Looking at L2 vocabulary knowledge dimensions from an assessment perspective – challenges and potential solutions

Henrik Gyllstad
Lund University

The heightened interest in L2 vocabulary over the last two or three decades has brought with it a number of suggestions of how vocabulary knowledge should be modelled. From a testing and assessment perspective, this paper takes a closer look at some of these suggestions and attempts to tease out how terms like model, dimension and construct are used to describe different aspects of vocabulary knowledge, and how the terms relate to each other. Next, the two widely assumed dimensions of vocabulary breadth and depth are investigated in terms of their viability for testing purposes. The paper identifies several challenges in this regard, among others the questionable assumption that multi-word units like collocations naturally belong in the depth dimension, and problems that follow from the complex and often ill-defined nature of the depth dimension. Suggestions for remedies are provided.

1. Introduction

Ever since Meara (1980) pointed out the then Cinderella-like status of vocabulary some three decades ago, the field of foreign and second language vocabulary (L2)[1] has seen a formidable explosion in terms of activity and the number of studies published. The dramatic yet welcome increase in research on vocabulary over the last 30 years has brought with it an increase also with regard to terminology. A striking example of the plethora of terms that may exist for a single concept, arguably some having more or less central meanings than others, can be seen in Wray's (2002) account of terms used to describe aspects of formulaicity, presented as Figure 1. As Wray points out, even though there are clear cases of conceptual duplication across the terms used, there are also cases of terms shared across different fields that do not refer to the same thing. Whether

---

1 Henceforth, the abbreviation L2 will be used to denote both a second and a foreign language.

this proliferation of terms relates to a parallel proliferation of constructs is a crucial issue for research on language testing and assessment.

Figure 1. Terms used to describe aspects of formulaicity (taken from Wray, 2002: 9).

amalgams – automatic – chunks – clichés – co-ordinate constructions – collocations – complex lexemes – composites – conventionalized forms – F[ixed] E[xpressions] including I[dioms] – fixed expressions – formulaic language – formulaic speech – formulas/formulae – fossilized forms – frozen metaphors – frozen phrases – gambits – gestalt – holistic – holophrases – idiomatic – idioms – irregular – lexical simplex – lexical(ized) phrases – lexicalized sentence stems – listemes – multiword items/units – multiword lexical phenomena – noncompositional – noncomputational – nonproductive – nonpropositional – petrifications – phrasemes – praxons – preassembled speech – precoded conventionalized routines – prefabricated routines and patterns – ready-made expressions – ready-made utterances – recurring utterances – rote – routine formulae – schemata – semipreconstructed phrases that constitute single choices – sentence builders – set phrases – stable and familiar expressions with specialized subsenses – stereotyped phrases – stereotypes – stock utterances – synthetic – unanalyzed chunks of speech – unanalyzed multiword chunks - units

Cronbach and Mehl define a construct as "some postulated attribute of people, assumed to be reflected in test performance" (1955, p. 283). A person may at any time possess such an attribute, either fully or to some degree, or not possess it. A complementary definition of the term construct is supplied by Chapelle, who states that "a construct is a meaningful interpretation of observed behavior" (1998, p. 33). If applied to the field of vocabulary assessment, then a test-taker's scores on a vocabulary test constitute the observed behaviour that is to be interpreted meaningfully, and by extension, the scores are assumed to indirectly reflect some kind of mental ability or knowledge, in this case knowledge about words. A construct is thus a form of knowledge or an ability that can be observed and/or measured, and as such it is of course essential to the scientific study of any kind, since it enables a scientific community to label knowledge and/or abilities, to define clearly what they are, to potentially break them down into several interrelated sub-abilities, and to relate them to other constructs. However, contention is no doubt part and parcel of any thriving academic discipline, and although the evolution of a construct often involves competing definitions and perspectives, it becomes problematic in the long run if constructs are not clearly and properly defined, and if some degree of homogeneity is not reached. In the worst case it could hinder a further understanding of the field.

In the remainder of this paper, I will first take a look at some of the central terminology used for describing knowledge and abilities in the field of L2 vocabulary acquisition, primarily from a testing and assessment perspective. I will discuss how the terminology is used, identify potential problems, and suggest remedies to these when possible. I will then discuss the origins and applications of the influential and widely-used dimensions of vocabulary breadth and depth, particularly in relation to some of the challenges that researchers face when using these for assessment purposes. In doing this, I will also propose remedies to overcome some of the more persistent challenges.

