Stellenbosch Papers in Linguistics Plus, Vol. 43, 2014, 45-61
doi: 10.5774/43-0-169

The dimensional approach to vocabulary testing: What can we learn from past and present practices?

Déogratias Nizonkiza and Karien van den Berg (née Hattingh)

School of Languages, Potchefstroom Campus, North-West University, South Africa
Email: deo.nizonkiza@nwu.ac.za; karien.hattingh@nwu.ac.za

Abstract

Vocabulary constitutes an important component of language and its study has attracted the interest of second-language (L2) and foreign-language (FL) teachers and applied language researchers, booming in the 1990s (cf. for example Ellis 1992, Read 2000). Among other things, this interest has been characterised by the attention paid to testing learners' knowledge of vocabulary. The dimensional approach to vocabulary knowledge as proposed by Henriksen (1999), i.e. vocabulary size, depth, and receptive-productive knowledge/skills, has influenced test design for measuring L2/FL vocabulary acquisition. This article aims to describe the major vocabulary tests along the vocabulary dimensions and highlights what testing under this approach has contributed to the teaching of vocabulary. To this end, it reviews some major L2/FL vocabulary tests alongside the above dimensions, focusing on the pedagogical consequences that followed testing. The review shows that testing has not been an end in itself. The extensive investigation of vocabulary size has led to standardisation of methods, as well as insight into how to determine the amount of vocabulary needed at different learning stages. Furthermore, it has influenced the development of course materials for fostering vocabulary growth. However, testing depth and productive knowledge still lags behind. Despite progress made in this regard, scholars have not succeeded in measuring the two dimensions in a standardised manner, nor have they determined the extent of depth and productive knowledge associated with different learning stages. Given the importance of speaking and writing (i.e. productive use rather than mere comprehension), suggestions for future directions are discussed.

Keywords: vocabulary dimensions, vocabulary size, vocabulary depth, productive vocabulary, receptive vocabulary, testing vocabulary

1. Introduction

Vocabulary is an essential building block of language and, as such, it makes sense to measure learners' knowledge of it (Schmitt, Schmitt and Clapham 2001:55). Vocabulary, and the assessment thereof, has indeed moved from being peripheral to being regarded as an integral component of L2/FL proficiency (see Daller, Milton and Treffers-Daller 2007; Meara 2002;

Zareva, Schwanenflugel and Nikolova 2005, among others) and therefore tends to feature prominently in placement tests, for example.

Various researchers – such as Richards (1976), Nation (1990), Chapelle (1998), Meara (1996) and Henriksen (1999) – have proposed taxonomies to characterise vocabulary knowledge, demonstrating that this phenomenon is complex and multifaceted. Research on vocabulary has followed two different but related tracks (Zareva 2010), namely the "separate trait model" (Nation 1990, Richards 1976) and the "global trait model" (Chapelle 1998, Henriksen 1999, Meara 1996, Read 2000). While the global trait model – also known as the dimensional approach – suggests studying vocabulary at a more global level, for instance studying one or two aspects that may characterise the lexicon as a whole, the separate trait model suggests studying all the constituting elements of vocabulary knowledge separately. Though the lack of consensus as to a general definition of vocabulary competence remains, there currently seems to be consensus that vocabulary is investigated and tested under the dimensional approach that distinguishes between three dimensions: size, depth, and receptive-productive knowledge (see section 3.1).

That said, vocabulary size remains the most widely researched and tested dimension (cf. Daller et al. 2007; Ishii and Schmitt 2009; Milton 2009; Read 2007; Schmitt, Ng and Garras 2011). This is problematic because it means that vocabulary knowledge is being investigated and measured in terms of only one part of a whole, despite the apparent agreement that the construct comprises more than just that. Measuring size cannot suffice in describing vocabulary knowledge (Ishii and Schmitt 2009, Read 1993, Wesche and Paribakht 1996) and therefore poses problems for validity in terms of construct-underrepresentation. Schmitt et al. (2011:106) support this view in saying that "vocabulary knowledge can be conceptualised not only as the number of lexical items known (vocabulary size), but also as how well these items are mastered (vocabulary depth or quality)". In other words, it is not enough to determine the extent of a learner's vocabulary size; we also have to evaluate its depth.

With the above insights in mind, this paper reviews the current testing practices and their related pedagogical consequences, aiming to offer avenues for exploration and reflection on "where to from here", to borrow Granger and Meunier's (2008) phrase. The fundamental goal that guides the discussion is to understand the pedagogical consequences that followed testing vocabulary under the dimensional approach. To this end, the following questions will be addressed: 1) what insights have we gained from testing vocabulary under the dimensional approach?, and 2) to what extent has this new approach influenced the teaching of vocabulary? In answering these questions, we wish to establish whether there is indeed a need to broaden the focus of vocabulary testing beyond vocabulary size, or whether it is sufficient to continue focusing on just this one dimension of vocabulary teaching and assessment. We believe that greater emphasis should be placed on depth and productive vocabulary knowledge, as well as on methods for teaching and assessing them to provide at least a balanced perspective on vocabulary knowledge in which all dimensions are recognised. We demonstrate this by first taking a look at what has been done in vocabulary testing in recent decades, and then at what it has brought us, before making suggestions for future research ventures.

