LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

CONCEPTUAL REVIEW ARTICLE

Size and Depth of Vocabulary Knowledge: What the Research Shows

Norbert Schmitt
University of Nottingham

When discussing vocabulary, a distinction is often made between size of vocabulary (number of known words) and depth of knowledge (how well those words are known). However, the relationship between the two constructs is still unclear. Some scholars argue that there is little real difference between the two, while regression analyses show that depth typically adds unique explanatory power compared to size alone. Ultimately, the relationship between size and depth of vocabulary knowledge depends on how each is conceptualized and measured. In an attempt to provide an empirical basis for exploring the size–depth relationship, this critical synthesis identifies studies that contain measures of both size and depth. Based on a number of different conceptualizations of depth, various patterns emerged. For higher frequency words and for learners with smaller vocabulary sizes, there is often little difference between size and a variety of depth measures. However, for lower frequency words and for larger vocabulary sizes, there is often a gap between size and depth, as depth measures lag behind the measures of size. Furthermore, some types of word knowledge (e.g., derivative knowledge) seem to have generally lower correlations with size than other types.

Keywords vocabulary size; vocabulary depth; vocabulary measurement; meaning; formulaic language; form–meaning link

Introduction

The mental lexicon is a complex phenomenon, and the exact nature of lexical knowledge has always perplexed researchers and teachers. This is not surprising as a lexicon can hold many thousands of words, each with numerous links of various kinds to the other words in the lexical network. Moreover, the links between different words are often difficult to explain clearly, thus making

Correspondence concerning this article should be addressed to Norbert Schmitt, University of Nottingham, University Park, Nottingham NG7 2RD, UK. E-mail: norbert.schmitt@nottingham.ac.uk

DOI: 10.1111/lang.12077

research into these links difficult. For example, a good red wine might be associated with words like *red, full-bodied, complex, spicy, satisfying,* and *ripe*, but explaining how these words are stored in the mental lexicon and how they are related to each other in various ways is not straightforward. Descriptions of the mental lexicon are further complicated by the fact that each word does not usually exist on its own, but rather is typically a part of a word family with numerous related members (e.g., *joy, joyful, joyous, joyfully*), of a lexical set (*emotion, joy, ecstasy*), and of formulaic language (*get/have no joy from something,* "have no success in getting something you want").

In grappling with these complexities, vocabulary specialists have developed a number of descriptive frameworks. One of the best known is the distinction between *size* or *breadth* of vocabulary knowledge (in simple terms, how many words are known) and *depth* or *quality* of vocabulary knowledge (i.e., how well those words are known) (Anderson & Freebody, 1981). This distinction has been widely taken up (e.g., Read, 2004), but there is some debate about its usefulness. In particular, empirical evidence typically shows a strong correlation between the two aspects, which has led some scholars (e.g., Vermeer, 2001) to assert that there is no conceptual distinction between size and depth. On the other hand, regression analyses usually show that depth measures have unique explanatory power in additional to size measures (e.g., Qian, 1999, 2000). This would suggest that the distinction is valid, pointing to two discrete constructs.

Certainly, from teaching experience, it seems justified to think of size and depth as two separate constructs. Teachers may well have had some students in their second-language classes who knew a relatively small number of words, but knew them quite well. This may be due to a particular study approach, where the students studied the words in textbooks, looked them up in dictionaries, looked for them in their readings, and practiced them over and over. This intensive treatment approach may lead to a good understanding of words, but if the time spent on these few words is excessive, the students may well lack the time to study a wide range of words. Similarly, some learners may use a second language in a very constrained work situation, for example, a taxi driver who uses the second language largely on the job. They may have only a small specialized vocabulary, but might be able to use the words they do know quite well.

Using different vocabulary learning strategies may lead to a very different kind of learning. The use of word lists comes to mind. If students learn the meanings of words from a word list (perhaps using a translation equivalent) in order to take a quiz the next day, they may well never follow up on this initial state of knowledge. Studying the words in isolation without contextual

elaboration limits the students to learning only something about the word form, something about the meaning, and some linkage between the form and meaning. To the extent that the students are able to retain some of these form–meaning connections (and this may be difficult without consolidation), then the students may have a relatively larger vocabulary size, but would know relatively little about these words, and probably would not be able to use them to any great degree.

These examples suggest that it is possible to know a little about a larger number of words or to know a great deal about a smaller number of words. That is, size and depth do not necessarily grow in a parallel manner. However, the interesting question is how size and depth typically develop in practice. Although size and depth may be strongly imbalanced for some learners, for most learners the relationship is likely to exist somewhere between these two extremes. This is an empirical issue, and this article will approach it by reviewing a large number of studies that include both a vocabulary size measure (or a reasonable proxy thereof) and one or more depth measures. Although many of these studies did not focus on the comparison between size and depth, this critical synthesis will use their data and results in an attempt to provide an empirically based description of the relationship between vocabulary size and depth.

Conceptualizing Vocabulary Size and Depth

Size of vocabulary knowledge is relatively straightforward to conceptualize, as it is basically counting known lexical items (typically operationalized as knowledge of the form–meaning connection), and most measurement and discussion of vocabulary to date have focused on size. (However, it is not so simple to measure—see below.) In contrast, there are a large number of overlapping ways in which depth of knowledge can be conceptualized. The diversity of depth conceptualizations makes it extremely difficult to know how to approach depth from a theoretical perspective. Thus, the conceptualizations discussed in this article will be largely driven by what measures of depth have been used in the research and can relate to knowledge of individual lexical aspects (e.g., knowledge of multiple polysemous meaning senses) as well as more holistic mastery (e.g., having a rich associative network formed around a word). As such, alternative conceptualizations such as the developmental approach outlined by Read (2000) and Henriksen's (1999) three-dimension description of lexical knowledge will not be covered because I could find no studies that

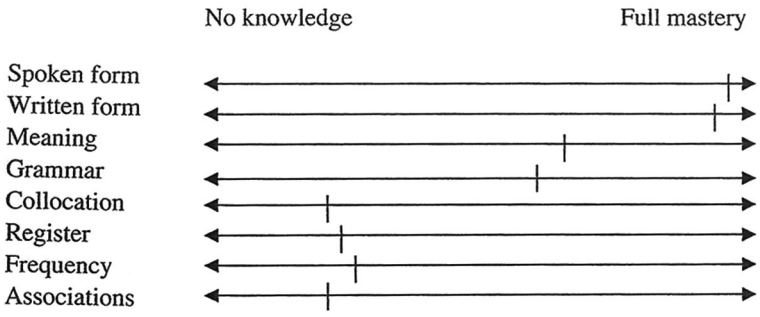

Figure 1 Developing knowledge of a word (Schmitt, 2010a, p. 38).

used them to measure depth of knowledge in ways that can be employed for this article's purpose of empirically comparing vocabulary size and depth. It is important to note that the depth conceptualizations I did find in the research are not necessarily mutually exclusive, and many share a considerable degree of commonality.

A widespread way of conceptualizing vocabulary depth of knowledge is by breaking it down into its separate elements, which has been described as a component or dimensions approach (Read, 2000). The genesis of this approach is usually traced back to an article by Jack Richards (1976) in *TESOL Quarterly*, where he discussed several assumptions about knowing vocabulary. Paul Nation refined this approach and his 2001 listing is still considered the best specification of the range of so-called word knowledge aspects to date, with each having receptive (R) and productive levels of mastery (P) (Table 1).

The richest depth could be seen as mastery of all these word knowledge aspects, but knowledge of individual aspects (collocation, derivative forms, polysemous meaning senses) can also be seen as contributing to depth of knowledge. Of course, these word knowledge aspects are not mastered in a dichotomous known/not known manner. Rather, they are likely to be developmental in nature, although each of the aspects probably develops at different rates, as illustrated in a hypothetical graph of what developing word knowledge might look like after a number of learning exposures (Figure 1). (See Lindsay and Gaskell (2010) for a psycholinguistic perspective on learning vocabulary over time.)

Some of these types of word knowledge (e.g., form, meaning) are particularly amenable to intentional study, which can lead to high learning rates and typically results in the ability to explicitly access this knowledge (i.e.,

Table 1 What is involved in knowing a word

Form	spoken	R	What does the word sound like?
		P	How is the word pronounced?
	Written	R	What does the word look like?
		P	How is the word written and spelled?
	word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express this meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	Collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	constraints on use (register, frequency . . .)	R	Where, when and how often would we expect to meet this word?
		P	Where, when and how often can we use this word?

(Nation, 2001, p. 27)

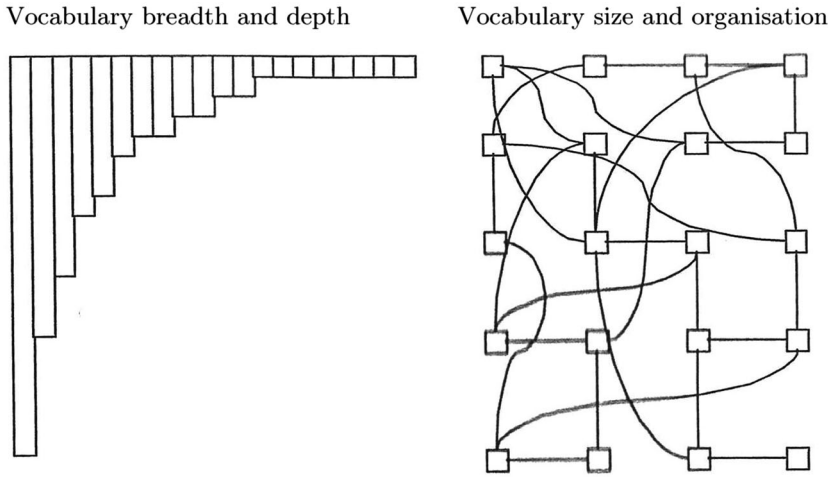

Figure 2 Two ways of looking at vocabulary (Meara and Wolter, 2004, p. 89).

declarative knowledge). Conversely, many word knowledge aspects are related to vocabulary in contextualized use. For example, a word's collocations, register constraints, and frequency may well vary in different usage contexts. These types of contextualized word knowledge are typically implicit in nature (i.e., intuitive and difficult to explain) and are probably mainly gained from massive exposure to a language, essentially by developing statistical probabilities of use (e.g., Ellis, 2012). Thus, the word knowledge aspects that are more amenable to explicit study are likely to be mastered before those word knowledge aspects that require exposure to numerous diverse contexts. Also, it is usually the more explicit aspects, particularly the form–meaning link, that are measured in size tests, while the depth tests tend to focus on more implicit aspects of word knowledge, such as associations and collocations.

