and what they wanted to find out next. These reflections fed into their goal setting for their ongoing research and vocabulary development so that the vocabulary notes that they made became closely linked to the self-directed projects that they did. At the end of each cycle, the students gave 20–30 minute presentations to each other in small groups, and reviewed their performance over the whole cycle before they set new goals for the next project.

As part of gaining greater control over their vocabulary practices, early on in the academic year, the classmates had a guided discussion of their vocabulary beliefs and experiences. Here they also read other students’ vocabulary histories (Barfield, 2012) before writing their own. A vocabulary history (VH) is similar to a language learning history (Benson & Nunan, 2005; Murphey, 1997; Pavlenko, 2001) in which a learner narrates their personal story from when they began to learn English through to the present before articulating their future learning plans and goals. In a VH, however, the learner gives much greater consideration to how they have tried to learn vocabulary, by detailing for example their positive and negative experiences.

### 3 Results and Discussion

This study focuses on one student, Ichiro (pseudonym), a second-year International Business and Law major taking an elective English course over his first two years of university. Below I have reconstructed Ichiro’s VH from different reflections and self-assessments that he wrote:

Ichiro never really enjoyed learning English in high school. He studied English only for university entrance exams just because he had to. “I hardly had any opportunities to use English,” he comments, and, although he felt he wasn’t good at English, he liked watching movies without Japanese subtitles.

In his first year at university Ichiro felt ashamed about speaking English with his classmates, even though he had chosen to take the elective English course. He knew the key to making progress was not to be ashamed, but felt negative about using and learning English. Often he felt like giving up. At the start of the second year, he couldn’t imagine himself using English in the future, and his initial goal was “to come to feel I like English”. If he could remove his sense of shame and start feeling positive about learning English, he believed he would be interested in learning English and then he would be able to imagine himself using English in the future and find new goals for himself.

Through junior and senior high school I hadn’t learned English actively… I learned English vocabulary only from sentences which I read in English classes. When I found vocabularies whose meaning I don’t know, I looked up words in an English-Japanese dictionary. But I did only it. I didn’t write down the words, neither look the sentences...
Ichiro’s VH is situated within particular educational practices that have had a highly negative effect on his ability to use the language for his own purposes and to engage with others through English. His previous vocabulary experiences were ‘other-determined’ (Benson, 2011, pp. 84–87, 113–114) and focused on expanding the quantity of his lexical knowledge, rather than on his determining for himself how to develop the quality of his lexical competence. The criteria that he previously used to choose vocabulary were those of difficulty and newness, not whether the items were useful, important, exciting, or interesting for his own purposes. The criteria of ‘new’ and ‘difficult’ did not work well for him.

Once at university, Ichiro’s criteria for vocabulary selection gradually started to shift more towards using vocabulary for his own purposes. This process of change was protracted, and it began with a change in his dictionary use:

Getting into the university, my way learning English vocabulary changes. First, I try not to use an English-Japanese (E-J) dictionary as possible. Instead, I try to use an English-English (E-E) dictionary when I find English words which I don’t know. Using an E-E dictionary, I can learn not only the meaning of the words, but also which context I should use the words in. It takes me more time than using an E-J dictionary, but it gives me more things I can learn.

For Ichiro, the shift in practice towards use goes together with the notion of learning vocabulary actively and developing his way of making vocabulary notes so that he could consolidate his lexical knowledge. He became aware that he could find different ways to help remember the vocabulary that he noted down. He also began to believe that using English actively would help him learn vocabulary. However, while he was aware of the potential benefits of such changes, these changes still appeared somewhat focused towards a school knowledge perspective of vocabulary learning—that is, a greater need for memorization than use:

Next, I try to use English vocabulary I learned actively. When high school, I only read English sentences, not used English in writing and speaking. But I realized that it’s more useful to use English actively in learning English vocabulary. I don’t make notes of English vocabulary. But I think I should write English words which I don’t know, and should look it over again. I have a poor memory, so I have to devise one way and another to learn English vocabulary, and I will try hard.

Ichiro was motivated to develop his vocabulary practices, but it took him several months longer, as we will see below, to adapt his vocabulary practices around different issues that he researched within a particular classroom community of content-based learning.

In early June 2011, as Ichiro started to research Amnesty International and the abolition of the death penalty, his way of recording vocabulary (as shown in Figure 1) tended towards the inclusion of a great quantity of different lexical items in both Japanese and English. While his notes also showed a concern with recording items with related meanings, a listing or chaining effect can be discerned in the way
that associated lexical items are included; thus, it seems as if Ichiro was collecting rather than organizing vocabulary at this point.

Later in June 2011, his way of recording vocabulary, as shown in Figure 2, became more organized and focused on paraphrase and collocation, but Ichiro still seemed concerned with recording a large quantity of lexical items.

I didn’t have useful ways of learning and using vocabulary. So I referred to the prints (= class handouts on different ways of learning vocabulary) and my classmates’ ways. First I wrote down the meanings of the words, and I just wrote down some collocations. I have not found the best for me to learn vocabulary. My goal for the third cycle is to find the best way for me and make use of it.

In the third research project in the autumn semester, as Ichiro looked at Corporate Social Responsibility issues to do with the tobacco industry, some subtle shifts took place in his way of recording vocabulary:

Figure 1. An example from Ichiro’s way of recording vocabulary in early June 2011.

Vocabularies about “law”

- defendant 被告 (→ 被告弁護人 defends attorney)
  - the person in a trial who is accused of committing a crime
- trial 裁判, 公判
  - bring him to trial = bring him up for trial = put him on trial
  (裁判: judgement, justice, case, suit, lawsuit)
  - 民事訴訟 → civil suit
  - 刑事訴訟 → criminal suit
- 「裁判－」
  1) ～ 官: judge, judiciary
  2) ～ 汎: lawsuit
  3) ～ 手続き: legal procedure

Figure 2. An example from Ichiro’s way of recording vocabulary in late June 2011.

Vocabulary Learning and Instruction, 1(1), 10–19.
Figure 3 lets us see how Ichiro had reduced the number of connections that he recorded for a particular item. With the introduction of relationships of equivalence and opposition (i.e. cause [great—serious] disruption—minimal/minimum), he achieved a much greater degree of organization between the elements in this mini lexical network. This points to a growing qualitative dimension to decision-making and sense of self-direction on Ichiro’s part. Just over a month later Ichiro’s way of organizing his vocabulary notes changed again, as shown in Figure 4. Ichiro was now categorizing different possible combinations into groups and clearly organizing his vocabulary notes by meaning.

![Image of Figure 3: An example from Ichiro's way of recording vocabulary in October 2011.]

![Image of Figure 4: An example from Ichiro's way of recording vocabulary in November 2011.]

Ichiro put his changing goals and developing practice in these terms:

My goal is to explain difficult words in easy words to make my partner understand it, and to make me understand it. So, when I find the words that I don’t know in my research, I use E-E dictionary and understand the meaning in English. I focus on ‘new’ words and key words in my research. Even if the key words are not so difficult, I look up the words in E-E dictionary, and find some connection words. It makes easy for me to explain the important points in my research. The biggest change is that I use E-E dictionary and collocations dictionary. I hadn’t used these dictionaries, used only E-J or J-E dictionaries. I found that E-E dictionaries and collocations dictionaries are very interesting. I do not only look up in dictionary, but try to use them when I write some reports and makes notes as much as possible. By doing so, I think I am developing my vocab [sic] more quickly.

Ichiro’s self-evaluation lets us understand how the development of his vocabulary practices was bound up with his making sense not only for himself but also for other students in the class, as he explained and discussed his research with them. Explaining content knowledge to himself and to others is the central mediating process by which he realized the need to reconstruct and extend his knowledge of...
vocabulary that he already knew. In short, Ichiro started developing different connections for key vocabulary to “explain the important points” in his research.

A number of important factors can be identified in Ichiro’s continuing lexical story. Ichiro was involved in the development of his vocabulary practices, and his involvement came from taking responsibility for determining for himself ways that work best for him. This involvement was closely connected to his interest in his research and being able to explain clearly to others his developing knowledge of the issue that he was researching. Vocabulary development was no longer a solitary struggle for Ichiro: it was connected to specific processes of knowledge construction situated within a particular community of research with his peers. Ichiro was also able to think critically about developing different vocabulary practices for the communicative purposes that he had within that classroom community. He moved beyond dictionary study as the main tool of his vocabulary development: he saw that using vocabulary for other purposes within his university studies offered a positive direction to take, too.

The final part of this paper focuses on how we can interpret these changes in Ichiro’s vocabulary development in relation to wider questions of lexical restructuring and network building. The different changes in Ichiro’s way of making vocabulary notes point to lexical restructuring as a process of semantization (Henriksen, 1999). Initially, in Figure 1, Ichiro’s attempts to organize his vocabulary development seem predominantly (and associatively) organized by his L1 conceptual knowledge. In his vocabulary notes, there are numerous individual associative links, but they are organized in a list all at the same level of generality. Whatever the common semantic bond between them is, this bond remains unclear and has no hierarchical organizing power. In the second example of Ichiro’s vocabulary practices (see Figure 2) one important change is how Ichiro has now started to include definitional information and synonyms in his vocabulary notes. Here Ichiro focuses on recording a large number of near-equivalent items (ask/call for something, demand/request/urge something), and it seems that this stage is driven by the need for him to paraphrase to others important ideas from his research. The third example (as shown in Figure 3) shows Ichiro creating relationships between collocations in his vocabulary notes, based on equivalence and opposition (e.g. cause great/serious disruption—minimal/minimum). There are now fewer items in his vocabulary notes, but these items reveal a great degree of systematic organization to the previous two stages. In the final example (see Figure 4), there are signs that Ichiro is organizing his vocabulary notes according to (multiple) semantic features, judging from the way he categorizes the collocates of trend. There are also instances of semantic polarity in this network, and its components are grouped according to abstracted categorizations that appear separated from everyday concrete experience.

The development of Ichiro’s vocabulary practices can be interpreted as both lexico-semantically and conceptually driven. This allows some interesting connections to be made to a Vygotskian model of conceptual development where word meaning has a central role in mediating knowledge development. Vygotsky (1934) proposed that there are three major phases by which children, through interaction with a more expert other, organize different elements into groups. These three phases might be broadly described as random categorization, associative grouping, and higher-level abstract conceptual organization. The very first phase involves what are
called ‘heaps.’ Heaps are random subjective groupings of elements where there is no discernible principle of organization at work. The second phase involves what Vygotsky called ‘complexes.’ Complexes show some degree of conscious objective organization in that they involve groupings of elements according to associative bonds between them. Initially, these associations are based on changing bonds of similarity. Later they may be organized by contrastive bonds where elements are grouped by some shared differing feature. These groupings however are not stable, as the principle of organization used by the individual is likely to shift at some point. Vygotsky also noted that it is difficult for a child to generalize from complexes to think in terms of abstract concepts: “In a complex, there is no hierarchical organization of the relations between different traits of the object. All attributes are functionally equal” (Vygotsky, 1934, p. 115). The final type of complex is called a ‘potential concept’ or ‘pseudoconcept.’ This kind of grouping of elements is the bridge towards the third phase of mature conceptual thinking. In a pseudoconcept the child is able to group elements in a concept that appears to be the same as what an adult might do, but crucially the child is unable to “abstract, to single out elements, and to view the abstracted elements apart from the totality of the concrete experience in which they are embedded” (Vygotsky, 1934, p. 135). The child’s thinking is context bound, in other words. However, by interacting with an expert other (an adult or language instructor, for example), the child begins to understand the sense of a particular concept and to start to generalize. For Vygotsky, it is talking with others that is the central tool in enabling an individual to abstract from concrete experience and generalize towards hierarchically organized modes of conceptual thinking.

Putting Ichiro’s stages of vocabulary practice development side by side with particular Vygotskian perspectives, we can see some unexpectedly striking parallels, as shown in Table 1.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Grounded development of vocabulary practices</th>
<th>Possible parallels to Vygotskian conceptual development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1: associative grouping</td>
<td>Numerous individual associative links, but they are organized in a list all at the same level of generality</td>
<td>Heaps: random subjective groupings</td>
</tr>
<tr>
<td>Stage 2: equivalence driven by paraphrase</td>
<td>Numerous individual collocational links; no clear (delimited) grouping of collocates</td>
<td>Complexes: groupings by associative bonds, with greater degrees of organization</td>
</tr>
<tr>
<td>Stage 3: Increasing abstraction of multiword units</td>
<td>Reduced number of connections; organization by equivalence and opposition</td>
<td>Pseudo concepts: groupings embedded in concrete experience</td>
</tr>
<tr>
<td>Stage 4: semantization and restructuring</td>
<td>Categorization and organization by clearly labelled groups</td>
<td>Scientific concepts: abstracted from concrete experience and hierarchically organized</td>
</tr>
</tbody>
</table>

Table 1. Grounded Theorization of the Development of Vocabulary Practices in Content-Based Learning
There is no doubt that Ichiro brings greater degrees of observable organization to the development of his vocabulary practices over time. While it is possible that we may see broadly similar stages of development in other learners’ practices, realized in different idiosyncratic ways, we must be careful not to assume that all learners will necessarily go through these stages in a linear fashion at equal rates of development. The value of this theorization, in so far as it does hold, is probably best seen in terms of sensitizing us to potential rather than actual fixed stages of development that different young adult learners may go through at different times in different ways.

4 Conclusion

The analysis of Ichiro’s VH and practices helps us realize that the more learners start to become engaged in using the language to express their own meanings in communication with others, the more likely they are to break out of single atomistic ways of developing their lexical knowledge. Vocabulary histories like Ichiro’s let us understand lexical development as (potentially) closely connected to knowledge construction and associated patterns of conceptual development. On the basis of the claims made in this paper, these patterns seem likely to emerge where a learner:

- can exercise strong control over the content of their learning;
- is engaged in constructing new knowledge of the world through the research-based projects that he or she does;
- keeps developing (and, most importantly, reflecting on the development of) their vocabulary practices over a quite long period.

In the case of Ichiro, we have been able to identify some possible parallels between changes in his vocabulary notes and a Vygotskian view of concept development. By continuing to take a fine look at different learners’ stories and their changing vocabulary practices and shifting goals over time, we may hopefully come to appreciate other unexpected dimensions to learners’ second language vocabulary development.

Acknowledgements

My thanks go to Alison Stewart, John Shillaw, Mike Nix, and Tin Tin Htun for commenting on earlier drafts of this article.

References


The frequency model of vocabulary learning and Japanese learners

Dale Brown
Nanzan University
doi: http://dx.doi.org/10.7820/vli.v01.1.brown

Abstract
The frequency model of vocabulary learning, the idea that words are learnt broadly in order of their frequency, is routinely applied in language teaching, testing and research. There has, however, been little research actually confirming it. This paper reports on a small-scale study which investigated the extent to which the vocabulary knowledge of a group of Japanese university students follows the model. Forty-nine low-intermediate proficiency participants took a 100-item Yes/No test containing 20 words from five frequency bands. It was found that, as expected, for the group as a whole, knowledge of the words was related to frequency, with the proportion of words known falling as frequency declined. Following Milton’s study, however, an analysis of the results of individual participants revealed that around 20% showed different patterns of knowledge and did not follow the frequency model. One question, however, is whether the frequency information on which the Yes/No test is based is the best approximation of these learners’ experience of English. A re-examination of the data in terms of JACET8000 levels found that once more the group as a whole followed the frequency model, and in addition that more of the individual participants conformed to the model. The study thus demonstrates the importance of using frequency data that is relevant to the learners in question and provides confirmation that frequency is a key determiner in the learning of vocabulary. It seems that teachers can reasonably make use of frequency information to support their students’ learning.

Keywords: vocabulary learning; word frequency; frequency model; JACET8000.

1 Background
The frequency model of vocabulary learning is the idea that words are learnt broadly in order of their frequency. This model is central to ideas about principled vocabulary teaching, is a key consideration in the design and writing of textbooks and is used on a daily basis by teachers as they consider which words are likely to prove challenging in their classes, even if the only frequency resource used is their own intuition. In vocabulary testing, the frequency model is involved in practically every well-known test: indeed it is hard to think of a vocabulary test that does not take account of the model in some way, whether in the selection of items, controlling the difficulty of contexts or analysing responses. In vocabulary research, and indeed language learning research more generally, frequency is routinely controlled for in accordance with the model. More generally, frequency is recognised as a key driver of all aspects of language learning, and indeed of human learning in general (Ellis, 2002).
Perhaps, because the idea of the frequency model is so ubiquitous and seems so obvious, there is surprisingly little research actually confirming it. Milton (2007) traces the idea back to Palmer (1917) and notes how other authors have accepted it without question. Nation (2006b) cites Read (1988) and Laufer, Elder, Hill, and Congdon (2004) as having found that learners’ scores drop on tests as you move from higher to lower frequency levels; while Schmitt (2010) also cites Read along with Schmitt, Schmitt, and Clapham (2001) similarly. None of these studies, however, set out to explicitly examine the frequency model.

