Assessing the Depth of Second Language Vocabulary Knowledge*

Fransiscus Xaverius Mukarto

Abstract

This paper reviews the constructs of vocabulary knowledge, some of the established measures of the depth of L2 vocabulary knowledge, and proposes three measures to assess the depth of meaning dimension of L2 vocabulary knowledge and one measure of the breadth of meaning dimension of L2 vocabulary knowledge. The depth of meaning dimension refers to the knowledge of the syntactic and semantic features which make up the meaning of a word, and therefore, the knowledge of the meaning boundary of a word. The three measures are Forward Translation Recognition Matrix, Sentence Completion Recognition Matrix, and Acceptability Judgement. Meanwhile, the breadth of meaning of a word refers to the multiple meaning senses of a word. One measure, i.e. measure of breadth of meaning, and its variant are proposed. It is suggested that the proposed measures be critically reviewed, developed and improved for the sake of future research on the depth of vocabulary knowledge.

***Keywords**: vocabulary knowledge, meaning dimension.*

Introduction

Research in second language vocabulary acquisition has received considerable attention in the last two decades (Huckin & Coady, 1999: 182). By then, research in SLA focused primarily on how L2 learners acquire grammatical subsystems (syntax) and grammatical morpheme (morphology); it had barely touched the acquisition of vocabulary (Ellis, 1985: 5-6). Most research on L2 vocabulary acquisition to date, however, has focused on "estimates of vocabulary size or 'breadth' measures rather than on the depth of vocabulary knowledge of specific words or the degree of such knowledge (Wesche & Paribakht, 1996: 13), on the growth of L2 lexicons and on the number of words gained or forgotten over time (Schmitt, 1998:282). Despite the merits of such research or measures, one obvious limitation is that "they do not measure how well given words are known (Read, 1988 as quoted in Wesche & Paribakht, 1996: 13). As such measures fail to assess the quality of the depth of vocabulary knowledge, they cannot be used to track the acquisition development of given words. The lack of research on the depth of L2 vocabulary knowledge might have resulted from the lack of research instruments for measuring the depth of L2 vocabulary knowledge. This paper is aimed to address this issue by proposing potential instruments to measure such knowledge.

* *Presented at the 38th RELC International Seminar, SEAMEO Regional Language Centre, Singapore.*

Vocabulary Knowledge

In order to design a measure of the depth of vocabulary knowledge, one needs to have a clear construct of what constitutes vocabulary knowledge. Several attempts to make an exhaustive list of components that make up vocabulary knowledge have been made. Cronbach (1942 as cited in Bogaards, 2000; in Wesche & Paribakht, 1996: 28) distinguished five aspects or criteria of word knowledge: (1) generalization (knowing the definition), (2) application (knowledge about use), (3) breadth of meanings (knowing different senses of a word), (4) precision of meaning (knowing how to use the word in many different situations), and (5) availability (being able to use the word productively).

Richards (1976: 77-89), three decades later, proposed several aspects or assumptions of vocabulary knowledge. According to Richards, knowing a word means (1) knowing its relative frequency and its collocation, (2) knowing the limitation imposed on its use, (3) knowing its syntactic behavior, (4) knowing its basic forms and derivations, (5) knowing its association with other words, (6) knowing its semantic value, and (7) knowing many of the different meanings associated with the word. Nation (1990) adopted Richards's assumptions of word knowledge, he added the receptive & productive knowledge and several other components and reorganized them. He categorizes the components of lexical knowledge into form (spoken and written), position (grammar and collocation), function (frequency and appropriateness) and meaning (concept and associative) as presented in table 1.

Table 1: *Components of word knowledge (Nation, 1990: 31)*

Form		
Spoken form	R	What does the word sound like?
	P	How is the word pronounced?
Written form	R	What does the word look like?
	P	How is the word written and spelled?
Position		
Grammatical position	R	In what patterns does the word occur?
	P	In what patterns must we use the word?
Collocation	R	What words and types of words can we express before and after the word?
	P	What words or types of words must we use with this word?
Position		
Frequency	R	How common is the word?
	P	How often should the word be used?
Appropriateness	R	Where would we expect to find this word?
	P	Where can this word be used?
Meaning		
Concept	R	What does the word mean?
	P	What word should be used to express this meaning?
Association	R	What other words does this word make us
	P	think of?
		What other words could we use instead of this one?

Vocabulary knowledge is, therefore, complex in nature. There are various aspects or dimensions of word knowledge that L2 learners have to acquire and various tasks that they have to perform in the acquisition process of L2 lexicons. Considering the various tasks and dimensions in vocabulary learning, Nation (1990: 32) observes that "knowing a word as it is described (in table 1) applies to only a small proportion of the total vocabulary of a native speaker." Learning even an L2 word or a lexical item is a complex task. Naturally, learners' knowledge of a word is not binary in nature, nor is it an all or nothing phenomenon. The consensus is that vocabulary acquisition is incremental in nature and that the acquisition of certain word knowledge dimensions occurs concurrently (Schmitt, 1998: 283). It ranges from false familiarity with word forms to the ability to use a word correctly in free production (Færch, Haastrup & Phillipson, 1984), or as Palmberg (1987) puts it, from recognition of potential vocabulary to the ability to use it.