The conceptualization of vocabulary as components views vocabulary depth as knowledge of individual lexical items. However, Meara and Wolter (2004) quite rightly point out that "[v]ocabulary size is not a feature of individual words: rather it is a characteristic of the test taker's entire vocabulary" (p. 87). They illustrate this point with a comparison of two ways of looking at vocabulary knowledge (Figure 2).

In the first way, lexical items to the left learned earlier have greater depth, while more recently learned items to the right have less depth of knowledge.[1] Meara and Wolter point out that, in this view, items are unrelated to each other, and learning one has no effect on learning another. This is unlikely to

be true, and so they promote a more interrelated network model, where greater linkage between items leads to better connectedness and organization. In this conceptualization, developing depth is seen as greater lexical organization. This has typically been assessed by the ability to either recognize or provide a greater number of associations and/or more typical ones (e.g., *common* → ordinary, shared, boundary, name). Meara (1997) suggests that lexical organization might even be at the root of receptive–productive mastery: Items with the right kind of connection would become productive, while those lacking such connections would remain at a receptive level.

Another general approach to conceptualizing depth relates to what learners can do with the lexical item. A very widespread distinction following on from this approach is receptive versus productive mastery of an item (sometimes referred to as passive and active mastery, respectively). Receptive mastery entails being able to comprehend lexical items when listening or reading, while productive mastery entails being able to produce lexical items when speaking or writing. This dichotomy has great ecological validity, as virtually every language teacher will have experience of learners understanding lexical items when listening or reading, but not being able to produce those items in their speech or writing. Unsurprisingly, studies have generally shown that learners are able to demonstrate more receptive than productive knowledge (e.g., Laufer & Paribakht, 1998; Waring, 1998), but the exact relationship between the two is less than clear. In contrast to Meara's (1997) association idea above, Melka (1997) suggests that receptive and productive mastery lie on a developmental continuum and that knowledge gradually shifts from receptive mastery toward productive mastery as more is learned about the lexical item.

Read (2000) notes another interesting threshold issue: "Is there a certain minimum amount of word knowledge that is required before productive use is possible?" (p. 154). To date there has been little research to inform this key issue. Most research has compared the ratios between receptive and productive vocabulary, but very few studies have explored the type and amount of lexical knowledge necessary to enable productive use of individual lexical items. However, the component word knowledge perspective is suggestive. For receptive purposes, knowing the form–meaning link may be enough. The form is given in the speech or writing, and users must recognize that form and then recall the meaning to be attached to it. All of the rest of the word knowledge information (part of speech, derivative forms, collocations, etc.) are already provided in the context. Knowing these other aspects would undoubtedly aid comprehension, but the form–meaning link by itself would probably be enough to extract meaning in most cases. However, when users wish to produce the lexical item,

they already have the meaning they wish to express in their heads, but must know all of the word knowledge aspects to produce the item appropriately in the context they are creating. From this perspective, productive mastery is more difficult or advanced than receptive mastery because (1) more word knowledge components are required and (2) many of these components are contextual in nature (e.g., collocation, register constraints) and take a long time to develop.

Another possible way of thinking about what learners can do with lexical items is how fluently or automatically the items can be used in each of the four skills (reading, writing, listening, and speaking).[2] Vocabulary has often been used as an expedient (perhaps easily measureable) linguistic aspect with which to measure fluency/automaticity/speed of retrieval or judgment in psycholinguistic experiments (e.g., Segalowitz, Watson, & Segalowitz, 1995; Segalowitz, Segalowitz, & Wood, 1998), but fluency has just begun to be addressed in research that focuses specifically on vocabulary itself, such as Harrington (2006), Harrington and Carey (2009), Pellicer-Sánchez and Schmitt (2012), and Read and Shiotsu (2010). Fluency is important because it moves the conceptualization of lexical proficiency onward from simple knowledge to the ability to use that knowledge in both comprehension and production (i.e., to implicit or procedural control). As the essence of vocabulary mastery is the ability to use lexical items fluently in communication (not the ability to talk about them metalinguistically), fluency is a key aspect of lexical proficiency. As it increases over time (e.g., Segalowitz et al., 1998), fluency can be conceptualized as part of depth (although Daller, Milton, & Treffers-Daller [2007] view it as an independent dimension in addition to size and depth).

We have seen in this section that depth is a rather loose construct that can be conceptualized in a variety of ways, and there are certainly other conceptual possibilities that I did not cover here. In fact, one reviewer commented that depth is "about the wooliest, least definable, and least operationalisable construct in the entirety of cognitive science past or present." Still, the degree of quality/employability/depth of the items in one's mental lexicon surely matters, and so this notion is an important one that needs to be engaged with. However, it is difficult to think of one conceptual approach that can capture all of the aspects of depth discussed above. Schmitt (2010b) cites the lack of an overall theory of vocabulary acquisition as one of the prominent gaps in the field, and one reason for this lacuna is the difficulty in capturing all of the multifarious aspects of lexical knowledge under one theory. Given the theoretical difficulties, this article takes a research-first approach, which will hopefully help illuminate the issues and be a step toward building a better understanding of the nature of depth of vocabulary knowledge.

Measuring Vocabulary Size and Depth

The various ways of conceptualizing depth (word knowledge components, lexical organization, receptive and productive mastery, and fluency) are not discrete; rather they are overlapping and interrelated. The boundaries between each are fuzzy, and they blend into one another depending on how one defines and operationalizes each conceptualization. The reason why there is currently no measure of depth as a whole is partly due to depth being a very broad construct that cannot simply be measured in a single test or even practically in a battery of tests. This complexity makes it difficult to get good measures of either size or depth, and the relationship between size and depth depends to a large degree on how each is measured. With certain measures, breadth and depth may appear very similar. Selection of other measures might lead to a large gap between size and depth being indicated. This makes an understanding of the measures themselves critical to an understanding of the relationship between size and depth. A good example of this is the oft-cited paper by Vermeer (2001), who suggests that, for all practical purposes, size is indistinguishable from depth. But is this correct? This can only be determined by analyzing the measures he used.

Vermeer (2001) studied 50 children learning Dutch as either their first (L1) or their second language (L2). There were two size tests: one in which words were spoken and the child had to choose the correct meaning from four pictures (multiple-choice meaning recognition) and one in which the child had to indicate the meaning in some way (meaning recall). Unsurprisingly, the recognition size scores were much higher than the recall size scores (about twice higher for L1 and about four times higher for L2), although the two tests correlated strongly at $r = .80$ (L1) and .75 (L2). This difference in scores demonstrates that measuring vocabulary size is not straightforward, but depends on the measures used. The depth test required children to explain various semantic aspects of 10 concrete nouns, for example, "What is a ___?, What does a ___ usually look like?, Can you tell us some more about a ___?" The recognition size test correlated with the depth test at $r = .85$ (L1) and .76 (L2). The meaning recall test correlated with the depth test at $r = .51$ (L1) and .72 (L2). Vermeer rightly highlights these consistently strong correlations, but they are not surprising when one considers that the depth test only tapped into deeper semantic knowledge of a single meaning sense, so all tests (both size and depth) were essentially various types of meaning tests. Thus, it was almost inevitable that they would produce the very similar results. Vermeer

used the similarity of results as a basis for stating that there is no conceptual distinction between size and depth. This may be true if depth is conceptualized narrowly as knowledge of a single meaning sense. However, Vermeer might have found completely different results if he had measured vocabulary depth as knowledge of other meaning senses (polysemy) or knowledge of other aspects of word knowledge (e.g., collocation or derivative knowledge, which may well be much more difficult to acquire). This is especially true if he had required productive mastery of the aforementioned aspects. As it is, he almost certainly overinterpreted his results, which I argue stem from all the instruments tapping single meaning senses, as demonstrating the essential sameness of vocabulary size and depth in general. The key point is that, in discussions of size and depth, it all comes down to how each construct is conceptualized and measured.

Length constraints do not allow a discussion here of the vocabulary measurement issues which will underlie this size–depth critique. However, a brief overview is provided in the Supporting Information online (including an explanation of vocabulary tests commonly used), and a fuller discussion is available in Schmitt (2010b).