2. Central terminology used in research on L2 vocabulary acquisition and assessment – models, dimensions and constructs

As was pointed out in the previous section, the heightened interest in L2 vocabulary has entailed an increase in the number of constructs that have been proposed and used. Recent examples connected to vocabulary size tests, i.e. tests of the number of words in a language for which a learner has at least a basic form-meaning knowledge, are written receptive vocabulary size (Meara & Buxton, 1987), controlled productive vocabulary size (Laufer & Nation, 1999) and aural receptive vocabulary size (Milton & Hopkins, 2006). These three examples have a parent construct ('vocabulary size') as a common denominator, but are more specific by adding terms that narrow the construct down even further, e.g. 'receptive', 'productive', 'aural', and 'written'. This is obviously a good thing, as the added specificity makes it clearer what kind of knowledge is targeted. Interestingly, even though the notion of construct is arguably very central when describing vocabulary knowledge and its assessment, the term itself is not always used specifically in the literature. Instead, the term *dimension* often appears when L2 vocabulary researchers discuss acquisition and assessment matters. Here are some examples of 'dimensions' proposed in the literature on L2 vocabulary acquisition.

- Henriksen (1999), in describing a model of lexical development:
  a) partial to precise knowledge, b) depth of knowledge, and c) receptive to productive use ability.
- Meara (2005), in describing a model of lexical competence/skill:
  a) vocabulary size, b) vocabulary organization, and c) vocabulary accessibility.
- Daller et al. (2007), in describing a learner's vocabulary knowledge in "lexical space":
  a) lexical breadth, b) lexical depth, and c) lexical fluency.

The first thing to note about the three proposals is that they all assume three dimensions, perhaps either true to a geometrical definition of space assuming length, breadth and depth, or simply giving support to the proverb that says that all good things come in threes. As to the first dimension (a) of the three models, it could be seen to deal with the same underlying process, namely the building of a repository of vocabulary items. What is characteristic of this dimension is that it has more to do with quantity than quality. Learners are shown to know x number of words, but this knowledge is minimally seen as a basic form-meaning mapping. Meara's (2005) *vocabulary size* and Daller et al.'s (2007) *lexical breadth* are very similar in this sense, whereas my understanding of Henriksen's (1999) *partial to precise knowledge* dimension is that she refers to the development of individual word knowledge, and that she emphasizes that the acquisition process is not an all-or-nothing activity. There are differences among authors as regards the second dimension (b), too. Daller et al. see *lexical depth* largely from a word knowledge framework perspective. Based on Nation's (2001) (see Table 2) descriptive approach to what aspects are involved in knowing a word, depth is seen as those aspects that go beyond the basic form-meaning mapping, e.g. concepts and referents, associations, collocations and constraints on use. Meara's second dimension is called *vocabulary organisation*, and it is conceptually different to that of Daller et al. Meara envisages vocabulary organisation as the structured, lexical network that makes up a learner's mental lexicon. The focus here is on the links between words in this network and on how, from a more holistic perspective, they can inform us about the network as a whole. The fundamental difference between these first two approaches will be further discussed later on in this chapter. Henriksen's dimension, called *depth of knowledge*, may sound closer to that of Daller et al., but in fact she discusses it more in terms of network building in line with Meara's conception of vocabulary organisation. When it comes to the third dimension (c), the versions proposed by Daller et al. and Meara are conceptually close. The former call it *lexical fluency* and state that it is intended to define "how readily and automatically a learner is able to use the words they know and the information they have on the use of these words" (Daller et al., 2007, p. 8). This may involve the speed and accuracy with which word forms can be recognised receptively or retrieved for expressing targeted meanings when speaking or writing (productive vocabulary). Meara's version, called *vocabulary accessibility*, is said to have to do with "how easily you can manipulate the words you know" (Meara, 2005, p. 271), which is likely to imply both receptive and productive aspects, even though Meara's development of tests of this dimension has focused largely on receptive recognition skills. Henriksen's version is called *receptive to productive use ability*, which is argued to be a continuum, describing "levels of access or use ability" (1999, p. 314). Thus, there is a clear conceptual overlap between the three dif-

ferent versions, but it is also evident that the authors describe these dimensions in different ways and propose different ways to operationalise them.