Bogaards (2000) observes that L2 learners may learn the following dimensions: form (spoken & written), meaning (acquired in an incremental fashion), morphology (conditions on derivation and compounding), syntax (applying the right rules to the right word or lexical unit, particularly in the learning of verbs, i.e. the argument structure and types of arguments required, and adjectives), collocates (what word may go with what words), and discourse (such as style, register, appropriateness of particular senses of a word). The complexity of vocabulary knowledge poses a very difficult problem for researchers wishing to assess the depth of learners' L2 vocabulary knowledge and to track the development of the acquisition of given words. Schmitt (1998: 282) observes that although there has been virtual explosion of vocabulary studies, "at the moment we have only the broadest idea of how acquisition might occur. We certainly have no knowledge of the acquisition stages that particular words might move through". In the following sections, several measures of the depth of vocabulary knowledge in the literature are reviewed.

Existing Measures of the Depth of Vocabulary Knowledge

The number of existing measures of the depth of vocabulary knowledge in the research literature is relatively small when compared with the breadth measures of vocabulary knowledge. Schmitt (1998: 284 quoting Read 1997) notes two approaches in measuring the depth of vocabulary knowledge: the developmental approach and the dimensional approach. In the developmental approach, scales are used to describe the stages of the acquisition of a word. Meanwhile, the dimensional approach measures the level of acquisition of the various components of word knowledge discussed in the previous section.

1. Vocabulary Knowledge Scale

One measure in the developmental approach that has received significant attention as reflected by the research literature is the Vocabulary Knowledge Scale (VKS) designed by Wesche and Paribakht (Schmitt 1998: 284). According to Wesche and Paribakht (1996) and Read (2000: 132-138), the VKS is a generic instrument, in the sense that it can be used to measure any set of words. It uses five scales to capture certain stages in the initial development of core knowledge of given words. The VKS combines self-report and performance items to elicit the self-perceived and demonstrated knowledge of specific words in written form. It consists of two types of scales: one for tapping learners' perceived knowledge of given words and the other for scoring the responses. The scale ratings range from 1 representing

complete unfamiliarity to 5 representing the ability to use a word with grammatical and semantic accuracy in a sentence. VKS, however, cannot be used to measure very small increment in word knowledge, nor can it be used to tap sophisticated knowledge of given words and to tap the knowledge of various associative meanings.

The VKS elicitation scale (Figure 1) and the VKS scoring categories (Figure 2) are presented below.

Figure 1: VKS elicitation scale self-report categories (Wesche & Paribakht, 1996: 30)

Self-report categories	
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 the word in a sentence: _____ (If you do this, please also do section IV.)

FIGURE 2: VKS scoring categories: Meaning of scores (Wesche & Paribakht, 1996: 30)

Self-report categories	Possible scores	Meaning of scores
I.	1	The word is not familiar at all.
II.	2	The word is familiar but its meaning is not known.
III.	3	A correct synonym or translation is given.
IV.	4	The word is used with semantic appropriateness in a sentence.
V.	5	The word is used with semantic appropriateness and grammatical accuracy in a sentence.

Meanwhile, in the dimensional approach, researchers have attempted to design several assessment instruments, notably 'word associates' (henceforth WA) test developed by Read (Read, 2000; Wesche & Paribakht (1996), Euralex French Tests (Bogaards, 2000) and interviews (Schmitt, 1998) and Sentence Completion Test (Ijaz, 1986).

2. Word Associates Test

After undergoing several revisions during its development, the latest version of the word association test designed by Read is aimed at measuring the depth of learners' knowledge of a particular class of words only, i.e. adjective. The test consists of 50 test items and it tests learners' ability to identify whether or not

there is any syntagmatic (collocational), paradigmatic (synonymous), and analytic (part-whole or whole-part) relationship between a stimulus word and each of the eight other words presented as choices. Out of these alternatives, four are distractors. The eight associates and distractors are divided into two groups. Consider the example below:

sudden

beautiful	quick	surprising	thirsty

change	doctor	noise	school

The words in the left box are adjectives. The relationship between the associates (*quick* and *surprising*) and the target word is either paradigmatic (synonymous) or analytic. The words in the right box are all nouns, and the relationship between the target word *sudden* and its associates (*change* and *noise*) is syntagmatic.

3. The Euralex French Tests

Each Euralex French test consists of 60 items, 40 of which present words with some kind of relationship while the remaining 20 present words with no relationships whatsoever (dummy items). The tests are designed to test high level of vocabulary knowledge as the words used in the test are low frequency words. Therefore, these tests are not appropriate to track the development of vocabulary knowledge of high frequency words and most learners of French as a second or foreign language. The test requires the test takers to decide whether there is any kind of association between the pairs of words. The instructions read 'You have to decide if you can see an obvious connection between the two words'. The test taker should give "Yes" or "No" responses. Sample items are as follows:

1. [ ] pied: grue 2. [ ] sarcler: bois.

The aspects of knowledge assessed include meaning relationships (synonym and hyponym), selection restrictions (verbs/ nouns, free associates), fixed expressions (expressions and compound words) and cultural aspects. Although the tests were originally designed to measure the vocabulary knowledge of learners of high proficiency level, the test can be modified to test the vocabulary knowledge of learners with lower language proficiency.

4. Interviews

The interview has also been used as a means for measuring the depth of vocabulary knowledge (Schmitt, 1998 and Read, 2000: 178-180). In the interview procedure, Read presents students with a selection of words and open-ended questions to elicit various aspects of their knowledge of each word. A sample test item of the written form of the interview procedure for the word 'interpret' is as follows:

Figure 3: Test sheet for the word "interpret"

TO INTERPRET	
1.	Write two sentences: A and B. In each sentence, use the two words given. A. interpret experiment B. interpret language

2.	Write those words that can fit in the blank. to interpret a(n) _____ i _____ ii _____ iii _____
3.	Write the correct ending for the word in each of the following sentences: Someone who interprets is an interpret..... . Something that can be interpreted is interpret..... . Someone who interpret gives an interpret..... .