2. A brief history of vocabulary testing

The history of vocabulary testing allows us to track the changes across four periods (Read 1997). Starting before the 1970s, the traditional period, intuitive or subjective in nature, focused on essay writing and prose translations. Though rote learning of vocabulary lists was a typical teaching strategy at the time, its main function was to provide learners with resources for translation purposes. During the 1970s, the integrative stage entailed testing vocabulary by means of objective test items or multiple choice in standardised tests. Individual linguistic items such as words were not assessed beyond the limited context of the sentence. These tests had a strong lexical focus, though the focus was not on vocabulary as such (Spolsky 1995). Overall language proficiency was the main concern. The early 1980s saw the third stage, characterised as communicative, following an integrative approach by combining various sub-skills into one test battery (Madsen 1983). As such, vocabulary was tested as part of reading, writing and/or speaking. Thus, prior to the 1990s, vocabulary was a neglected component of language proficiency in as far as teaching and research are concerned.

In the late 1980s, Read and Nation (1986) drew attention to some of the prevailing issues in testing L2/FL vocabulary, such as estimating size, the criteria for knowing a word, distinguishing basic and specialised words and their sampling, construction and use of diagnostic tests, testing words in isolation, and the distinction between receptive and productive knowledge. This was the first venture in the debate on testing vocabulary in its own right, sparking renewed interest in its importance from applied linguistic researchers and L2/FL teachers in more recent years (Ellis 1992, Read 2000). This fourth period was mainly characterised by explorations into depth knowledge and the overall state of a learner's mental lexicon, aiming to test vocabulary in order to measure lexical competence per se and not any other skill or language competence in general. In other words, the focus shifted from testing words to measure language skills, to testing word knowledge, i.e. testing vocabulary in its own right. Researchers therefore ventured to validate vocabulary size tests (e.g. Nation's (1990) Vocabulary Levels Test), develop deep word knowledge tests (e.g. Read's (1993) Word Associates Test), and develop productive tests (e.g. Laufer and Nation's (1995) Lexical Frequency Profile; Laufer and Nation's (1999) Productive Vocabulary Levels Test). Vocabulary tests can now be associated with three vocabulary dimensions, namely size, depth, and receptive-productive knowledge (Henriksen 1999 and Read 2000, 2007).

The following section defines the notion of 'vocabulary dimensions' and considers the recent focus on and methods of assessing vocabulary. Different tests, namely vocabulary size tests, depth tests, and productive tests, are discussed briefly to illustrate current practices. We would like to draw attention to the lack of focus on productive vocabulary especially, despite the variety revealed when discussing today's practice. The pedagogical consequences of these tests determine the weight or emphasis placed on testing the different dimensions of vocabulary knowledge. Following the discussion of the tests, these consequences are evaluated in order to aid our understanding of the value that the current vocabulary assessment approach has brought. This evaluation emphasises the imbalance between the dimensions and the need for broadening our focus to include productive vocabulary more prominently in assessment and teaching. Concluding remarks address some issues for future consideration based on the discussion.

3. Current practices

3.1 The dimensional approach to vocabulary

This discussion follows Henriksen's (1999) classification of vocabulary knowledge dimensions (1999) into three components, namely partial to precise knowledge (related to vocabulary size), shallow to deep knowledge, and receptive to productive knowledge.

Vocabulary size is also known as the breadth dimension of vocabulary knowledge, referring to the number of words one knows (Henriksen 1999, Meara 1996). Conceptual meaning constitutes its main concern (Schmitt 1994). For a long time, vocabulary size has received considerable attention in research and it continues to be popular in various methods of assessment, as discussed below.

Depth knowledge relates to how well a word is known (Henriksen 1999). Also referred to as quality of lexical knowledge, it concerns aspects of deep word knowledge, i.e. paradigmatic, syntagmatic and analytic relations of a word. Read (1993:359) defines these aspects of depth knowledge as follows:

Paradigmatic: Two words are related because they are synonyms or at least similar in meaning, perhaps with one being more general than the other. Examples are: *edit – revise*; *team – group*. It should be noted, however, that the paradigmatic aspect is not defined in consistent terms. While Read (1993) restricts paradigmatic knowledge to synonyms, Read (2004) extends its meaning to include synonyms as well as superordinates. Furthermore, Read's (1993) definition seems to be narrow as it does not accommodate some aspects, such as antonyms, that are important word associations and that could be part of paradigmatic knowledge.

Syntagmatic: The two words are collocates that often occur together in a sentence. Examples: *edit – film*; *team – scientists*.

Analytic: The associate represents one aspect or component of the meaning of the stimulus word and is likely to form part of its dictionary definition. Examples: *edit – publishing*; *team – together*.

L2/FL researchers commonly agree on this approach (cf. Greidanus, Bogaards, Van der Linden, Nienhuis and Dewolf 2004, and Zareva et al. 2005, among others), although they still diverge regarding what exactly depth entails (Zareva et al. 2005). The dimensional approach assumes connectivity of words in the mental lexicon and considers word knowledge development as a network building process (Haastrup and Henriksen 2000, Meara and Fitzpatrick 2000, Meara 2009, Meara and Wolter 2004, Read 2004). Under this approach, deeper word knowledge could be the result of the number of words linked to a specific word, and the strength of such links (Meara and Fitzpatrick 2000, Meara and Wolter 2004). For Meara (2009), depth knowledge depends on the interaction between individual words, which results in a better organisation of the mental network as proficiency grows. However, he claims that this organisation shows general knowledge rather than details as to how much each word is known. Furthermore, while the view on the interaction between individual words expressed by Meara (2009) could be equated with the number of words a particular word is linked to, he seems to disregard the strength of the links elsewhere, such as in Meara and Fitzpatrick (2000) and Meara and Wolter (2004).