The use of the term *dimension* raises the question as to what the relation is between this term and the term *construct*. It seems that in some cases in the literature construct and dimension are used more or less synonymously, whereas in other cases they are used hierarchically in a hyponymic relation, with dimension as a hypernym and construct as its hyponym. There are also cases of the converse relation, for example in Henriksen (1999), where construct is the superordinate (hypernym) term and dimension the subordinate (hyponym). Another term that is used in this context is *model*. Hierarchically, a model can be seen as a set of propositions that clarify how different constructs relate to each other. Meara (2005) talks about his three dimensions as being part of a model of vocabulary skills, while Henriksen (1999) proposes a model of lexical competence. Daller et al. (2007) do not use the term model when discussing their multi-dimensional space, but it is interesting to note that the name of the volume in which their text is published is called *Modelling and Assessing Vocabulary Knowledge.* The terms *model, dimension* and *construct* might be seen as co-existing at different hierarchical levels, albeit with some restrictions. Thus, I would like to propose that a model may consist of several dimensions, which in turn may comprise various constructs. A dimension can also be a construct, so long as type of knowledge or ability referred to is clearly defined – and by extension – measurable through some sort of test or assessment. If it is not, then the use of dimension rather than construct is more suitable. Furthermore, a dimension can consist of several constructs, just as a construct in principle can be divided into two or more 'sub-constructs'. An example of this would be the dimension of vocabulary size, which can also be said to be a construct. In order to accommodate more detailed descriptions of vocabulary knowledge, e.g. *aural receptive vocabulary size* (Milton & Hopkins, 2006) or *controlled productive vocabulary size* (Laufer & Nation, 1999), it is possible to treat these as two sub-constructs within the construct (and dimension) of vocabulary size. From an assessment perspective, researchers ought to define constructs with precision. One way of doing this is by following Bachman's (1990, p. 40-45) three-stage analysis:

   a. the construct needs to be defined theoretically;
   b. the construct needs to be defined operationally;
   c. procedures must be established for the quantification of observations.

The theoretical definition (a) is a specification of the relevant characteristics of the ability we intend to measure, and its distinction from other similar constructs. If there are several subcomponents to a construct, then the interrela-

tions between these must be specified. When it comes to the operational definition of the construct (b), this process involves attempts to make the construct observable. To a great extent, the theoretical definition will govern what options are available. For example, the theoretical definition of the construct 'listening comprehension' suggests an operationalisation as a task in which information must be decoded aurally in some fashion. With respect to the third stage (c), our measurement should be quantified on a scale. If applied to vocabulary depth (see the section below), with many subcomponents argued to be part of this construct, it is then very important to try to pin down how they relate to each other. To the best of my knowledge, this has not been done. On a theoretical level, Schmitt (2010b) has intuitively hypothesized how the different word knowledge aspects of Nation's (2001) framework (see Table 2) relate to each other developmentally, but these hypotheses need to be empirically tested.

Having discussed the use of terminology in L2 vocabulary knowledge modelling, I will now turn to discussing the viability of two of the most influential dimensions in the field, vocabulary breadth and vocabulary depth, in order to see if they can be treated as constructs.

3. Vocabulary breadth and vocabulary depth: two influential dimensions and some inherent issues and challenges

3.1. The definitions of vocabulary breadth and depth

Two of the most prominent dimensions used in L2 vocabulary research are 'vocabulary breadth' and 'vocabulary depth'. As was made clear in the previous section, competing terms exist (e.g. 'size' instead of 'breadth'), but the breadth and depth terminology can be traced back to a paper by Anderson and Freebody (1981), where breadth and depth are referred to as "aspects". For now, this term will be used as in the authors' original wording. I will later come back to how it relates to dimension and construct. Anderson and Freebody use the two aspects in a discussion about the role of vocabulary knowledge in reading comprehension, and they state clearly at the beginning of their paper that what they are interested in is "knowledge of word meanings" (1981, p. 77). This is how they define the two aspects (Anderson & Freebody, 1981, pp. 92-93)

> It is useful to distinguish between two aspects of an individual's vocabulary knowledge. The first may be called "breadth" of knowledge, by which we mean the number of words for which the person knows at least some of the significant aspects of meaning. … [There] is a second dimension of vocabulary knowledge, namely the quality or "depth" of understanding. We shall

assume that, for most purposes, a person has a sufficiently deep understanding of a word if it conveys to him or her all of the distinctions that would be understood by an ordinary adult under normal circumstances.

These two aspects of vocabulary knowledge have indeed been influential and widely used. Not surprisingly, though, they have also been the subject of some criticism.