A detailed account of the interview procedure used by Read can be found in Read (1989).

Meanwhile, Schmitt used the interview procedure in a longitudinal study involving three advanced adult university students with different L1 backgrounds. The study aimed to track their acquisition of 11 words and their acquisition development over the period of one year. The aspects of word knowledge under study were spelling, association, grammatical information, and meaning. A detailed account of the procedure can be found in Schmitt (1998). The advantage of this procedure is that it allows researchers to tap as much information as required. The negative side is that it is time-consuming and therefore limits the number of research subjects.

Proposed Measures of the Depth of Vocabulary Knowledge

Assessment instruments are designed to measure certain constructs, i.e. particular kinds of knowledge or ability that given assessment instruments are designed to measure (Read, 2000: 95). The instruments that I propose in this paper were designed with the following objectives in mind: to assess the depth of vocabulary knowledge. The depth of vocabulary knowledge covers the depth and breadth of the meaning dimension of a given word. The depth of meaning refers to the knowledge of both the syntactic and semantic features that constitute the core meaning of a word while the breadth of meanings refers to the multiple meaning senses of a word.

1. Measures of the Depth of Meaning Dimension of Word Knowledge

Key syntactic and semantic features of a word can generally be found in dictionary definitions. For example, Oxford Advanced Learners' Dictionary defines the word "murder" as "the illegal deliberate killing of a human being" (noun) and "to kill sb illegally and deliberately" (verb). From the definition one can identify some key features: the nature of the action is *illegal* and *deliberate*, and the object is a *human being*. Finer semantic features of the word "murder" can be identified by contrasting it with other words sharing the common meaning sense *kill*: "kill", "assassinate", etc. By contrasting them, one can identify the features that constitute the meanings of these words. For example, the words "murder" and "assassinate" require a human agent and object while "kill" may take a *human* or *non-human* agent or object. A human agent for "murder" correlates well with the other fea-

tures of the word which indicate that the killing action is *deliberate* and *illegal*, while *non-human* agent does not. Meanwhile, although both "murder" and "assassinate" take only human objects, the nature of objects in both verbs differs: the objects of the word "assassinate" are usually prominent political figures.

From his research on semantic features, Poedjosoedarmo (1989: 80-81) lists a number of features that one needs to consider when analyzing the meaning of a verb: agent (human or non-human, animate or inanimate, number, and sex), objects (with or without objects, human or non-human, animate or inanimate, size, weight, etc), time, place, process, frequency, motives or reasons, result or effect, beneficiary, direction, instrument used, etc. Consider the following example, contrasting the Indonesian verbs:

(1) *membawa* (to carry),
(2) *gendong* (to carry something or someone on the back or hip, supported by the waist or one's arms, often with the help of a cloth sling),
(3) *gotong* (to carry something heavy that two or more people must cooperate),
(4) *jinjing* (to carry something light and relatively small in size in one's hand),
and
(5) *pikul* (to carry something heavy divided into two and each put on the end of a supporting stick on one's shoulder).

The word *membawa* is the cover or generic term for all these verbs. Let us now contrast some of these verbs with each other. In *gendong*, the object is placed on the back or hip of the person carrying it, meanwhile in *pikul* the part of the body that is used to carry is the shoulder, not the back or hip. The words *gendong* and *gotong* contrast in the number of agents: one in *gendong* and more than one in *gotong*. The words gotong and jinjing contrast in the number of agent and the nature of objects: while the object of *gotong* is something heavy, the object of *jinjing* is something light and usually small in size.

The next construct to be described is meaning mapping. A map, according to the Cambridge International Dictionary of English (1975: 863), means a drawing of (part of) the earth's surface showing the shape and position of different countries, political borders, natural features such as rivers and mountains, and artificial features such as roads and buildings. The Oxford Advanced Learners' Dictionary uses the expression "representation on paper" for the term "drawing" in the Cambridge definition. There is some parallelism between the term "map" that I use and the literal definition above; both shares the features "representation", "features" and "boundaries". What I mean by the term "meaning mapping" here is, then, the representation in the mind of a word meaning: both the syntactic and semantic features within the meaning boundary of a word which make up the meaning of that word (depth of meaning) and the multiple meaning senses of a word (breadth of meaning). If the claims that vocabulary acquisition is incremental (Schmitt, 1998) and that vocabulary acquisition is a process of continuous lexical disambiguation (Sonnaiya 1991: 273) are right, the number and kinds of features within the meaning boundary of a word and the meaning boundary itself may change to eventually approximate the meaning of the word mapped by a native speaker. For example, a learner at one stage of the acquisition of the word "murder" may map the feature *± human* agent within the meaning boundary of the word. In this case, the mapping of the word meaning is inaccurate because the feature *–human* agent which actu-

ally is outside the meaning boundary is incorrectly mapped as within the meaning boundary. At a later stage he or she may realize that this word takes only *+human* agent and exclude the *–human* agent from within the meaning boundary.