The third dimension of vocabulary knowledge is receptive-productive. It is referred to as a control dimension (Henriksen 1999) and constitutes the continuum where word knowledge progresses from being receptively understood to being productively used. Receptive and productive knowledge is often equated with passive and active knowledge, or thought of in terms of whether it is used for listening and reading, or speaking and writing. However, passive skills often require some degree of active anticipation (Milton 2009:13). Laufer and Goldstein (2004:405) note that there is "no consensus as to whether this distinction is dichotomous or whether it constitutes a continuum". It is important to point out here that we recognise this distinction to be one of degree rather than absolute – either the one or the other – but for the purpose of this discussion, the distinction is considered in dichotomous terms.

Until now, vocabulary tests have focused on measuring either breadth or depth knowledge, mainly because it is so difficult to measure both at the same time. At the vocabulary size level, a firm connection has been established between having "good vocabulary knowledge" and "good comprehension" (Shen 2008:135), but the kind of vocabulary knowledge referred to is receptive rather than productive. Receptive knowledge is generally understood to aid comprehension more than it does use, while productive knowledge plays more toward use. Use is a problematic aspect, more so than comprehension, for L2/FL learners when they have to write essays and term papers.

Receptive vocabulary size can give some indication of the size of productive knowledge (Webb 2008), and so it may be tempting to rely on it alone. We would like to advise against depending on this approach because the relationship between receptive and productive lexicons is not straightforward, though they are thought to be inter-related and an increase in one is generally considered to imply an increase in the other (Milton 2009). Estimates of proportions between the two vary widely, illustrating the point. Waring (1997) indicates that the proportion between the two is around 50%, i.e. productive knowledge could be half of what receptive knowledge is in proportion. Milton (2009), however, notes that productive knowledge can be anywhere between 50% and 80% of receptive knowledge. We believe that with such considerable variation, one can hardly be sure of drawing valid conclusions in relation to productive knowledge levels based on values indicating vocabulary size level.

Furthermore, from a methodological point of view, teaching vocabulary productively may be more important for the learners as it is likely to result in more positive outcomes. Nation and Chung (2009:546) identify four major effects of teaching multiword units in particular (stressing that they have their receptive equivalents): "students' grammatical accuracy, as well as their fluency will improve; utterances will become more native-like; and learners will be able to communicate early on in their acquisition process". In keeping with the idea of teaching words productively, Nizonkiza (2012), as well as Nizonkiza and Van de Poel (in press), for instance, suggest adapting McCarthy and O'Dell's (2005) collocation web model, which, as they put it, could facilitate retention as it allows learners to create and recreate the word webs in their minds. This suggestion is also supported by Handl (2009).

A legitimate question worth raising then is whether or not this aspect of vocabulary knowledge (i.e. depth knowledge) and especially productive knowledge, does not deserve more attention than it has received. We particularly have in mind multi-word units that have proven to characterise maturity in writing (Gledhill 2000, Paquot 2008) and which help achieve native-like fluency (Pawley and Syder 1983, Wray 2002), but are also very problematic for L2/FL learners (Nesselhauf 2005, Laufer and Waldman 2011). If the answer is a definite ‘yes’, then why has this dimension not been explored in more depth? Furthermore, why have tests not become more focused on productive knowledge? To answer these questions, the following sections briefly consider current vocabulary assessment practices in terms of testing vocabulary size and depth and productive knowledge.

3.2 Testing vocabulary size knowledge

Vocabulary size is measured only receptively, with the most popular tests being the Vocabulary Levels Test (VLT), the Vocabulary Size Test (VST) version based on the VLT, and the Eurocentres Vocabulary Size Test (EVST). These three tests are widely used and well-documented, and there is evidence available regarding their validity as assessment instruments for their intended purpose (Shen 2008:136). Test formats are typically yes/no-items (EVST), matching (VLT), and multiple choice (VST), with word selection often guided by frequency measures (Read 2007). These tests gained popularity because they are often administered electronically, making them easy to design, score and administer (Nation 2001, Read 2000, Schmitt 1994), while offering the possibility of testing a large number of words – particularly the EVST (Schmitt 1994) – with instant results (Nation 2001).

Despite the popularity of these tests, they have been criticised. Read (1997, 2000) criticises the VLT test on the grounds that it measures a low level of word knowledge because the ability to associate a word with its definition or phrase with the same meaning does not reveal the extent to which the word is known. Laufer and Paribakht (1998) note that the yes/no format of the EVST does not allow for distinction between items for which learners have full or partial knowledge. Read (2007) considers vocabulary size as a superficial indication of the extent to which a word is known, suggesting that knowing a word entails knowing the words with which it is associated. Daller et al. (2007:7) support this view, stating that “lexical breadth is intended in essence, to define the number of words a learner knows regardless of how well he or she knows them”. Furthermore, these tests allow for guessing; scores are adjusted based on non-words in the case of EVST[1], the results of which are based on self-report, which may lead to substantial under- or overestimation of participants’ vocabulary size (cf. Beglar 2010, Read 2000, Schmitt 1994, Schmitt et al. 2011). Thus, reliability and validity are threatened (Meara 1996).

In essence, the type of knowledge demonstrated by these tests is questionable in terms of validity evidence, and results may be of little value for communicative teaching practice. Despite the popularity of vocabulary size tests, it seems that many challenges remain. It is therefore worthwhile considering the potential benefits that assessment practices focusing on depth and productive knowledge hold and support.