Firstly, as was pointed out by Read in his account of the term depth (2004), Anderson and Freebody's definitions leave us with a number of unclear terms. For example, in relation to "depth", it is not clear what is meant by "distinctions". Also, it raises the question as to what "an ordinary adult" is and what "normal circumstances" are. My own reading of Anderson and Freebody (1981) is that what they mean by distinctions when outlining the depth aspect is in effect <u>meaning</u> distinctions. This is arguably clear in the passage following the one where breadth and depth are initially defined (Anderson & Freebody, 1981, p. 93):

> […] the meaning a young child has for a word is likely to be more global, less differentiated than that of an older person. With increasing age, the child makes more and more of the adult distinctions.

The interpretation that the term "distinctions" refers to meaning distinctions is furthermore strengthened by a later passage, where a study by Gentner (1975) is reported. In this study, children were asked to act out, with the help of dolls, transactions based on given directions involving verbs like *buy, sell, spend, give* and *take.* The children acted out *buy* and *sell* as if they were essentially *take* and *give,* thus disregarding the money transfer element that is inherent in the meaning of the former verbs. It could thus be argued that what Anderson and Freebody originally meant by vocabulary depth was the repertoire of meanings and subtle sense distinctions that a word can convey. However, in Read's (2004)

Table 1. The application of the term depth in L2 vocabulary acquisition research (based on Read, 2004: 211-212).

1. Precision of meaning	(the difference between having a limited, vague idea of what a word means and having a much more elaborated and specific knowledge of its meaning)
2. Comprehensive word knowledge	(knowledge of a word, not only its semantic features but also orthographic, phonological, morphological, syntactic, collocational and pragmatic characteristics)
3. Network knowledge	(the incorporation of the word into a lexical network in the mental lexicon, together with the ability to link it to – and distinguish it from – related words)

account of how the term depth had been operationalised up to the early 2000s, there are three applications of the term. The additional two are seen as points 2 and 3 in Table 1.

It is clear from the above descriptions that it is only the first application called 'Precision of meaning' that is consistent with how Anderson and Freebody (1981) originally defined depth of word knowledge. The second operationalisation outlined by Read is that of comprehensive word knowledge. Here, as the name implies, a sizeable number of aspects are involved in knowing a word. One of the most recent and influential descriptions of such aspects is that of Nation (2001), shown here as Table 2. It is beyond the scope of this paper to go into a detailed description of Nation's framework, but one thing is relevant. Typically, the aspects called 'spoken' and 'written' under the heading 'Form', together with 'form and meaning' under the heading 'Meaning' are seen as breadth aspects, whereas the remaining ones in the table are usually considered depth aspects. This means that knowledge of word parts, word associations, grammatical functions and collocations are usually considered depth of word knowledge aspects, an assumption I will return to later in this chapter.

Table 2. Description of "what is involved in knowing a word", from Nation (2001: 27).

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 recognisable in this word?
		P	What word parts are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	concepts 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?

R = receptive knowledge, P = productive knowledge

The third operationalisation according to Read is network knowledge. The assumption behind network knowledge is that newly learned words are stored in a network of already known items. One of the main proponents of this interpretation is Paul Meara and associates (see e.g. Meara & Wolter, 2004; Wolter, 2005; Meara, 2006), but Henriksen subscribes to this view as well, as we saw earlier in this chapter.

3.2. Critical views of breadth and depth

A point of criticism that has been levelled at the use of breadth and depth has to do with their being fundamentally different constructs, and thus not really comparable. For example, Meara and Wolter (2004) have argued that vocabulary breadth, or vocabulary size, as they prefer to call it, is a construct that is a measure of a learner's entire vocabulary, since scores on a particular number of words are extrapolated to give an indication of an overall size score, given that the selection of test items is valid. As such, vocabulary size is not a characteristic of individual words. Vocabulary depth, on the other hand, is typically seen as a characteristic of individual words, where extrapolation is not possible, or at least very difficult.