Two studies that did examine the frequency model are Aizawa (2006) and Milton (2007). Aizawa tested 350 Japanese university students on their knowledge of items in the eight 1000-item levels of the JACET8000 list (JACET, 2003). Aizawa’s results (Figure 1) show the frequency model working well for the most frequent bands, as the number of items known declines step-by-step over the first four levels. Beyond this, however, other factors seem to come into play.

One explanation for Aizawa’s results is the nature of word frequency. Figure 2 shows a frequency curve, based on data from the Corpus of Contemporary American English (Davies, 2008). As the curve makes clear, the items comprising the first 1000 lemmas are much more frequent than those in the second thousand, but as we move further to the right, the differences between frequency bands become much smaller.

Milton (2007) similarly looked at the vocabulary profiles of learners, using the X_Lex test (Meara & Milton, 2003) with 227 Greek learners of English. He found that in terms of the overall results the frequency model worked well, but looking at individual results revealed a more complex picture. Around 60% of individuals produced profiles in line with the frequency model, while the remainder did not. Two alternative profile patterns were identified: a level 2 deficit, in which $1 > 2 < 3$, which accounted for around 25% of participants; and a structural deficit, in which $1 < 2 > 3$, which accounted for around 10%.

The above studies provide us as teachers and researchers with useful information on the frequency model. Aizawa (2006) shows that there are limits to the frequency model, but it works well over the first few thousand words; Milton
(2007) reveals that the model applies to the majority of learners, but not to all learners. This finding is important from a pedagogical perspective as it could mean that assumptions regarding learners’ knowledge are faulty for a sizable number of students. The current study, which is a small-scale partial replication of Milton’s work, thus seeks to look into how well the frequency model applies to a group of Japanese university students.

2 Method

The data in this study were collected as part of a larger project looking at the lexical problems faced by learners when reading. Data were collected from a highly homogenous group of 49 learners, all non-English majors studying in two different, but related, departments at a Japanese university. All have progressed through the Japanese school system, reaching similar levels of general academic achievement and similar levels of English proficiency.

The 49 participants took a Yes/No test created for the wider research project. This test included 20 words from each of the first five 1000-word family bands in Nation’s (2006a) BNC-based wordlists. The 20 words from each list were selected by first placing the headwords from each 1000-item list in alphabetical order, numbering them, and selecting every 50th word with the count beginning at a different number between 1 and 50 for each of the five lists. In addition to the 100 real words, 20 false words were taken from Milton’s (2009) first sample X_Lex test. These were included in order to encourage the learners to think carefully about whether they knew each word. The 120 items were presented in random order. The test had good reliability (Cronbach’s $\alpha = 0.82$) and relatively few false words were selected, an average of 1.49 per participant, for a false alarm rate of 7.45%.

3 Results

Table 1 provides descriptive statistics for the Yes/No test. With regard to the results in terms of the frequency bands, the group results reveal the expected pattern whereby the first 1000 items are best known and each subsequent band of words is
There is a reasonably consistent fall in the number of words known between each level, with only level 4 appearing to deviate from the general pattern. Statistical tests confirm these impressions. As the results for the fifth frequency level were not normally distributed, the non-parametric Friedman’s test was used and found the differences between the levels to be significant ($x^2 = 171.42$, asympt sig = 0.000). Wilcoxon Signed-Ranks tests were used as post hoc tests to compare the differences between the frequency levels. The alpha was set at 0.05, adjusted with a Bonferroni correction to 0.005. All the bands were found to be significantly different from each other with the exception of levels 3 and 4.

### Table 1. Participant Results ($k = 100$, $n = 49$)

<table>
<thead>
<tr>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
<th>Low</th>
<th>High</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>76.10</td>
<td>7.853</td>
<td>43</td>
<td>49</td>
<td>92</td>
<td>0.822</td>
</tr>
</tbody>
</table>

Note: Reliability = Cronbach’s.

### Table 2. Results for Each Frequency Band ($k = 20$ per band, $n = 49$)

<table>
<thead>
<tr>
<th>Band</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.61</td>
<td>0.640</td>
<td>2</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>17.10</td>
<td>1.699</td>
<td>9</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>14.82</td>
<td>2.098</td>
<td>11</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>14.76</td>
<td>2.488</td>
<td>12</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>9.82</td>
<td>3.046</td>
<td>13</td>
<td>4</td>
<td>17</td>
</tr>
</tbody>
</table>

Turning to the individual results, we find similarities with Milton, with a large majority of the participants, 37 (76%), following the normal pattern. The level 2 deficit pattern is seen in the results of five participants, 10% of the total, while a structural deficit is seen in only one case. Compared with Milton, there are a relatively large number of participants displaying other patterns: 6 (12%). However,
in all cases this is because equal numbers of words from different levels were indicated as known. In two of these cases, in fact, the participants indicated all 20 level 1 words and all 20 level 2 words to be known. Counting these as examples of the normal pattern, it accounts for 80% of the participants.

4 Discussion

In this group of Japanese learners the frequency model of vocabulary learning would appear to operate quite strongly. This is reassuring, considering the importance of the model in the field, but also somewhat surprising. As Milton (2007) notes, “frequency information drawn from a wide variety of native speaker sources may not be relevant to foreign language learners who are not exposed to this sort of language but have only textbooks to draw on” (p. 49), while Nation (2004) comments that “the British National Corpus (BNC) is predominantly a corpus of British, adult, formal, informative language, and most English learners in primary and secondary school systems are not British, are children, and need both formal and informal language for both social and informative purposes” (p. 4). Nation attempted to overcome this problem by basing his BNC wordlists on the spoken component of the BNC alone, but the description of the language above remains largely accurate even of this spoken component.

Why then does the frequency model appear to operate so strongly? One factor may be that, at least at the relatively high frequency levels under consideration here, most of the frequent words are the same whatever variety of English one examines. That is, whether one looks at British English or otherwise, formal or informal language, adult or child language, informative or social language, the most frequent words may be, to a considerable extent, the same. Another possibility is that the participants’ exposure to English, which has largely taken place through their formal school education, has predominantly focused on adult, formal, informative language, even if probably not focused on British English.

Nevertheless, it is likely that the BNC is not the best representation of the participants’ exposure to English. Besides the foreignness of the BNC to the participants as a group, the BNC, or any general corpus, cannot be taken as a representation of any individual’s experience of the language. In addition, Nation’s wordlists are in any case not purely based on frequency; range and dispersion also being taken into account when drawing up the lists, further muddying the waters. In trying to examine the validity of the frequency model, I am using the BNC wordlists as an approximation of the frequency of words in the participants’ experience of English. It is then possible that the frequency model in fact operates more strongly than is shown in the above figures. Indeed, as we can never measure the frequency of each individual’s exposure to the words of a language, we can never judge the full extent of the frequency model. Nevertheless, it may be possible to find better approximations of these participants’ experience of English. As we are looking at Japanese learners, an obvious potentially better approximation is the JACET8000 list (JACET, 2003). This list, while also in part based on the BNC, tries to give a sense of how English is experienced in Japan, and so takes account of the language of school textbooks, popular tests of English, children’s books and TV and film scripts.
Looking at the words in the Yes/No test and where they appear in the JACET8000 list is revealing, as Figure 4 shows. One fundamental difference between the two is immediately clear: Nation’s lists are of word families while the JACET8000 list seems to be of lemmas, hence the much higher numbers in the latter. There still appears to be, nevertheless, an overall sequence of steps across the bands, with band 4, once again, the only band seeming to not clearly follow the pattern. This is confirmed by the median JACET8000 rank of the items in each band, shown in Table 3, and also by a correlation, Kendall’s \( r = 0.60, n = 85, p < 0.001 \), between the BNC-based bands and the JACET8000 rank. Still, the figure makes it clear that while the band 1 words are quite tightly bunched together in terms of their JACET8000 ranking, the words in the other bands are much more spread out and there is a great deal of overlap between the bands. This raises the question of whether a Yes/No test based on the JACET8000 list would show stronger evidence of the frequency model. Indeed, the Yes/No results correlate more strongly with the JACET8000 ranks (Kendall’s \( r = -0.594, n = 85, p < 0.001 \)) than with the BNC-based frequency bands (\( r = -0.510, n = 100, p < 0.001 \)). Unfortunately, it was not possible to construct and administer a new test as part of this project. What can be done, however, is to look at the percentage of the items known from different JACET8000 frequency levels. Of the 100 test items, 23 are in JACET level 1 (counting the three items on the JACET Plus250 list as level 1 items), 13 in level 2, 17 in level 3 and 11 in level 4. There are only four words from JACET level 5, so this level and those above it are not considered in this analysis. Figure 5 (see also Table 4) presents the percentage of these words indicated as known. As before, there are the familiar steps down across the frequency levels, though levels 1 and 2 are close to full knowledge. Turning once more to the individual results, the ceiling effects mean that a number of participants indicated full knowledge of all 36 level 1 and level 2 words. Counting these participants as also displaying the normal frequency profile, it accounts for 42 (86%) of the 49 participants. The level 2 deficit profile is seen in only a single participant, while the structural deficit appears in four participants. Among these five participants, in four cases one more word checked
would have made their profile normal. The remaining two participants indicated full knowledge of all 53 words in the first three JACET levels.

Not all of the 100 items are included in this figure: three items appear in the separate JACET Plus250 list (a list of basic items such as numbers, days of the week and months, frequent country names, frequent abbreviations and frequent irregular past tense and past participle forms); 12 do not appear in the JACET8000 list at all. These 15 items do not have a JACET8000 rank and so do not appear in the figure.

<table>
<thead>
<tr>
<th>Frequency band</th>
<th>Median rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>355.5</td>
</tr>
<tr>
<td>2</td>
<td>1634.5</td>
</tr>
<tr>
<td>3</td>
<td>3400.5</td>
</tr>
<tr>
<td>4</td>
<td>3802</td>
</tr>
<tr>
<td>5</td>
<td>7451.5</td>
</tr>
</tbody>
</table>

Note. Items appearing in the JACET Plus250 list were treated as being ranked 1 and items not in the JACET8000 list as ranked 8001.

Figure 5. The percentage of words known from the first four JACET8000 levels.

Table 4. Results for the First Four JACET Levels (n = 49)

<table>
<thead>
<tr>
<th>JACET level</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
<th>Low</th>
<th>High</th>
<th>Mean percentage</th>
<th>SD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (k = 23)</td>
<td>22.82</td>
<td>0.441</td>
<td>2</td>
<td>21</td>
<td>23</td>
<td>99.20</td>
<td>1.919</td>
</tr>
<tr>
<td>2 (k = 13)</td>
<td>12.41</td>
<td>0.934</td>
<td>4</td>
<td>9</td>
<td>13</td>
<td>95.45</td>
<td>7.181</td>
</tr>
<tr>
<td>3 (k = 17)</td>
<td>12.82</td>
<td>2.224</td>
<td>14</td>
<td>3</td>
<td>17</td>
<td>75.39</td>
<td>13.080</td>
</tr>
<tr>
<td>4 (k = 11)</td>
<td>6.98</td>
<td>2.106</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>63.45</td>
<td>19.149</td>
</tr>
</tbody>
</table>

*Vocabulary Learning and Instruction, 1(1), 20–28.*
5 Conclusion

The above analyses suggest first and foremost that the frequency model of vocabulary learning is indeed very strong among this group of participants. The majority of those whose frequency profiles do not follow the normal pattern only narrowly deviate from that pattern. Furthermore, it may be the case that those learners who do not appear to follow the frequency model have somehow had differing experiences of English. That is to say, it is possible that their knowledge does follow the frequency model, but the frequencies of items in their language experience differ from that of others. As was noted earlier, we can only judge the frequency model by working with approximations, and, again as the analyses suggest, the quality of those approximations has a bearing on the results. The $X_{Lex}$ test Milton used in his research draws on both general frequency information and word lists designed for EFL (Milton, 2007, p. 50), but there must be some possibility that for some of his Greek participants who did not follow the normal pattern, those lists were simply not a good representation of their experience of English.

Regarding language teaching and learning, this study confirms two points. First, we should try to use frequency data that makes sense for the learners in question. For teachers of Japanese learners, the JACET8000 list appears to offer advantages, as it was designed to do so. Second, when dealing with the highest frequency levels, we can be fairly confident in assuming that frequency correlates with the likelihood of a word being known. Certainly, there are other factors that affect vocabulary learning, and some items of low frequency are quickly picked up by learners while some high frequency items often remain unknown. For example, looking at the items in this study, in terms of the BNC-based frequency bands, items such as \textit{client} (Band 1), \textit{chuck} (Band 2) and \textit{legislate} (Band 4) were marked as known by relatively few participants compared with other items in the same band. Similarly, among the items in the first four JACET8000 levels relatively few participants indicated that they knew items such as \textit{inch} (level 2), \textit{ministry} (level 3) and \textit{ass} (level 4). Yet despite there being other factors affecting the learning of words, the overall results demonstrate that the effects of frequency are powerful. For us as teachers, frequency can provide extremely helpful guidance whether we are engaged in small-scale tasks such as checking the vocabulary load of a reading text or selecting words for explicit teaching or large-scale tasks such as planning a course. As Milton (2009) concludes: “the importance of frequency in vocabulary learning is as near to a fact as it is possible to get in L2 acquisition” (p. 242).

References


A Study of Learners’ Intuitions Behind the Use of Utterance Verbs in English

Yoshiaki Sato and Aaron Batty
Keio University
doi: http://dx.doi.org/10.7820/vli.v01.1.sato.batty

Abstract

Verbs of utterance are some of the most fundamental verbs in the English language, yet their usage patterns are exceptionally diverse. Learners of English should be able to use these words correctly and comfortably, but without an understanding of their core meanings acquisition of their various patterns of use can be daunting. The present research investigates the differences between English learners’ and native speakers’ intuitions regarding the utterance verbs “speak,” “talk,” “say,” and “tell.” The participants were 80 users of English in four proficiency groups (Low, Mid, High, and native). The participants were polled via questionnaire on their intuitions regarding various uses of the four utterance verbs. Data were analyzed and compared with descriptive statistics and t tests. Although the intuitions of learners of increasing proficiency increasingly resembled those of NS, the verbs “speak” and “talk” posed special problems, indicating a lack of understanding of these verbs’ core meanings. Language educators are recommended to pay particular attention to these verbs’ more idiomatic uses (e.g. “talk politics”) to address these deficiencies.

Keywords: utterance verbs; semantically-interrelated verbs; division of labor; constructional ranges; core meaning; lexical hypothesis; lexical acquisition; systematic teaching.

1 Background

Utterance verbs such as “speak,” “talk,” “say,” and “tell” are among the list of English lexical items generally expected to be mastered in the early stages of learning, even for English as a foreign language (EFL) learners. All these four utterance verbs are among the first 1000 items in the General Service List (West, 1953). It may not be as easy to assume, however, that EFL learners will be able to use and comprehend, with clear distinction and to their fullest, each of those semantically interrelated words.

When using utterance verbs in English, it is essential that the user of the language judge whether the content of an utterance can be the direct object of the verb (e.g. “She said ‘no’”), whether two noun phrases representing a person and the content of the utterance can come in a row right after the verb (e.g. “I told you the truth”), or whether the verb should be used as an intransitive one with the help of a prepositional phrase (e.g. “I want to talk with you over the matter”). These kinds of constructions vary from one verb to another; consequently, in most cases it is a formidable task for EFL learners to master these constructions (Hornby, 1975).
Following this line of thought, in learning the basic utterance verbs it is critical that learners grasp the essence of those words, that is, the lexical core of each verb, which can explain and make explicit the division of labor among them. By mastering the core of each of the semantically interrelated verbs in the conceptual field of utterance, and learning to differentiate between those verbs in reference to the core of each verb, learners will be able to learn the correspondence between the concept and its referents, as well as the constructional ranges of each verb, relating to and differentiating from each other. Here, it can become an empirical enquiry as to how learners internalize these verbs in comparison with NSs. Indeed, there has not been much empirical research done to ascertain how well these utterance verbs are learned by Japanese EFL learners, and where they tend to face difficulty with words of this type. Thus, this research aims at answering with quantitative data to the following research questions.