The following sections will explore the size–depth relationship by reviewing as much extant research as possible which has both (1) a size measure (or a reasonable proxy) and (2) some measure that relates to one of the depth conceptualizations mentioned above. This overview will focus on L2 studies, although it will include a few relevant L1 studies. Care will be taken to consider what aspects were measured in both the size and depth measures. As most size measures focus on the form–meaning link, most of the variation in the studies rests in the depth measurements. Consequently, this critical synthesis will be divided according to the various depth conceptualizations. I have divided these into seven main categories, which will form the organization of this article from this point forward according to whether vocabulary size versus depth are conceptualized as:

1. receptive versus productive mastery
2. knowledge of multiple word knowledge components
3. knowledge of polysemous meaning senses
4. knowledge of derivative forms (word family members)
5. knowledge of collocation
6. the ability to use lexical items fluently
7. the degree and kind of lexical organization

Vocabulary Size and Depth Conceptualized as Receptive Versus Productive Mastery

Melka (1997) surveyed several studies that claim the difference between receptive and productive mastery is rather small; one of these estimates that 92% of receptive vocabulary is known productively. Takala (1984) suggests the figure may be even higher. Other studies suggest that there is a major gap between the two: Laufer (2005) found that only 16% of receptive vocabulary was known productively at the 5,000 frequency level and 35% at the 2,000 level. Other studies conclude that around one-half to three-quarters of receptive vocabulary is known productively (Fan, 2000; Laufer & Paribakht, 1998). The inconsistency of these figures highlights the difficulties and confusion involved in dealing with the receptive–productive issue. One problem is the lack of an accepted conceptualization of what receptive and productive mastery of vocabulary entails. It is probably best seen as skills based, that is, defined as the ability to recognize and understand words in a reading passage versus the ability to produce them in writing. However, I could find no research that followed this approach and also included a size measure for comparison. Most research addresses receptive and/or productive knowledge of only the form–meaning link. Even here, scores of receptive and productive vocabulary can vary considerably depending on the type of measurement used. Waring (1999) found that a difficult receptive test format could even lead to lower scores than a relatively easy productive one. Another perennial problem is that virtually all receptive measures involve some variation of multiple-choice format, which provides the possibility to score correctly through guessing, but productive formats typically do not (see Webb, 2008). Unfortunately, most researchers do not consider the effects of guessing, and all of the results below need to be viewed with this caveat in mind. For all of these reasons, the following discussion will give careful consideration to each study's method of measurement.

Knowing the Form–Meaning Link to Recognition and Recall Levels of Mastery

Laufer and Paribakht (1998) gave Israeli learners of English as a Foreign Language (EFL) and Canadian students of English as a Second Language (ESL) the Vocabulary Levels Test (VLT) (form recognition) and the Productive Vocabulary Levels Test (PVLT) (form recall) (see Appendix S1 in the Supporting Information online for details of these tests). The Israelis had a form recognition size of a little over 4,000 word families and a form recall size of a little over 3,000 families, and the students in Canada had sizes of about 6,500 families

Table 2 The ratio between form recall and form recognition (recall ÷ recognition)

Frequency level	EFL%[a]	ESL%[b]	German%[c]	Japanese%[d]
2,000	93.5	84.4	84.1	55.7
3,000	75.9	58.3	66.4	31.5
5,000	62.0	63.0	46.4	15.5
10,000	46.0	44.0	–	–

[a]Laufer and Paribakht Israeli participants [b]Laufer and Paribakht Canadian participants [c]Tschirner German participants [d]Waring Japanese participants

and 4,000 families, respectively. Form recall and form recognition correlated strongly ($r = .72 – .89$), but the form recognition scores were always higher than the form recall scores. The ratio of form recall to form recognition (i.e., PVLT ÷ VLT) was 77% for the Israeli students and 62% for the Canadian students. But this varied considerably according to frequency level and EFL versus ESL status (Table 2). The recall–recognition gap was small at the 2,000 frequency level, at only about 7–16 percentage points. However, by the 10,000 level, the gap increased to 54–56 percentage points. Overall, as the frequency level decreases, the recognition–recall gap increases. This means that learners are more likely to have both form recognition and form recall mastery at the higher frequencies (i.e., smaller gap) and less likely to have form recall mastery at the lower frequency levels (i.e., form recognition mastery only and a wider gap). Thus form recall lags behind form recognition. Tschirner (2004) found congruent results using the same tests with 142 1st-year German university students of English language and literature (Table 2). Waring's (1998) results with Japanese EFL learners were also similar (Table 2).

Nemati (2010) gave the VLT and PVLT to 100 Indian ESL learners in Grades 8–12. Overall, mastery of form recall seemed difficult to achieve for these learners, with the form recognition scores being about five times higher than the form recall scores. There were correlations of $r = .62–.86$ between the VLT scores at the various frequency levels (i.e., VLT 2,000 vs. VLT 3,000; VLT 2,000 vs. VLT 5,000, VLT 3,000 vs. VLT 5,000). The correlations among the PVLT levels were $r = .58–.72$. It is interesting that these correlations were stronger than the correlations ($r = .27–.38$) between the VLT and PVLT scores at each individual frequency level (i.e., VLT 2,000 vs. PVLT 2,000; VLT 3,000 vs. PVLT 3,000; VLT 5,000 vs. PVLT 5,000). That is, the form recognition scores were more closely related to each other across the three frequency levels than they were to the form recall scores at their own particular frequency

level. The same is true for the form recall scores. This suggests that the two degrees of mastery of form (recognition vs. recall) really are two quite different constructs.

Comparing Form Recall and Meaning Recall

The Computer Adaptive Test of Size and Strength test (CATSS; Laufer & Goldstein, 2004; see Appendix S1 of the Supporting Information online) allows the comparison of knowledge of form versus meaning. Laufer and Goldstein gave it to 435 Hebrew- and Arabic-speaking high school and university students. The vocabulary size of the participants was at least 3,000 word families on the meaning recall format. The form and meaning recall scores correlated at $r =$ .53, but the ratio of form recall to meaning recall (form recall ÷ meaning recall) varied according to frequency level: 2,000: 35.0%; 3,000: 30.4%; 5,000: 16.0%. Thus the form recall–meaning recall gap increased as the frequency decreased in a similar manner to the above studies. It therefore follows that form recall is more difficult than meaning recall, and so form recall can be seen as a greater degree of depth than meaning recall. In fact, form recall seemed to be relatively difficult to master for these participants, as they knew only a minority of the words to this level of mastery, even at the highest frequency levels.

Using the same CATSS format but without using the initial letter as a cue in the form recall gap, Webb (2008) gave matching form recall and meaning recall tests to 83 Japanese university EFL students. The form recall–meaning recall ratio was 77% with a strict spelling criterion, but 93% when a lenient spelling criterion was used. Webb found that the ratio increased with decreasing frequency of the target words with strict scoring (at about 1,000 = 88% [All participants], 93% [Upper group], 84% [Lower group]; at about 2,000 = 73%, 86%, 65%; at about 3,000–5,000 = 65%, 75%, 58%), but that there was little difference with lenient scoring. Webb's trends are in line with Laufer and Goldstein (2004), whose results show that the gap increases as words become less frequent, and it is apparent that this is true for relatively more and less proficient learners. However, Webb also shows that this depends on the degree of precision required for form recall. If the exact form is required (strict spelling), then form recall knowledge does not keep up with meaning recall knowledge. But if partial knowledge of the form is considered enough (lenient spelling), then the gap basically disappears. In other words, these learners seemed to have a general idea of the spelling for L2 words they knew the meaning for, but in a large number of cases (approximately between 10% and 40% of cases, depending on the frequency) they did not know the exact spelling.

Knowledge of Written Versus Spoken Word Forms

The above results used written tests and so apply only to the written form. While it is debatable whether knowledge of either written or spoken form should be considered deeper than knowledge of the other, it seems reasonable to consider knowledge of both forms as deeper than knowledge of only one. However, there have been relatively few studies addressing oral vocabulary. Those that do exist have taken three distinct approaches to measurement.

Milton and various colleagues have opted for Yes/No tests (also called checklist tests), which can be interpreted as self-report meaning recall tests (see Appendix S1 in the Supporting Information online for background on Yes/No tests). Using Meara's (n.d.) written (X_Lex) tests, and matching spoken (A_Lex) Yes/No tests, Milton and Hopkins (2006) studied 88 Greek and 38 Arabic-speaking learners of English with a wide proficiency range. The X_Lex and A_Lex scores correlated at $r = .68$. The Greek and Arabic speakers had similar written English vocabulary sizes (Greek: 2,699; Arabic: 2,553), but differed greatly in their spoken scores (Greek: 2,017; Arabic: 2,823). Thus, the Arabic speakers had a relatively larger spoken vocabulary size than written one (although statistically nonsignificantly so), while the reverse was true for the Greek participants ($p < .01$). Milton (2009) provided an additional analysis and discussion of Milton and Hopkins (2006) and concluded that, as vocabulary size increases, the gap between written versus oral knowledge widens. For vocabulary sizes of $\leq$ 2,000 lemmas, the number of families known phonologically (as indicated by the A_Lex test) are greater than the number known orthographically (X_Lex test) (i.e., oral size > written size), while for vocabulary sizes of >2,000 families, the number of families known orthographically are greater than the number known phonologically (i.e., written size > oral size). Milton also concluded that this trend obtained for Farsi speakers, as evidenced by Milton and Riordan's (2006) study. Thus, based on the two Milton studies, it would appear that more advanced learners know more sight vocabulary than oral vocabulary by a considerable margin. The relationship between phonologic and orthographic lexical knowledge also depends on the L1. Arabic speakers were slower to develop their orthographic knowledge than the Greeks, but perhaps this is not surprising given that the Greek script is far more similar to the English script than the Arabic script. This L1 difference is also shown by the correlations between the A_Lex and X_Lex scores (Greek: $r = .81$; Arabic: $r = .65$).