3.3 Testing vocabulary depth knowledge

Depth tests have been designed following one of two main approaches: a developmental approach, e.g. the Vocabulary Knowledge Scale (VKS; Wesche and Paribakht 1996), or the dimensional approach, e.g. the Word Associates Test (WAT; Read 1993). The VKS measures

[1] Test-takers are required to demonstrate if they know the word or not by pressing the ‘yes’ or ‘no’ key upon reading the word on the computer screen. In order to adjust the scores of test-takers who might overrate their knowledge, the list contains non-words, i.e. one non-word item for every two real words.

vocabulary knowledge on a scale of stages of word acquisition. It is a self-report test which was developed as a vocabulary acquisition construct with the conviction that knowing a word is a gradual process which can be placed on a continuum from *unknown* to *known* with stages of knowing in-between. It presents a list of words to participants who are required to rate their knowledge of these words on the following scale:

I. I don't remember having seen this word before.
II. I have seen this word before, but I don't know what it means.
III. I have seen this word before, and I think it means _______. (synonym or translation).
IV. I know this word. It means _______. (synonym or translation).
V. I can use this word in a sentence: _______. (Write a sentence.) (If you do this section, please also do Section IV.)[2]

The WAT focuses on the aspects of deep word knowledge as distinguished earlier. It consists of an item known as a stimulus word (SW) which is followed by a set of eight words (original version) or six words (later versions). Half of the words, referred to as associates, have a relationship at the paradigmatic, syntagmatic (collocations), or analytic levels with the SW; the other half, referred to as "non-associates", does not. The test-takers' role is to identify the associates. The test has known a relative proliferation and constitutes a useful tool for research on vocabulary knowledge (Greidanus et al. 2004, Read 2007). Originally designed in English for specific purposes and L2 contexts, the test has been extended to and adapted for basic vocabulary and L1 studies (Read 2004). Tests such as Gyllstad's (2007) COLLEX (collocating lexis) and COLLMATCH (collocate matching) are examples of a different approach where vocabulary depth is measured in terms of a constituent element, such as collocations.

Both the VKS and WAT tests have been found to be reliable and valid measures of depth, and to correlate with overall proficiency. Akbarian (2010) also found correlations between depth and breadth measures of the WAT and the VLT. The VKS correlates with the general proficiency measure Test Of English as a Foreign Language (TOEFL; Zareva et al. 2005), while the WAT distinguishes between learners at different learning stages (Nizonkiza 2011, Schoonen and Verhallen 2008). Furthermore, validation studies proved the WAT to be an excellent tool for evaluating the extent to which learners know words and their associates (Schmitt et al. 2011, Schoonen and Verhallen 2008), covering both meaning and collocations (Schmitt 2010).

The VKS, however, does not seem to represent the learning stages (i.e. the process or sequence of learning words) and the estimate of vocabulary knowledge as claimed by the test designers (Henriksen 1999, Meara 1996, Read 2000). The test can also be criticised for relying on unverified self-report and the scales are internally inconsistent (Read 1993, 1998). The validity of the WAT may be threatened by learners guessing which words they associate with the stimulus word (Greidanus et al. 2004; Read 1993, 1998), which may result in overestimating learners' knowledge (Schmitt et al. 2011). Batty (2012) also expresses concern that the aspects measured may belong to different dimensions of vocabulary knowledge. Despite the criticisms, it should be noted that the WAT is considered more valuable for research purposes and in turn provides insights for classroom application.

[2] Scale found in Schmitt (2010).

Following its proven importance as a component of overall proficiency, collocation has received considerable attention recently (cf. Gyllstad 2007, Pawley and Syder 1983, Wray 2002) and is the only one tested independently among the different components of depth. Gyllstad's (2007) COLLEX and COLLMATCH are tests mostly referred to in the literature, and are both tests of collocations at the receptive level. They are quick to administer and can be used as proficiency indicators (Gyllstad 2007, 2009), with scores from these tests correlating significantly with overall depth and size as well. Even though these tests have been validated, the only type of collocation tested consists of verb-noun combinations, which raises the question of generalisability of the results as acknowledged by Gyllstad (2007, 2009) himself.

3.4 Tests of productive knowledge

Productive knowledge consists of controlled productive knowledge on the one hand and free productive knowledge on the other.

Controlled productive knowledge is measured by means of the Productive Vocabulary Levels Test (PVLT). It was developed by Laufer and Nation (1999), based on the VLT, in order to test the controlled productive ability, which refers to:

> […] the ability to use a word when compelled to do so by a teacher or researcher, whether in an unconstrained context such as a sentence writing task, or in a constrained context such as a fill in task where a sentence context is provided and the missing target word has to be supplied.
>
> (Laufer and Nation 1999:37)

The test distinguishes between high and low proficiency learners and is very practical as it is fast and easy to administer, mark and interpret (Laufer and Nation 1999). However, as some test items require more word knowledge and more use of contextual information, it may be questionable what the test measures as a whole (Meara and Fitzpatrick 2000:21; cf. also Read 2000, Schmitt 2010).

Free productive knowledge has been measured mainly through lexical richness and association tasks. Laufer and Nation's (1995) Lexical Frequency Profile (LFP) is one of the best known frequency-based free productive knowledge tests of vocabulary, while Meara and Fitzpatrick's (2000) Lex30 is a popular test that uses association tasks (Schmitt 2010). Tests used for this purpose also tend to be easy to administer, with some evidence available concerning validity and reliability (see Laufer 2005 for LFP, and Fitzpatrick and Clenton 2010 and Walters 2012 for Lex30), providing support to various degrees. Furthermore, both tests are believed to discriminate between proficiency levels. However, the LFP does not distinguish between well-known and partially known words, and in particular how words combine in lexical phrases (i.e. collocations), which is a characteristic feature of good writing (Gledhill 2000, Schmitt 1994). Furthermore, Meara and Fitzpatrick (2000) find that a valid measure of free productive vocabulary should require a huge amount of text, which is difficult to get even from native speakers as their texts often consist of a small set of highly frequent words. Therefore, the test does not help one to assess the overall size of participants' productive vocabulary knowledge. That said, assessing overall productive vocabulary size is no small task, and the LFP can provide comparative results which are relatively valid in certain contexts.