2 Research Questions

1. What proves to be difficult for adult Japanese EFL learners, when they learn the utterance verbs “speak,” “talk,” “say,” and “tell?”
2. Are there differences between learners at different proficiency levels in their degree of utterance verb acquisition?
3. Are there differences in intuition between learners and NSs behind the use of these utterance verbs?

3 Method

A questionnaire comprising 29 items on usages of the utterance verbs was administered to three learner groups of different proficiency levels: High (Japanese university students who had studied in an English-speaking country or area; \( n = 8 \)), Mid (Japanese university students who had been studying English only in Japan; \( n = 36 \)), and Low (Japanese high school students; \( n = 22 \)). Examples of items are shown in the Appendix. Data were also gathered from NS (\( n = 14 \)), who agreed to answer the same questionnaire via an attached file through e-mail.

4 Results and Discussion

Out of the 44 Japanese university students who answered the questionnaire, eight were returnees who had studied in an English-speaking country or area (USA, UK, Australia, and Hong Kong) for more than 12 months (mean: 4 years and 7 months). Comparing the average scores on Part 1 among the eight returnees, the other 36 university students, and the 22 high-school students indicated that the division of the learner participants into three proficiency groups would be meaningful. Table 1 shows the average scores on Part 1 for these three learner groups and the group of native speakers (NS) (the full score being 20). Table 2 shows the rates of the items answered correctly by each of the four groups.

The NS group scored 100\% on 17 out of the 20 questions, which seems sufficient to show that all the items in this section are natural expressions in English, and that they are acceptable for most NS. The High group (returnees) had higher
rates than the Mid (university) and Low (high school) learner groups in general, with all eight returnees scoring 100% on items 8 and 15–19. However, the returnees’ scores went down on items 1, 6, 13, and 20, all of which call for “talk” as the right answer, and also on items 5 and 11, both of which call for “speak” as the correct answer. Remarkably, on item 5, the group of returnees scored lower than the other two learner groups. Figure 1 shows the percentages of correct answers for each of the four participant groups on the items focusing on each of the four utterance verbs.

Table 1. Descriptive Statistics for Total Scores on Part 1

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS (n=14)</td>
<td>19.79</td>
<td>0.43</td>
</tr>
<tr>
<td>High (n=8)</td>
<td>14.75</td>
<td>3.41</td>
</tr>
<tr>
<td>Mid (n=36)</td>
<td>10.50</td>
<td>2.84</td>
</tr>
<tr>
<td>Low (n=22)</td>
<td>7.95</td>
<td>3.33</td>
</tr>
</tbody>
</table>

Table 2. Percentage of Correct Answers to Part 1 Gap-Fill Questions

<table>
<thead>
<tr>
<th>Sentence</th>
<th>NS (%) (n=14)</th>
<th>High (%) (n=8)</th>
<th>Mid (%) (n=36)</th>
<th>Low (%) (n=22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) I couldn’t talk her into coming with us</td>
<td>100</td>
<td>50</td>
<td>36</td>
<td>41</td>
</tr>
<tr>
<td>(2) He spoke up for the union’s strike</td>
<td>100</td>
<td>75</td>
<td>56</td>
<td>23</td>
</tr>
<tr>
<td>(3) I got caught because somebody told on me</td>
<td>100</td>
<td>75</td>
<td>39</td>
<td>32</td>
</tr>
<tr>
<td>(4) What does that sign say over there?</td>
<td>100</td>
<td>88</td>
<td>64</td>
<td>59</td>
</tr>
<tr>
<td>(5) The newspaper editorial spoke highly of the new mayor</td>
<td>93</td>
<td>13</td>
<td>36</td>
<td>27</td>
</tr>
<tr>
<td>(6) Never talk back to your mother</td>
<td>100</td>
<td>38</td>
<td>28</td>
<td>23</td>
</tr>
<tr>
<td>(7) What did you say your name was?</td>
<td>100</td>
<td>88</td>
<td>58</td>
<td>45</td>
</tr>
<tr>
<td>(8) Don’t tell me a lie</td>
<td>100</td>
<td>100</td>
<td>89</td>
<td>59</td>
</tr>
<tr>
<td>(9) Carl said to Linda, “I’ll be back by seven”</td>
<td>93</td>
<td>63</td>
<td>72</td>
<td>55</td>
</tr>
<tr>
<td>(10) I can tell by the smell that it’s garlic and basil</td>
<td>100</td>
<td>83</td>
<td>64</td>
<td>41</td>
</tr>
<tr>
<td>(11) His actions speak for his true feelings</td>
<td>100</td>
<td>50</td>
<td>19</td>
<td>32</td>
</tr>
<tr>
<td>(12) I finally talked her out of a new computer</td>
<td>100</td>
<td>88</td>
<td>28</td>
<td>36</td>
</tr>
<tr>
<td>(13) The receptionist talked me through the application form</td>
<td>100</td>
<td>38</td>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>(14) It is time for all of us to speak out against racism</td>
<td>100</td>
<td>75</td>
<td>69</td>
<td>36</td>
</tr>
<tr>
<td>(15) I wouldn’t say no to a beer right now</td>
<td>100</td>
<td>100</td>
<td>86</td>
<td>36</td>
</tr>
<tr>
<td>(16) It’s easy to tell the expensive wine from the cheap one</td>
<td>100</td>
<td>100</td>
<td>74</td>
<td>27</td>
</tr>
<tr>
<td>(17) Speak up! We can’t hear you clearly</td>
<td>100</td>
<td>100</td>
<td>75</td>
<td>59</td>
</tr>
<tr>
<td>(18) My brother told me not to drive so fast in the rain</td>
<td>100</td>
<td>100</td>
<td>69</td>
<td>73</td>
</tr>
<tr>
<td>(19) The tour guide said to be back on the bus in 15 minutes</td>
<td>93</td>
<td>100</td>
<td>58</td>
<td>50</td>
</tr>
<tr>
<td>(20) Now you’re talking!</td>
<td>100</td>
<td>38</td>
<td>19</td>
<td>23</td>
</tr>
</tbody>
</table>
Additionally, t-tests were performed on the gap-fill responses in order to investigate the differences between the learner groups and the NSs; all the results were significant, except for the uses of “say” and “tell” by the High group (see Table 3).

Between “speak” and “talk,” the lower scores were on items calling for “talk.” Even for the returnees, the six items on which they scored lowest (1, 5, 6, 11, 13, and 20) included as many as four items that called for “talk” as the right answer (1, 6, 13, and 20). These tendencies are not at all observed with the NS.

One may wonder what makes it so difficult to acquire the usages of “talk.” For one thing, the participants may not have learned the use of idiomatic expressions such as “talk back” (6), “talk someone into (doing) something” (1), “talk someone through something” (13). For another, “talk” is generally taught as or supposed to be an intransitive verb, and thus this knowledge, whether it is explicit or implicit, might have served as a restraint against selecting this verb where it can act transitively (1 and 13). This might explain, for instance, why more than half of the learner participants (50% of the returnees, 81% of the middle level and 58% of the low level, respectively) opted for the wrong answer “tell” on item 13. In contrast to “talk,” “tell” is usually taught as or supposed to be a transitive verb, which can take a person as its object.

This verb selection mechanism may remind one of “syntactic bootstrapping” in L1 acquisition (Karmiloff-Smith, 1992; Landau & Gleitman, 1985; Naigles & Hoff-Ginsberg, 1995) in that learners judge the use of a verb based on its syntactic properties. However, “syntactic bootstrapping” is a strategy for taking advantage of the syntactic information of a lexical item to acquire the meaning of the word,
resulting in the integration between the meaning and the construction ranges available to the word. In contrast, the strategy that some of our learner participants might have adopted would have been a mere dependence on the constructional pattern of a verb, isolated from its lexical core. Admitting that an L2 lemma contains both the semantic and the syntactic sides (Jiang, 2000; Levelt, 1989), there would be few correct judgments concerning the constructions available to that L2 lemma if such assessments were based merely on syntactic patterns lacking any kind of inherent connection to the lexical core of the word in question. Such overdependence on constructional patterns results in learners failing to learn the meanings as well as the construction ranges for unknown words.

Along this line of discussion, Levelt (1989) notes “nothing in the speaker’s message will by itself trigger a particular syntactic form. ... There must always be mediating lexical items, triggered by the message.” This “lexical hypothesis” was meant to explain a part of the mechanism for speaking, but it also has relevance here in English as a second language (ESL) lexical acquisition. This hypothesis suggests that a syntactic pattern of a word will not come by itself, but rather it arises as a result of using the word in a certain construction that will help to match it to some part of the message to be conveyed. If this is the case, it then follows that learners are advised to refer to the semantic side, i.e. the lexical core of a word (especially, a verb in our case) before they expect to master possible constructions using the word.

All the discussion here seems to suggest that in learning to use such basic utterance verbs as “speak,” “talk,” “say,” and “tell,” there lies some difficulty in lexical acquisition which may not be simply overcome by exposing learners to a sufficient quantity of authentic L2 input for a certain period of time. To overcome such difficulty would also require a systematic teaching or learning strategy based on the core of each of the verbs.

A brief look at the results for Part 3 of the questionnaire will make the point made above clearer. In Part 3, the NS tended to give a clear “yes” or “no” answer to a given collocation containing one of the four utterance verbs. All the learner groups lacked such a clear-cut response tendency. Following is a brief discussion focusing on “talk + noun phrase.” Figure 2 shows a rough visual scale image of the acceptance rates of the four groups (NS, High, Mid, and Low) for the collocation patterns of “talk.”

As has already been pointed out, “talk” is generally used as an intransitive verb, but it can also work transitively, with an object following it. Even in that case, such noun phrases will be accepted as long as they do not run counter to the core of the verb. The core of “talk” is “verbally interact,” and the typical deciding factor will be verbal negotiation or interaction. Therefore, such noun phrases as are capable of representing a sort of space for verbal negotiation or interaction will have the semantically legitimate right to come immediately after “talk.” This explains collocations like “talk business,” “talk baseball,” “talk nonsense,” and “talk religion.” Indeed, most of the native participants accepted these collocations as natural. In this usage, the noun phrase stands for a kind of interaction space for verbal negotiation, but not for any space belonging exclusively to one of the two in dialogue. This is why “your family” or “our problem” would not be suitable for collocating with “talk.” To use these noun phrases after this verb, one must convert them into prepositional phrases like “about your family” or “about our problem.”
However, the learner groups, including the returnees, had a number of individuals who wrongly accepted such collocations as acceptable. The correct possible collocations of “talk,” on the other hand, were missed by most of the learners.

5 Conclusion

This research aimed to address empirical questions of what points are difficult for adult Japanese EFL learners when learning the four basic utterance verbs “speak,” “talk,” “say,” and “tell”; whether there is any difference in how well they manage to learn these four verbs according to their proficiency level (learning stage); and to what degree learners’ mental lexicons of these four utterance verbs differ from NSs’ intuitions. The data gathered and analyzed showed clearly a problematic situation shared by all of the learner groups, including the group consisting of the eight returnees, which was not observable in the NS group. It was suggested that for Japanese EFL learners, among the four utterance verbs, “talk” and “speak” may be more difficult to learn than “say” and “tell.” “Talk,” especially, caused a number of problem situations. Possible scenarios suggested to explain this include “the idiom effect” as well as the effect of teaching or learning methods (strategies) that depend exclusively on the syntactic properties (i.e. whether a given verb is transitive or intransitive) for making judgments about verbal constructions. This leads to the pedagogical implication of this research. That is, it is advisable to devise ways to teach the constructional ranges of a verb based on its lexical core, rather than instructing the syntactic information of the verb, isolated from its core meaning. Work is continuing on this project, and a follow-up study with a larger number of participants should be able to shed further light on this issue.

References


Appendix: Examples of Parts 1–3 of the questionnaire.

Appendix

<table>
<thead>
<tr>
<th>Part 1 [gap-filling: 20 items]</th>
</tr>
</thead>
<tbody>
<tr>
<td>I couldn’t [          ] her into coming with us.</td>
</tr>
<tr>
<td>1. speak          2. talk          3. say          4. tell</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part 2 [acceptance judgment: 5 items]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Meaning that money helps you out in most hard situations) Money [          ].</td>
</tr>
<tr>
<td>not acceptable          most acceptable</td>
</tr>
<tr>
<td>speaks</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>talks</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>says</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>tells</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part 3 [collocation judgment: 4 items]</th>
</tr>
</thead>
<tbody>
<tr>
<td>talk + [noun or noun phrase]</td>
</tr>
<tr>
<td>business</td>
</tr>
<tr>
<td>our problem</td>
</tr>
</tbody>
</table>
Utilizing Student-Generated Pictures for Formative Vocabulary Instruction

Charles J. Anderson

Kyushu Sangyo University
doi: http://dx.doi.org/10.7820/vli.v01.1.anderson

Abstract

Contemporary vocabulary learning strategies focus on the learner. This may overlook the effect good classroom instruction, such as formative feedback, can have on acquisition. Formative feedback is strongly correlated with positive learning outcomes because it provides explicit information the learner can use to move beyond what is known and towards a learning goal. Drawing pictures of target vocabulary is one activity that supports vocabulary acquisition while also promoting more formative feedback. Research indicates that the drawing of pictures can deepen understanding and improve recall. Furthermore, students' illustrations also facilitate the delivery of more specific feedback than more conventional vocabulary acquisition strategies, provided teachers use them effectively. This activity aids less proficient university students by facilitating better feedback and deepening existing vocabulary knowledge and improving recall.

Keywords: vocabulary learning; vocabulary acquisition; pictures; drawing; illustration; learning strategies.

1 Background

Numerous methods for effectively teaching vocabulary have been proposed in the second language acquisition (SLA) literature over the past 20 years. Schmitt (1997) groups them into two broad categories: discovering strategies for initial learning, and consolidation strategies for remembering and deepening understanding of previously encountered words. A majority of these strategies fall into the more important second category, and most have been shown to be effective. Relying only upon these consolidation strategies, however, risks disregarding other potentially beneficial features of classroom instruction that have been shown to strongly correlate with positive learning outcomes. Of particular interest is the role formative feedback can play in facilitating learning. In meta-analyses, formative feedback has been repeatedly shown to be one of the greatest predictors of learning outcomes (Hattie, 2009). Formative feedback is any information the learner can use to move beyond what is known, towards what they desire to know. However, until now the literature has not discussed teaching methods that facilitate this process. This paper proposes a classroom activity which could facilitate formative feedback from teachers to students, picture drawing, and details its use in a classroom setting over the course of a semester.
1.1 Using Picture Drawing for Formative Feedback in the Classroom

One vocabulary learning strategy that provides repeated opportunities for formative feedback is drawing pictures in support of vocabulary acquisition. This activity asks students to attempt to capture the meaning of the target English word in a simple, quickly drawn picture. This strategy has been positively identified in the literature (Schmitt, 1997), but little research has investigated its efficacy. It is hypothesized that illustrations, similar to the keyword method (Nation, 2001), can improve recall by providing learners with mnemonics, but unlike keywords it has the added benefits of requiring less training, is quick, flexible, and can be used with any student who can put pencil to paper. Perhaps most importantly, it can provide teachers with clear information about learners’ understanding of target words. Drawing pictures can also support vocabulary learning by repurposing existing L1 strategies, including personalization, elaboration, deeper semantic and lexical understanding as well as meta-cognitive approaches to study. These are strategies many low-proficiency students lack (Mochizuki, 1999), but can use, provided they receive adequate support.

An additional benefit of picture drawing is that it promotes more formative feedback from teacher to student. It allows the teacher to determine what aspects of word meaning the learner has obtained and what is yet to be mastered, which facilitates more effective feedback and learning. Studies have shown the teacher leading an activity often has a larger learning effect than any particular strategy (Timperley & Alton-Lee, 2008). It is of course considerably more difficult to identify what makes a good teacher than a good task (Korthagen, 2004), but a major feature of effective instruction closely linked to positive learning outcomes is the use of formative feedback (Petty, 2006). Formative feedback is more effective than summative feedback (such as test results) because it provides explicit and specific instruction that enables a learner to engage with an activity, monitor their progress and, upon completion, move towards a new goal and increased self-regulation (Black & William, 1998). The effectiveness of appropriate formative feedback to learning is reflected in consistently high effect sizes across numerous meta-analyses that see it as one of the most important contributors to learning (Hattie, 2009). If such feedback is critical to effective learning, then vocabulary acquisition strategies facilitating more effective feedback should promote positive learning outcomes. Picture drawing activities can facilitate such outcomes.