The question is how to measure or tap such knowledge? It seems that the existing measures of the quality of vocabulary knowledge discussed above are not sensitive enough to measure this particular knowledge because they were not designed to measure such a construct. Therefore, measures which are sensitive enough to test such knowledge need to be designed. The measures should, construct wise, be designed to tap learners' mapping accuracy of the word meanings by asking students to indicate whether certain semantic and or syntactic features are within or outside the meaning boundary of the word. Asking a test taker to identify whether certain set of features are inside or outside the meaning boundary is a big challenge to a test designer as most learners who take the test are not linguists. It is unrealistic to ask such questions as, "Does the word 'murder' take a non-human inanimate agent?" because such a question contains specific terms that test takers may not understand. The question is how to design instruments sensitive enough to measure such a construct.

In designing such measures, an assessment instrument designer needs to go through the following procedures:

1. Select a set of words the knowledge of which will be measured.
2. Identify the semantic and syntactic features or components that constitute the meaning of the words. This can be done by way of contrastive analysis as is illustrated above, i.e. by contrasting a selected word with other words which have paradigmatic relation with the word or which are to some extent synonymous with the word. Such words are usually within the same semantic field. For example, the word "murder" takes two arguments (agent and object), the agent and object are living human beings, the action is illegal but is done deliberately.
3. After the words have been selected and the syntactic and semantic features identified, the words need to be presented in context. The question, as raised by Read (2000), is how much context a word needs. As a guiding principle, the syntactic and the semantic features of the words will determine how much context is needed. Let us take the word "murder" again as an example. This word is a verb requiring two arguments, so we need to provide at least two types of sentences: intransitive (one argument) and transitive (two arguments) sentences. Another feature, i.e. the agent and object are human being, requires sentential contexts in which the agents and objects are of various kinds: human being, non-human being (animate and inanimate). Look at the following sentential contexts as illustration.

   (1) Drugs can _____.
   The context contains one argument and the argument or the agent is inanimate non-human being. This feature is outside the meaning boundary of "murder".
   (2) This tiger _____ a farmer.
   The context contains two arguments but the first argument, an animate non-human agent, is outside the meaning boundary of the tested word.
   (3) This man robbed a taxi driver and _____ him.

The context contains two arguments, both living human being. As nothing specific is said about the nature of the action, the context implies that the action is deliberate and illegal. Therefore, the set of features is within the meaning boundary of the word.

4. Select the format of the measures. The format of the measuring instruments can be the forward translation recognition matrix (FTRM), the sentence completion recognition matrix (SCRM), or acceptability judgement (AJ). The following section will discuss each of them.

a. Forward Translation Recognition Matrix (FTRM)

Forward Translation Recognition Matrix is a self-report assessment instrument used to measure the depth of meanings of a set of verbs within given semantic fields. Verbs are chosen because verbs play a central role in sentences. They are relational in nature because they relate the arguments in the argument structure. In addition, verbs have the potential to cause more problems to L2 learners than either nouns which serve more or less as labels and adjectives. Forward translation is a term used to refer to the translation of L1 into L2, while the translation of L2 into L1 is called backward translation. In translation production, one has to come up with the translation of the presented word, whereas in translation recognition he or she has to choose from a number of options the best translation of the presented word (Mukarto, 2001). As its name suggests, the instrument is used to measure only the receptive vocabulary knowledge. Consider figure 4, a sample test item which measures the vocabulary knowledge of five verbs within the semantic field KILL. Knowledge of the five verbs is measured.

Figure 4: Sample test item of FTRM measuring the depth of vocabulary knowledge

He robbed a taxi driver and _____ him.
*Ia merampok sopir taksi dan **membunuh**nya.*

No.	Verbs	YB	KB	TT	KS	YS
A	assassinated					
B	executed					
C	killed					
D	murdered					
E	slaughtered					

Where **YB** stands for *Yakin Benar* (definitely appropriate or correct)
KB stands for *Kelihatannya Benar* (seems appropriate or correct)
TT stands for *Tidak Tahu* (do not know)
KS stands for *Kelihatannya salah* (seems inappropriate or wrong)
YS stands for *Yakin Salah* (definitely inappropriate or wrong)

The sample item consists of (a) an English sentence to complete, serving as sentential context for target or tested verb, (b) and an Indonesian sentence equivalent with the English sentence, but contains the lexical prompt, i.e. a verb to

translate, printed in bold, and (c) a matrix. The matrix consists of seven columns. The first column contains alphabetical letters. These letters, when combined with the test item number, will make the number of test item, for example 4A, 4B, 4C, etc. for purpose of ease in data tabulation. The second column contains the tested verbs. The tested verbs are arranged in alphabetical order, for example "assassinate" followed by "execute", "kill", "murder" and "slaughter". The number of rows may vary, depending on the number of tested verbs and distrator(s), if any. The other five columns, when combined with the rows, form the cells in which the subjects mark their answers.

The matrix system is adopted for two reasons: economy and practicality. First, it is economical in the sense that it can test the same set of semantic features for all the tested verbs. Earlier versions of the test used many sentential contexts, depending on the syntactic and semantic features identified, to measure the knowledge of just one verb. As a result the number of test item was large and the measure was neither economical nor practical. The current matrix system allows the use of the same sentential contexts providing certain syntactic and semantic features to measure the knowledge of all the selected verbs. Second, it is practical because the subjects have to read only a small number of sentential contexts and the order of the tested words is predictable. It allows them to work on the test more quickly, and thus lowering the fatigue factor.