Even though vocabulary breadth (or vocabulary size) is not without its problems as a construct, it has desirable measurement characteristics. With its ratio scale, assessment scores start at zero and range up to thousands, even tens of thousands for advanced learners of a language. One of the inherent problems with vocabulary size, however, is linked to the old question of what a word is. In order to try to come up with estimates of someone's vocabulary size, it is important to decide and state clearly if the unit of counting in word frequency lists is word form, lemma or word family. Of course, except perhaps for beginner learners, it normally makes sense to work with lemmatized lists. Once learners have understood the inflectional system of a language, especially for receptive knowledge, they can fairly straightforwardly link different forms of a verb (*play, plays, playing*) or a noun (*house, houses*) together, at least when it comes to non-morphologically complex languages like English. Another approach to word frequency lists is to use the concept of word families. Word families are normally defined as "a headword, its inflected forms, and its closely related derived forms" (Nation 2001, p. 8). Even though it makes some sense to use word families from a learning burden point of view, it is questionable to assume that once a member of a word family is known, all the other members will be known too, perhaps without ever having seen some of them. Bogaards (2001) has rightly warned against this assumption (see also Cobb & Horst, 2004; Schmitt & Zimmerman, 2002), lamenting the fact that no empirical evidence has been presented to properly support its validity. Bogaards uses this example in his criticism, arguing that the following uses of the form *level*, as a consequence, should then not be problematic to L2 learners in terms of understanding (2001, p. 322-323):

(a) a high <u>level</u> of radiation
(b) on a <u>level</u> with
(c) a <u>level</u> teaspoon
(d) have a <u>level</u> head
(e) to <u>level</u> a part of the town
(f) death is a <u>leveler</u>
(g) a <u>leveling</u> staff
(h) an <u>unlevel</u> surface

It is clear that the polysemy and the derivational patterns of the form *level*, as illustrated in (a) – (h) above, may still pose a problem to learners of English, just like Bogaards implies. However, it should be noted that it might be the case that receptive understanding is still easier than productive knowledge in this regard. Thus, understanding the concept of *an unlevel surface,* in the sense that the prefix *un-* negates the adjective *level* in the context of surfaces, is arguably more straightforward than being able to produce a derivative word form expressing that same meaning. For example, how should a learner know which prefix to use for negating *level* from the range of alternatives, for example *in-, dis-, non-* or *un-*?

3.3. Two specific challenges to the viability of breadth and depth

In addition to the points of criticism accounted for above, there are two further challenges to the constructs of vocabulary breadth/size and depth, namely:

a) the ubiquity of lexical items larger than one single orthographic word,
b) the multi-faceted nature of the depth construct.

The first challenge is the ubiquity of lexical items larger than one single orthographic word. Below, a number of examples of such items, all part of the vocabulary of English, are juxtaposed with a single orthographic word.

break	single orthographic word
break up	phrasal verb
lunch break	compound noun
break a record	collocation
break a leg	idiom

The first three examples should be fairly uncontroversial, but the difference between a collocation and an idiom is perhaps not so straightforward. In this analysis, the sequence *break a leg* is an idiom since it is not possible to understand its meaning by adding up the meanings of the individual components,

i.e. *break a leg* is non-compositional. However, this sequence can also evoke a more literal reading, to denote the fracture of a bone that someone might suffer in an accident. In this reading, the sequence would be what Howarth (1996) refers to as a free combination. Likewise, the sequence *break a record* has two possible readings, too. One of them denotes the more literal process of someone destroying a vinyl record, as played on turntables. This would then also be called a free combination. However, the other reading would be called a collocation, since one of the components (words) of the sequence is used in a figurative, de-lexical, or technical sense, in this case the verb *break*. It stands to reason that lexical items like these are very important for second language learning. The point here is that some of them behave like single orthographic words – certainly the compound noun, but arguably the phrasal verb and perhaps the collocation and idiom as well. If this is the case, then they should be made part of the vocabulary inventory and included in a frequency list where single orthographic words would reside jointly with multi-word items (see Cobb, this volume and Henriksen, this volume). As a case in point, Shin and Nation (2008) have presented an analysis, based on the 10-million-word spoken part of the British National Corpus (BNC), in which as many as 84 collocations occurred with such high frequency that they would make it into the top 1,000 single word types of the spoken corpus. It should be noted here that Shin and Nation's use of the term collocation mainly resides in one of two traditions of collocation research, called the frequency-based tradition, the other being the phraseological tradition (see Nesselhauf, 2004; Gyllstad, 2007; Barfield & Gyllstad, 2009 for accounts of these). The 84 collocations of the first frequency band include for example *you know*, *I think*, and *come back*. Furthermore, as many as 224 collocations would make it into the second 1,000 word type band of the corpus (see Table 3). As argued by Shin and Nation (2008), a large number of collocations would qualify for inclusion in the most frequent single word bands, if no distinction was made between single words and collocations. This argument seriously challenges the construct of vocabulary size.