2 The Activity

This activity has been used successfully in low English proficiency (as measured by the TOEIC Bridge Test) Japanese university conversation classes. Before starting to draw pictures students need knowledge of, or access to, both the words under study and their L1 equivalents. Students were first required to study 40 words weekly and complete a vocabulary notebook containing the words, two of their most common L1 equivalents and example sentences for the highest-frequency words. After completing this preliminary step the picture drawing strategy was introduced. Training students on how to use the strategy was relatively simple, with a brief introduction outlining the approach. Students were then encouraged to reflect on how they acquired L1 logographic fluency to demonstrate how the strategy could support their current vocabulary study methods. Students were then provided with a variety of examples. They were next asked to identify good and poor
exemplars and to support their opinions. Examples of symbolic shorthand were identified and promoted through examples that showed how different parts of speech might be represented visually. A demonstration of how to draw appropriate pictures using a think-aloud procedure was employed to ensure students understood the goal of the exercise: to use pictures to facilitate recall and deeper lexical understanding. Much of this consisted of what Carless (2007) has labeled “pre-emptive formative feedback” and was intended to avoid problems previously encountered and provide guidance to improve production of effective pictures.

The worksheet containing space for an English word, the Japanese translation and picture was then handed out. Students first attempted the task in the following order: (1) Write the target English word. (2) Write one unknown or poorly understood Japanese definition. (3) Draw a picture to promote recall and understanding of the target English word and the L1 translation. Extensive individualized feedback was provided to all students while engaged in the task to ensure students understood the strategy and how to implement it.

In order to gauge understanding, the L1 and L2 meanings were obscured and students were asked to produce the meaning with only the picture as a cue. Students who were unable to recall the L1 and L2 meanings were encouraged to create better pictures through more personalization or the addition of more detail. Students who were able to only recall the L1 meaning were asked to rehearse the L2 meaning more as they drew the picture. Students who were able to recall both the L1 and L2 meanings were encouraged to elaborate on the word’s meaning, part of speech or usage. Words that were consistently difficult to recall or draw were identified and a think-out-loud procedure was used to model potential solutions to the entire class.

Students drew pictures of the remaining 40 words for homework. The following week a short 15-item, low-stakes test containing the previously studied words was then administered to provide students and teachers with summative feedback on progress. Students who performed poorly on these tests received a further round of formative feedback on their previous drawings. To ensure student participation, appropriately completed homework received 2% per assignment (2% × 10 = 20%) of the final grade. Incomplete homework received no score, but was required to be completed in order to receive a final passing grade.

2.1 Facilitating Feedback

All feedback provided was focused on improving the student’s ability to use pictures to recall target English words. Students who were able to successfully recall both the target word and the L1 equivalent were then encouraged to use the target word productively. Students who were able to recall the L1 meaning but not the L2 target were encouraged to process the task using more L2. Students who were unable to recall the target L2 word, nor its L1 equivalent were encouraged to engage more deeply in the task and were provided with more instruction. Students who provided incorrect answers were asked to refine their picture with more detail. An examination of students’ drawings and teacher–student interactions should illustrate how formative feedback was used to support the activity:

An important first step in promoting deeper vocabulary acquisition is to ensure students have established a strong L1–L2 link (Schmitt, 2008). Monitoring was conducted by covering the target word and translation and asking students to
recall the target word, which in itself is a valuable learning strategy (Nation, 2001). A quick glance at student A’s pictures (Figure 1a–f) allowed for a quick evaluation of depth of engagement with the task. Pictures indecipherable to the teacher such as Figure 1a are a good focus for feedback. In this instance, student A was able to successfully recall the word “true” and a follow-up question as why recall was possible elicited student A’s love for abstract art and how it felt “true” which demonstrated proper engagement with the activity.

Comparing multiple pictures also provides evidence of students’ deeper lexical knowledge which is seen as a potential contributor to vocabulary learning (Oxford & Crookall, 1990). Student B’s illustrations (Figure 1g and h) show grammatical marking, whereas student A’s illustrations (Figure 1b and f) show none. A recall check confirmed student B was able to recall both words in their correct L2 adjectival form, explaining that the “stars” in both pictures represented “ideas not things.” Student A after recalling Figure 1b as “run,” not “quick,” explained it looked like a person running. Each student’s response to the retrieval cue stimulated a deeper teacher–student discussion into how their pictures might be improved to provide better retrieval and deepen grammatical understanding. As student B had demonstrated the ability to process pictures for grammar, the feedback focused on extending this processing to other words by asking “Could you use this strategy for other kinds of words?” This interaction encouraged B to develop symbols for verbs and adverbs, later adding detail (not shown) to Figure 1i and j. Student A’s feedback was more reflective in tone; responding to the question of “Why was ‘surprised’ (Figure 1f) easier to recall than ‘quick’ (Figure 1b)?” Student A suggested erasing the extra detail in “quick” to make it look more like “surprised.” This in turn led to a comparison of Figure 1d–f and a discussion of how words might be visually tagged for grammar. In this way students’ responses, correct and incorrect, in conjunction with pictures, provided both student and educator an opportunity to discuss more efficient task processing, which was potentially beneficial for student learning and teaching practice.

Figure 1. Examples of student-generated pictures.

Vocabulary Learning and Instruction, 1(1), 37–43.
Pictures facilitate delivering motivating, formative feedback in support of improved meta-cognitive awareness (Schmitt, 1997) relatively easily. When student B was unable to recall either the Japanese or English meaning of Figure 1i, the following interaction resulted:

Teacher: “Can you describe the picture?”
Student: “The man simply spoke.”
Teacher: “Try and add detail. What is he simply talking about”
Student: “He is talking about the simply (sic) results.”
Teacher: “Why simple? What results do you find simple?”
Student: “Because graph(sic).”(sic)
Teacher: “Oh do you think graphs display information simply?”
Student: “Yes.”
Teacher: “OK. Try to personalize it” “Who is showing the graph?”
Student: “This teacher is displaying simply.”(sic)
Teacher: “Who is he displaying it simply to?”
Student: “To students”
Teacher: “OK. Now create a mental image of a teacher simply explaining a graph.
Student: (Thinks)
Teacher: “Now can you remember ‘simply’?”

When the student answered affirmatively the teacher was free to move on. A quick re-check later demonstrated the student’s fast and accurate recall and evidence of engagement.

Pictures easily support the addition of more detail to help contextualize semantic elaboration (Oxford & Crookall, 1990), or integrate the word into an existing schema (Sökmen, 1997). Initially, the picture shown in Figure 1j consisted of only a stick figure and the “sweat.” A quick retrieval check displayed confusion between “hot” and “pressure” which led to the following interaction:

Teacher: (Pointing to the picture in Figure 1k) “What is this word in Japanese?”
Student: “あつい” (Hot).
Teacher: “Why do you think so?”
Student: “He looks hot.”
Teacher: (Revealing the Japanese and English) “What’s the difference between “hot” and “pressure”?
Student: (No answer)
Teacher: “When do you feel pressure?”
Student: “When I do homework.”
Teacher: “Why not draw when you feel pressure?”

As a result of this interaction, student C included books on either side to represent homework, and a subsequent check showed no further confusion between “hot” and “pressure”.

Often students who displayed an ability to quickly recall the Japanese meaning but failed to recall the target English word reported thinking about the meaning while drawing, but when questioned further indicated they were thinking about the Japanese meaning. The suggestion to think about and repeat the word in English prior to and while drawing the picture appeared to result in better recall, but more research will be needed to confirm this observation.

Anderson: Student-Generated Pictures 41
Vocabulary Learning and Instruction, 1(1), 37–43.
These exchanges are just a small sample of the formative feedback students received while engaged in the drawing activity. While the author acknowledges other activities, such as writing personalized sentences may also increase opportunities for feedback, pictures are unique because all students, regardless of L2 proficiency, can express the degree to which they understand a word. It can also be argued that the drawing in Figure 11 provides better evidence of comprehension of the word “release” than an equivalent example sentence such as the man was happy when he was released from jail. Pictures also provide more insight into the individuals’ learning process with accordingly increased opportunities for feedback on task processing, comprehension, and directions for future study.

3 Conclusion

The drawing of pictures to facilitate better word learning is a relatively simple but powerful and underutilized learning strategy. However strategies alone are not enough; the role of an efficient teacher and engaged students cannot be underestimated (Gu, 2003). What is then left is the need for a level-appropriate, engaging strategy that also serves to facilitate teacher–student feedback. Drawing pictures meets these criteria, and offers educators and learners a novel, yet effective activity with which to support the complex and difficult task of acquiring word knowledge.

References


In this paper I would like to briefly overview vocabulary testing literature and discuss the four empirical studies conducted by Jeffrey Stewart, Rie Koizumi, Aaron Batty, and Tatsuo Iso, after placing them in the framework of the three dimensions of vocabulary knowledge: size, depth, and lexical accessibility.

Keywords: three dimensions of vocabulary knowledge: size, depth, and lexical accessibility; passive/receptive vocabulary knowledge; active/productive vocabulary knowledge.

1 Tests of Vocabulary Size

The first dimension of vocabulary knowledge (Daller, Milton, & Treffers-Daller, 2007) is size or breadth. Researchers have developed a variety of measures to assess vocabulary size. Initially a number of attempts were made to estimate English native speakers’ vocabulary sizes by sampling words from a dictionary and challenging test-takers to give definitions of those words. Subsequently more sophisticated methods of testing vocabulary have been proposed: The Vocabulary Levels Test (VLT) (Nation, 1990, 2001; Schmitt, Schmitt, & Clapham, 2001); The Eurocentre Vocabulary Test (Meara & Jones, 1990); The Mochizuki test (Mochizuki, 1998); The Vocabulary Size Test (Nation & Beglar, 2007). All these tests are considered to measure test-takers’ receptive vocabulary sizes.

On the other hand, very few attempts have been made to assess test-takers’ productive vocabulary sizes. Laufer and Nation (1999) developed a productive version of the VLT which challenged test-takers to complete a word in a sentence context whose first few letters were given as a cue. It is possible to estimate test-takers’ productive vocabulary sizes with the Productive Levels Test, even though the VLTs were intended to diagnose the frequency levels learners needed to study (Nation, 1990). Another attempt to measure productive vocabulary size is Lex 30 (Meara & Fitzpatrick, 2000). In this test, participants give as many associations as they can think of from 30 stimulus words. The number of associations beyond the most frequently used 1000-word level is counted and used as an indicator of the test-taker’s productive vocabulary size. Laufer and Goldstein (2004) proposed to measure vocabulary knowledge strength because learners differ in their vocabulary knowledge along the two axes, active/passive and recall/recogn. Active recall knowledge corresponds to productive vocabulary in that a learner is able to produce a second language word on his or her first language word cue. Passive recognition knowledge may correspond to what multiple-choice vocabulary size tests measure, i.e., receptive vocabulary. Mochizuki (2007) developed a computer program vocabulary test, the J8VST, which measures Japanese English as a foreign language (EFL) learners’
vocabulary sizes in four modes: L2 recall; L1 recall; L2 recognition; and, L1 recognition. It is possible to estimate productive vocabulary size by using the J8VST.

There have been several studies that compared L2 learners’ receptive and productive vocabulary sizes using the receptive and productive Levels Test (Fan, 2000; Laufer, 1998; Waring, 1997). Webb (2008) criticized these studies for comparing the two dimensions of vocabulary knowledge by using measures that favor the estimation of receptive knowledge. He used recall tests for both measuring receptive and productive vocabulary sizes and found that the scores on the two vocabulary tests were almost the same when a sensitive scoring method was employed, although the scores on the productive vocabulary test were lower than those of the receptive one when strict marking was used.

What is at issue now is that there are no established productive vocabulary size tests that can estimate test-takers’ productive vocabulary in a valid, reliable, and efficient manner. An L1 to L2 recall test seems to be a valid and reliable way of testing productive vocabulary but responses must be marked manually. A recall test in a computer program, like the J8VST, can be administered more efficiently than the one marked manually, but it requires an enormous amount of time to develop.

With productive vocabulary measurement in this state, Jeffrey Stewart’s study is an epoch-making breakthrough. He has developed a productive vocabulary test, The KSU Active Multiple-Choice Test, or “Active MC” test, which requires test-takers to recall an L2 word based on an L1 definition prompt. The second and third letters of the target word as well as a blank for each of the remaining letters are also provided as a hint and to avoid other possibly correct answers. The test-takers simply mark the first letter of the target word on an answer sheet. Stewart found an extremely high correlation (0.93) between this new test and a conventional recall test of the same items. The most appealing feature of this new test is that it can be marked by an optical mark reader. It releases researchers from troublesome manual marking, saving great amounts of time and effort. It will definitely increase the number of studies that address the productive vocabulary of English as a second language (ESL) and EFL learners.

Stewart’s Active MC format tests learners’ productive vocabulary in the discreet, selective, and context-independent dimensions of the vocabulary test categorization framework proposed by Read (2000). Researchers have also been investigating learners’ productive vocabulary in the other poles of the Read’s framework: the embedded, comprehensive, and context-dependent dimensions. They have developed indices that are intended to measure the lexical richness, lexical diversity, or lexical density of learners’ writing or speech (e.g. Laufer & Nation, 1995). Probably the type-token ratio (TTR) is the best-known measure of lexical diversity though it has the drawback of being affected by text length. Consequently, other indices have been proposed to overcome this defect: the Guiraud index; D; and, the measure of textual lexical diversity (MTLD). Rie Koizumi investigated the reliability of these four measures of lexical diversity when measuring texts of differing lengths. She found that among the four measures, MTLD was least affected by text length. Koizumi also concluded that texts should be at least 100 words long for lexical diversity investigation. These findings hold promise for researchers investigating L2 learners’ speaking or writing skill development. They are now able to measure the lexical diversity of learners’ speaking or writing performances.
at relatively early stages of language development because they require samples of learner production as short as 100 words.

However, it may be argued that researchers should take samples produced on a variety of topics when examining the lexical diversity of learner production, because lexical diversity is likely to vary according to the topic under consideration. For instance, learners can produce a number of different subordinate words on a topic they are familiar with. A learner who loves dogs, for instance, may produce various names like Scotch terrier, Pomeranian, Siberian Husky, etc. on the topic of dogs. If the same learner has little interest in birds, however, she/he will likely keep using the common word *bird* on that topic because of her/his ignorance of bird types. This learner’s lexical diversity index of a text on dogs will be much higher than that of a text on birds. Thus, it is important to collect samples of texts produced on a variety of topics in order to show the general development of lexical diversity among learners.

It should also be noted that lexical diversity is but one of the properties of text production: there are other properties that characterize text quality such as cohesion and coherence, organization, grammar, etc. In the case of beginners who produce texts as short as 100 words, it may be argued that lexical diversity does not have a high priority: pronunciation, delivery, and size may be more important in speech; grammar, cohesion and coherence, and organization may be more crucial in writing.

Despite the above, Koizumi’s findings still lend strong support for the measures of lexical diversity as a tool for investigating productive vocabulary in the embedded, comprehensive, and context-dependent dimensions.

2 Tests of Vocabulary Depth

We have been talking about the vocabulary size dimension, especially productive vocabulary. Now let us turn to the depth dimension. In my opinion, it was Harold Palmer, often considered the “father of British applied linguistics” (Stern, 1983, p.100), who first drew attention to one aspect of vocabulary depth, collocation. He not only selected 3000 headwords, which he argued together with their commonest derivatives would cover 95% of the contents of all ordinary English texts (Palmer, 1931), but also made a tentative list of English collocations for technicians so that they could apply it to textbook compilation (Palmer, 1933). This shows his understanding of the importance of vocabulary depth knowledge and innovative insight into vocabulary teaching. Other researchers also advanced our understanding of the components of word knowledge in the first half of the 20th century. Along with Palmer, Michael West and Laurence Faucett, both experienced teachers and researchers, and E.L. Thorndike, a statistical linguist, agreed on seven criteria for vocabulary selection for ESL learners: (1) word frequency; (2) structural value (all structure words included); (3) universality (words likely to cause offence locally—excluded); (4) subject range (no specialist items); (5) definition words (for dictionary-making, etc.); (6) word-building capability; and (7) style (“colloquial” or slang words—excluded) (Howatt, 1984, p. 256). Using these criteria they produced the so-called “Carnegie Report” on vocabulary selection in 1936, which West later published as A General Service List of English Words (GSL; West, 1953).
These works by Palmer (1931, 1933) and West (1953) demonstrate that the researchers understood the necessity of teaching vocabulary depth knowledge as well as increasing learners’ vocabulary sizes. In the same year as the GSL was published, Dolch and Leeds (1953) emphasized the importance of measuring less common usages of words. They argued that vocabulary tests should measure different senses of a word because understanding the sense a speaker or writer intends to convey is crucial in successful communication. This argument seems, to the best of my knowledge, to be the first to address the necessity to measure vocabulary depth and degrees of learners’ developing vocabulary knowledge.