A subject has to answer such questions in two steps: first, recognize or identify whether or not the options are English equivalent(s) of the Indonesian word prompt "*membunuh*" and second, determine whether the word is used in an appropriate context. If s/he is certain or believes that one option, "execute" for example, is not equivalent to the Indonesian word prompt, then s/he will check or tick the appropriate cell **YS**, indicating that she is certain that it is not equivalent. In case she is uncertain but thinks that it is not equivalent, she ticks the appropriate cell indicating her answer, **KS**. However, if she finds that the English word, for example "kill" or "murder", is equivalent to the Indonesian lexical prompt, s/he will have to check whether or not the word meets the collocational constraint imposed by the provided context. If s/he thinks both English words are equivalent and satisfy the collocational constraint or the syntactic and semantic constraints, s/he will check the cell **YB**, indicating that they are equivalent to the prompt and satisfy the syntactic and semantic constraints. If s/he knows that the word may be equivalent, for example "assassinate", but it does not satisfy the collocational constraints, then s/he has to tick the cell **YS** indicating that the word is not acceptable in the sentential context. However, if s/he is doubtful whether or not the English word is equivalent and/or whether or not the word meets the syntactic and semantic constraints, then s/he checks the cell **KB**. In the case the subject is not familiar with the L2 word and does not know whether or not it is acceptable, s/he checks the cell **TT**.

A set of FTRM has been designed and used to measure the Indonesian EFL learners' depth of meaning dimension or the semantic mapping accuracy of twelve verbs within two semantic fields KILL and BREAK and to find out the patterns of semantic mapping development from the low intermediate level to the advanced level. The set consists of 32 sentential contexts: 16 contexts each semantic field. Three proficiency groups of 40 Indonesian EFL learners each were involved in the study. The three groups were the low intermediate, high intermediate and advanced groups. Results of the study will be reported elsewhere.

b. Sentence Completion Recognition Matrix (SCRM)

SCRM is similar to FTRM in terms of the designing procedures, format and scoring. The difference is that in the SCRM there is no L1 prompt to translate. The Indonesian translation of the sentential context may also be omitted, depending on the English proficiency level of the subjects. Construct wise, SCRM is probably simpler but better than FTRM. In SCRM the subjects observe only the syntactic and semantic features contained within the sentential context. Meanwhile, in FTRM the subjects have to map the L1 word prompt into the L2 and the L1 prompt may limit the number of correct options because the variables to be observed by the subjects is not only the syntactic and semantic features provided by the sentential context, but also the syntactic and semantic features contained within the L1 word prompt. Likewise, the lexical processing in SCRM is also shorter than that in FRTM as subjects do not have to refer to the L1 prompt and match the L2 words with the L1 prompt. A sample test item for the SCRM is provided below.

Figure 5: Sample test item of SCRM assessing the depth of vocabulary knowledge

He robbed a taxi driver and _____ him.
Ia merampok sopir taksi dan _____-nya.

No.	Verbs	YB	KB	TT	KS	YS
A	assassinated					
B	executed					
C	killed					
D	murdered					
E	slaughtered					

The scoring, meaning of scores, and conversion of scores in SCRM are the same as in FTRM so there is no need to discuss them again here. The directions or instructions, however, differ to some extent, particularly with the mapping of the L1 word prompt into the L2. In addition, while the FTRM has been pilot-tested, the SCRM has not.

The use of the sentential context in the learners' mother tongue is optional, depending on the proficiency level of the research subjects. The sentential contexts in the L1 are deemed necessary if and only if the subjects include low proficiency group(s) so that the subjects may not understand the English sentential contexts. If they do not understand the sentential contexts, they will not know the salient semantic features contained within the contexts. In turn, the validity of the test may be questioned.

One set of SCRM, which is a variant of FTRM, has been developed. It has been pilot-tested but has never been used. As a variant of the FTRM, the SCRM differs from the FTRM only in the absence of the Indonesian word prompts in the sentential contexts.

c. Acceptability Judgement

A variant of the FTRM and SCRM is acceptability judgement measure, also a self-report assessment instrument. The procedures used in designing and scoring the test are similar to the ones used in the designing of SCRM. The difference is that this instrument is used to measure the depth of vocabulary knowledge in which the individual verbs tested do not belong to the same semantic field, for example, "kill", "break", and "carry". Another difference is that no matrix is required. This instrument requires the subjects to indicate whether the verb used in the sentential contexts match the set of features contained within the sentential context or vice versa. The number of options provided to the subjects was the same as that in SCRM and FTRM and the format is similar. To measure the knowledge of the verb "carry", in which the feature *direction* is important, the sentential contexts to present the verbs may look like the following:

(4) When on duty, he always **carries** a pistol.
(5) "Tom, can you **carry** the hammer here," the mechanic said.
(6) "Tom, can you **carry** the newspaper to your father in the veranda?' mother said.

In the sentential context (6), the direction of the activity is generic, there is no specific direction, and therefore the verb "carry" is acceptable or appropriate. In (7), the direction is toward the speaker and the verb "carry" is not appropriate or acceptable. Likewise, the verb "carry" is not appropriate, because the direction is away from the speaker.

One potential problem to consider is the number of sentential contexts, which is large, because a target word may require a large number of contexts. This may result in fatigue on the part of the research subjects.

2. Measure of the Breadth of Meaning Dimension of Word Knowledge

The instrument used to assess the learners' receptive aspect in the acquisition of the multiple meaning senses of English verbs, ranging from their typical meanings to their least typical meanings is, henceforth, called the Measure of Breadth of Meaning or MBM. The meaning of a word is considered to be receptively acquired by a learner if s/he can identify that a word is appropriately used is a (sentential) context and can identify its meaning in its existing context. Context is necessary because it determines what a target word means. The meaning inventory of verbs can be found in dictionaries. Most frequent verbs usually have more meaning senses than less frequent verbs. The meaning senses selected to be measured may be the major meaning senses of the verbs (as in Schmitt 1998) or they may range from the most common meaning to the least common ones.