Table 3. The number of collocations that would potentially qualify into single word frequency bands of English (table taken from Shin & Nation, 2008: 345).

Collocations	84	224 (308)*	259 (567)*	324 (891)*	3807 (4698)*
Single word frequency bands	1st 1000	2nd 1000	3rd 1000	4th 1000	5th 1000

* The number in brackets shows the cumulative number of collocations.

If we accept the assumption that lexical items such as collocations are part of everyone's vocabulary, then we need to start thinking of ways of incorporating lexical items larger than single words into measures of vocabulary size. The reason why this has not yet been done is probably because it is fraught with all sorts of problems. It is very likely that the vocabulary size construct based on single orthographic words will maintain its validity for years to come because of its desirable measurement characteristics. However, attempts at creating measures of vocabulary size where the nature of word usage – as illustrated by Shin and Nation's study – is addressed should be well on their way (see e.g. Martinez & Schmitt, 2012, and chapter by Cobb, this volume).

Another consequence of this discussion is that it is not clear whether collocations and collocation knowledge should reside in the vocabulary depth construct. For many researchers who follow Nation's (2001) descriptive framework of word knowledge (see Table 2), aspects except for basic form and meaning knowledge are typically treated as depth components (see e.g. Read, 2000; Jiang, 2004; Milton, 2009; Schmitt, 2000, 2010a). In my own work on developing English collocation tests (Gyllstad, 2007, 2009), I have been reluctant to call my two test formats – COLLEX and COLLMATCH – depth tests. Both test formats are receptive recognition measures of verb + noun collocations such as *pay a visit*, *do justice* and *keep a diary*. The reason for my reluctance is that I have not seen any convincing arguments yet for why they should be measures of depth. True, if one subscribes to the idea that any test that measures either form knowledge or form-meaning knowledge of single words is a size test, and everything else is a depth test, then it follows that collocation tests would be depth tests. However, I think this is an over-simplification.

This is also clearly connected to the second major challenge to the dichotomy breadth/depth: the multi-faceted nature of the depth construct, as it is conventionally used. Typically, the following aspects of word knowledge are listed under the heading depth, in its comprehensive word knowledge interpretation:

- meaning knowledge beyond the most frequent,
  dictionary-based meaning of a word
- word associations
- collocations
- word parts
- grammatical functions

These aspects of depth are quite disparate, which makes the definition of depth as a single construct and its subsequent operationalisation very difficult. As Milton (2009) rightly points out, depth has not been sufficiently and unambiguously defined (Milton, 2009, p. 150):

> The difficulties in measuring qualities, such as depth, start with the definitions of this quality. We lack clear, comprehensive and unambiguous definitions to work with and this challenges the validity of any test that might fall within this area. […] Without a clear construct, it is impossible to create a test that can accurately measure a quality whatever that quality is.

I have two additional points to make here. First of all, the coining of depth as a dimension has been valuable in pushing the thinking and theorizing in the field forward. However, it only makes sense to call it a dimension; as a construct, it is arguably far too vague and elusive. Secondly, one important approach to ascertaining the viability of a construct is through empirical investigation, and the most straightforward way of doing this is through correlation studies. A considerable number of studies have indeed been carried out to investigate the relation between breadth and depth (e.g. Qian, 1999; Nurweni & Read, 1999; Vermeer, 2001; Meara & Wolter, 2004; Wolter, 2005; Gyllstad, 2007). Qian (1999) used the Vocabulary Levels Test (VLT) (Nation 2001) as a size measure and found correlations between scores on that test with scores on the Word Associates Test (WAT) (Read, 1993, 1998) as a depth measure at $r = .82$, based on data from 74 L1 Korean and L1 Chinese ESL college and university students, predominately 18-27 year-olds. Nurweni and Read (1999) administered both a receptive vocabulary size measure and a WAT format depth measure to 350 L1 Indonesian ESL first-year university students, and they observed a correlation of $r = .62$ for the whole group. In a subsequent analysis, in which the 350 students were subdivided according to scores on a general proficiency exam, they observed a correlation of $r = .81$ for high level students (10% of the whole group); $r = .43$ for mid level students (42% of the whole group); and $r = .18$ for low level students (48% of the whole group). Vermeer (2001), testing 50 L1 and L2 Dutch kindergarten 5-year-olds, arrived at correlations ranging between $r = .70$ and .83 between a receptive vocabulary size measure and an association task depth measure. Meara and Wolter (2004) found a modest level of correlation between scores on a test of overall vocabulary size and scores on a vocabulary depth test ($r = < .3$), based on data from 147 Japanese learners of English. This depth test, called V_Links, is argued to be a test of lexical organisation, following the lexical network interpretation of depth (Read, 2004). The result was taken as support for the view that size and organisation are "more-or-less independent features of L2 lexicons" (Meara & Wolter, 2004, p. 93). Wolter (2005), putting different versions of V_Links to the test, found similarly low, or even inverse (though not significant), correlations with vocabulary size. Wolter concludes that there is evidence to suggest that vocabulary organisation, as measured by V_Links (versions 2.0 and 4.0), and vocabulary size may develop orthogonally (2005, p. 208).