4. Practical implications of current test practices

The consequences that arose from the testing practices discussed above constitute the fundamental question that has guided this review. These consequences are considered here in order to determine whether we are in fact testing what matters most.

The most widely tested dimension or type of vocabulary knowledge, namely size, is mainly measured receptively. Using these tests has helped researchers and L2/FL practitioners gain insight into vocabulary growth, leading to important pedagogical consequences. Originally, vocabulary size tests were mainly designed to evaluate students' reading comprehension (Laufer 1998, Read 2000, 2007), diagnose their vocabulary problems, and examine their achievements in vocabulary learning (Schmitt 1994). For instance, commenting on the VLT, Nation (2001:21-22) states that "the test was designed to quickly let teachers find out whether learners need to be working on high-frequency or low-frequency words, and roughly how much work needs to be done on these words".

In time, research evidence confirmed that testing vocabulary size in effect tests overall proficiency, i.e. there is a correlation between vocabulary size and overall level of proficiency (cf. Beglar 2010, Meara 1996, Meara and Buxton 1987, Meara and Jones 1988, Nation 1990, Nation and Beglar 2007). This allowed researchers to determine how much vocabulary is needed at different learning stages. Even though figures still vary, researchers tend to agree that a minimal threshold of 5,000 word families are needed for understanding lectures at the undergraduate level, and an optimal one of 8,000 word families for reading authentic texts without external support (Laufer and Ravenhorst-Kalovski 2010, Nation 2006, Schmitt 2008, Schmitt and Schmitt 2012). As a result, these tests were standardised. Nation's (2009) 4000 Essential English Words also resulted from this type of practice and research, and offers a valuable tool to inform teachers.

Receptive tests of vocabulary depth knowledge commonly seem to indicate that depth knowledge grows alongside overall proficiency. However, none of them has been standardised. COLLEX and COLLMATCH, each testing one aspect of depth, give no suggestions as to how exactly to use the tests to quantify how much is known at which learning stage. The tests can only be used for research purposes and not for pedagogical applications. The same holds for the VKS which does not seem to point to any direct pedagogical consequences either.

The WAT, which proved to be popular among researchers, is the most widely used depth test (Ishii and Schmitt 2009, Schmitt 2010, Schmitt et al. 2011) and has pointed to possible pedagogical consequences. For instance, Schoonen and Verhallen (2008) adapted Read's (1993) WAT for primary school pupils (grades 3-6) in order to empirically test the feasibility of assessing their deep word knowledge. They used a word web layout where children were required to connect the stimulus words to their associations by means of lines. Their study confirmed the word web layout as an efficient and reliable method to assess children's depth knowledge and thus could be used for teaching depth knowledge even at low proficiency levels.

However, with validation studies conducted in relation to depth tests pointing to both weaknesses and strengths, the research and pedagogical consequences as discussed only constitute basic groundwork for further research rather than for wider pedagogical applications

(Schmitt et al. 2011). This confirms Milton's (2009) observation that testing depth knowledge has raised more questions than it has answered.

As far as the productive vocabulary tests, i.e. PVLT (controlled) and LFP and Lex30 (free) are concerned, these tests have been validated and proven to distinguish between levels of proficiency. However, like depth tests, none of them has been standardised and we are therefore confronted with the problem of quantifying the size of learners' knowledge in relation to their learning stages. The lack of standardised tests means that data are not consistent enough to provide reliable results (Daller et al. 2007). Nation (2007:42) also emphasises this, stating that "measures of language use currently cannot tell the size of a learner's vocabulary, productive or otherwise, but they indicate how skilful the learner is in drawing on vocabulary knowledge to perform productive tasks". The major complication is methodological, in terms of capturing the construct of productive knowledge in the best way possible. Without a clear construct, it is difficult to estimate the size of vocabulary that learners can use productively, as is possible for breadth knowledge (Milton 2009).

5. Discussion

Despite the inconsistencies and challenges mentioned in relation to testing vocabulary depth and productive knowledge, we can say that the different instruments that test them point to the same observation – that growth in vocabulary relates to overall proficiency, echoing findings at the vocabulary size level. This insight can contribute towards taking this debate a step further. It implies that the more proficient learners are likely to know well and use more words productively, suggesting that once they are standardised, the tests could be used to quantify the degree to which words are known (well) and used, and at which level. This could be an important step towards determining how much to bring to the learners' attention depending on the learning stage, the same way this is done in relation to the vocabulary size dimension.

The prolonged emphasis on vocabulary size assessment is understandable, considering the advancement that this research has brought as well as the correlation established between this dimension and overall language proficiency. However, mastering vocabulary in use requires mastery of other dimensions as well. Following the dimensions approach to vocabulary testing, there appears to be an imbalance as far as focus is concerned. Based on this discussion, we recommend standardising vocabulary depth and productive tests, and using them to inform pedagogical practice. This is of course no small venture, but provides an opportunity for future research. The weight given to the different dimensions has greatly influenced the teaching implications following testing practices. Given the importance of language at the productive level, this dimension should be given due attention in the hope of arriving at applications similar to those achieved with regard to vocabulary breadth. This will require collective endeavour, as well as addressing issues that still raise controversy among scholars. Rather than saying that too much emphasis has been placed on vocabulary size testing, this paper contributes to the discussion by demonstrating that there is both a need for broadening our focus, as well as room to do so, to include other dimensions of vocabulary knowledge. It would be irresponsible to rest assured, based on the fact that a correlation has been established between one dimension of vocabulary knowledge and overall proficiency, and assume that we need not do more.

