Frequency and Vocabulary Types

Syarifuddin

English Language Teaching (ELT)
Universitas Negeri Malang
Email: syarifuddinalmenaki@gmail.com

Abstract

Vocabulary is a fundamental component of language use, and research repeatedly suggests the dominance of vocabulary knowledge over four language abilities such as listening, reading, speaking and writing. For many learners, vocabulary growth is of their major and even lasting concern in learning a second language. Vocabulary knowledge is not only about knowing enough words but more importantly also about knowing enough of the right words i.e. the words that the learners may often meet and use in typical discourse they involve. In relation to this, teachers as well as material writers are required to do a cost-benefit analysis to select what vocabulary deserves inclusion and should get most attention in the language program. Using this matter as a starting point, this article aims to present vocabulary types according to frequency levels and to overview the nature of each type and how each should be dealt with in the language program.

1. Introduction

What matters about vocabulary learning is not only the need for learners to know enough words (i.e. to have sufficient vocabulary sizes) but also the need for them to know enough of the right or most useful words, that is, the words they may frequently meet and use in an authentic language use or in typical discourse they involve. For this reason, teachers and material writers need to do a cost-benefit analysis of whether any particular vocabulary items deserve instruction or inclusion. For the purpose of helping teachers and material writers in making such analysis, vocabulary scholars and researchers have broken vocabulary into different types or categories.

The most influential and widely used vocabulary types are Nation’s (2001) four-part categorization. Basing his idea on the criteria of frequency (i.e. number of times a word occur in a text) and range (i.e., number of texts a word occur in), he divides vocabulary occurring in any text into four categories: (1) high-frequency vocabulary, (2) academic vocabulary, (3) technical vocabulary, and (4) low-frequency vocabulary. Another more recent categorization has been proposed by Schmitt and Schmitt (2014). They revisit Nation’s (2001) four-part categorization, and argue that the four-part categorization is no longer defensible as a pedagogical basis in the light of recent research on vocabulary field. In it, they suggest new
boundaries for both high- and low-frequency vocabulary, propose a new label for the vocabulary existing between the high- and low-frequency vocabulary, and in result break vocabulary into three categories: (1) high-frequency vocabulary, (2) mid-frequency vocabulary, and (3) low-frequency vocabulary.

2. **High-Frequency Vocabulary**

High-frequency vocabulary contains vocabulary which occurs so frequently in all texts, and makes up the majority of running words of the texts. Therefore, this vocabulary is extremely useful for learners especially in the early stage of learning a language. Learning this vocabulary will provide learners with the highest return for their learning efforts since that they will meet and use it often. Nation and Meara (2010) strongly suggest any words within this category to be the first and main vocabulary goal for learners, and encourage both teachers and learners to do as many efforts as possible to ensure these words are learned well.

Although this vocabulary makes up the majority of running words in all texts, it consists of a relatively small number of words. This is considerably a fortunateness for many learners. In English, the high-frequency vocabulary has traditionally been thought to consist of around 2,000 most frequent word families (Schmitt and Schmitt, 2014) which make up around 80% of the running words in written texts and around 90% of spoken texts (Matsuoka, 2012). The origin of the 2,000 figure, as Schmitt and Schmitt (2014) note, is largely from the influence of West’s (1953) *General Service List* (GSL) and the reinforcement by research over 50 years old.

For many decades, the most frequent 2,000 word families has been a widely cited boundary for the high-frequency vocabulary of English in much literature. However, Schmitt and Schmitt (2014) recently reassess this traditional boundary. On the basis of multiple perspectives which include frequency and coverage, acquisition studies, the amount of vocabulary necessary for English usage, the ranges of graded readers, and the ranges of dictionary defining vocabulary, they argue that the high-frequency vocabulary of English should extend up to the most frequent 3,000 word families. At this point, we may suggest that the high-frequency vocabulary comprise around the 2,000 – 3,000 most frequent word families, which will provide approximately 80-90% coverage of texts.

The prominent examples of the high-frequency vocabulary lists of English include West’s (1953) *General Service List* (GSL), the British National Corpus (BNC) first 3,000 word frequency list, the British National Corpus and Corpus of Contemporary American
English (BNC-COCA) first 3,000 word frequency list, and the Oxford 3,000 headwords. Of these lists, by far the most influential and widely used wordlist is West’s GSL. The GSL is a list of around 2,000 of most frequent word families of English selected from a corpus containing approximately 2.5 million running words which was manually compiled in the pre-computer era in the early twentieth century (Browne, 2013, 2014a, 2014b). The criteria used for its word selection were the combination of quantitative and qualitative criteria among which frequency was the most dominant and even the most important (Nation, 2004).