However, some of the findings by Milton and colleagues have been called into question in a second approach to investigating depth in oral vocabulary by Van Zeeland (2013), who used a meaning recall interview for her measure. This

probing interactive measure should give a more accurate indication of learners' underlying knowledge than a self-assessment measure such as the Yes/No format. Van Zeeland measured the written and spoken vocabulary knowledge of advanced L2 learners of English from various L1 backgrounds. Results showed a stronger correlation between written and spoken word knowledge than found by Milton and Hopkins (2006) ($r = .85$ vs. .68). The relationship between learners' knowledge of written and spoken vocabulary furthermore remained constant as overall scores increased, meaning that the growing advantage of written over spoken knowledge, as reported by Milton, was not found. These results suggest that knowledge of vocabulary in the two modes may be more closely related than suggested by Yes/No test results. Interestingly, van Zeeland also analyzed the accuracy of learners' self-assessment of their written and spoken word knowledge by means of a Yes/No test covering the same vocabulary as the interview. This showed that learners' self-assessment was considerably less accurate in the spoken than in the written mode. This questions the use of the Yes/No format for assessing spoken vocabulary knowledge. The discrepancy in results between van Zeeland and Milton and colleagues might thus possibly be explained by weaknesses in the A_Lex test (no validity evidence provided). However, they might equally well be explained by proficiency. Van Zeeland's participants were advanced postgraduate students studying at a British university, while Milton and Hopkins' (2006) participants were language school students with a mean vocabulary size of below 3,000 word families. It may well be that more advanced learners tend to have relatively balanced spoken and written vocabularies, while lower-level students are prone to the type of imbalanced vocabularies found by Milton and colleagues.

The third approach employs translation. As part of a larger study, Shimamoto (2000) measured meaning recall knowledge in an L1 translation test comprising 50 L2 target words (of unspecified frequency) in isolation in written and spoken modes. The 134 Japanese university students were able to translate 55.6% of the spoken words and 65.2% of the written words, indicating that these learners knew more written forms than spoken forms.

Summary

It is clear that lexical knowledge is not uniform across the receptive–productive distinction. The differences in mastery of all of the lexical aspects discussed in this section indicate they are known to different degrees of depth. It is easier to recognize and understand a word's written form than to recall that form from memory. Furthermore, the ability to recall a word's written form precisely when one knows its meaning is more difficult than the ability to recall a word's

meaning when given its form. The gap between both written form recall versus recognition and form versus meaning recall is smaller at the higher frequency levels, presumably because the high amount of exposure to these words helps to fill in the more difficult types of knowledge. But as frequency decreases, the gap increases as the more difficult knowledge types lag behind. However, there is also some evidence that learners are able to recall an approximate written form for most words they know the meaning of. There is relatively little evidence concerning knowledge of spoken forms and what evidence exists is mixed, but there are initial indications that, while less proficient learners might have imbalanced vocabularies, advanced learners tend to have developed written and spoken vocabularies of similar sizes. However, if there is some imbalance at advanced proficiencies, it is likely to be in favor of written vocabulary. Word frequency seems to have some effect, and at high frequency levels, learners may know more words in spoken than written form. At the lower frequency levels, learners may know more words in the written form, although the L1 seems to be a factor that affects this.

Vocabulary Size and Depth Conceptualized as Knowing Multiple Word Knowledge Aspects

Another way of conceptualizing depth of knowledge is some degree of mastery over one or more of the word knowledge aspects illustrated in Table 1. A limited number of studies have included test batteries that address a number of these aspects concurrently, albeit usually with very small numbers of target items. (This is a common limitation across most depth measures.) In one of the earliest, Schmitt (1998) carried out a detailed longitudinal study of the lexical knowledge of three university postgraduate ESL students. He measured various types of word knowledge of 11 words: spelling recall, derivative recall, productive associations, and polysemous meaning recall. As the proportion of meaning knowledge increased, so did the association and derivative scores. Similarly, as the association proportion increased, so did the derivative and meaning scores. The same trend applies when comparing association and meaning scores to grammar scores. Overall, this paints a picture of the four types of word knowledge improving (rather slowly) in a roughly parallel manner.

Chui (2006) studied 186 Hong Kong university ESL students using the PVLT and a five-part depth test. The depth test covered 20 headwords from Coxhead's (2000) Academic Word List (AWL) and required students to (1) recognize their part of speech, (2) recall their meaning, (3) recognize their collocations, (4) produce an assigned derivative form, and (5) construct a

meaningful sentence with these words. She found that the average total vocabulary size was relatively low for university study, at around 3,350 word families (although see Appendix S1 in the Supporting Information online for more on the PVLT and its limitations). Furthermore, although AWL words are necessary for academic study, the mean scores on the depth test showed real deficiencies in these ESL students' knowledge: word class 87.6%, meaning 64.5%, collocation 57.4%, and derivatives 52.3%. The correlations between the AWL size (form recall) and AWL depth was higher for derivatives ($r = .78$), collocations ($r = .69$), and sentence production ($r = .69$) than for meaning ($r = .53$) and word class ($r = .53$). This is not surprising because form recall is the most difficult degree of mastery of the form–meaning link, and so correlates with the more difficult contextual word knowledge aspects (derivatives, collocations) and the ability to use the word productively. Word class and meaning recall were better known in absolute terms and thus seem to be easier word knowledge aspects to acquire and may be more easily learned than contextual word knowledge regardless of one's overall vocabulary size.

Shimamoto (2000) gave the VLT and the following four depth measures for 50 words to 134 Japanese university students: (1) meaning recall from a spoken prompt, (2) meaning recall from a written prompt, (3) recognition of paradigmatic associations (e.g., synonyms or hyponyms), and (4) collocation recognition. For all students, the mean vocabulary size was 3,036 word families, and this size correlated with the four depth tests as follows: spoken meaning recall ($r = .76$), written meaning recall ($r = .77$), paradigmatic associations ($r = .81$), and collocation ($r = .73$). Overall, the correlations were similar and strong. The students were then split into three vocabulary size categories: smaller (mean 2,600 families), medium (3,200), and larger (3,700). The relative relationship between the four depth categories remained the same for each size group (paradigmatic > collocation ≈ written meaning recall > spoken meaning recall). The correlations between size and depth varied between $r = .56$ and .68 for the 2,600 group, and between $r = .37$ and .57 for the 3,700 group (but were inexplicably statistically nonsignificant for the 3,200 group). Overall, the relationships between size and the various depth measures were relatively strong, but seemed to be weaker for learners with relatively larger vocabularies. So as vocabulary size increased, depth of word knowledge seemed to lag behind.

In a series of studies, Webb (2005, 2007a, 2007b), explored the acquisition of 10–20 nonwords by Japanese EFL learners from very small amounts of exposure. The noteworthy feature of these studies is the extensive battery of tests used to measure a variety of word knowledge aspects, both in terms

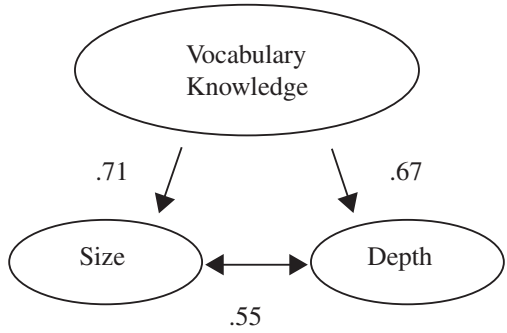

Figure 3 Structural equation model of vocabulary knowledge (extract from a larger model; Tseng and Schmitt, 2008, p. 381).

of recall and recognition: form–meaning link, spelling, word class, paradigmatic association, and syntagmatic association. Although his methodology cannot inform directly about the relationship between size and depth, it can provide insights into the relationships between various types of depth knowledge, at least at the initial acquisition stage. Although there is some variation in how well each word knowledge aspect was acquired, in all studies there is a remarkably even profile of the word knowledge aspects, with no aspect being clearly learned more or less well than the others (although the syntagmatic associations did tend to have slightly lower scores). Overall, knowledge of the various word knowledge aspects seems to correspond well to knowledge of the form–meaning link. This finding is supported by Koizumi (2005), who also studied Japanese learners. Her form recall test correlated with her depth tests as follows: derivation ($r = .78$), antonyms ($r = .81$), and collocation ($r = .67$).

It is thus clear that size and word knowledge aspects are related. Tseng and Schmitt (2008) demonstrated this in another way by using structural equation modeling (SEM) with Chinese L1 university students. Their size measure was the combined scores of the VLT 2,000, 3,000, and 5,000 levels, while their depth measure was the combined scores of a meaning recognition test of polysemy, a form recall test, and a collocation recognition test. They found that the composite size measure loaded on the construct of Vocabulary Knowledge at $r = .71$ and the composite depth measure at $r = .67$ and that they correlated with each other at $r = .55$ (Figure 3). In a later study using similar measures and participants, Tseng (2011) found similar loadings of $r = .76$ (size) and $r = .66$ (depth) for lower performers, but even higher loadings of $r = .88$ for both

size and depth for the high performers. This means that, although related, both size and depth are strong independent components of lexical knowledge.

All of the studies reviewed confirm that vocabulary size and the degree of mastery of various word knowledge aspects are related, with the Tseng and Schmitt (2008) and Tseng (2011) SEM models showing both loading on the construct of overall vocabulary knowledge at very similar levels. However, discerning clear patterns beyond this is difficult. Some studies show size and mastery of word knowledge aspects improving in a roughly parallel manner, but how size is measured may make a difference. Size as measured by form recall might be more strongly related to various contextualized word knowledge aspects than size as measured by the easier meaning recall. The relationship between size and depth may be weaker at larger vocabulary sizes than at smaller sizes, as mastery of word knowledge aspects lags behind knowledge of the form–meaning link for increasing numbers of words. All of these results need to be viewed cautiously in light of the typically small number of depth measurement items used in the studies.