Research needs to build on what has been achieved so far in order to cast light on future directions to take and thereby contribute toward insights about all the dimensions of vocabulary knowledge. To succeed in this venture, the problematic conceptualisation of vocabulary depth, as well as the distinction between receptive and productive knowledge, needs attention. We agree with Vermeer (2001), as well as Milton (2009:150), with the latter noting that "without a clear construct, it is impossible to create a test that can accurately measure and quantify a quality whatever that quality is".

Daller et al. (2007:7) focus on the concept of 'depth' and identify knowledge of collocations, connotations and preferred associations as areas in particular need of clear and simple characterisation in a way that makes quantification and testing possible. Attempting to include all the components that depth consists of in one test battery would be a challenge indeed, but is not the fundamental issue here. Rather, identifying a component or several components which may best represent the others is what matters more, and testing only this/these then seems warranted.

The body of research tends to support the notion of redefining vocabulary depth knowledge as a whole, as current definitions seem to relate disconnected elements (Daller et al. 2007; Batty 2012; Read 2000, 2004), questioning the traditional conceptualisation of depth. Milton (2009) and Batty (2012) – according to whom depth may consist of disconnected elements, such as collocations, associational knowledge, constraints on use, polysemy, etc. – support this view. For them, it is hard to believe that all these elements belong to a single dimension and the question that arises then is which of these should remain together and which ones should be set apart.

Schmitt (2010:241) supports an alternative, suggesting that a combined approach may be the best option. This would imply using both measures of the quantity as well as the quality of lexical knowledge, either in a contained study or in consecutive ones "whose results can be linked for greater understanding". Both receptive and productive knowledge would have to be tested and the results interpreted in an integrated manner. Ishii and Schmitt (2009) can be given credit for attempting the first step in this direction, using four tests: a vocabulary size test (Japanese VLT), a test of multiple meaning senses for words, a test of derivative word forms, and a test of lexical choice between near synonyms. They suggested interpreting the results in an integrated manner, i.e. setting vocabulary size as the baseline and drawing up tables showing students' performance on depth tests, comparing scores among peers with similar vocabulary size**.** Developing productive variants of the widely used receptive tests, or developing receptive tests based on productive ones can also help in this regard. This would offer possibilities for obtaining comparable results and drawing relevant conclusions, which may provide new insights for teaching vocabulary.

As discussed above, the construct of productive knowledge may not be as clear as we may think and should also be reconsidered. For instance, none of the tests of free productive knowledge considers collocations. In both LFP and Lex30, credit is given to students for using infrequent words, which makes sense in a way because a more proficient user is likely to know more words that are infrequent. However, this does not give any indication as to whether or not this makes them good users of the language insofar as collocation use characterises proficiency and native-like fluency in both comprehension and production (cf. Boers and Lindstromberg 2009, Pawley and Syder 1983, Wray 2002 among others), collocations not being necessarily infrequent. Nation (2001) advises against including collocations in tests of vocabulary breadth, unless they are core idioms, in which case "there is no obvious relationship between parts and the whole". Multi-word units, such as *I would like a__(burger and chips)___*, otherwise comprise words that share a relationship between the meaning of the parts and that of the whole. Based on such a line of argument, we believe that it makes sense to include collocations (in this sense) in the construct for vocabulary depth and productive knowledge.

Finally, such reconsideration of vocabulary and its assessment may affect how it is used in other contexts as well. The shift in focus to assessing vocabulary in itself, mentioned earlier, may be considered a turning point, but it remains useful and valuable as an indicator of general language proficiency. Examples in the South African context include the Test of Academic Literacy Levels (TALL; cf. e.g. Van Dyk and Weideman 2004) used at the North-West, Pretoria, and Stellenbosch universities, and that administered by the Alternative Admissions Research Project (AARP; cf. e.g. Cliff and Yeld 2006, Cliff and Hanslo 2009) at the University of Cape Town. Both of these tests aim to establish the preparedness of students for the academic environment as far as interacting with texts and engaging in critical discourse for academic purposes are concerned. Vocabulary in context is listed as one contributing factor to the academic literacy construct. Its assessment manifests in more ways than one. The construct (academic potential) and aim of these tests are different to those considered in this discussion (vocabulary knowledge). However, future development and reconsideration of vocabulary assessment, as discussed here, may also impact on the development of such tests. A first step in this direction could be establishing a relationship between TALL and/or AARP scores and both vocabulary size and depth as measured by either of the tests described in the present study.

6. Conclusion

Vocabulary testing has informed pedagogical practice for some time, relying mostly on receptive vocabulary size. We have argued for a broader consideration of vocabulary dimensions under the current approach. This study is valuable in that it demonstrates the importance of further investigating the depth dimension in terms of productive lexical knowledge. Also, it emphasises the potential usefulness of such endeavours in demonstrating the way forward in vocabulary assessment and in offering potentially valuable insight and pedagogical implications.