The GSL has been an important resource for teachers and material writers for many decades. However, it also has been criticized over years for several problems primarily related to the principle on which it was created and to the corpus from which it was drawn. For example, Browne (2014a) highlights three issues: (1) the corpus from which the list was extracted is considered to be out-of-date for current English usage today (most of the texts were published before 1930), (2) the corpus is also too small in size according to modern standards (the original corpus contains only 2.5 million running words), and (3) the list did not clearly define what constitute a word. Richards (1974) also identifies a number of inconsistencies in the selection of the GSL wordlist. He exemplifies that it contains certain words from the semantic field of animals such as bear, elephant, and monkey, but excludes others such as lion, tiger, and fox. From the perspective of current corpus linguistic research, Brezina and Gablasova (2015) has also questioned the combination of objective and subjective criteria involved in compiling the wordlist, which has potentially brought a large element of subjectivity into its final product.

In response to the aforementioned problems identified with West’s original GSL, vocabulary researchers have recently made an attempt to create an update of the list. By far, two carefully constructed studies have reported on their results. First, Browne and his colleagues released an update of West’s GSL known as the New General Service List (NGSL). The NGSL contains around 2,800 high-frequency lemmatized words which is claimed to have fulfilled the following goals (Browne, 2013; 2014a; 2014b):

1. to update and expand the size of the corpus used (273 million words) compared to the limited corpus behind the original GSL (about 2.5 million words), with the hope of increasing the generalizability and validity of the list.
2. to create an NGSL of the most important high-frequency words useful for second language learners of English which gives the highest possible coverage of English texts with the fewest words possible.
3. to make a NGSL that is based on a clearer definition of what constitutes a word.
4. to be a starting point for discussion among interested scholars and teachers around the world, with the goal of updating and revising the list based on this input (in much the same way that West did with the original interim version of the GSL).
 Generated on the basis of the principles which combined between quantitative and qualitative measures, the NGSL provides the following strengths in comparison to the original GSL: (1) it was extracted from a larger, more balanced and modern corpus, that is, the 273 million-token sub-corpora carefully selected from the 2 billion-token Cambridge English Corpus (CEC), (2) it clearly defines what constitutes a word within the list (known as “modified lemmas”), and (3) it contains less lemmas with higher coverage across a range of different texts. This wordlist is available for free access or download at www.newgeneralservicelist.org. The associated resources integrating the list for analytical tools are also provided in this website.

Second, almost at the same time of Browne’s NGSL publication, another update of the original GSL was published by Brezina and Gablasova (2013). To avoid confusion, Brezina and Gablasova’s new GSL will be referred to as Other New General Service List (ONGSL) hereafter. The ONGSL was obtained by examining the vocabulary overlap among four corpora (LOB, BNC, BE06, and EnTenTen12) in the top 3000 frequent words based on purely quantitative criteria which include: (1) frequency, (2) dispersion, and (3) stability of a lexical item across different corpora. The total size of the four corpora were over 12 billion running words (tokens). The final product of their study was the new GSL (i.e. ONGSL) which consists of 2,494 lemmas and has a coverage of around 80% of the source corpora. However, it is worth noticing that in a number of comparative analyses carried out by Browne (2014a, 2014b), the NGSL provides higher and better coverage than both the original GSL and the ONGSL do across a variety of different texts.

3. Academic Vocabulary

Academic vocabulary is the next word frequency band for learners to focus on after the high-frequency vocabulary. This vocabulary consists of non-high-frequency vocabulary which occurs frequently in academic texts no matter what subject areas are being concerned (Nation, 2001). Therefore, this vocabulary is essentially useful for learners wishing to study at tertiary level where English is used as a medium. It is the most important vocabulary to understand a variety of academic texts.

The most influential list of the academic vocabulary is Coxhead’s (2000) Academic Word List (AWL) containing around 570 word families that are not in the most frequent 2,000 word families in West’s GSL, and occur so frequently across a wide range of academic texts. The list was made by examining the frequency (i.e. the number of times a word occur
in texts) and the range (i.e. the number of different texts where a word occurs in) of words across a variety of academic texts from a variety of subject areas which include Humanities, Commerce, Law, and Science (Coxhead, 2000). The whole corpus contains approximately 3 million running words or tokens. Typically, the AWL provides up to 10% coverage of the running words of academic texts, and is worth studying for learners in senior secondary school and university (Nation 2004).

Coxhead’s AWL was developed by means of West’s GSL as its baseline for identification of what constitute academic vocabulary from general vocabulary. This means that which words to include into the final list was under the control of West’s GSL. As pointed out previously, the original GSL has been updated for current language use. Therefore, there is also a need to update the original AWL by means of the updated version of the original GSL as the general vocabulary baseline.

This attempt has been taken by Browne and his colleagues. They have developed a New Academic Word List (NAWL) to work in conjunction with the NGSL in the same way that Coxhead's AWL worked in conjunction with West’s original GSL. The NAWL is a list of 963 lemmas carefully selected based on frequency, dispersion, and appropriateness from a carefully selected academic corpus containing about 288 million running words (Browne, Culligan and Phillips, n.d.). The complete list is available at http://www.newacademicwordlist.org.