Vocabulary Size and Depth Conceptualized as Knowledge of Polysemous Meaning Senses

In the previous section, I looked at depth of vocabulary knowledge as multiple vocabulary knowledge aspects, but it is also possible to conceptualize depth as mastery of a single aspect. More importantly for this synthesis, many studies have used a single aspect for their measure of depth. For the most part, the earlier section on reception and production focused on form and meaning and the link between the two. But depth of meaning knowledge can also be seen as knowledge of the multiple meaning senses of polysemous words.[3] Although knowledge of these meaning senses can be discrete and declarative in nature, there is also a contextual facet, as polysemous words can mean different things depending on the context and we cannot understand the word without context indicating which meaning sense is in play.[4]

In an already mentioned study, Schmitt (1998) followed three advanced L2 postgraduate university students (with a TOEFL score of about 550) over the period of 1 academic year and elicited their knowledge of the multiple meanings of 11 polysemous target words (e.g., *abandon, illuminate, trace*). Overall, knowledge of the multiple meaning senses increased, and was retained more than it was forgotten, but none of the students had a complete knowledge of all the possible word senses available for the target words. Crossley, Salsbury, and McNamara (2010) studied the spontaneous production of polysemous words

over a 12-month period from six beginning L2 learners from an Intensive English Language Program (IELP; mean TOEFL score = 358), analyzing the vocabulary produced during 30–45-minute interviews. They found that the students began to produce more polysemous words in their first trimester, but that they only began to extend the core meanings of these polysemous words (*know, name, place, play, think*, and *work*) in the second and third trimesters. Although no vocabulary size figures were reported in either study, the participants' general language proficiency was reported to have improved in each, and so we can assume their vocabulary size probably increased as well. To the extent that this assumption holds, the increased vocabulary size relates to increased knowledge and/or use of polysemous meaning senses.

Verhallen and Schoonen (1993) studied 40 L1 Dutch and 40 Turkish L2 Dutch children at the ages of 9 and 11. In extensive highly structured individual interviews, they elicited a comprehensive range of knowledge of the various meaning aspects for the Dutch equivalents of the following six words: *nose, predator/beast of prey, alarm clock, secret, book,* and *hair*. Although no descriptive statistics were reported, the Dutch children produced a greater percentage (58.6%) of the 8,833 total meaning aspects produced in the study (e.g., you smell with a nose, breathe through it, it has two holes, it is pointed) than the Turkish children (41.4%). Likewise, higher proficiency children produced more (52.5%) than lower proficiency children (47.5%). The 11-year-olds produced 53.85% compared to 46.15% for the 9-year-olds. If we assume that children with L1 Dutch, higher proficiency, and older age have relatively larger vocabularies, then larger vocabulary size relates to more comprehensive semantic knowledge of the target words.

The above studies do not have direct measures of vocabulary size, but the relationship between increasing knowledge of multiple meanings and improving language proficiency (with the assumption of concurrent vocabulary size gains) strongly suggests that there is a relationship between vocabulary size and knowledge of multiple meanings. However, the study designs do not allow a solid empirical comparison (e.g., a correlation figure), and so the strength of this relationship remains unspecified.

Vocabulary Size and Depth Conceptualized as Knowledge of Derivative Forms

Another individual word knowledge aspect that can be considered an indicator of vocabulary depth is knowledge of affixes and how these relate to the various derivative members of a word family. Learning and appropriately using a

complete word family can be difficult for learners (*nation, nationalize, national, international, internationalize*, etc.; e.g., Schmitt & Zimmerman, 2002), and misuse is a common source of usage error (**The oil spill is a nation disaster*). Indeed, native English-speaking adolescents do not master some affixes until into high school, so large vocabulary size does not guarantee affix knowledge, as it comes late even in the L1 (Nagy, Diakidoy, & Anderson, 1993).

In L2 contexts, the modest relationship between size and derivative knowledge is shown by low to moderate correlations in three studies carried out in Japan. Schmitt and Meara (1997) studied high school and university EFL learners at the beginning and end of an academic year. They found that vocabulary size (VLT) correlated with verb suffix recall at $r = .27$ (T1) and .35 (T2), and with verb suffix recognition at $r = .37$ (T1) and .41 (T2). The Japanese learners averaged a vocabulary size of 3,900 word families at T1 and 4,230 at T2, but this increase was not reflected in appreciably stronger size–suffix correlations in the T2, as it was for size–association correlations (see the section on association below). Similarly, Mochizuki and Aizawa (2000) studied 403 high school and university EFL students with a modified VLT (mean: 3,769 words) and a test of 29 affixes that measured recognition of prefix meaning (e.g., *anti-* = opposed) and knowledge of suffixes' word class function (*-able* = adjective form). Size correlated with knowledge of prefixes at $r = .58$, suffixes at $r = .54$, and combined at $r = .65$. Noro (2002) studied 90 university students with VLT 2,000 and 3,000 levels and a morphology test which required knowledge of the meaning of the affix and its word class (homeless = without a home and it changes a noun into an adjective). The students were divided into high- and low-vocabulary size groups (high = a minimum vocabulary of 2,500 families). For all 90 participants, the VLT scores correlated with the morphology scores at $r = .69$. For the lower-size group, the VLT and morphology scores correlated more strongly ($r = .54$) than for the higher-size group ($r = .42$). This suggests that, as vocabulary size grows, the gap between size and derivative morphology knowledge widens, but this conclusion is limited by the relatively small vocabulary size of even Noro's high group. Overall, the correlations in these studies are lower than for many other conceptualizations of depth found in this review, but it must be noted that the vocabulary sizes of the participants in all three studies were quite low, and one must wonder if the correlations might be stronger for learners with higher vocabulary sizes.

Kieffer and Lesaux (2008) also found moderate correlations between vocabulary size and derivation scores in an immersion environment. They studied 87 Spanish ESL students in the United States who were assessed in both Grade 4 and Grade 5. Size was measured by the Peabody Picture Vocabulary Test

(PPVT; Dunn & Dunn, 1997), in which learners choose a picture from four choices that corresponds to a word presented orally (meaning recognition). The morphology test provided a derivative form (e.g., *complexity*) and learners were asked to extract the base word (e.g., *complex*) to complete a sentence (e.g., *the problem is* ___). Vocabulary size and this form of morphology knowledge correlated at $r = .53$ (Grade 4) and $r = .46$ (Grade 5). Kieffer and Lesaux (2012a) used the same instruments when following 77 of the above students as they transitioned to middle schools and observed the following correlations: $r = .50$ (Grade 4), .51 (Grade 5), .57 (Grade 6), and .56 (Grade 7). That is, although there was almost always improvement in both size and derivative morphological knowledge over the period, the relationship between vocabulary size and derivation knowledge remained relatively stable at modest levels for these students through 4 years of schooling. In terms of rate of gain, the correlation was $r = .67$, indicating that the students with more rapid growth in one also demonstrated more rapid growth in the other, although the relationship was not particularly strong. In a third study, Kieffer and Lesaux (2012b) inserted measures of synonyms, multiple meaning senses, and semantic associations into a SEM model for their construct of vocabulary size and similar measures of derivative knowledge as above for their construct of morphological awareness. They found a correlation of $r = .76$ between the two constructs, showing a stronger relationship than in their 2012a study, but it must be noted that their vocabulary size construct is really closer to semantic word knowledge than size.

Given natives' protracted acquisition of affixes, it is not surprising that ESL learners have trouble learning the various derivative forms and that this knowledge is not in place until later in lexical development. Thus, it seems that suffix or derivation knowledge does not develop fully until a relatively large vocabulary is in place, although the modest correlations between size and derivations reviewed in this section show that larger sizes do not guarantee mastery of morphology. Milton (2009) reviewed Mochizuki and Aizawa's (2000) results and concluded that a vocabulary size of 3,000–5,000 families is necessary for affixes to be mastered receptively, but that even at 5,000 families some affixes may not be known well (i.e., *ante-, counter-, in-, inter-, -ish, -ity, -less, -y*).

Vocabulary size and derivative knowledge are related, but not as strongly as some other word knowledge aspects, with correlations generally falling in the .50s or lower. This is exemplified by Schmitt and Meara (1997), who used both suffix recognition and recall measures. It is not so surprising to find a low correlation between productive derivative knowledge and vocabulary size ($r = .27, .35$), but even the correlations for receptive derivative knowledge and

size were only $r = .37$ and .41. As with the other word knowledge aspects discussed, the relationship between size and derivative knowledge may be weaker at larger vocabulary sizes than at smaller sizes, as mastery of derivative morphology lags behind knowledge of the form–meaning link for the larger number of words. This is shown by only a moderate relationship ($r = .67$) between the rate of growth of size and morphological knowledge. A possible explanation for the lower correlations is that, unlike other word knowledge aspects, knowledge of affixes is part of English's morphology system. Although it may often be difficult to know which affixes go with which words (e.g., Schmitt and Zimmerman's 2002 results) some affixes are probably predictable enough to lead to the lower size–depth correlations that were found. In other words, the lower correlations with vocabulary size may mean simply that the more transparent elements of the morphology system had already been acquired before learners reached an advanced stage of language proficiency, as indicated by vocabulary size.

Vocabulary Size and Depth Conceptualized as Knowledge of Collocation

There are not many studies that have directly compared vocabulary size and mastery of collocation (see Schmitt [2010b] for an overview of formulaic language in general and Henriksen [2013] for an overview of the acquisition of collocations). However, we can get an initial indication of the size–collocation relationship from studies that have compared collocation knowledge with L2 language proficiency, because many studies have shown that vocabulary size correlates substantially with language proficiency. For example, correlations of $r = .43$–$.79$ were produced in the DIALANG[5] project (see Schmitt, 2010b, Table 1.1). Laufer and Goldstein (2004) found that knowing the form–meaning link of words accounted for 42.6% of the total variance in participants' class grades according to a regression analysis. Milton, Wade, and Hopkins (2010) found that orthographic vocabulary size (X_Lex) correlated with IELTS scores at $r = .68$, and phonological size (A_Lex) at $r = .55$.