The review of prominent vocabulary tests provided here reveals a strong emphasis on receptive knowledge and little on productive knowledge, despite the fact that this aspect poses more serious challenges to L2/FL learners in becoming communicatively competent. We underscored the pedagogical consequences that arose from testing practices and conclude that much that is positive has come from it. Also, testing vocabulary under the dimensional approach has not been an end in itself. However, these consequences are mainly based on testing receptive lexical breadth only. There is room for broadening our scope and learning more because, like vocabulary size, vocabulary depth and productive knowledge seem to correlate with overall proficiency. We suggest that it is not enough to depend on assessing vocabulary size alone and that the correlation established should not be considered sufficient for validity of tests per se. Rather, we recommend that, as a starting point, researchers should expand the focus to include depth knowledge alongside vocabulary size. Furthermore, productive knowledge in addition to receptive vocabulary knowledge should be a main concern, particularly to inform pedagogical practice in order to enhance learners' vocabulary knowledge across dimensions. Seeing that vocabulary depth and productive tests have been validated, they now need to be standardised in order to be useful for research and pedagogical purposes. This should, however, be preceded by addressing the issue of the lack of a clear construct of what exactly depth and productive knowledge comprise. As a first step in this direction, we could consider adopting Schmitt et al.'s (2011:106) suggestion that productive knowledge is indeed part of deep word knowledge.

References

Akbarian, I. 2010. The relationship between vocabulary size and depth for ESP/EAP learners. *System* 38(3): 391-401.

Batty, A. 2012. Identifying dimensions of vocabulary knowledge in the Word Associates Test. *Vocabulary Learning and Instruction* 1(1): 70-77.

Beglar, D. 2010. A Rasch-based validation of the Vocabulary Size Test. *Language Testing* 27(1): 101-118.

Boers, F. and S. Lindstromberg. 2009. *Optimizing a lexical approach to instructed language acquisition*. New York: Palgrave Macmillan.

Chapelle, C. 1998. Construct definition and validity inquiry in SLA research. In L.F. Bachman and A.D. Cohen (eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press. pp. 32-70.

Cliff, A.F. and M. Hanslo. 2009. The design and use of 'alternate' assessments of academic literacy as selection mechanisms in higher education. *Southern African Linguistics and Applied Language Studies* 27(3): 265-276.

Cliff, A.F. and N. Yeld. 2006. Test domains and constructs: Academic literacy. In H. Griesel (ed.) *Access and entry level benchmarks: The National Benchmark Tests Project.* Pretoria: Higher Education South Africa. pp. 19-27.

Daller, H., J. Milton and J. Treffers-Daller (eds.) 2007. *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

Ellis, R. 1992. *Second language acquisition and language pedagogy*. Clarendon: Multilingual Matters.

Fitzpatrick, T. and J. Clenton. 2010. The challenge of validation: Assessing the performance of a test of productive ability. *Language Testing* 27(4): 538-555.

Gledhill, C.J. 2000. *Collocations in science writing*. Tubingen: Gunter Narr Verlag.

Granger, S. and F. Meunier (eds.) 2008. *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins Publishing Company.

Greidanus, T., P. Bogaards, E. Van der Linden, L. Nienhuis and T. Dewolf. 2004. The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards and B. Laufer (eds.) *Vocabulary in second language*. Amsterdam: John Benjamins Publishing Company. pp. 191-208.

Gyllstad, H. 2007. Testing English Collocations. Unpublished PhD dissertation, Lund University.

Gyllstad, H. 2009. Designing and evaluating tests of receptive collocation knowledge: COLLEX and COLLMATCH. In A. Barfield and H. Gyllstad (eds.) *Researching collocations in another language*. New York: Palgrave Macmillan. pp. 153-170.

Haastrup, K. and B. Henriksen. 2000. Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics* 10(2): 221-240.

Handl, S. 2009. Towards collocational webs for presenting collocations in learners' dictionaries. In A. Barfield and H. Gyllstad (eds.) *Researching collocations in another language*. New York: Palgrave Macmillan. pp. 69-85.

Henriksen, B. 1999. Three dimensions of vocabulary development. *Studies in Second Language Acquisition* 21(2): 303-317.

Ishii, T. and N. Schmitt. 2009. Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal* 40(1): 5-22.

Laufer, B. 1998. The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics* 19(2): 255-271.

Laufer, B. 2005. Lexical frequency profiles: From a Monte Carlo analysis to the real world. A response to Meara 2005. *Applied Linguistics* 26(4): 582-588.

Laufer, B. and Z. Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning* 54(3): 399-436.

Laufer, B. and I.S.P. Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16(3): 307-322.

Laufer, B. and I.S.P. Nation. 1999. A vocabulary size test of controlled productive ability. *Language Testing* 16(1): 33-51.

Laufer, B. and S. Paribakht. 1998. The relationship between passive and active vocabularies: Effects of language learning contexts. *Language Learning* 48(3): 365-391.

Laufer, B. and G.C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* 22(1): 15-30.

Laufer, B. and T. Waldman. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61(2): 647-672.

Madsen, P. 1983. *Techniques in testing.* Oxford: Oxford University Press.

McCarthy, M. and F. O'Dell. 2005. *English collocations in use*. Cambridge: Cambridge University Press.

Meara, P. 1996. The dimensions of lexical competence. In G. Brown, K. Malmkjaer and J. Williams (eds.) *Competence and performance in language learning*. Cambridge: Cambridge University Press. pp. 35-53.

Meara, P. 2002. The rediscovery of vocabulary. *Second Language Research* 18(4): 393-407.

Meara, P. 2009. *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins.

Meara, P. and B. Buxton. 1987. An alternative to multiple choice vocabulary tests. *Language Testing* 4(2): 142-154.