On balance then, except for the studies by Meara and Wolter, breadth and depth seem to correlate highly with each other, which raises questions about their viability as independent constructs. Based on his own investigations of breadth and depth, Vermeer concluded that (2001, p. 222):

> Breadth and depth are often considered opposites. It is a moot point whether this opposition is justified. Another assumption is that a deeper knowledge of words is the consequence of knowing more words, or that, conversely, the more words someone knows, the finer the networks and the deeper the word knowledge.

Vermeer's caveat is thus that one should not assume *a priori* that breadth and depth are poles.

In order to illustrate in detail some of the challenges implied by using size and depth empirically, I will briefly account for a study (taken from Gyllstad, 2007) which aimed at finding validation support for two tests of collocation, the aforementioned COLLEX and COLLMATCH tests. The purpose was to see whether the collocation tests gravitated more towards vocabulary size or vocabulary depth when correlated with tests widely assumed to be size and depth tests, respectively. Scores from 24 Swedish learners of English on five different tests were gathered. The learners ranged from upper secondary school students to third term university students. The five tests used are shown in Table 4. The analysis yielded very high correlations between the test scores from vocabulary size (VLT) and vocabulary depth (WAT) at $r = .93$. The collocation tests (COLLEX, COLLMATCH) correlated at $r = .90$ with vocabulary size (VLT) and at $r = .85-.90$ with the vocabulary depth measure (WAT).

Table 4. Tests used in a validation study investigating how collocation knowledge relates to the vocabulary size and depth constructs (based on Gyllstad, 2007).

Test	Brief description	Source
COLLEX	A 50-item test of receptive collocation knowledge	Gyllstad (2007)
COLLMATCH	A 100-item test of receptive collocation knowledge	Gyllstad (2007)
Vocabulary Levels Test (VLT)	Version 1, 150 items (vocabulary size)	Nation (2001); Schmitt (2000)
Word Associates Test (WAT)	A 320-item test (vocabulary depth)	Read (1998)
CAE Reading Comprehension Test	43 items	Cambridge ESOL Examination

The question is, what does all this tell us? The collocation tests correlated highly with vocabulary size and almost equally highly with vocabulary depth. At the same time, the size and depth measures in turn correlated highly with one another. A common way of interpreting high correlations is to assume that the variables that are involved are closely related or even the same thing. From a testing perspective, Norbert Schmitt (personal communication) has argued for the fact that every size test is in fact also a depth test. What he seems to mean by this is that for any given word in a size test, test-takers must have some sort of depth of word knowledge of that word in order to fulfill the test task. This presupposes, of course, a view of depth where word knowledge starts with a rather incomplete and partial level of knowledge, for example mere form recognition or very tentative and uncertain meaning knowledge. Most researchers, however, assume that basic form-meaning knowledge is part of the vocabulary breadth/size knowledge construct, and that depth is what comes beyond this basic knowledge.

An analysis that could shed light on the potential difference between the assumed constructs is multiple linear regression (see Bachman, 2004). It would for example be possible to try to estimate how much of the variation in a set of reading comprehension scores can be explained by vocabulary size scores. Then, as a second step, the variable of vocabulary depth would be entered into the regression model in order to ascertain whether the percentage of explained variance would increase. If that is the case, then vocabulary depth could be argued to bring an added, unique contribution to the variance in reading comprehension scores. As a case in point, Qian (1999) found that his measure of depth of vocabulary knowledge added a further 11% to the prediction of reading comprehension scores, over and above the prediction afforded by vocabulary size. A final remark that needs to be made here, though, is that we must look critically at the test instruments themselves. For example, in my own study (Gyllstad, 2007) and several of the studies reported above, including that of Qian (1999), a version of the Word Associates Test (WAT) (Read, 1993, 1998) was used. Some of the words featuring in the WAT are fairly low-frequency items, and vocabulary size is therefore suspected to have a considerable influence on test-takers' performance. A closer look at some of the words featured in the specific WAT test version used in Qian (1999) and Gyllstad (2007) confirms this. For example, target words like *ample*, *synthetic* (both 6K), and *fertile* (7K), together with associate words like *cautious* (5K) and *plentiful* (8K) are clearly not high-frequency words. This confounds the two variables and arguably explains at least part of the observed high correlations between vocabulary size and vocabulary depth scores.