As a method of measuring the degree of a learner’s vocabulary knowledge, Dale (1965) proposed a four-stage self-report scale:

Stage 1: “I never saw it before.”
Stage 2: “I’ve heard of it, but I don’t know what it means.”
Stage 3: “I recognize it in context — it has something to do with . . .”
Stage 4: “I know it.”

In this scale learners self-report their knowledge of individual words in four stages. Results reveal how firmly a learner thinks he knows each word. The scale was intended for L1 English learners. Later, Paribakht and Wesche (1993) proposed a similar scale, the Vocabulary Knowledge Scale, for L2 learners with the final stage asking if a learner was able to use the word productively.

There have been a number of attempts to measure some aspects of vocabulary depth knowledge. For example, Ordonez, Carlo, Snow, and McLaughlin (2002) employ definition tasks that they claim elicit not only paradigmatic but also syntagmatic knowledge of a word and thus measure depth of vocabulary knowledge. In investigation of other components of word knowledge, researchers have invented their own measurements, for example: affix and association (Schmitt & Meara, 1997); synonym and collocation (Mochizuki, 2002); antonym, derivation, and collocation (Koizum, 2005). These are mostly one-time studies that are not replicated or modified in other studies and thus the testing methods are not developed or improved, let alone, standardized.

Among a number of attempts to measure depth of vocabulary the Word Associates Test (WAT) (Read, 1993; 1998) has been most widely employed by researchers (e.g. Qian, 1999, 2002). Although the WAT appears to test synonyms and collocates, one study showed that the test might measure only one dimension; i.e. vocabulary (Schoonen & Verhallen, 2008). This is what Aaron Batty addressed in his study. He showed that the WAT best fitted a bifactor model that presumes the test measures synonym and collocate factors, as well as a general vocabulary factor. This model is based on the assumption that all test items are affected by the general vocabulary factor in addition to half of the items being affected by the synonym factor and the other half by the collocate factor. So Batty revealed that the WAT measures not only the general vocabulary knowledge but also the synonym and collocational aspects of vocabulary.

One thing I am particularly concerned about with the WAT is a high correlation coefficient with a vocabulary size test. Qian (2002) reported a high correlation, 0.88, between his WAT and the VLT (Nation, 1990) scores. This means that the two tests share more than 77% of their variance. This raises doubts on...
the validity of the WAT as a measure of vocabulary depth: it may measure the size dimension more than the depth dimension. It might be regarded as another measure of vocabulary size rather than a measure of vocabulary depth. Batty reported that the highest factor loadings were for vocabulary g-factor, with half of the synonym items and only two collocate items loading higher on the corresponding factors than on the g-factor. Because the vocabulary g-factor and the synonym factor are considered to be major components of vocabulary size tests, the WAT can be primarily seen as a vocabulary size test with a minor component of vocabulary depth.

So what makes the WAT a vocabulary size test? In my opinion it is the synonym section. Although it tests different senses of adjectives, it may be argued that learners’ knowledge of different senses of polysemous words is highly correlated with their vocabulary size in that both are based on their knowledge of word meaning. Thus, it seems on the surface that the WAT measures the depth aspect of polysemous meanings but it actually measures the vocabulary size dimension.

From this it may be concluded that although Batty’s study indicates that the WAT is best fitted with a bifactor model, it would seem more appropriate to regard it as another test of vocabulary size. Therefore, we should make further efforts to create a measure of vocabulary depth. Or should we review the construct of vocabulary knowledge as Vermeer (2001) claimed that there was no conceptual difference between breadth and depth measures of vocabulary?

3 Tests of lexical access

The third dimension of vocabulary knowledge is lexical accessibility or fluency. Although lexical accessibility has been vigorously investigated in psycholinguistics, only a limited number of researchers have addressed the issue in applied linguistics (Aizawa & Iso, 2010; Coulson, 2005; Kadota, 2010; Meara, 2005). These studies share a common feature in that they used self-designed computer programs to measure lexical access time instead of established software such as SuperLab. Coulson (2005) used a computer program called Q_Lex to measure lexical access time. This test challenges test-takers to find an English word in a string of letters and measures the time between when a string appears on the computer display and when a test-taker clicks the button to indicate he/she has found the word in the string. Aizawa and Iso (2010) reported their attempts to create a computer program, the Lexical Access Time Test (LEXATT), and found that measuring the time duration while a test-taker is pressing a key is a reliable way of lexical access time measurement. Tatsuo Iso subsequently revised LEXATT as LEXATT2 for more accurate measurement. Kadota (2010) applied the psycholinguistic method of measuring priming effects to measuring lexical access time. He developed a computer program, the Computer-Based English Lexical Processing Test (CELP), which shows a test-taker a prime and then a target word and measures how fast the test-taker judges whether the prime and the target word are semantically related or not. Although studies show that there are significant correlations between CELP scores and English skill test scores (Hase & Shiki, 2011; Nakanishi & Sugiura, 2011), no studies have been conducted to validate CELP as a test of measuring lexical access speed.

In such circumstances LEXATT2 will likely become a standard tool for lexical access measurement for two reasons. First, it is a valid test of lexical
access time. Less proficient test-takers responded more slowly to stimulus words than their more proficient peers and took more time in responding to longer words while proficient test-takers reacted in similar times irrespective of word length. These findings can be seen as evidence of the validity of LEXATT2 for measuring lexical access time. Second, LEXATT2 is extremely practical. Learners can take the test on-line and it requires only 10–15 minutes to complete. So it is a highly practical tool for lexical access time measurement compared with conventional recognition speed tests.

Although LEXATT2 has been validated by Iso, further validation by comparing its reaction times with those of SuperLab would establish its role as a test of lexical access time.

4 Conclusion

The four studies I have examined in this paper had their roles in the framework of vocabulary testing. They have contributed a great deal to the development of each of the vocabulary research fields. I would like to close this paper with suggestions to the four researchers for further studies. First, Jeffrey Stewart could make the most of his productive vocabulary size test by using it to investigate relationships between ESL learners’ productive skills and productive (active) vocabulary sizes. Though there are a number of studies investigating relationships between learners’ skills and receptive vocabulary sizes, very few studies have addressed how their productive vocabulary sizes are related to their productive skills. Stewart’s new productive vocabulary size test should promote explorations into this area. Second, Rie Koizumi could explore how ESL learners’ speeches and writings may be accounted for by different factors such as lexical density, vocabulary size, grammar, cohesion and coherence, and discourse organization. It would be a great discovery to find the extent to which different factors account for productive skills. Third, Aaron Batty could continue examining what the WAT actually measures. Although he found that the WAT fits the bifactor model best, the test may be just another vocabulary size test. He could identify what causes it to behave more like a size test (what would happen if low-frequency words were removed?) rather than a depth test, and then go on to create a valid vocabulary depth test. Finally, Tatsuo Iso could probe relationships between English skills and lexical access time using LEXATT2. It would contribute a great deal to vocabulary acquisition research if it was revealed what role lexical access plays in L2 skill performances.

It is important to further investigate vocabulary testing in the vocabulary knowledge framework so that learners’ vocabulary knowledge can be better diagnosed. It is also necessary to try to integrate tests of different dimensions so that results can show test-takers’ vocabulary knowledge as a whole and better predict their language proficiency.

References


Vocabulary Learning and Instruction, 1(1), 44–52.


Hase, N., & Shiki, O. (2011). What the data from CELP test tell us: Relationship between CELP test and reading measurements for fluency. In N. Hase (Chair), Exploring the roles of lexical processing in second language proficiency: Using the newly developed vocabulary processing test. Symposium conducted at the JACET 50th Commemorative International Convention, Seinan Gakuin University, Fukuoka, Japan.


*Vocabulary Learning and Instruction, 1*(1), 44–52.


A Multiple-Choice Test of Active Vocabulary Knowledge

Jeffrey Stewart
Kyushu Sangyo University
doi: http://dx.doi.org/10.7820/vli.v01.1.stewart

Abstract

Most researchers distinguish between receptive (passive) and productive (active) word knowledge. Most vocabulary tests employed in second language acquisition (SLA), such as the Vocabulary Levels Test (VLT) and Vocabulary Size Test (VST), test receptive knowledge. This is unfortunate, as the multiple-choice format employed on most receptive tests inflates estimates of vocabulary size, and there are clear theoretical advantages to focusing instead on productive knowledge, which is associated with greater strength of knowledge as well as written and oral communication skills. This is in large part due to the logistical problems associated with such tests, as the full-word answers given must either be entered online or hand-marked. This paper will describe a multiple-choice format test of active vocabulary knowledge, in which learners confirm their knowledge of an English word by selecting its first letter. As there are 25 possible options, odds of guessing the correct answer by chance are reduced to 0.04. Findings of the study include that word difficulty estimates and scores are highly correlated to those of conventional, full-word active tests (>0.90), and that test reliability is higher on the proposed format than on that of a receptive test of the same words.

Keywords: vocabulary acquisition; productive vocabulary knowledge; language testing.

1 Background

Most researchers distinguish between receptive and productive word knowledge (e.g. Meara, 1990). Receptive, or “passive” knowledge, associated with the skills of reading and listening, involves retrieving meanings of words once they are presented, implying comprehension of input. Productive, or “active” knowledge, however, is associated with speaking and writing, and involves retrieval of form once the word is required in written or spoken contexts.

“Passive” tests of second language vocabulary knowledge are numerous, and include the Vocabulary Levels Test (VLT; Nation 1990) and the Vocabulary Size Test (VST; Nation & Beglar, 2007). Such tests typically operationalize assessment of receptive knowledge with a multiple-choice format, presenting word forms and asking learners to select the correct meanings from a list of options.

“Active” tests of knowledge, in which learners are asked to provide the word forms themselves, also exist, though appear to be used less frequently in practice. A drawback of such tests is that written answers typically must be hand-scored. In addition to being time-consuming, this procedure can also lead to inconsistencies between raters regarding accepted responses. To address these issues, such tests are
typically computer-delivered to expedite marking, and/or require correct spelling on answers. Though technology continues to become more widespread in educational institutions, at present the employment of computers in the testing of large groups still often poses logistical problems for educators who wish to test learners’ vocabulary knowledge in applied contexts. Consequently, researchers and test-makers tend to employ multiple-choice formats, which are most suitable for assessment of receptive knowledge. This is unfortunate, as research indicates that although active knowledge is more difficult to test, the construct may play an important and distinct role in diagnosing language proficiency. Laufer and Goldstein (2004) administered tests of both active and passive vocabulary knowledge, and found that active tests of vocabulary were of considerably greater difficulty than those of passive knowledge. Consequently, the researchers concluded that the active test measured a construct of vocabulary “strength”, which appeared to develop at higher stages of proficiency. This active construct of vocabulary knowledge could be of particular value in the assessment of sub-skills of language proficiency that extend beyond the receptive processing of input. In Japan, the bulk of English education involves the comprehension and translation of written texts, with few opportunities to produce language (see Barfield, 2012, this issue). While pressure to pass entrance exams has created a culture in which Japanese students are adept at writing multiple-choice tests, it is frequently observed that even students capable of high marks on English exams have considerably more difficulty expressing their own thoughts in spoken or written forms. Pedagogy that stresses the production of studied vocabulary could have greater washback on spoken and written proficiency.

A further consideration is the degree to which the multiple-choice format commonly employed in receptive-knowledge tests compromises measurement. Multiple-choice tests have become popular due to their practicality and cost-effectiveness rather than their statistical properties, as multiple-choice answer sheets can be collected and quickly scanned by a computer using optical mark recognition software. Yet despite these practical advantages, they remain most suitable for assessment of lower-order skills (Phelps, 1996), and pose various challenges to measurement. It has been acknowledged that guessing on multiple-choice tests can adversely affect test reliability (Zimmerman & Williams, 1965), and the multiple-choice format employed on vocabulary tests such as the VLT has been shown to inflate estimates of learner vocabulary size (Stewart & White, 2011).

Ideally, then, it would be desirable to find a vocabulary test format that held not only the theoretical and statistical advantages of a test of active recall, but also the practical advantages of a multiple-choice format that can be processed using a scanner and OMR software, without requiring either a computer for each student or a complex hand-marking procedure carried out by human raters on collected paper tests. The purpose of this paper is to describe such a test format, and compare its reliability and difficulty to both a conventional active recall test and to a popular measure of receptive vocabulary knowledge, the VST.

1.1 The KSU Active Multiple-Choice Test Format

The KSU Active Multiple-Choice Vocabulary test (hereafter referred to as the “Active MC”) is a pencil-and-paper, OMR-scannable multiple-choice test of active vocabulary knowledge. The novel multiple-choice format is made possible
with Remark OMR software, which allows custom design of answer sheets. Prompted by L1 definitions, the part of speech, the second and third letters in the word and the number of letters contained in the entire word, students are asked to provide the first letter of the target word by selecting the correct letter from multiple-choice bubbles that list every letter in the English alphabet (with the exception of the letter x, which is removed for brevity). Student instructions for the test format and an example item are depicted in Figure 1.

Without the L1 definitions, the target word in the example shown in Figure 1 is essentially unrecognizable from the hints given about its form. This is by design. As the intention of the test is to assess active recall of L2 vocabulary, it is essential that the hints provided prompt recall of the word form, rather than recognition. The additional information simply prevents the selection of synonyms that could also correspond to the L1 translations provided.

2 Research Questions

This format could provide a variety of advantages for test-makers. As conventional distractors are not employed, the process of item-writing is greatly simplified, yet the format retains the advantages of a conventional paper-and-pencil multiple-choice test, as the format is machine-readable, and tests can be scored by computer. However, despite these advantages, a number of questions remain before the format can be used with confidence:

*To what degree does selection of a word’s first letter correlate with knowledge of the entire word?*

The reader may note that simply providing the first letter of a word may not be equivalent to providing the entire word. While the correlation between the two forms of knowledge is unlikely to be perfect, we wish success on the first task to approximate success on the second as closely as possible.

*How do statistical reliability and estimates of word difficulty under the proposed format differ to those of conventional test formats?*

This format also holds the potential of reducing the measurement error associated with conventional multiple-choice formats: though options are given, the greatly extended number of choices restricts the probability of a correct answer due to random guessing to 0.04. Even if test takers become sophisticated in their strategies and choose more frequently occurring letters, the number of plausible options remains far higher than conventional formats. To test this hypothesis, however, it is necessary to compare test reliability to that of a test of the same words that uses a conventional multiple-choice format.
3 Results

3.1 Correlation to a Conventional Test of Active Knowledge

In order to compare scores on the proposed Active MC format to scores on a conventional active recall test in which the entire word must be provided (henceforth referred to as the “Active Full Word” format), second year English conversation students at a Japanese university (n = 205) wrote a test of the second 1000 most common words in English (k = 40) employing the Active MC format. Upon completion, students were asked to then write the words in full. These full answers were then hand-marked. Only words written with correct spelling were marked as correct, as this practice approximates the scoring procedure of popular tests of active recall such as Laufer’s CATSS test (Laufer & Goldstein, 2004). Tests in which answers were not provided for one of the formats were omitted from the sample. The resulting scores from the two formats were then compared.

The Active MC test had a mean of 33.70 points with a standard deviation of 5.88. The Active Full Word test had a slightly lower mean of 30.71 points, with a standard deviation of 7.67. The two test formats were highly correlated at 0.93, despite a somewhat restricted range of scores. This suggests that while the relationship is not perfect, the proposed Active MC test format closely approximates word knowledge under the active recall construct of vocabulary knowledge (Figure 2).

![Figure 2. Correlations between Active MC format (x) and full-word active recall format (y).](image)

3.2 Comparing Reliability and Estimates of Word Difficulty Between Formats

The first three 1000-word levels (k = 30) of a Japanese-bilingual version of the VST (Nation & Beglar, 2007) was adapted to the active recall format by using the correct Japanese definitions on the VST as prompts for recall of word. This created two tests of the same words, allowing for comparisons of the two formats.