Thus, language proficiency can provide an indirect indication of the relationship between size and mastery of collocation. Bonk (2001) gave a collocation battery to 98 low/intermediate to advanced ESL university students and used a reduced 49-item TOEFL test for proficiency. The TOEFL scores correlated with the collocation scores at a relatively strong $r = .73$. Also, collocation items loaded on a different factor in a factor analysis than did the general-proficiency TOEFL items. Likewise, Nizonkiza (2012) found that TOEFL scores correlated

with a form recall collocation test (*Her appointment will fi ____ the gap created when the marketing manager left.*) at $r = .84$. For spoken language, Boers, Eyckmans, Kappel, Stengers, and Demecheleer (2006) found that the number of correctly formed formulaic sequence types (many of which were collocations) in an interview correlated moderately with oral proficiency ratings ($r = .33$–$.61$). These studies show that better mastery of collocation seems to go with higher language proficiency (with its presumed larger vocabulary size).

However, it is obviously better to measure size directly than rely on a language proficiency proxy. Gyllstad (2007) studied 24 Swedish high-proficiency upper-secondary school and university learners of English. He measured their vocabulary size with the VLT and their knowledge of collocations with COLLEX 5, a 3-option form recognition test of 50 items and COLLMATCH 3, a yes/no collocation judgement test of 100 items. He found that the learners' vocabulary size and their knowledge of collocation correlated very strongly at $r = .90$. If the learners had a very large vocabulary size of 10,000 families, then they scored over 90% on both of the collocation tests. Even with much smaller vocabularies, they still knew most of the collocations: 5,000 families about 85% of the collocations and 3,000 families about 70%. Thus, his generally high-proficiency Swedish learners had high levels of collocation recognition knowledge regardless of their vocabulary size. It is probably worth noting that Swedish is closely related to English, and it is an open question to what degree this facilitated the high scores compared to the generally advanced proficiency.

However, Laufer and Waldman (2011) reviewed the literature and concluded that the problem with collocations is not recognition, but using them properly, that is, productive mastery. They cite numerous studies that show that there are collocation errors even when there are no grammatical errors or errors with individual words. This lack of productive knowledge is well demonstrated by Levitzky-Aviad and Laufer (2013). They looked at 290 essays written by Israeli EFL students in Grades 6, 9, and 12 and in 1st year of university. There were very few verb–noun or adjective–noun collocations (1–1.5 per 200-word sample on average), so even the university students did not produce many of these collocations. (Collocations were identified manually and confirmed by whether they occurred in one of three native corpora.) It must be noted that the vocabulary size estimates from the form-recall "active knowledge" version of the CATSS test measure (ACATSS) were low (up to around 2,850 word families for the university students), although they showed steady growth over the four education levels. There were no significant correlations between size and collocation, but this is probably because so few collocations were produced.

Overall, at low vocabulary sizes, it appears difficult to produce collocations in free writing.

This is in contrast to Siyanova and Schmitt (2008), who analyzed the 2.5 million-word Russian subcorpus of the *International Corpus of Learner English* (ICLE, http://www.uclouvain.be/en-cecl-icle.html), and found a large number of appropriate collocations in the Russian university students' compositions. Although the vocabulary size of the ICLE writers is not reported, it is possible to speculate that the Russian university students may have had larger vocabulary sizes than the very low sizes reported for the Israeli students in the Levitzky-Aviad and Laufer (2013) study. If so, it suggests that the ability to produce appropriate collocations may be related to vocabulary size. This suggestion is supported by an unpublished Swansea Ph.D. dissertation by McGavigan (2009) summarized by Milton (2009, pp. 151–155). This was a study with Greek learners of English on another category of formulaic language, idioms. According to Milton, McGavigan found that the vocabulary size of 100 learners correlated with their knowledge of idioms on a gap-fill form recall test at $r = .64$. Overall, Milton concludes from McGavigan's data that these learners needed 3,000 word families before they could acquire many idioms, although even the best performing learners did not know many of them (<25% of the idioms tested).

Mastery of collocations seems to be robustly related to general language proficiency, which is strongly suggestive of its relationship to vocabulary size. There is some direct evidence that students can have good knowledge of collocations at a range of medium to large vocabulary sizes (3,000–10,000 word families) if this knowledge is measured with recognition tests. However, it seems much more difficult to develop mastery of collocations to a level where they can be produced in free writing, in terms of both amount and accuracy. Here vocabulary size seems to make more of a difference, with more advanced learners seemingly able to produce more, and more appropriate, collocations than lower-level students.

Vocabulary Size and Depth Conceptualized as Lexical Fluency

It is uncontroversial that the ability to process language quickly is a component of more advanced language proficiency, and the ability to access the vocabulary in one's lexicon quickly and accurately plays a part in all four language skills. For example, with the average rate of speech at 150+ words per minute, speakers have about 200 to 400 milliseconds to choose and say a word (de Bot, 1992). There has always been an interest in fluency of language use in psycholinguistics, but almost all of this research has been laboratory based

and focused on areas such as bilingualism (e.g., Kroll, Michael, Tokowicz, & Dufour, 2002). Recently, there has been some interest in training fluency in L2 language pedagogy, largely driven by work of Norman Segalowitz and colleagues and synthesized in Segalowitz (2010) (see Segalowitz et al., 1995; Segalowitz et al., 1998). However, I could find only two studies that directly related lexical fluency to vocabulary size (but see Harrington [2006] and Harrington & Carey [2009] for work on fluency and lexical proficiency). Laufer and Nation (2001) used a modified computerized VLT format, but showing only one stem at a time on a screen, instead of three. The response times of 454 Israeli university students were measured while they completed the test. This Vocabulary Receptive Speed Test (VORST) showed that there was an increase in lexical access speed at a vocabulary size of around 5,000 word families and that the larger the vocabulary size, the faster the access speed. Laufer and Nation also compared access speed at the four frequency levels. The correlations between size and speed were strongest at the 10,000 level ($r = .67$) but were also reasonably strong at the 5,000 level ($r = .50$). Overall, the more frequent bands were answered faster, but it seemed to take a relatively large vocabulary size (possibly 5,000) before faster test answering speed came into play across the board. However, when Miralpeix and Meara (2014) gave 145 Spanish university students X-Lex and Y-Lex Yes/No size tests and timed semantic lexical decision tasks of whether target words were animate or inanimate, they found no correlation between vocabulary size and speed.

It is difficult to reconcile these contradictory findings, but it may have something to do with the measures used in the two studies just reviewed. The size–speed relationship obtained with the VORST, where learners had to match word forms with meaning, but did not appear when learners had to make an animate/inanimate lexical decision. It may be that the speed of different types of tasks are differentially related to vocabulary size. It must also be noted that both of these speed measures are based on making judgments about isolated words, and so it is unclear how these relate to the ability to fluently employ vocabulary in the four skills.

Vocabulary Size and Depth Conceptualized as Lexical Organization

The most common conceptualization of vocabulary depth in research has been lexical organization, simply because the most-used measures of depth have been the two Word Associate Format (WAF) tests developed by Read (1993 and 1998 versions) or variations of these formats (see Appendix S2 of the

Supporting Information online for information on these tests). These tests are intended to measure both semantic and collocational associations. A look at the summary of these studies in Table S1 of the Supporting Information online shows that the majority of correlations between size (usually measured by the VLT) and association depth typically fall in the $r = .7–.8$ range, although studies using other measures sometimes report somewhat lower correlations. This was true for learners with very small vocabularies (e.g., around 2,000 word families; Akbarian, 2010: VLT–WAF = .86) and for learners with advanced vocabularies (e.g., Gyllstad, 2007: VLT-WAF = .85 and .89). It was also true for a number of different L1 populations. Based on these studies, we can see that lexicons do seem to become more organized as vocabulary size grows, at least to the extent that word associations and the tests measuring them reflect lexical organization. Thus depth conceptualized as lexical organization appears strongly related to vocabulary size.

A number of these studies have included additional analyses that can enrich the above overall conclusion. One of these is regression analysis, which can determine how much of the variation in the dependent variable (usually reading comprehension) size and depth can independently account for (as expressed in the regression's r^2 value). Typically size was the strongest predictor, with WAF depth adding a significant, but rather small, additional contribution: 5% (Qian, 1999); 4% (Farvardin & Koosha, 2011); 3.5% (Noro, 2002); and 2% (Huang, 2006). However, Qian (2002) found that the WAF explained the greatest amount of reading comprehension scores, and size added an additional 8%. Similarly, Mehrpour, Razmjoo, and Kian (2010) found WAF the stronger predictor of reading comprehension. Whether size or lexical organization proves the strongest predictor in these studies, the key lesson is probably that the additional contribution by the other one is relatively small and that size and lexical organization can be considered essentially equivalent in predicting reading comprehension. It is an interesting, but unexplored, question whether the two would equally predict other kinds of language use.

Some studies have used SEM to compare a number of size and depth measures. One example is Zhang (2012), who studied 172 Chinese adult EFL Masters students studying in China. He found that the VLT and WAF correlated at $r = .52$; however, in his SEM model, there was a latent variable labeled Vocabulary Knowledge, and the VLT loaded at $r = .86$ on it, while the WAF scores loaded at $r = .60$. So size seemed to have a stronger weighting on overall vocabulary knowledge than depth. Similarly, Tseng and Schmitt (2008) found that size loaded somewhat more strongly ($r = .71$) than depth ($r = .67$) on the trait Vocabulary Knowledge.