4. Concluding remarks

In this chapter, I have discussed the terminology used in modelling vocabulary knowledge, especially in relation to assessment purposes. In particular, the uses and referents of terms like *model*, *dimension* and *construct* have been addressed. Although a certain degree of terminological variation is bound to exist in all scientific disciplines, rigour and consensus are equally desirable. I have proposed that a distinction be made between dimension and construct, and that constructs must be defined clearly following procedures suggested by e.g. Bachman (1990). Furthermore, by taking a closer look at the two influential dimensions of vocabulary breadth and depth, I have argued that vocabulary depth has been valuable in furthering the thinking in the field, but its ill-defined, cover-all nature makes it inappropriate as a construct to be used in assessment procedures. I have also highlighted some of the inherent problems of using breadth and depth in vocabulary assessment, such as the ubiquitous existence of multi-word units and the question of their potential inclusion in the breadth dimension.

References

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.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Barfield, A., & Gyllstad, H. (2009). Introduction: Researching second language collocation knowledge and development. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language - Multiple interpretations* (pp. 1-18). Basingstoke: Palgrave Macmillan.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition*, *23*(3), 321–343.

Cobb, T., & Horst, M. (2004). Is there room for an academic word list in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 15-38). Amsterdam: Benjamins.

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

Cronbach, L.J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). Editors' introduction: Conventions, terminology and an overview of the book. In H. Daller, J. Milton, & J. Treffers-Daller (Eds), *Modelling and assessing vocabulary knowledge* (pp. 1-32). Cambridge: Cambridge University Press.

Gentner, D. (1975). Evidence for the psychological reality of semantic components: The verbs of possession. In D. A. Norman, D.E. Rumelhart & the LNR research group (Eds), *Explorations in cognition* (pp. 211-246). San Francisco: Freeman.

Gyllstad, H. (2007). *Testing English collocations: Developing tests for use with advanced Swedish learners*. PhD Thesis. Lund: Lund University. Available on-line at: http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=599011&fileOId=2172422

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

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

Howarth, P. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Lexicographica Series Maior 75. Tübingen: Max Niemeyer.

Jiang, N. (2004). Semantic transfer and development in adult L2 vocabulary acquisition. In P. Bogaards & B. Laufer (eds.), *Vocabulary in a second language* (pp. 101-126). Amsterdam: Benjamins.

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

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied linguistics*, *33*(3), 299-320.

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching and Linguistics; Abstracts 13*, 221-246.

Meara, P. (2005). Designing vocabulary tests for English, Spanish and other languages. In C. Butler, M. Gómez-González & S. Doval Suárez (Eds), *The dynamics of language use* (pp. 271-286). Amsterdam: Benjamins.

Meara, P. (2006). Emergent properties of multilingual lexicons. *Applied Linguistics, 27*(4), 620-644.

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

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: Museum Tusculanum Press.

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

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *The Canadian Modern Language Review, 63*(1), 127-147.

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

Nesselhauf, N. (2004). What are collocations? In D.J. Allerton, N. Nesselhauf, & P. Skandera (Eds.), *Phraseological units: Basic concepts and their application* (pp. 1-21). Basel: Schwabe.

Nurweni, A. & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes, 18*(2), 161-175.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review, 56*(2), 282-308.

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

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 & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 209-227). Amsterdam: Benjamins.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2010a). *Researching vocabulary – A vocabulary research manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N. (2010b). Key Issues in teaching and learning vocabulary. In R. Chacón-Beltrán, C. Abello-Contesse, & M. Del Mar Torreblanca-Lopéz, (Eds.) *Insights into non-native vocabulary teaching and learning* (pp. 28-40). Bristol: Multilingual Matters.

Schmitt, N. & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly, 36*(2), 145-171.

Shin, D. & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339-348.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217-235.

Wolter, B. (2005). *V_Links: A new approach to assessing depth of word knowledge*. Unpublished PhD thesis. University of Wales, Swansea.

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