In designing MBM, an assessment instrument designer needs to go through the following procedures:

1. Select the words or verbs to be assessed.
2. Select the meaning senses to be measured. Consult a number of good dictionaries for the inventory of word or verb meanings.
3. Select the sentential contexts to present the meaning senses of the verbs.
4. Design (the format of) the assessing instrument.

The MBM that I propose takes the following format: (1) the sentential contexts to present the verbs with their particular meaning senses (2) an answer sheet where the subjects give their judgement on the acceptability or the appropriateness of the verbs in their sentential contexts and the meaning senses of the verbs. Sample of MBM test items and answer sheet are presented in figure 6.

Figure 6: Samples of test items and answer sheet

Sample of test items

1. Write the meaning of the word printed in bold. 2. Is the word used in appropriate context?	
1.	The milk has **gone** sour.
2.	He **goes** to school by bus.
3.	My car **went** beautiful.
4.	Pink and orange don't **go**.
5.	What time does the last train **go**?

Sample of answer sheet

No	Appropriate?			Meaning
	Yes	No	Don't Know	
1.				
2.				
3.				
4.				
5.				

The target verb is presented in a number of contexts, i.e. five contexts in the example above, depending on the number of meaning senses to be assessed. In sentential context (SC) 1, the verb "go" takes an adjective as its complement and so does SC3 which serves as a distractor. In SC2 the verb takes a prepositional phrase as its complement. Meanwhile, in SC 4 and SC5 the verb does not take any complement or intransitive. SC1, SC2, SC4 and SC5 have different meaning senses.

The answer sheet is in the form of a matrix. The matrix consists of three parts: test item numbers, appropriateness and meaning of the tested verbs. There are three possible answers for the appropriateness section: appropriate, not appropriate, and no knowledge of appropriateness. The subjects write the meaning of the verbs in the form of the translation equivalents in the subjects' language.

One set of the MBM has been designed to measure the measure the breadth of the meaning dimensions of ten English verbs. Four meaning senses of each word are tested, ranging from the core meaning senses to the less core ones. One distractor is provided for each verb. The developed test was used in a study involving three proficiency groups of Indonesian EFL learners. The subjects were the same subjects as in the FTRM above. The test was used to measure the breadth of the meaning dimension acquired receptively by three proficiency groups and to figure out the developmental patterns of the acquisition of the breadth of meaning dimension of the target English verbs. Results of the study will also be reported elsewhere.

3. Scoring Procedures

a. Scoring Procedures in the Measures of the Depth of Word Meaning

Measures of the depth of meaning of words are used to elicit data on the semantic mapping accuracy of the English verbs. In the FTRM the subjects' tasks include

1. determining whether the English verbs in the matrix is an acceptable translation of the Indonesian verb prompt considering the semantic features or information contained within the English verbs and those within the provided sentential context, and
2. indicating the level of mapping confidence in his or her answers. A subject may confidently indicate that certain semantic features are within or outside the meaning boundary of a word. However, he or she might also be doubtful, not knowing for sure whether they are within or outside the meaning boundary not because he or she does not have the knowledge but perhaps because such features are not fully integrated in his or her lexical entry yet.

SCRM differs from FTRM in that no translation recognition is required from the learners' mother tongue to English. This seemingly simple difference may bring about significant difference in the acceptability of the answers given by the subjects because in the SCRM the answers are constrained only by the features contained within the sentential context while in the FTRM the answers are constrained both by the sentential contexts and by the learners' L1 word prompts.

Considering the subject's tasks and the nature of data elicited, a number of scoring systems are adopted, depending on the purpose of the assessment. They include nominal, ordinal and interval scores.

Table 2: Nominal scores and their meanings

Nominal Score	Meaning of Score
5	(1) Correct semantic mapping, (2) High level of mapping confidence
4	(1) Correct semantic mapping, (2) Low level of mapping confidence
3	No idea
2	(1) Incorrect semantic mapping, (2) Low level of mapping confidence
1	(1) Incorrect semantic mapping, (2) High level of mapping confidence
0	No response

Table 2 present the nominal score used to code the subjects' responses. The nominal scores range from 0 to 5. To illustrate how the coding is to be done, consider the following example of the scoring of a subject's answer to sample stimulus items in figure 7.

Figure 7: Sample of test item in sentence completion recognition matrix

He robbed a taxi driver and _____ him.
Ia merampok sopir taksi dan _____nya.

No.	Verbs	YB	KB	TT	KS	YS
A	assassinated					
B	executed					
C	Killed					
D	murdered					
E	slaughtered					

The score 5 is given if the semantic mapping is correct and the level of mapping confidence is high. In the sample item above, a score of 5 is given if the subject ticks the cell **YB** for "murder" and **YS** for "execute". A score of 4 is given if the subject ticks the cell **KB** for "kill" and **KS** for "assassinate", indicating that the mapping is correct but the level of mapping confidence islow. A score 3 is given if the subject indicates that s/he does not have any idea whether or not the mapping is correct by ticking the cell **TT.** A score 2 was given when the subject ticks the cell **KS** for "kill" and **KB** for "execute". The level of mapping confidence is low and the mapping is incorrect. A score 1 is given if the subject believes that the mapping was correct but it is actually incorrect, e.g. ticking **YB** for "assassinate" or **YS** for "kill". A score of 0 is given when there is no response from the subject.