Meara, P. and T. Fitzpatrick. 2000. Lex30: An improved method of assessing productive vocabulary in an L2. *System* 28(2000): 19-30.

Meara, P. and G. Jones. 1988. Vocabulary size as a placement indicator. In P. Grunwell (ed.) *Applied linguistics in society*. London: Centre for Information and Language Teaching and Research. pp. 80-87.

Meara, P. and B. Wolter. 2004. V_Links: Beyond vocabulary depth. In D. Albrechtsen, K. Haastrup and B. Henriksen (eds.) *Angles on the English-speaking world 4*. Copenhagen: Museum Tusculanum Press. pp. 85-96.

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

Nation, I.S.P. 1990. *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

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

Nation, I.S.P. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review* 63(1): 59-82.

Nation, I.S.P. 2007. Fundamental issues in modeling and assessing vocabulary knowledge. In H. Daller, J. Milton and J. Treffers-Daller (eds.) *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press. pp. 35-43.

Nation, I.S.P. 2009. *4000 Essential English words, books 1-6*. Seoul: Compass Publishing.

Nation, I.S.P. and D. Beglar. 2007. A vocabulary size test. *The Language Teacher* 31(7): 9-13.

Nation, I.S.P. and T. Chung. 2009. Teaching and testing vocabulary. In M.H. Long and C.J. Doughty (eds.) *The handbook of language teaching*. Malden: Wiley-Blackwell. pp. 543-559.

Nesselhauf, N. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins Publishing Company.

Nizonkiza, D. 2011. The relationship between lexical competence, collocational competence, and second language proficiency. *English Text Construction* 4(1): 113-146.

Nizonkiza, D. 2012. The Relationship Between Lexical Competence, Collocational Competence, and L2 Proficiency. Unpublished PhD thesis, University of Antwerp.

Nizonkiza, D. and K. Van de Poel. (In press). Teachability of collocations: The role of word frequency counts. *Southern African Linguistics and Applied Language Studies*.

Paquot, M. 2008. Exemplification in learner writing: A cross-linguistic perspective. In S. Granger and F. Meunier (eds.) *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins Publishing Company. pp. 101-119.

Pawley, A. and F.H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards and R.W. Schmidt (eds.) *Language and communication*. London: Longman. pp. 191-227.

Read, J. 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10(3): 355-371.

Read, J. 1997. Assessing vocabulary in second language. In C. Clapham and D. Corson (eds.) *Encyclopaedia of language and education (volume 7): Language testing and assessment*. Dordrecht: Kluwer Academic Publishers. pp. 90-98.

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

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

Read, J. 2004. Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards and L. Laufer (eds.) *Vocabulary in second language*. Amsterdam: John Benjamins Publishing Company. pp. 209-227.

Read, J. 2007. Second language vocabulary assessment: Current issues and new directions. *International Journal of English Studies* 7(2): 105-125.

Read, J. and I.S.P. Nation. 1986. *Some issues in the testing of vocabulary knowledge.* Paper presented at LT + 25: A language testing symposium in honour of John B. Carroll and Robert Lado, 11-13 May 1986, Quiryat Anavim, Israel.

Richards, J.C. 1976. The role of vocabulary teaching. *TESOL Quarterly* 10(1): 77-90.

Schmitt, N. 1994. Vocabulary testing: Questions for test development with six examples of tests of vocabulary size and depth. *Thai TESOL Bulletin* 6(2): 9-16.

Schmitt, N. 2008. Review article: Instructed second language vocabulary learning. *Language Teaching Research* 12(3): 329-363.

Schmitt, N. 2010. *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave Macmillan.

Schmitt, N. and D. Schmitt. 2012. *A reassessment of frequency and vocabulary size in L2 vocabulary teaching. Plenary speech.* Available online: http://journals.cambridge.org (Accessed 15 February 2012).

Schmitt, N., C.J.W. Ng and J. Garras. 2011. The word associates format: Validation evidence. *Language Testing* 28(1): 105-126.

Schmitt, N., D. Schmitt and C. Clapham. 2001. Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing* 18(1): 55-88.

Schoonen, R. and M. Verhallen. 2008. The assessment of deep word knowledge in young first and second language learners. *Language Testing* 25(2): 211-236.

Shen, Z. 2008. The roles of depth and breadth of vocabulary knowledge in EFL reading performance. *Asian Social Science* 4(12): 135-137.

Spolsky, B. 1995. *Measured words*. Oxford: Oxford University Press.

Van Dyk, T. and A. Weideman. 2004. Switching constructs: On the selection of an appropriate blueprint for academic literacy. *Journal for Language Teaching* 38(1): 1-13.

Vermeer, A. 2001. Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency input. *Applied Psycholinguistics* 22(2): 217-234.

Walters, J.D. 2012. Aspects of validity of a test of productive vocabulary: Lex30. *Language Assessment Quarterly* 9(2): 172-185.

Waring, R. 1997. Graded and extensive reading. Questions and answers. *The Language Teacher* 27(5): 9-12.

Webb, S. 2008. Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition* 30(1): 79-95.

Wesche, M. and S. Paribakht. 1996. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review* 53(1): 13-40.

Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zareva, A. 2010. Multicompetence and L2 users' associate links: Being unlike nativelike. *International Journal of Applied Linguistics* 20(1): 2-22.

Zareva, A., P. Schwanenflugel and Y. Nikolova. 2005. Relationship between lexical competence and language proficiency. *Studies in Second Language* 27(4): 567-595.