The active recall test was given to a sample of first and second-year English conversation students at a Japanese university (n = 119). The tested group’s scores on the TOEIC Bridge Test ranged from 90 to 140. The group was asked
to select the first letters of the tested words and to then write the words in full. These full answers were then hand-marked with the entire answer examined rather than simply the first letter, resulting in a conventional measure of active vocabulary knowledge (‘Active Full Word’ format). Once again, only words written with correct spelling were marked as correct, and tests in which answers were not provided for one of the formats were omitted from the sample.

This allowed for direct comparisons between active recall formats, but a challenge of the study was that the VST test format provides the word form and requests a definition, whereas the active recall formats provide the definition and request the word form. Therefore, it was not possible to administer both tests to the same students, as one test would reveal the answers of the other. Consequently, the Japanese-bilingual VST-format test was given to a second group of students of comparable proficiency \((n = 125)\). Correlations, item difficulties and split-half reliabilities between the resulting three test formats were then compared. Means, standard deviations and Cronbach alpha reliabilities for the three tested formats are listed in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>VST</td>
<td>19.54</td>
<td>3.11</td>
<td>0.60</td>
</tr>
<tr>
<td>Active MC</td>
<td>12.00</td>
<td>3.48</td>
<td>0.70</td>
</tr>
<tr>
<td>Active Full Word</td>
<td>7.73</td>
<td>3.77</td>
<td>0.81</td>
</tr>
</tbody>
</table>

At 19.54 points with a Cronbach alpha value of 0.60, the VST-format test had the highest mean and the lowest reliability. The Active Full Word test of the same words had a dramatically lower mean of only 7.73, though reliability was substantially higher, with a Cronbach alpha of 0.81. The primary test format under investigation, the Active MC, ranks between the two both in terms of difficulty and internal reliability.

Score correlations between the VST format and Active format tests were not possible, as the two formats were written by different groups of participants. However, it is possible to correlate item difficulties for each word, to determine if word difficulties are consistent between test formats (see Table 2). Though lower between the original VST and the Active Recall formats, the correlation between the Active Recall tests was again very high, at 0.90.

<table>
<thead>
<tr>
<th></th>
<th>VST</th>
<th>Active MC</th>
<th>Active Full Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>VST</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active MC</td>
<td>0.61</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Active Full Word</td>
<td>0.63</td>
<td>0.90</td>
<td>1</td>
</tr>
</tbody>
</table>

In order to examine relationships between test formats and estimated word difficulty in detail, item facilities for the tested words under all three formats are graphed in Figure 3. An item facility of 1.00 indicates that all respondents answered correctly; an item facility of 0 indicates none did.

*Stewart: Multiple-Choice Active Vocabulary Test* 57

_Vocabulary Learning and Instruction_, 1(1), 53–59.
There are clear differences between difficulty estimates under the VST format and the two Active formats. For the first 10 items, which test the first 1000 most common words in English, the VST and Active formats behave similarly. But from the second 1000-word level, results begin to diverge. With the exception of “reward” which was very difficult under all formats, few words on the VST format demonstrated item facilities below 0.2, indicating that chance guessing on the multiple-choice VST format does, in fact, affect item difficulty estimates. Indeed, some words that almost no students could guess under the Active format, such as “nil”, still had high item facilities under the VST format. This suggests that when learners encounter a word they do not know under the VST format, they tend to guess, inflating estimates of words known receptively.

![Figure 3. Word difficulty estimates under Active MC, Active Full Word and VST test formats.](image)

The Active MC and Full Word item facilities correlated highly at 0.90, but differences between the two revealed hazards of the proposed Active MC format. One word, “Microphone”, was markedly easier under the Active MC format. This is likely because the Japanese word for “microphone” is simply an abbreviation of the English term (“マイク”, or “Mic”), and therefore, respondents could simply select the letter “m” without knowing the word’s full form. This indicates that the Active MC format is not appropriate for testing of loan words.

There were, however, situations in which word difficulty estimates could arguably be said to be more accurate under the Active MC format than the Active Full Word format, because the Active MC format did not penalize learners for incorrect spellings. The word “dinosaur” serves as an example: though under the Active Full Word format few students knew it, an examination of responses revealed that a high proportion likely did, but simply could not master the word’s more challenging spelling. Some alternative responses included “dinosaw”. Though these answers were technically incorrect, it is likely that in spoken contexts the learner would be considered to have mastery over the words, albeit with differences from native speaker pronunciation.

4 Conclusions

The proposed Active MC test format returned high correlations (≥0.90) with the conventional, Active Full Word format. Though the Active MC format’s
reliability of 0.70 was lower than the Active Full Word reliability of 0.81, it was also markedly higher than that of the VST format (0.60). This suggests that although the Active Full Word test format remains the most statistically reliable, in contexts in which hand-marking tests is not feasible, the Active MC format provides an alternative which, in addition to being highly correlated to conventional active vocabulary knowledge estimates, is easy to construct, easy to mark, and offers higher reliability than a conventional multiple-choice test of the same words.

Though the Active Full Word format was slightly more difficult than the Active MC format, it is unclear how much of this is due to guesswork on the Active MC format, and how much is due to the additional difficulty imposed on the conventional active recall format by requiring test takers to correctly spell all words. As provision of words’ correct spellings is a necessary condition of many such tests, it is possible that these tests can underestimate vocabulary knowledge on spoken productive measures, by conflating spelling with knowledge. A closer analysis of the Active Full Word format is required to determine if this is the case.

References


Relationships Between Text Length and Lexical Diversity Measures: Can We Use Short Texts of Less than 100 Tokens?

Rie Koizumi
Tokiwa University
doi: http://dx.doi.org/10.7820/vli.v01.1.koizumi

Abstract
Lexical diversity (LD) measures have been known to be sensitive to the length of the text, and numerous revised LD measures have been proposed. This study aims to identify LD measures that are least affected by text length and can be used for the analysis of short L2 texts (50–200 tokens). This study compares the type-token ratio, Guiraud index, D, and measure of textual lexical diversity (MTLD) to assess their degree of susceptibility to text length. Spoken texts of 200 tokens from 20 L2 English learners at the lower-intermediate-level were divided into segments of 50 to 200 tokens and the text length impact was examined. It was found that MTLD was least affected by text length, and that it should be used with texts of at least 100 tokens.

Keywords: lexical diversity; text length; type-token ratio; Guiraud index; D; measure of textual lexical diversity (MTLD); speaking performance.

1 Introduction
How do L2 learners use vocabulary knowledge in their speaking and writing? Research on vocabulary use, or, as it has generally been termed, lexical richness, is a vital but under-researched topic (Schmitt, 2010; Skehan, 2009). Among several aspects of lexical richness, lexical diversity (LD, also known as lexical variation) refers to “the range and variety of vocabulary deployed in a text by either a speaker or a writer” (McCarthy & Jarvis, 2007, p. 459). Numerous measures of LD have been proposed (see Malvern, Richards, Chipere, & Durán, 2004 for a list), but perhaps the best known measure is the type-token ratio (TTR), in which the number of different words a learner writes in a text is divided by the total number of words in order to determine the degree of variation.

However, an acknowledged drawback of many such measures is that they have been known to be sensitive to the length of the text analyzed. Effects of text length on LD measures should be avoided as much as possible because they generate construct-irrelevant variances and misleading results. Numerous revised LD measures have been proposed to remedy this risk, including the Guiraud index (Guiraud, hereafter; Guiraud, 1960), D (Malvern et al., 2004), and the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010), to name a few.

Previous studies (Hess, Sefton, & Landry, 1986; Malvern et al., 2004; McCarthy & Jarvis, 2007, 2010) show that like TTR, Guiraud (values) are substantially affected by text length, whereas D and MTLD (values) are affected...
only to a small degree. However, previous studies on D and MTLD examined the impact of length of texts of 100 tokens or more (McCarthy & Jarvis, 2007, 2010). Consequently, use of D and MTLD could still be problematic for analysis of texts under 100 tokens. This is unfortunate, as evaluating the quality of such short texts is sometimes necessary in studies of L2 production and vocabulary.

The present study investigates whether LD measures are usable for texts of <100 tokens, in order to determine if these measures can be used for analysis of shorter texts produced by learners. This study compares four LD measures (TTR, Guiraud, D, and MTLD) for analyzing short texts of 50 to 200 tokens and addresses the following research question: Which LD measures are least affected by text length across a range of 50 to 200 tokens?

In the next section, these four LD measures and their characteristics are summarized.

2 TTR, Guiraud, D, and MTLD

Brief explanations and examples of how to calculate the four LD measures are provided below.

Here is a spoken text from a male junior-high-school student describing a picture and comparing two pictures. His entire text consisted of 253 tokens, and 50 tokens were selected:

This mountain is very big and the scene is very good and there is a lake. Lake is very beautiful and beside there is a tree and over there I see the big mountain. There is a very beautiful." And this Taro’s room. Before is Taro’s name on textbook but

The exemplar text has 25 types and 50 tokens. We can analyze this text by using any of the following four measures:

(a) **TTR**: the number of different words (types)/all words produced (tokens). In this example, the TTR is 0.50 (i.e. 25/50).
(b) **Guiraud**: the number of types/the square root of the number of tokens (types/√tokens). In this example, the Guiraud is 3.54 (i.e. 25/√50).
(c) **D**: Sample 35 tokens randomly from a text analyzed like this: This, but, this, mountain, on, There, is, Before, big, the, very, is, good, beautiful, And, name, there, a, Lake, Taro’s, very, and, there, a, textbook, and, there, see, big, mountain, is, a, room, is, Taro’s.

Once a word is selected, it is not selected again (i.e. without replacement). This randomly selected 35-token segment has TTR of 0.57, with 20 types and 35 tokens. Then, another segment of 35 tokens is sampled and the TTR is calculated. Sampling up to 100 times should be repeated and the mean TTR for the hundred 35-token segments (0.57) should be computed. Similarly, producing the mean TTR for hundred 36-token segments (0.56), the mean TTR for hundred 37-token segments (0.54), and so on, up to the mean TTR for hundred 50-token segments (0.50) should be continued. Figure 1 plots 16 means for each token segment and shows the empirical TTR curve. Identify the D value for which the theoretical curve best fits the empirically derived TTR curve. D for the 50-token text was 12.00.

---

*Koizumi: Relationships Between Text Length and Lexical Diversity Measures* 61

*Vocabulary Learning and Instruction, 1(1), 60–69.*
D can be computed using vocd (McKee, Malvern, & Richards, 2000) or D_Tool (Meara & Miralpeix, 2007).

(d) **MTLD**: “the mean length of sequential word strings in a text that maintain a given TTR value” (0.720; McCarthy & Jarvis, 2010, p. 384). Count (x) the number of times the text reaches TTR of 0.72 or below, from the beginning of the text through to the end. As seen in Table 1, TTR is calculated to obtain (x) for a segment from the beginning by increasing words successively. For example, the tenth word, *very*, is the eighth type, with TTR of 0.80. A segment up to the fourteenth word, *is*, has TTR of 0.72 or below (i.e. 0.71). This word is the last one of the first word string (factor) of the text. Then, the next word, *a*, becomes the first word of the next string. This time, the third word, *lake*, has TTR of 0.67, but this segment is not counted because it has less than 10 words. McCarthy (2005) decided not to consider “factors of less than ten words, . . . as they may only represent a brief syntactical or rhetorical, textual blip” (p. 106). The second word string ends with the 37th word, *a*, with a TTR of 0.70. The third string starts with *very* but does not reach 0.72 (0.92); however, the value of 0.92 reaches 27% of the trajectory from 1.00 to 0.72.
Table 1. Example of measure of textual lexical diversity Forward Calculation

<table>
<thead>
<tr>
<th>Type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Token</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>TTR</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.89</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>good</em> and <em>there</em> <em>is</em> a <em>Lake</em> <em>lake</em> <em>is</em> very <em>beautiful</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Token</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>TTR</td>
<td>0.82</td>
<td>0.75</td>
<td>0.77</td>
<td>0.71</td>
<td>1.00</td>
<td>1.00</td>
<td>0.67</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td><em>and</em> <em>beside</em> <em>there</em> <em>is</em> a <em>Tree</em> <em>and</em> <em>over</em> <em>there</em> <em>I</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Token</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>TTR</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.86</td>
<td>0.88</td>
<td>0.89</td>
<td>0.80</td>
<td>0.82</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td><em>see</em> the <em>big</em> <em>mountain</em> <em>there</em> <em>is</em> a <em>very</em> <em>beautiful</em> <em>and</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Token</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>TTR</td>
<td>0.79</td>
<td>0.80</td>
<td>0.81</td>
<td>0.82</td>
<td>0.78</td>
<td>0.74</td>
<td>0.70</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td><em>this</em> <em>Taro’s</em> <em>room</em> <em>before</em> <em>is</em> <em>Taro’s</em> <em>name</em> <em>on</em> <em>textbook</em> <em>but</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Token</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>TTR</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.89</td>
<td>0.90</td>
<td>0.91</td>
<td>0.92</td>
<td>0.92</td>
<td></td>
</tr>
</tbody>
</table>

Note. TTR, type-token ratio. Underlined = TTR of 0.72 or below.
(i.e. \(1.00-0.92\)/\(1.00-0.72\) = 0.08/0.28). The remaining segment, which does not arrive at 0.72, is taken into consideration to enhance the reliability of MTLD (McCarthy, 2005). Therefore, \((x)\), the number of times the text reaches TTR of 0.72 or below, was two plus 0.27 (i.e. 2.27). The mean number of words required is computed using the number of tokens divided by \((x)\) as a formula; in the example, it is 21.03 (i.e. 50/2.27). Similarly, the calculation is made backward from the last word, but, to the first word, this, which produces another value (27.41). The two values derived from forward calculation and from backward calculation are averaged; the example has a value of 24.70.

MTLD can be derived easily with the use of the Gramulator (Version 5.0; McCarthy, 2011) as follows:

(a) Start the Gramulator and push the [Start] button. Figure 2 appears.
(b) Click on the folder already saved, which appears below “Browse to desired folder.” Texts in the folder are shown in “Folder Contents.” Click on the [Click to Retain this Corpus] button to add the selected folder to the corpus.
(c) Click on the folder name in “Click corpus to load.” Then, the folder is selected and shown in the bottom right (e.g. “Corpus 1 is Gramulator”).
(d) Go to [Modules] \(\rightarrow\) [Assessment] \(\rightarrow\) [The Evaluator]. Figure 3 appears.
(e) Select the folder in “Select Corpus” in the right. Select “Measures” in “Index Banks” in the left. Select “MTLD” in the middle. (You can also obtain HD-D and Maas here; see McCarthy & Jarvis, 2010).
(f) Go to [Process] \(\rightarrow\) [Analyze]. Three types of MTLD values are derived on the bottom: MTLD (Raw) is a raw MTLD value described above. MTLD (Znar) and MTLD (Zsci) show z-scores; both are standardized based on norms of texts (either narrative or science) derived from the Touchstone Applied Science Associates (TASA) corpus (P.M. McCarthy, personal communication, November 9, 2011; see McCarthy, 2010, for TASA).

Figure 2. Example of measure of textual lexical diversity (MTLD) calculation using the Gramulator (1).
Guiraud, D, and MTLD were developed to reduce effects of text length. The three measures and TTR are all intended to assess LD, with larger values indicating more lexically diverse texts. Guiraud is a simple transformation of TTR by taking a square root of a denominator and adjusting the values in such a way that long texts are not too disadvantaged.

D and MTLD use computers but differ in principles. D is more firmly based on TTR, although random selection and curve fitting reduce the impact of text length. However, MTLD uses TTR as a cutoff point to inspect the text length for which a speaker/writer can maintain a certain level of LD. Another feature of MTLD is that MTLD “considers LD at the textual level” (McCarthy, 2005, p. 93). It takes textual patterns into account and analyzes all the words in the text from the first word to the last word and from the last to the beginning word (i.e. sequentially). McCarthy (2005) argued that “maintaining the text structure rather than sampling the text provides a more authentic measure of diversity” (p. 88).

3 Method

3.1 Language samples

Language samples from 20 Japanese learners of English at junior and senior high schools were used. A tape-mediated speaking test was employed to elicit utterances on familiar topics (e.g. self-introduction and picture description), with a total of 225 seconds of speech per learner. Of 38 students who produced 200 or more tokens, 20 representative speakers were selected (M = 236.55 tokens, SD = 30.79, minimum = 202, maximum = 283). Then, each learner’s 200 tokens were separated into a text.