It should be noted, however, that these raw scores (0 – 5) are of nominal scale. They do not represent ordinal or interval scale. For statistical purposes, these raw scores have to be converted depending on the purpose of the statistical analysis. Conversion from the nominal scores to the ordinal ones is used to compute the aggregated semantic mapping accuracy.

To find out the aggregated semantic mapping accuracy, i.e. the sum total of all the accurate mapping (positive values) and inaccurate semantic mapping (negative value) integrated within the L2 lexicon, the raw scores are converted into ordinal scores as in table 3.

Table 3: Conversion table for overall semantic mapping accuracy

Raw Scores	Meaning of Score	Con-vert. Score
5	(1) Correct semantic mapping, (2) High level of mapping confidence	2
4	(1) Correct semantic mapping, (2) Low level of mapping confidence	1
3	No idea	0
2	(1) Incorrect semantic mapping, (2) Low level of mapping confidence	-1
1	(1) Incorrect semantic mapping, (2) High level of mapping confidence	-2
0	No response	0

Aggregated semantic mapping accuracy represents the degree or depth of meaning knowledge of a particular word. The raw scores are weighted differently because of the differences in the accuracy and the level of mapping confidence in the L2 lexicon.

Interval scores are used to find out the proportion of accurate or inaccurate semantic mapping. Two types of score conversions may be used. The first concerns only on the semantic mapping accuracy while the level of mapping confidence is ignored. The second concerns both the accuracy and the level of mapping confidence.

Table 4 illustrates how the raw scores are converted to find out the proportion of the accurate semantic mapping, disregarding the level of mapping confidence. In this case the raw scores are converted into interval scale.

Table 4: Conversion table for accurate semantic mapping

Raw Scores	Meaning of Score	Convert. Score
5	(1) Correct semantic mapping, ~~(2) High level of mapping confidence~~	1
4	(1) Correct semantic mapping, ~~(2) Low level of mapping confidence~~	1
3	No idea	0
2	(1) Incorrect semantic mapping, ~~(2) Low level of mapping confidence~~	0
1	(1) Incorrect semantic mapping, ~~(2) High level of mapping confidence~~	0
0	No response	0

To find out the proportion of correct or accurate semantic mapping, the raw scores 4 and 5 are converted to 1 while the other raw scores are converted to 0. Likewise, to find out the proportion of the inaccurate semantic mapping, the raw scores 1 and 2 are converted to 1 while the other scores are converted to 0.

When the level of confidence which indicates the level of integration of the semantic features within the L2 lexicon is counted, the second type of score conversion as in table 5 is used.

Table 5: Score conversion for accurate semantic mapping with high level of mapping confidence

Raw Scores	Meaning of Score	Convert. Score
5	(1) Correct semantic mapping, (2) High level of confidence	1
4	(1) Correct semantic mapping, (2) Low level of confidence	0
3	No idea	0
2	(1) Incorrect semantic mapping, (2) Low level of confidence	0
1	(1) Incorrect semantic mapping, (2) High level of confidence	0
0	No response	0

Table 5 illustrates how the raw scores are converted when the purpose is to find out the proportion of the accurate semantic mapping with a high degree of confidence. The nominal score 5 was converted to interval score 1, while the others to 0. If the purpose was to find out the proportion of inaccurate semantic mapping with low level of mapping confidence, the raw score 2 was converted to 1 and the others to 0.

b. Scoring Procedures in the Measures of the Depth of Word Meaning

In the breadth of meaning dimension, knowledge of word meaning is operationally characterized by knowledge of the conceptual meaning of the word, i.e. indicated by correct or acceptable translation of the word, and knowledge of the use of the word. As in the Forward Translation Recognition Matrix, three types of scoring or scales were used: raw scores (nominal scale), scores based on the

weighting of the aspects of the breadth of meaning for non-parametric tests (ordinal scale), and converted scores for use in parametric tests.

Table 6 illustrates the coding and scoring of the assessed constructs or aspects of the meaning senses of the verbs, i.e. use and conceptual meaning.

Table 6: Coding of responses and its score conversion in MBM

Breadth of Meaning Knowledge		Combination	Raw Scores	Weighting		Ordinal Scores
Concept	Use			Concept	Use	
Correct (+C)	Correct (+U)	+C ,+U	8	3	1	4
Correct (+C)	No response (0)	+C, 0	7	3	0	3
Correct (+C)	Incorrect (-U)	+C, -U	6	3	-1	2
No response (0)	Correct (+U)	0,+U	5	0	1	1
No response (0)	No response (0)	0, 0	4	0	0	0
No response (0)	Incorrect (-U)	0, -U	3	0	-1	-1
Incorrect (-C)	Correct (+U)	-C,+U	2	-3	1	-2
Incorrect (-C)	No response (0)	-C, 0	1	-3	0	-3
Incorrect (-C)	Incorrect (-U)	-C, -U	0	-3	-1	-4

The first two columns under the heading "Breadth of Meaning Knowledge" covers both the conceptual meaning and the uses of the tested words. The order in the first column moves from accurate to inaccurate mapping, while in the second column from correct to incorrect use. The fourth columns contain the code or raw scores, which are nominal in nature. The fifth and sixth columns contain the weighting of the responses by the subjects. The two aspects of the breadth of meaning are assigned different weighting: three (3) for correct conceptual meaning and one (one) for correct use on the one hand minus three (-3) for incorrect conceptual meaning and minus (-1) for incorrect use. Meanwhile, zero for no response which indicates no knowledge is assigned 0. Conceptual meaning and use are given different weight because in receptive vocabulary knowledge, the uses of words do not have as important role as the conceptual meanings of words. In productive vocabulary knowledge, however, both are equally important.