Some studies have been able to compare the size and depth correlations across learners of varying proficiencies or vocabulary sizes. Henriksen (2008) compared the size and depth of Danish EFL students at three different grade levels. The VLT and a productive association test correlated at $r = .85$ (Grade 7), .69 (Grade 10), and .55 (Grade 13). The VLT and a receptive matching association test correlated at $r = .72$ (Grade 7), but were statistically nonsignificant for Grades 10 and 13.[6] Thus the relationship between size and association is stronger at younger grades than more advanced ones. Because the students had increasing vocabulary sizes at all three grades, we can also interpret the results to show a stronger size–association relationship for smaller vocabulary sizes than larger ones. That is, as vocabulary sizes increased (with greater language proficiency), the association scores did not keep up. Noro (2002) also used groups of varying proficiencies, dividing her Japanese EFL university students into high (minimum of 2,500 families) and low groups. For all 90 participants, the VLT scores correlated with the WAF scores at $r = .69$. For the lower-size group, the size and association scores correlated more strongly ($r = .51$) than for the higher-size group ($r = .36$). These results would appear to indicate that it is easier to grow vocabulary size (as measured by a form recognition test) than association knowledge (as measured by either receptive or productive tests). That is, the form–meaning link is easier to acquire than the lexical organization shown by associations.

Unfortunately, there is just as much evidence for an opposing conclusion. Nurweni and Read (1999) divided their lower-level participants into three groups according to their general achievement in English. The higher proficiency group produced a Translation and WAF correlation of $r = .81$, the middle group a correlation of .43, and lower group a correlation of .18. Thus, the higher the vocabulary size, the closer the relationship between size and lexical organization. Schmitt and Meara's (1997) year-long study of Japanese high school and university EFL learners yielded correlations of vocabulary size with association recall of $r = .49$ (T1) and .62 (T2), and with association recognition of $r = .39$ (T1) and .61 (T2). As already mentioned, the Japanese learners averaged a vocabulary size of 3,900 word families at T1 and 4,230 at T2. Although the size and association correlations were moderate, they seemed to gain strength in the T2 in step with the modest increase in vocabulary size. Akbarian (2010) found a similar trend for his lower-level learners. At the moment, it is difficult to interpret this conflicting evidence and come to conclusions about the strength of the relationship between size and organization as vocabulary size grows.

Greidanus, Bogaards, van der Linden, Nienhuis, and de Wolf (2004) gave both form recognition and form recall size tests along with the WAF. The form recognition scores correlated with the association test at $r = .70$, and the form recall at .81. Thus, the ability to recall the forms of target words was somewhat more strongly related to the ability to recognize various associations than the ability to recognize the word forms. In other words, lexical organization seems most strongly related to the highest level of form–meaning knowledge (form recall), as shown by Laufer and Goldstein (2004) and Laufer, Elder, Hill, and Congdon (2004).

Vocabulary size (as measured by the VLT) and lexical organization (as measured by the WAF) are strongly related to each other. Regression analyses show that the independent contribution of one is relatively negligible once the contribution of the other has been accounted for. However SEM models suggest that if one has greater weighting for overall vocabulary knowledge, it is likely to be vocabulary size. At the moment, it is impossible to say whether the size–organization relationship strengthens or weakens as vocabulary size increases.

General Discussion and Conclusion

So, what are we to make of the empirical evidence? Do size and depth behave as separate constructs as ventured in the introduction, or are they essentially the same construct as Vermeer (2001) suggested? This review has shown that the answer depends on how one conceptualizes, and consequently measures, both size and depth. I agree with Read (2004) that there is currently no true measure of depth, and whatever conceptualization or measure is used, it will only ever tap into limited facets of the overall quality of understanding of a lexical item. The size–depth relationship depends on various factors such as the size of the learner's lexicon, the frequency level of the target words measured, and the learner's L1. For higher frequency words, and for learners with smaller vocabulary sizes, there is often little difference between size and a variety of depth measures. However, for lower frequency words and for larger vocabulary sizes, there is often a gap between size and depth, as depth measures lag behind the measures of size. Furthermore, some types of word knowledge (e.g., derivative knowledge) seem to have generally lower correlations with size than other types.

How one views the size–depth relationship should depend on one's purpose of use. If one wishes to discuss the nature of vocabulary in general, particularly

with practitioners, then the distinction is useful. Vocabulary has become mainstream and is now a major topic in language teaching research (Ellis, 2009). The message about the need for a large vocabulary size to be able to function well in English seems to be taking hold (e.g., Nation, 2006; Schmitt, 2008). However, that message by itself is insufficient, as learners need to know words well in order to use them productively, appropriately, and fluently. The size–depth distinction is useful when talking to practitioners to drive home the need for rich, sustained instruction and input in order to develop knowledge beyond the simple memorization of the form–meaning links.

However, if one's purpose is to characterize vocabulary knowledge in more precise terms, as when theorizing, or when designing and interpreting research or assessments, depth is probably too vague a term to be useful. There are too many conceptualizations of depth to use one cover term for all of them, as it seems clear from this review that the various methods of measuring depth are tapping into different (but related) constructs.

This review used an empirical approach partly because there is no overall theory of vocabulary knowledge and acquisition. So, how do the results inform theory building? One thing that has become clear to me is that there can be no clear distinction between size and depth. Size by definition is the number of lexical items known to some criterion level of mastery. But the criterion will always be some measure of depth. In practice, it is normally a variation of the form–meaning link, but could just as well be something like the ability to comprehend the item fluently and accurately when reading. Perhaps the only measure of size that does not depend on any particular conceptualization of depth would be the number of items for which a learner has at least some partial knowledge of *at least* one aspect of depth. However, this definition of size brings its own difficulties, as trying to determine the lowest threshold of partial knowledge is extremely problematic: Is it the ability to answer a meaning recognition test item, some intuition that the word has been seen before but without knowledge of its meaning (e.g., the Vocabulary Knowledge Scale; Paribakht & Wesche, 1997), or even an N400 response on an ERP apparatus (e.g., Osterhout, McLaughlin, Pitkänen, Frenck-Mestre, & Molinaro, 2006)? Defining the exact nature of partial knowledge is challenging for theorists and probably even more so for language assessors.

Another outcome is that virtually all aspects of vocabulary knowledge seem interrelated. This makes it difficult to discuss any particular conceptualization of depth in isolation. If we take Nation's (2001) listing of word knowledge aspects

(Table 1), each one will likely be related to the others in various ways, and studies have demonstrated correlations between a number of them (e.g., Schmitt & Meara, 1997: between derivative forms and word associations). This makes it difficult to theorize depth as anything but the combined interrelationships between word knowledge aspects. But then, each type of word knowledge also has characteristics like receptive versus productive mastery and fluency of use, adding to the complexity. It is difficult to see how all of these strands can be captured concurrently by any theoretical explanation, but I suspect that the most promising approach would revolve around lexical organization: The degree that any item is integrated into the rest of the mental lexicon would be considered its depth, with any link of any sort (e.g., synonym, derivative, collocation) to any other lexical item adding to the overall depth. Unfortunately, our understanding of lexical organization is not yet advanced enough to pursue this direction in a tangible way (see Fitzpatrick [2006] for the current state of an association-based approach).

This review also has implications for vocabulary assessment. The first follows from the discussion above about the size versus depth criterion. The most widely used vocabulary tests are size tests, and they typically describe their results as the number of words known, but do not define what this actually entails. Test users are left to interpret the scores as they wish, for example, words that learners can understand in reading or words that learners can use in their writing. (Nobody interprets the scores as simply words that learners can answer on a vocabulary test.) In the future, test developers need to explicitly state what correct answers on their tests entail and what degree of depth they represent. This is because different criteria of mastery will potentially lead to quite different size estimates.

This applies even at the form–meaning link level of mastery. Most size tests use items focusing on some variant of the form–meaning link, but Laufer and Goldstein (2004) and Laufer et al. (2004) clearly show that these have a hierarchy of difficulty. So, for instance, size tests based on meaning recognition item formats will likely produce higher size estimates than those based on form recall item formats. Testers thus need to consider which form–meaning level they wish to use, and explicitly state to the end user how this should guide their score interpretations.

One of the things that became clear when doing the research for this review is that few of the depth tests were validated to any great extent. There was typically little consideration for validity or reliability issues, and the number of target items was usually very small. The state of development of depth tests is

such that many of the conclusions in this article can only be tentative. Only with much more robust tests can some of the issues raised in this review be resolved. Thus development of better measures of the quality of lexical knowledge should be prioritized.

Read (2004) and Milton (2009) suggested that it may be time to dispense with the general notion of depth altogether and concentrate on more specific measures of the quality of vocabulary knowledge that are tuned more finely to specific research questions. While the notion of depth might remain a valuable tool to speak about vocabulary knowledge in general terms, Read and Milton's suggestion seems the logical step forward for the precision needed for research and assessment. The best recommendation seems to be to specify which type of post-initial learning is being targeted in future research and then discuss findings only in terms of that particular type of knowledge.

Final revised version accepted 14 April 2014

Notes

1 Although in most cases an earlier time of learning is likely to be related to greater depth, it is also possible that some words are inherently less difficult than other words regardless of when they have been learned. This may be because they have fewer depth features (fewer meanings, fewer grammatical irregularities, etc.) or they may be cognates of L1words. The relationship between depth and time of learning may be less straightforward in these cases.

2 Depth can also be conceptualized as how well lexical items can be employed in the four skills. It is clear that vocabulary size is a key facilitator of reading, listening, writing, and speaking skills (e.g., Alderson, 2005). However, there is little research comparing depth of individual lexical items and the ability to employ those items in the skills, and so this strand is not taken up in this review.

3 Depth of meaning knowledge can also be conceptualized as the complexity of a single meaning sense. For example, Salsbury, Crossley, and McNamara (2011) looked at the degree of concreteness, imagability, meaningfulness, and familiarity of words produced in spontaneous speech. Unfortunately for our purposes, these and other such studies have not included size measurements for comparison.

4 Learning multiple meanings for the same word form (*illuminate*: shed light on; *illuminate*: draw colorful illustrations in books) may not be much different than learning different words each with their own meaning. If this is true, it might be better to view knowledge of multiple meaning senses as a size measure instead of a depth measure.