3.2 Analyses

Using the parallel sampling method (e.g. Hess et al., 1986), a text of 200 tokens from one participant was split into 25 segments: four segments of 50 tokens,
three segments of 60 tokens, two segments of 70, 80, 90, and 100 tokens, and one segment of 110, 120, 130, 140, 150, 160, 170, 180, 190, and 200 tokens.

Each segment was analyzed using four LD measures: TTR, Guiraud, D, and MTLD. The count of types and tokens was performed using KWIC Concordance for Windows (Version 5.0; Tsukamoto, 2011). D was computed using D_Tool (Version 2.0; Meara & Miralpeix, 2007). MTLD (raw score) was calculated using Gramulator (Version 5.0; McCarthy, 2011). Base forms and their inflected forms were considered different types in this study; for example, table and tables, and play, plays, played, playing were counted as different.

Subsequently, values from equal-sized segments were averaged. To examine the research question, five token ranges were separately examined using a one-way repeated measures ANOVA and an effect size of partial eta-squared ($\eta_p^2$): (a) 50 to 100, (b) 100 to 150, (c) 150 to 200, (d) 100 to 200, and (e) 50 to 200. Statistical assumptions were checked and handled accordingly. The alpha level was set to 0.0025 (0.05/20) because 20 ANOVAs were performed.

4 Results and Discussion

Table 2 shows that there were significant and large effects of text length on TTR, Guiraud, and D when text length changed across 50 to 100 tokens, 100 to 150 tokens, and 50 to 200 tokens (e.g. $\eta_p^2 = 0.75, 0.49$, and 0.85 for TTR; see Appendix for descriptive statistics of each measure for each segment). In the range of 150- to 200 tokens, the effects of text length were also significant for both MTLD and D.

Table 2. ANOVA Results for Text Length Effects ($N=20$)

<table>
<thead>
<tr>
<th></th>
<th>50-100</th>
<th>100-150</th>
<th>150-200</th>
<th>100-200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F$ df $\eta_p^2$</td>
<td>$F$ df $\eta_p^2$</td>
<td>$F$ df $\eta_p^2$</td>
<td>$F$ df $\eta_p^2$</td>
</tr>
<tr>
<td>TTR</td>
<td>57.54 3.02, 57.43 $&lt;0.001$ 0.75</td>
<td>17.98 2.15, 40.78 $&lt;0.001$ 0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guiraud</td>
<td>81.33 2.95, 56.02 $&lt;0.001$ 0.81</td>
<td>50.68 2.28, 43.36 $&lt;0.001$ 0.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>6.49 2.66, 50.49 $&lt;0.001$ 0.25</td>
<td>10.43 2.48, 47.17 $&lt;0.001$ 0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTLD</td>
<td>2.06 2.34, 44.41 0.13 0.10</td>
<td>1.39 3.25, 61.69 0.25 0.07</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

MTLD, measure of textual lexical diversity; TTR, type-token ratio.
200-token segments and 100- to 200-token segments, text length significantly affected TTR and Guiraud to a substantial degree (e.g. $\eta_p^2 = 0.46$ and 0.64 for TTR).

In contrast, MTLD was not statistically significantly affected by text length, and the degree of impact was generally small for all five ranges ($\eta_p^2 = 0.10, 0.07, 0.04, 0.10$, and 0.17). Thus, MTLD was found to be the least affected by text length. However, it should be noted that even MTLD was sensitive to text length variations across a range of 50- to 100-token segments, 100- to 200-token segments, and 50- to 200-token segments ($\eta_p^2 = 0.10, 0.10$, and 0.17).

These results suggest that texts of 100 tokens should be considered the minimum requirement for measuring LD using MTLD. Another finding is that the text length effect was larger when texts have more token differences. In other words, a 150-token difference ($\eta_p^2 = 0.17$) produces larger effects than a 100-token difference ($\eta_p^2 = 0.10$), which has larger effects than a 50-token difference when texts have more than 100 tokens ($\eta_p^2 = 0.07$, and 0.04). Thus, comparisons should be made between texts of similar lengths (with an interval of 50 tokens, if possible).

The results of the present study generally concur with previous research, with the one major difference that D was affected to an excessive degree across ranges of 50 to 100 tokens, 100 to 150 tokens, 100 to 200 tokens, and 50 to 200 tokens. This shows that D is sensitive to text length when short texts up to 150 tokens are analyzed.

This study shows that text length affects MTLD to a limited degree if the texts analyzed consist of more than 100 tokens. Why does MTLD have such a characteristic in contrast to other LD measures? One reason may be that MTLD focuses on the text length required to arrive at a certain level of TTR, which the other measures do not, and such text length tends to be stable as long as an overall text has sufficient tokens (i.e. more than 100 tokens) and as long as texts are produced by the same person using similar topics. McCarthy (2005) argued that MTLD has high reliability, in the sense that shorter segments of a text generate values similar to those that a whole text produces, and that it has enough sensitivity to distinguish texts of different degrees of LD. The current study supports this argument.

5 Conclusions

This study inspected the impact of text length on four LD measures across 50- to 200-token texts and found that MTLD is least affected by text length, but that even this measure of LD should be used with texts of at least 100 tokens. This result indicates the usefulness of considering LD at the textual level.

For future research, analyses using spoken and written texts of multiple genres, with a larger number of language samples with more tokens, are needed to investigate the generalizability of this study, especially the utility of MTLD.

Acknowledgements

This research was partially supported by the Grant-in-Aid for Scientific Research (KAKENHI) of the Ministry of Education, Culture, Sports, Science and Technology in Japan (No. 22720216). I am deeply indebted to Yo In'nami and reviewers for their invaluable comments.

Vocabulary Learning and Instruction, 1(1), 60–69.
References


Note: Subsequent to the writing of this article, Gramulator (Version 5; McCarthy, 2011) is no longer available, as it has been replaced with Version 6. Unfortunately the link to Version 5 is dead, so a link to Version 6 is provided above.
Appendix. Descriptive Statistics of Lexical Diversity Measures (N = 20)

<table>
<thead>
<tr>
<th>Token</th>
<th>TTR</th>
<th>Guiraud</th>
<th>D</th>
<th>MTLD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>50</td>
<td>0.62</td>
<td>0.03</td>
<td>4.33</td>
<td>0.19</td>
</tr>
<tr>
<td>60</td>
<td>0.58</td>
<td>0.03</td>
<td>4.51</td>
<td>0.24</td>
</tr>
<tr>
<td>70</td>
<td>0.57</td>
<td>0.04</td>
<td>4.80</td>
<td>0.34</td>
</tr>
<tr>
<td>80</td>
<td>0.55</td>
<td>0.04</td>
<td>4.91</td>
<td>0.34</td>
</tr>
<tr>
<td>90</td>
<td>0.53</td>
<td>0.03</td>
<td>5.07</td>
<td>0.28</td>
</tr>
<tr>
<td>100</td>
<td>0.53</td>
<td>0.03</td>
<td>5.25</td>
<td>0.32</td>
</tr>
<tr>
<td>110</td>
<td>0.52</td>
<td>0.05</td>
<td>5.44</td>
<td>0.52</td>
</tr>
<tr>
<td>120</td>
<td>0.52</td>
<td>0.05</td>
<td>5.63</td>
<td>0.58</td>
</tr>
<tr>
<td>130</td>
<td>0.50</td>
<td>0.05</td>
<td>5.72</td>
<td>0.53</td>
</tr>
<tr>
<td>140</td>
<td>0.49</td>
<td>0.04</td>
<td>5.84</td>
<td>0.51</td>
</tr>
<tr>
<td>150</td>
<td>0.49</td>
<td>0.04</td>
<td>5.96</td>
<td>0.49</td>
</tr>
<tr>
<td>160</td>
<td>0.48</td>
<td>0.03</td>
<td>6.04</td>
<td>0.45</td>
</tr>
<tr>
<td>170</td>
<td>0.47</td>
<td>0.03</td>
<td>6.10</td>
<td>0.41</td>
</tr>
<tr>
<td>180</td>
<td>0.46</td>
<td>0.03</td>
<td>6.21</td>
<td>0.38</td>
</tr>
<tr>
<td>190</td>
<td>0.46</td>
<td>0.03</td>
<td>6.32</td>
<td>0.38</td>
</tr>
<tr>
<td>200</td>
<td>0.46</td>
<td>0.03</td>
<td>6.40</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Identifying Dimensions of Vocabulary Knowledge in the Word Associates Test

Aaron Batty
Keio University
doi: http://dx.doi.org/10.7820/vli.v01.1.batty

Abstract

Depth of vocabulary knowledge (DVK) (i.e. how much a learner knows about the words he knows) is typically conceptualized as a psychologically multidimensional construct, including various forms of word knowledge. Read’s Word Associates Test (WAT) is the most common test of DVK in the literature, assessing knowledge of words’ synonyms and collocates. Despite the fact that the WAT aims to measure two dimensions of vocabulary knowledge, no studies until now have investigated whether these dimensions are psychometrically distinct. The present study seeks to fill that gap. A known-reliable-and-valid WAT developed by David Qian was administered to 530 Japanese university English majors. Confirmatory factor analysis was employed to investigate the psychometric dimensionality of the WAT. It was discovered that a bifactor model, wherein the primary explanatory factor is a vocabulary g-factor, with additional, uncorrelated factors for synonym and collocate items, demonstrated the best fit. This finding implies that although these dimensions of DVK may be somewhat distinct, they are largely subsumed by general vocabulary knowledge.

Keywords: vocabulary; depth of vocabulary knowledge; word associates test; multidimensionality; structural equation modeling.

1 Background

1.1 Depth of Vocabulary Knowledge (DVK)

Vocabulary can be separated into two broad categories of knowledge: how many words one knows (vocabulary breadth) and how well one knows those words (vocabulary depth) (Nation, 1990; Richards, 1976). Operational definitions of DVK are numerous, but most conceptualize it as a dimensional construct (e.g. Henriksen, 1999; Hunston, Francis, & Manning, 1997; Read, 1993, 1998, 2000; Schoonen & Verhallen, 2008).

Some (e.g. Hunston et al., 1997; Miller & Fellbaum, 1991) claim that DVK lies in the semantic networks which connect a word with the other information necessary to truly understand and use it, a view shared by Henriksen (1999), who argues that in order to truly know a word, the learner must engage in semantic network building, i.e. creating intentional links between the target word and other words the learner knows, including morphological similarity, syntactic similarity and, of particular relevance to the present study, collocational similarity. It is partly this kind of knowledge depth that enables what Pawley and Syder term (1983)
“nativelike selection”—the ability for a speaker of a language to make the same lexical choices as a “native” speaker of that language.

Although many agree that collocational knowledge is critical to DVK, few vocabulary tests attempt to address it. The only widely known test format that does make such an attempt is the Word Associates Test (WAT) (Read, 1993, 1998), which, in addition to testing synonyms, attempts to incorporate collocational knowledge of words of tested words. However, its ability to treat collocational knowledge as a separate psychometric dimension has not been established. The goal of this study is to determine if the two dimensions measured by the test, knowledge of synonyms and knowledge of collocates, can be identified as distinct psychometric constructs.

1.2 The WAT

The test presents the examinee with a stimulus word followed by four possible synonyms and four possible collocates, from which the examinee is to choose four correct associates. The stimulus word is always an adjective and the collocates are always nouns modifiable by the stimulus. The correct answer may include one synonym and three collocates, two synonyms and two collocates, or three synonyms and one collocate. This uncertainty was added in an attempt to limit the effectiveness of guessing strategies (Read, 2000, p. 184). An example is shown in Figure 1. The newer WAT has been employed most visibly by Qian to investigate the link between DVK and reading performance (Qian, 2002), and later to evaluate the format as a possible addition to the TOEFL (Qian & Schedl, 2004).

1.3 Dimensionality of the WAT

Despite the fact that the WAT is intended to measure DVK conceptualized as a dimensional construct, very little research has been conducted on this aspect. Several researchers have scaled WAT data with the Rasch model (Rasch, 1960), either dichotomously (i.e. each correct answer is treated as a separate item) (e.g. Batty, 2006), or with the Rasch partial credit model (Masters, 1982), treating each stimulus word as an item worth four points (e.g. Batty, 2008; Read, 1998). In these cases, the items—whether dichotomous or polytomous—exhibited acceptable fit statistics, indicating that the items adhered closely to the Rasch model’s assumption of unidimensionality (Bond & Fox, 2007, p. 35). However, Rasch fit statistics can be deceptive, as they are sample dependent (Bond & Fox, 2007, p. 285). As such, increases in the number of items or observations will always result in improved fit to the Rasch model, which may obscure dimensionality. Factor analysis of an original, 1993 version of the WAT, found that a one-factor model exhibited better fit to the data than a two-factor model, although details are few (Schoonen & Verhallen, 2008). These findings all suggest that the WAT operates as a unidimensional
measure, despite its multidimensional theoretical underpinnings. However, to date, no one has carried out a principled and theoretically sound investigation of the dimensionality of the current WAT format.

There are three likely possibilities for the dimensional structure of the WAT. The first is a unidimensional structure, where a single vocabulary knowledge factor predicts all of the items, as illustrated by the path diagram in Figure 2.

![Figure 2. A simplified path diagram of a unidimensional model of the WAT.](image)

Another possibility is a two-factor model, wherein a synonym factor predicts the synonym items, and a collocate factor predicts the collocate items. In such a model, the two factors are correlated, as we can assume that both are types of vocabulary knowledge, and are therefore related (see Figure 3).

![Figure 3. A simplified path diagram of a two-factor model of the WAT.](image)

Another type of model pertinent to the WAT is the bifactor model (Holzinger & Swineford, 1937). In such a model, all the variables are assumed to load on a single general factor (or g-factor), while groups of variables additionally load on separate, smaller factors. In such a model, the subskill factors are distinct from the g-factor, rather than being subsumed by it. In the case of the WAT, such a model is acceptable from a theoretical standpoint, as it includes a single, overarching vocabulary dimension, in addition to distinct synonym and collocate factors which are pertinent only to them (see Figure 4). An advantage of such a model is that, if it displays appropriate model fit, reporting of the test’s subscores in addition to the total scores is justified.

1.4 Research Question

The goal of this study is to apply tests of dimensionality to the WAT to address the following research questions: Is the WAT a unidimensional measure of...
vocabulary knowledge, or do the subskills load on their own psychometric dimensions? If so, are the item types best modeled as correlated yet distinct factors, or as uncorrelated subskills subordinate to a general factor in a bifactor model?

2 Method

2.1 Instrument

The WAT form used in the present study was developed by Qian for the 2002 study discussed above. In that study, a reliability coefficient of 0.88 (N = 217) was observed and it correlated significantly with the results of a Nation Vocabulary Levels Test (Nation, 1990, 2001), an accepted and acceptable measure of vocabulary size (Read, 2000, p. 118). The instrument, therefore, can be assumed to be both reliable and concurrently valid. The 40 stimulus words were selected by Qian and are described as “general academic adjectives” (Qian, 2002, p. 525).

2.2 Sample

The participants were 530 first- and second-year English and international communication majors at a foreign-language university in Japan. The sample included a wide range of English language proficiencies, as evidenced by the university’s internal placement exam. The WAT was administered as an optional vocabulary test during normal class sessions, and results were delivered to the class instructors to be distributed to the students.

2.3 Data Analysis and Results

For the purpose of investigating dimensionality, each correct answer on the WAT was treated as a single dichotomous item, resulting in 160 possible items. To facilitate the creation of item parcels (see next section), the dataset was reduced to 145 items. The two synonyms and the three collocate items with the lowest point-biserial statistics were removed from the dataset, resulting in a dataset of
70 synonym and 75 collocate items. This reduced dataset was found to have a Cronbach alpha reliability coefficient of 0.89.

To ascertain whether the WAT conformed better to a one-factor or two-factor model, confirmatory structural equation modeling was employed using the Mplus software package (Muthén, Muthén, Asparouhov, & Nguyen, 2011). Mplus offers the weighted least squares with mean and variance adjusted (WLSMV) estimation method, which is widely recommended for analyses of categorical (e.g. binary) data (e.g. Brown, 2006, p. 388; Jasper, 2010). An initial two-factor analysis revealed a very high correlation between the factors ($r = 0.92$), indicating a high degree of multicollinearity in the model. Following the advice of Jasper (2010), the items were grouped into item parcels for further analysis. The items were ordered from least- to most-difficult as determined by Rasch analysis using Winsteps (Linacre, 2011) and parceled into bundles of five items and the item responses were summed.