To find out differences in the proportion of correct or incorrect mapping of the target verbs and to figure out the developmental pattern in the proportion of correct and incorrect mapping, the raw scores are converted to either 1 or 0. For example, to find out the proportion of the correct mapping of the target verbs, i.e. correct identification of meaning and use, the raw score 8 is converted to 1 and the others to zero.

The use of sentential context may sometimes cause some misunderstanding on the part of the subjects in answering the questions which may cause some difference in scoring the answers among raters. To avoid this, the directions given should clearly state that the focus of the assessment is on the use and conceptual meaning of the words, not on the morphosyntactic aspects of the words. For example, a subject may rate the use of "covered" in "She **covered** her eyes with her hands." as not acceptable because she or he thinks the correct form is "covers", not "covered". There is no way of detecting such a case in this format of the test. This, however, can be detected in the variant of this test, in which the subject is asked to supply the correct word or form if s/ he finds the word not appropriate.

Another potential problem is in scoring the answers. To avoid unnecessary differences among raters in determining the inter-rater reliability, the criteria for

correct or incorrect answers or translation should be stated clearly. In the pilot testing of the instrument it was found that

1. a subject may give more than one translations of the verb. In case a subject gives more than one translations, it should be seen whether all the translations are correct. If one or more translations are incorrect, it should be interpreted as incorrect. Giving more than one translation may indicate that the subject is doubtful of his or her first translation, especially when one or more of the translations is incorrect. If all the translations are correct, it may indicate that s/ he knows more than one translation for the word.
2. a subject may translate not only the meaning of the word but also the inflectional affix attached to the word. For example, a subject may translate the word "carry" in "The woman **carrying** the hand bag is my mother" as "yang membawa" in Indonesian, while other students may translate it as "membawa", without translating the meaning sense conveyed the affix "-ing". In such a case, both answers should be considered correct as the main meaning of the verb "carry" is correctly translated.

Concluding Remarks

I have reviewed the construct of vocabulary knowledge and a number of assessment instruments used to measure the depth of vocabulary knowledge. Each instrument has been designed to measure certain constructs and therefore is sensitive to measure the particular constructs it is designed for. Considering the need for measures of the depth and breadth of the meaning dimension of vocabulary knowledge and the lack of such measures, I have proposed three measures of the depth of meaning dimension of English verbs –FTRM, SCRM, and Acceptability Judgement– to assess the knowledge of the syntactic as well as the semantic aspects of vocabulary knowledge. The construct validity of the proposed tests has been described above. To assess the breadth of meaning of a word, particularly verbs, I have also proposed a Measure of the Breadth of Meaning (MBM) of verbs with multiple meaning senses and its variant (only briefly mentioned) which allow researchers to detect the possibility of incorrect judgement on the usage of the tested verb in a sentential context due to morphosyntactic factor. One set of test each has been designed for the FTRM, SCRM and MBM. The developed tests, particularly the FTRM and MBM, have also been used in a study involving three Indonesian EFL learners to assess both the depth and breadth of the meaning dimensions of a number of English verbs and to figure out the patterns of development in the acquisition of the verb meanings. Despite the fact that some of the measures have been developed, pilot-tested and improved, and used in a study, there is a constant need to critically review and develop all these proposed measures to establish their validity and reliability as well as to improve them for the sake of further research in the depth of vocabulary knowledge.

References

Ary, D., L.C. Jacobs, and A. Razavieh (1990). *Introduction to Research in Education* Forth Worth: Holt, Rinehart and Winston, Inc.

Bogaards, P. (2000).Testing L2 vocabulary knowledge at a high level: the case of the *Euralex French Tests. Studies in Second Language Acquisition* 21.

Ellis, R.1985. *Understanding Second Language Acquisition*. Oxford: Oxford University Press.

Faerch, C., Haastrup, K., & Phillipson, R. (1984). *Learner Language and Language Learning*. Clevendon, UK: Multilingual Matters.

Huckin, T. & J. Coady. 1999. Incidental vocabulary acquisition in a second language: a review. *Studies in Second Language Acquisition* 21: 181-193.

Ijaz, I.H. 1986. Linguistic and cognitive determinants of lexical acquisition in a second language. *Language Learning* 34, 401-451.

Nation, I.S.P. (1990). *Teaching and Learning Vocabulary*. New York: Heinle and Heinle.

Palmberg, R. (1987). Pattern of vocabulary development in foreign language learners. *Studies in Second Language Acquisition.* 11, 9-29.

Poedjosoedarmo, S. (1989*). Filsafat Bahasa*. Bandar Seri Begawan: Universiti Brunei Darussalam.

Read, J. (1989). Towards a deeper assessment of vocabulary knowledge. *ERIC Document Reproduction Service*, No. ED 301 048. Washington, DC: ERIC Clearing House on Languages and Linguistics.

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

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

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: a longitudinal study. *Language Learning* 48, 281-317.

Sonaiya, R. (1991). Vocabulary acquisition as a process of continuous lexical disambiguation. *IRAL* 29: 273-284.

Wesche, M. and T.S. Paribakht. (1996). Assessing second language vocabulary knowledge: depth vs. breadth. *Canadian Modern Language Review* 53, 13-39.