5 DIALANG is a European project for the development of diagnostic language tests in 14 European languages. It offers separate tests for reading, writing, listening, grammatical structures, and vocabulary in each of the languages (Alderson, 2005).

6 The researchers note that the receptive association task may have been too easy for the learners in the two higher grade levels. This may have led to the nonsignificant correlations for those grades.

References

Akbarian, I. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System*, *38*, 391–401.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). Newark, DE: International Reading Association.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*, 245–261.

Bonk, W. J. (2001). Testing ESL learners' knowledge of collocations. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development* (pp. 113–142). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Chui, A. S. Y. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching*, *16*, 1–23.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*, 213–238.

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*, 573–605.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editor's introduction. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 1–32). Cambridge, UK: Cambridge University Press.

de Bot, K. (1992). A bilingual production model: Levelt's speaking model adapted. *Applied Linguistics*, *13*, 1–24.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use? In S. T. Gries & D. S. Divjak (Eds.), *Frequency effects in language learning and processing* (Vol. *1*, pp. 7–34). Berlin, Germany: Mouton de Gruyter.

Ellis, R. (2009). Editorial. *Language Teaching Research*, *13*, 333–335.

Fan, M. (2000). How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal*, *31*, 105–119.

Farvardin, M. T., & Koosha, M. (2011). The role of vocabulary knowledge in Iranian EFL students' reading comprehension performance: Breadth or depth? *Theory and Practice in Language Studies*, *1*, 1575–1580.

Fitzpatrick, T. (2006). Habits and rabbits: Word associations and the L2 lexicon. *EUROSLA Yearbook*, *6*, 121–145.

Greidanus, T., Bogaards, P., van der Linden, E., Nienhuis, L., & de Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 191–208). Amsterdam: John Benjamins.

Gyllstad, H. (2007). *Testing English collocations*. Lund, Sweden: Lund University, Media-Tryck.

Harrington, M. (2006). The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA Yearbook*, *6*, 147–168.

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, *37*, 614–626.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*, 303–317.

Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Vocabulary and writing in a first and second language* (pp. 22–66). Basingstoke, UK: Palgrave Macmillan.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development—a progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 29–56). Amsterdam: John Benjamins.

Huang, H-F. (2006). *Breadth and depth of English vocabulary knowledge: Which really matters in the academic reading performance of Chinese university students*? Unpublished master's thesis. Montreal, Canada: McGill University.

Kieffer, M. J., & Lesaux, N. K. (2008). The role of derivational morphological awareness in the reading comprehension of Spanish-speaking English language learners. *Reading and Writing*, *21*, 783–804.

Kieffer, M. J., & Lesaux, N. K. (2012a). Development of morphological awareness and vocabulary knowledge in Spanish-speaking language minority learners: A parallel process latent growth curve model. *Applied Psycholinguistics*, *33*, 23–54.

Kieffer, M. J., & Lesaux, N. K. (2012b). Knowledge of words, knowledge about words: Dimensions of vocabulary in first and second language learners in sixth grade. *Reading and Writing*, *25*, 347–373.

Koizumi, R. (2005). *Relationships between productive vocabulary knowledge and speaking performance of Japanese learners of English at the novice level*. Unpublished Ph.D. dissertation. University of Tsukuba, Japan.

Kroll, J. F., Michael, E., Tokowicz, N., & Dufour, R. (2002). The development of lexical fluency in a second language. *Second Language Research*, *18*, 137–171.

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook*, *5*, 223–250.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*, 202–226.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436.

Laufer, B., & Nation, P. (2001). Passive vocabulary size and speed of meaning recognition. *EUROSLA Yearbook*, *1*, 7–28.

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*, 365–391.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*, 647–672.

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. *EUROSLA Monographs Series*, *2*, 127–148.

Lindsay, S., & Gaskell, M. G. (2010). A complementary systems account of word learning in L1 and L2. *Language Learning*, *60*(Suppl. 2), 45–63.

McGavigan, P. (2009). *The acquisition of fixed idioms in Greek learners of English as a foreign language*. Unpublished Ph.D. dissertation. Swansea University, UK.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 109–121). Cambridge, UK: Cambridge University Press.

Meara, P. (n.d). *Lognostics Web site*. Retrieved January 10, 2013, from http://www.lognostics.co.uk/

Meara, P., & Wolter, B. (2004). V_LINKS: Beyond vocabulary depth. In D. Albrechtsen, K. Haastrup, & B. Henriksen (Eds.), *Angles on the English speaking world 4* (pp. 85–96). Copenhagen, Denmark: Museum Tusculanum Press.

Mehrpour, S., Razmjoo, S. A., & Kian, P. (2010). The relationship between depth and breadth of vocabulary knowledge and reading comprehension among Iranian EFL learners. *Journal of English Language Teaching and Learning*, *53*(222), 97–127.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 84–102). Cambridge, UK: Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *Canadian Modern Language Review*, *63*, 127–147.

Milton, J., & Riordan, O. (2006). Level and script effects in the phonological and orthographic vocabulary size of Arabic and Farsi speakers. In P. Davidson, C. Coombe, D. Lloyd, & D. Palfreyman (Eds.), *Teaching and learning vocabulary in another language* (pp. 122–133). United Arab Emirates: TESOL Arabia.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a Foreign Language. In R. Chacón-Beltrán, C.

Abello-Contesse, & M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Bristol, UK: Multilingual Matters.

Miralpeix, I., & Meara, P. (2014). Knowledge of the written word. In J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 30–44). Basingstoke, UK: Palgrave Macmillan.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, *28*, 291–304.

Nagy, W. E., Diakidoy, I. A. N., & Anderson, R. C. (1993). The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. *Journal of Reading Behavior*, *25*, 155–170.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82.

Nemati, A. (2010). Active and passive vocabulary knowledge: The effect of years of instruction. *Asian EFL Journal*, *12*, 30–46.

Nizonkiza, D. (2012). Quantifying controlled productive knowledge of collocations across proficiency and word frequency levels. *Studies in Second Language Learning and Teaching*, *2*, 67–92.

Noro, T. (2002). The roles of depth and breadth of vocabulary knowledge in reading comprehension in EFL. *ARELE*, *13*, 71–80.

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, *18*, 161–175.

Osterhout, L., McLaughlin, J., Pitkänen, I., Frenck-Mestre, C., & Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: A means for exploring the neurocognition of second language processing. *Language Learning*, *56*(Suppl. 1), 199–230.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 174–200). Cambridge, UK: Cambridge University Press.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring yes-no vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*, 489–509.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*, 282–307.

Qian, D. D. (2000). *Validating the role of depth of vocabulary knowledge in assessing reading for basic comprehension* [TOEFL 2000 Research Report]. Princeton, NJ: Educational Testing Service.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*, 513–536.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355–371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41–60). Mahwah, NJ: Erlbaum.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 209–227). Amsterdam: John Benjamins.

Read, J., & Shiotsu, T. (2010). *Investigating the yes/no vocabulary test: Input modality, context, and response time*. Presentation given at the Language Testing Research Colloquium. Cambridge, UK.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*, 77–89.

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research*, *27*, 343–360.

Schmitt, N. (1998). Tracking the incidental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*, 281–317.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*, 329–363.

Schmitt, N. (2010a). Key issues in teaching and learning vocabulary. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 28–40). Bristol, UK: Multilingual Matters.

Schmitt, N. (2010b). *Researching vocabulary: A vocabulary rresearch manual*. Basingstoke, UK: Palgrave Macmillan.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*, 17–36.

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*, 145–171.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

Segalowitz, N., Watson, V., & Segalowitz, S. (1995). Vocabulary skill: Single-case assessment of automaticity of word recognition in a timed lexical decision task. *Second Language Research*, *11*, 121–136.

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics*, *19*, 53–67.

Shimamoto, T. (2000). An analysis of receptive vocabulary knowledge: Depth versus breadth. *JABAET*, *4*, 69–80.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, *64*, 429–458.

Takala, S. (1984). *Evaluation of students' knowledge of English vocabulary in the Finnish comprehensive school* (Rep. No. 350). Jyväskylä, Finland: Institute of Educational Research.

Tschirner, E. (2004). Breadth of vocabulary and advanced English study: An empirical investigation. *Electronic Journal of Foreign Language Teaching*, *1*, 27–39.

Tseng, W-T. (2011). *Modeling vocabulary knowledge: A mixed modeling approach*. Paper presented at 2011 Language Testing Research Colloquium, June 23–25. Ann Arbor, MI: University of Michigan.

Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*, 357–400.

Van Zeeland, H. (2013). *Second language vocabulary knowledge in and from listening*. Unpublished doctoral dissertation. University of Nottingham, UK.

Verhallen, M., & Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. *Applied Linguistics*, *14*, 344–363.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*, 217–234.

Waring, R. (1998). *Receptive and productive vocabulary: Do we know what we are talking about*? PACSLRF Conference, March 26. Tokyo, Japan.

Waring, R. (1999). *Tasks for assessing second language receptive and productive vocabulary*. Unpublished Ph.D. dissertation. University of Wales, Swansea, UK. Retrieved from http://www.robwaring.org/papers/phd/title.html

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*, 33–52.

Webb, S. (2007a). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research*, *11*, 63–81.

Webb, S. (2007b). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*, 46–65.

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*, 79–95.

Zhang, D. (2012). Vocabulary and grammatical knowledge in second language reading comprehension: A structural equation modeling study. *Modern Language Journal*, *96*, 558–575.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Measuring Vocabulary Size
Appendix S2: Measuring Depth of Vocabulary Knowledge
Table S1: Correlations between Size and Depth (Lexical Organization) Measures