4. **Technical Vocabulary**

There are some vocabulary reasonably common in a particular subject area but not so common in other areas. This type of vocabulary is referred to as technical vocabulary. This vocabulary consists of vocabulary that is very closely related to a topic or subject area of a text (Nation, 2001). Therefore, this vocabulary usually differs from subject area to subject area. It is important to realize that this vocabulary could come from either high-frequency vocabulary, academic vocabulary or low-frequency vocabulary.

The technical vocabulary typically covers 20-30 per cent of the running words of a specialized text (Chung and Nation, 2003). Its size may vary from around 1,000 to 5,000 word families (Nation, 2008 as cited in Matsuoka, 2012). For example, the technical vocabulary of anatomy would include words like xiphoid, vascular, neck, chest, skin, superior, posterior, transverse, cranial, disease and girdle (Chung and Nation, 2003). Some of this vocabulary are found in the high-frequency words, some may be in the academic
words and much of it consists of words occurring only in the fields of medicine and anatomy. This technical vocabulary is worth focusing for learners who are specializing in that fields.

5. Low-Frequency Vocabulary

In contrast to the high-frequency vocabulary, low-frequency vocabulary consists of vocabulary that occurs so infrequently in texts. This vocabulary is found to be the largest group of words among others, but covers only a small proportion of the running words of any text (Nation 2001; Schmitt and Schmitt 2014). It accounts for around 5% of the total running words in most texts (Chung and Nation, 2003). This vocabulary includes all the remaining words which are not in the high-frequency vocabulary, not in the academic vocabulary, and not in the technical vocabulary for a particular subject area (Nation, 2001).

In a more operational description, vocabulary scholars use a word frequency level band to define what words are in this group and to set its boundary. The traditional boundary was thought at the 10,000+ level, which means any vocabulary existing beyond the 10,000 frequency level. More recently, Schmitt and Schmitt (2014) propose in order to lower this boundary to the 9,000+ level, on the basis of a more recent corpus study by Nation (2006) which concluded that the most frequent 8-9,000 word families are sufficient to provide the vocabulary resources needed to be able to comprehend a wide range of English authentic texts.

Due to its high infrequent occurrence in text, teachers and learners should not spend too much time on the low-frequency vocabulary. This also suggests that the low-frequency vocabulary should not occupy the majority of vocabulary input in teaching materials, student coursebooks as well as teacher’s talks in classroom. It is not worth spending classroom time on such vocabulary. Rather, teachers should teach particular learning strategies to learners in dealing with such vocabulary. These strategies, for example, include guessing from context clues, using word parts to help remember words, using vocabulary cards and dictionaries (Nation, 2001)

6. Mid-Frequency Vocabulary

Mid-frequency vocabulary is a new category of words coined by Schmitt and Schmitt (2014). The previous sections have mentioned that the high-frequency vocabulary consists of the most frequent 3,000 word families, and that the low-frequency vocabulary starts from about the 9,000 frequency level. This suggests that there remains a great gap between the 3,000 and 9,000 levels which are not completely covered in both the academic and technical
vocabulary. Schmitt and Schmitt (2014) then introduce “mid-frequency vocabulary” to label this in-between vocabulary. Figure 1.1 shows the frequency continuum of three categories of vocabulary proposed by Schmitt and Schmitt (2014: 493):

In this conception, the mid-frequency vocabulary covers the following groups of words: the academic vocabulary which is frequent in all academic discourse, some technical vocabulary which is common in a particular subject area, and all the remaining of the in-between vocabulary which is not in the academic and technical vocabulary. Schmitt and Schmitt (2014) have highlighted some important benefits of learning the mid-frequency vocabulary. Among them is that the combination of high- and mid-frequency vocabulary will provide sufficient vocabulary resources necessary to be able to comprehend a wide range of authentic texts such as movies, TV programs transcripts, news broadcasts, newspapers, novels and magazines. Such combination typically provides around 95-98% coverage of the running words of the authentic texts. The figure of 95-98% is the lexical coverage threshold generally agreed among researchers in order to get an accepted comprehension of a text (Laufer and Ravenhorst-Kalovski, 2010; van Zeeland and Schmitt, 2012)

7. Final Words

The importance of the high-frequency vocabulary and its value to learners is clear. It contains words that provide lexical foundation to involve in daily basis communication. In this case, material writers (also teachers) are in a position to ensure such vocabulary available to learners, which means that the learners should get sufficient and favourable opportunities to learn such important vocabulary. Several computer programs to date have been available for lexical analysis such as www.lextutor.ca and http://www.er-central.com/ogte. Material writers and teachers can take advantage of such programs to analyse and compare teaching
materials with standard wordlist available, so that the inclusion of low-frequency vocabulary in the coursebooks or teaching materials could be minimized as much as possible.

8. References