The data were first fitted to a unidimensional (i.e. vocabulary g-factor) model. Model fit statistics is shown in Table 1. Overall model fit was less than satisfactory. The significant chi-square statistic indicated poor fit with the data (Brown, 2006, p. 81; Schumacker & Lomax, 2004, p. 82), although the chi-square statistic is often inflated by sample sizes of over 200, and so may not be a good indicator of fit for the present sample (Schumacker & Lomax, 2004, p. 100). The comparative fit and Tucker-Lewis indices (CFI and TLI, respectively) should ideally be approximately 0.95 or greater, and the root mean square error of approximation (RMSEA) should be 0.06 or less (Brown, 2006, p. 87). None of these are the case with the one-factor model. The data were then fitted to a two-factor model, wherein the synonym and collocate items load on correlated synonym and collocate factors, respectively. The fit of this model was very slightly superior to that of the one-factor model, but was still not ideal (see Table 1). Finally, the data were fitted to a bifactor model, as discussed above. This model was found to exhibit the best fit of all (see Table 1). Examining the standardized loadings on the three factors of this model (Table 2), it becomes clear that the highest item loadings overall were on the vocabulary g-factor, half of the synonym items loaded highest on the synonym factor, and only two of the collocate items loaded more highly on the collocate factor than on the g-factor.

**Table 1. Results of the SEMs of the Item-Parceled WAT dataset ($N = 530$)**

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 factor: vocabulary g-factor</td>
<td>1446.21*</td>
<td>0.840</td>
<td>0.828</td>
<td>0.073</td>
</tr>
<tr>
<td>2 factors: synonym and collocate,</td>
<td>1199.83*</td>
<td>0.877</td>
<td>0.867</td>
<td>0.064</td>
</tr>
<tr>
<td>correlated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bifactor model: vocabulary g-factor,</td>
<td>909.29*</td>
<td>0.916</td>
<td>0.902</td>
<td>0.055</td>
</tr>
<tr>
<td>synonym, and collocate</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: All chi-square statistics calculated via the DIFFTEST operation in Mplus 6.11 (Muthén et al., 2011), under WLSMV estimation.

*$p < 0.01.$

### 3 Discussion and Conclusion

This study sought to determine whether the WAT was best modeled as a unidimensional, two-factor, or bifactor test of vocabulary knowledge. The data were found to best fit the bifactor model, where the primary explanatory factor is a
single vocabulary g-factor, with additional, uncorrelated second-order subskill factors for synonym and collocate items.

These findings are of interest theoretically, as they suggest that knowledge of synonyms and collocates are distinct subskills of vocabulary knowledge. The raw-score correlation between the item types was low \( r = 0.61 \) relative to inter-item correlations within the collocate and synonym items alone (0.79 and 0.80, respectively), indicating that it may be meaningful for tests to provide distinct subscores for both synonym and collocational knowledge. Further development of items and tests to do this may be a fruitful avenue for future research.

In the case of the WAT itself, however, the collocation item types loaded somewhat weakly as a separate factor, calling the ability of this particular test to reliably assess collocational knowledge into question. Despite the items’ loadings on

<table>
<thead>
<tr>
<th>Item parcel</th>
<th>Vocabulary</th>
<th>Synonym</th>
<th>Collocate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syn. 1</td>
<td>0.350*</td>
<td>0.413*</td>
<td></td>
</tr>
<tr>
<td>Syn. 2</td>
<td>0.362*</td>
<td>0.243*</td>
<td></td>
</tr>
<tr>
<td>Syn. 3</td>
<td>0.550*</td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td>Syn. 4</td>
<td>0.427*</td>
<td>0.346*</td>
<td></td>
</tr>
<tr>
<td>Syn. 5</td>
<td>0.252*</td>
<td>0.570*</td>
<td></td>
</tr>
<tr>
<td>Syn. 6</td>
<td>0.249*</td>
<td>0.469*</td>
<td></td>
</tr>
<tr>
<td>Syn. 7</td>
<td>0.409*</td>
<td>0.459*</td>
<td></td>
</tr>
<tr>
<td>Syn. 8</td>
<td>0.572*</td>
<td>0.238*</td>
<td></td>
</tr>
<tr>
<td>Syn. 9</td>
<td>0.413*</td>
<td>0.314*</td>
<td></td>
</tr>
<tr>
<td>Syn. 10</td>
<td>0.357*</td>
<td>0.360*</td>
<td></td>
</tr>
<tr>
<td>Syn. 11</td>
<td>0.333*</td>
<td>0.496*</td>
<td></td>
</tr>
<tr>
<td>Syn. 12</td>
<td>0.417*</td>
<td>0.218*</td>
<td></td>
</tr>
<tr>
<td>Syn. 13</td>
<td>0.344*</td>
<td>0.498*</td>
<td></td>
</tr>
<tr>
<td>Syn. 14</td>
<td>0.601*</td>
<td>0.155*</td>
<td></td>
</tr>
<tr>
<td>Col. 1</td>
<td>0.409*</td>
<td></td>
<td>0.479*</td>
</tr>
<tr>
<td>Col. 2</td>
<td>0.502*</td>
<td></td>
<td>0.184*</td>
</tr>
<tr>
<td>Col. 3</td>
<td>0.529*</td>
<td></td>
<td>0.056</td>
</tr>
<tr>
<td>Col. 4</td>
<td>0.657*</td>
<td></td>
<td>-0.198*</td>
</tr>
<tr>
<td>Col. 5</td>
<td>0.448*</td>
<td></td>
<td>0.253*</td>
</tr>
<tr>
<td>Col. 6</td>
<td>0.315*</td>
<td></td>
<td>0.488*</td>
</tr>
<tr>
<td>Col. 7</td>
<td>0.380*</td>
<td></td>
<td>0.195*</td>
</tr>
<tr>
<td>Col. 8</td>
<td>0.541*</td>
<td></td>
<td>0.316*</td>
</tr>
<tr>
<td>Col. 9</td>
<td>0.527*</td>
<td></td>
<td>0.015</td>
</tr>
<tr>
<td>Col. 10</td>
<td>0.592*</td>
<td></td>
<td>-0.046</td>
</tr>
<tr>
<td>Col. 11</td>
<td>0.502*</td>
<td></td>
<td>0.245*</td>
</tr>
<tr>
<td>Col. 12</td>
<td>0.529*</td>
<td></td>
<td>0.093</td>
</tr>
<tr>
<td>Col. 13</td>
<td>0.605*</td>
<td></td>
<td>0.063</td>
</tr>
<tr>
<td>Col. 14</td>
<td>0.446*</td>
<td></td>
<td>0.389*</td>
</tr>
<tr>
<td>Col. 15</td>
<td>0.632*</td>
<td></td>
<td>-0.150*</td>
</tr>
</tbody>
</table>

Notes: Estimation method: WLSMV. The variances of the three factors were fixed at 1. \( N = 530 \).

\*p < 0.01.
distinct psychometric dimensions, the large vocabulary knowledge g-factor shared by both test sections muddles the interpretability of the two item types somewhat. Moreover, as the WAT is typically modeled as unidimensional (e.g. Batty, 2006, 2008; Nurweni & Read, 1999; Read, 1993, 1998), it raises the question of what is gained by its novel and potentially problematic design (e.g. guessing, examinee confusion; see Read, 1993, 1998; Schmitt, Ng, & Garras, 2011 for further discussion) over more robustly validated measures such as the Vocabulary Levels Test (Beglar & Hunt, 1999; Schmitt, Schmitt, & Clapham, 2001).

**Acknowledgements**

The author would like to extend special thanks to David Qian for providing a reliable and validated WAT, and Jeffrey Stewart for his valuable suggestions and contributions. Partial funding was provided by a grant from the Research Institute of Language Studies and Language Education at Kanda University of International Studies, Chiba, Japan.

**References**


*Vocabulary Learning and Instruction, 1*(1), 70–77.
Examining the Validity of the Lexical Access Time Test (LEXATT2)

Tatsuo Iso
Reitaku University
doi: http://dx.doi.org/10.7820/vli.v01.1.iso

Abstract

This study aimed to investigate the validity of the Lexical Access Time Test (LEXATT2). The first step was to examine the test results to determine if it was able to differentiate between participants with different English proficiency levels. The results were further analyzed to see if longer target words elicited slower lexical access times. The results of the 119 participants indicated promise in that LEXATT2 established an ability to distinguish proficiency levels to some extent. Further, it was found that LEXATT2 elicited slower lexical access time from the participants with lower English proficiency.

Keywords: lexical access; fluent reading; automaticity of lexical access; word recognition; meaning retrieval; lexical access time test; computerized language testing.

1 Introduction

In recent years, greater attention has been paid to automaticity of lexical access. Though the research on vocabulary knowledge has focused mainly on vocabulary size, having automatic access to words and their meanings is almost as important in performance. The importance has been recognized in reading research where automatic lexical access is one of the most fundamental processes for fluent reading (Grabe, 2009; Grabe & Stoller, 2002; Koda, 2005). Along with vocabulary size and organization, automaticity of lexical access is now considered a dimension of vocabulary knowledge (Meara, 2005).

Despite its importance, how we measure the automaticity of lexical access is at issue. To address this problem, this study aimed to introduce my latest attempt at building a test of the automaticity of lexical access as well as to validate the Lexical Access Time Test (LEXATT2).

2 Description of the LEXATT2

2.1 What it Measures

Lexical Access Time Test is a computer program which aims to measure the time that it takes to recognize a word and retrieve its meaning. As illustrated in Figure 1, lexical access time is defined as the difference between the reaction time for the verbal stimulus (RT (VL)) and the mean reaction time for nonverbal stimulus (MRT (NV)). Admitting that speed is merely one aspect of automaticity (Segalowitz & Segalowitz, 1993), LEXATT2 measures the speed of lexical access to keep the test as simple as possible.
2.2 How it Measures Lexical Access Time

There are two tasks in LEXATT2. The first is a nonverbal reaction time task. Five instances of nonverbal reaction time are measured and averaged to produce MRT (NV).

The second task consists of two stages: verbal reaction time stage and meaning confirmation stage. The verbal reaction time stage is the same as the nonverbal reaction time task except the stimuli are English words (see below for the target words). In the meaning confirmation stage that immediately follows the verbal reaction time stage, the test takers are asked to choose the meaning or synonym of the target word displayed in the previous stage. This stage is necessary to make sure the test taker retrieve a word meaning when reacting to the verbal stimulus, so that LEXATT2 is not measuring word recognition time.

2.3 Target Words

There are 45 target words in LEXATT2. They are high-frequency words, within the 2000 level of the JACET list of 8000 basic words (JACET, 2003). The target words are also limited to 4, 6, or 8 letters in length, and there are 15 words each. The actual target words are as follows: band, bite, boot, cave, cash, debt, sail, hang, seed, lord, mail, meat, pack, pole, tale, afford, actual, attend, export, cancer, charge, divide, fairly, formal, insist, lovely, bottle, manner, threat, creature, marriage, customer, identity, original, security, response, complain, exciting, positive, strength, maintain, division, presence, laughter, persuade.

3 The Study

3.1 Purpose

As previously mentioned, this study aims to investigate the validity of LEXATT2. For this purpose, the following research questions were asked.

1. Does LEXATT2 distinguish test takers of different English proficiency levels?
2. Does LEXATT2 elicit slower lexical access time for longer words? If so, how does the length of words interact with the test takers’ proficiency levels?
The assumption of the first question was that if the automaticity of lexical access was part of vocabulary knowledge, it would improve along with the proficiency. Therefore, LEXATT2 was expected to elicit shorter lexical access times from higher-proficiency groups. As for the second question, it was assumed that LEXATT2 would elicit longer lexical access times for longer words if the test was sufficiently accurate.

3.2 Participants

There were a total of 194 participants in the study. The participants were divided into four groups based on their proficiency levels: native speakers (NS), Japanese Teachers of English (JTE), higher-level university students (Stu-H), and lower-level university students (Stu-L). The cutoff point of the higher-proficiency students was 470 on Test of English for International Communication (TOEIC). Some of these subjects were eliminated after taking LEXATT2 due to the low percentages of correct answers (below 70% during the meaning confirmation stage).

Table 1 shows the distribution of the remaining 119 participants. As for the proficiency levels of the NS and the JTE groups, it was assumed that the NS group was the highest in proficiency followed by JTE, who were further assumed to have higher proficiency than the Stu-H. The two student groups were significantly different with regard to their proficiency levels based on TOEIC scores.

4 Results

Prior to test validation, the reliability of the test was calculated with correct/incorrect responses in the meaning confirmation stage. The Cronbach’s alpha was 0.80, and the test was considered reliable enough.

To answer the first research question, the four groups were compared in terms of the mean lexical access time (see Table 2). As illustrated in Figure 2, the lexical access time seemed to be shorter for the higher-proficiency group. A one-way ANOVA revealed that proficiency had a significant effect on the lexical access time.

Table 2. Group Means for Overall Lexical Access Time and for Four-, Six-, and Eight-Letter Words

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Lexical access time (ms)</th>
<th>SD</th>
<th>Four letters (ms)</th>
<th>Six letters (ms)</th>
<th>Eight letters (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS</td>
<td>27</td>
<td>287.37</td>
<td>184.43</td>
<td>282.47</td>
<td>283.85</td>
<td>295.41</td>
</tr>
<tr>
<td>JTE</td>
<td>22</td>
<td>357.00</td>
<td>199.65</td>
<td>349.41</td>
<td>342.79</td>
<td>376.80</td>
</tr>
<tr>
<td>Stu-H</td>
<td>29</td>
<td>446.00</td>
<td>341.31</td>
<td>406.23</td>
<td>570.10</td>
<td>798.05</td>
</tr>
<tr>
<td>Stu-L</td>
<td>41</td>
<td>537.15</td>
<td>341.31</td>
<td>406.23</td>
<td>570.10</td>
<td>798.05</td>
</tr>
</tbody>
</table>

Vocabulary Learning and Instruction, 1(1), 78–82.
time with $F(3, 115) = 4.92$, $p < 0.001$, $\eta^2 = 0.11$, and an LSD (least significant difference) post hoc test revealed that the differences between the NS and the Stu-H, the NS and the Stu-L, and the JTE and the Stu-L were all significant.

As for the second research question, which asked if the length of the target words affected the lexical access time, the results are also shown in Table 2 and Figure 3. A two-way ANOVA with mixed design revealed the interaction of the proficiency group and the word length was significant ($F(6, 230) = 15.58$, $p < 0.01$, $\eta^2_p = 0.29$). An LSD post hoc test showed that the lexical access time of the six-letter words by the Stu-L group was significantly slower than those of the other groups. This difference intensified with the eight-letter words.

Another tendency was discovered when the word-length effect was compared within the groups. While the NS’s and the JTE’s performance was unaffected by the length of the target words, the Stu-H and the Stu-L demonstrated significantly slower lexical access time for longer words, with the exception of the difference between the Stu-H group’s four- and six-letter words.

5 Discussion

The validity of LEXATT2 is discussed in light of the results shown above. With regard to the comparison of the overall lexical access time among proficiency
groups, LEXATT2 showed an ability to distinguish the highest proficiency group from lower ones. However, it was not successful in distinguishing two higher proficient groups, namely the NS and the JTE. Even so, this does not necessarily completely negate the validity of LEXATT2, as the difference in proficiency between these groups was obscure to start with. Further examinations with larger number of participants with wider ranges of tangible, not assumed, proficiency levels are needed.

In terms of the analysis of the effect of the word length, the results of the lower-level participants were in accordance with intuitive expectations. LEXATT2 elicited slower lexical access time for longer words. Such an effect, however, disappeared for the higher-level participants. One may argue that this threatens the validity of the test. Yet, it is highly possible that high enough proficiency mitigates the effect of word length in lexical access. Yet another possibility is that the high-proficiency participants in this study knew the target words so well that the word length did not affect the lexical access time. Including even longer words may bring about clearer pictures.

Although LEXATT2 was found to have high potential, the findings also suggest the need for improvements. As mentioned above, including longer words as stimuli may be necessary. Further, there is a need for an infallible system to prevent test takers from placing too much emphasis on speed in lieu of accuracy during lexical access. About 40% of the original participants (mostly university students) did not choose correct meanings of target words due to the fact that they did not pay attention to the meaning of the target word in the verbal reaction time stage. Admitting that there is room for aforementioned improvement, the results of this study seem to demonstrate the promise of LEXATT2 as a tool for measuring English learners’ lexical access speed.

References