6

Frequency as a Guide for Vocabulary Usefulness

High-, Mid-, and Low-Frequency Words

Laura Vilkaitė-Lozdienė and Norbert Schmitt

Introduction

If a person wants to acquire a second language, learning its vocabulary is definitely an important task. However, we cannot teach or learn all the words in a language, as there are simply too many of them. Therefore, some decisions need to be made, and some words have to be prioritized. The best way of choosing which words to teach or to learn depends on the purpose of learning. For example, if a person wants to become proficient in a specialized area (e.g., medicine), teaching a list of vocabulary items specific to that area might be the most useful approach (assuming a foundation of general English is already in place). If an academic tone in writing is desired, then working with words and phrases drawn from academic corpora might help the learner achieve this goal. But if the learning purpose is more general (e.g., to be able to read an article online or to be able to converse while traveling in a foreign country), then a way of selecting the most useful non-specialist vocabulary is necessary. For these "general" purposes, frequency has proven a very useful tool.

There are a number of reasons why frequency is important. First of all, the idea of the importance of frequency can be explained by the Zipf's law: "[w]ords occur according to a famously systematic frequency distribution such that there are few very high-frequency words that account for most of the tokens in text . . . and many low-frequency words" (Piantadosi, 2014). To put it another way, we can focus on a limited number of high-frequency words and achieve comprehension of most of the running words in any text. Therefore, unsurprisingly, frequency has a proven history in aiding language pedagogy: frequency lists have been used for decades for teaching the most useful general words (since West, 1953 and before), and lexical coverage studies have tried to establish how many words students need to understand in order to cope with English. Psycholinguistics studies also show that the frequency of words predicts their reading difficulty both in L1 and in L2 (van Heuven, Mandera, Keuleers, & Brysbaert, 2014). So frequency is not only a textual phenomenon which can be observed when analyzing corpora but it also has psychological validity. Finally, usage-based theories of language acquisition emphasize the effect of frequency when acquiring both individual words and tendencies of use of word sequences (Bybee,

2006, 1998; Ellis, 2002). Overall, frequency seems to have an effect on word processing and its acquisition.

For language pedagogy, there are two aspects regarding frequency to take into account:

- Frequent words are more important, as they are encountered more often than less frequent words (lexical coverage studies, Zipf's law).
- Words that are encountered more often have better chances of being learned.

In this chapter, we will only consider the first claim: high-frequency words are the most useful ones and they give learners the best value for their study effort. Thus, they need special attention in a language classroom.

Frequency is a good guiding criterion for word selection as it is very straightforward and objective. While knowing word frequency itself does not help much to decide on whether to teach a particular word or not, it can be used to divide words into groups (e.g., high-frequency, low-frequency) and to select a reasonable number of words to teach. An important question then remains where the best and the most meaningful cut-points for frequency bands should be, and what we should do with words labeled as high-frequency, mid-frequency, and low-frequency. In this chapter, we discuss these questions, as well as some limitations in the current frequency framework. Finally, we will offer some initial suggestions of where to move next.

Critical Issues and Topics

We will start by looking at the usefulness of frequency as a guiding criterion for choosing which vocabulary to teach, beginning with a brief discussion of the historical development of the idea of frequency in language pedagogy. We will then move on to the current understanding of high-, mid-, and low-frequency vocabulary.

Historical Development of Frequency in Pedagogy

While teaching and learning foreign languages has been relevant for thousands of years, there used to be no principled way to handle foreign language vocabulary, other than focusing on whatever words happened to occur in a text of interest. Furthermore, grammar has received the lion's share of attention in most traditional classrooms. Vocabulary started to be systematically approached only in the early 20th century, with a strand of lexical research attempting to make vocabulary easier by limiting it to some degree. This was known as the *Vocabulary Control Movement*.

In the early 1930s, K. Ogden and I.A. Richards developed a *Basic English*: a vocabulary of 850 words. It was supposed to be quickly learned and express any meaning that could be communicated in regular English. But the 850 words were so polysemous (with an estimated 12,000 meaning senses) that it was not really that much of a simplification, and so Basic English did not end up having much of an impact. Another approach in the Vocabulary Control Movement was to use systematic criteria to select the most useful words for language learning, developing principles of presenting common vocabulary first, and limiting the number of new words in any text. This approach was much more successful, and culminated in the *General Service List of English Words* (*GSL*) (West, 1953). The GSL was a list of about 2,000 words based on word frequency but also on structural value, universality, subject range, definition words, word-building capacity, and style (Howatt, 1984). The GSL had the

advantage of listing different parts of speech and different meaning senses, which made it much more useful than a simple frequency count.

With generative grammar and Chomskian ideas (e.g., Chomsky, 1986), the focus on vocabulary faded for a time. But in the later 20th century, computerized corpora became available which allowed the quick and reliable calculation of frequencies, and also the identification of patterns of vocabulary occurrence. Before the introduction of computerized corpora in the 1960s, the majority of linguistic studies were based on a small number of examples, quite commonly invented by a researcher (Hunston, 2012). The development of computers and the ability to collect, store, and analyze millions of word occurrences had a large influence on linguistics. The importance of frequency also became established with corpus research: linguists' attention has shifted from what is possible in language towards what is typical and used frequently (Barlow, 2011). Word counts have provided some very useful insights into the way the vocabulary of English works, and helped to rejuvenate interest in vocabulary issues.

Around the same time, Paul Nation led the way in focusing attention on vocabulary from the pedagogical perspective. He designed a program called RANGE (available on his website[1]) for analyzing the vocabulary of any text. The program divided vocabulary into 1,000, 2,000, and off-list items, so essentially it set in place a high-frequency/low-frequency dichotomy. In Nation's research (e.g., 2001a) high-frequency words were considered to be the first 2,000 most frequent word families, then there were academic words (initially the *University Word List*, and later the *Academic Word List* (Coxhead, 2000)). Other words that were not included in these established categories were labeled as low-frequency.

In 2014, Schmitt and Schmitt introduced an idea that had been floating around the vocabulary community for a while: the words beyond the 2,000 frequency band are important and should not be considered just off-list items. They suggested moving the boundary for high-frequency vocabulary to 3,000 word families, and introduced the term *mid-frequency vocabulary*.[2] The most commonly used current frequency framework revolves around Schmitt and Schmitt's high/mid/low-frequency categories. Therefore we will look at it in more detail in later sections.

Related Concepts

In order to talk about the current understanding of high-, mid-, and low-frequency words, two important related concepts need to be introduced and discussed. First, it is important to consider how a **word** is defined and what is counted as a word in any frequency count. We will discuss the differences between two commonly used counting units – **word families** and **lemmas**. Second, figures cited as thresholds for high- or low-frequency words are often adopted from **coverage** research. Therefore, it is important to briefly discuss this research as well.

Word Family and Lemma

It seems to be intuitively clear that researchers should agree on a unit of counting if they want to have reliable frequency counts across different studies (Bauer & Nation, 1993). At the moment, most coverage research and frequency lists are based on *word family* as a counting unit. Word family is a unit that includes both the base form of the word (such as, *work*) and its inflections (e.g., *worked*, *works*), as well as its main derivations (e.g., *worker*). The idea of the word family was introduced in order to systematically approach vocabulary in language

pedagogy (Bauer & Nation, 1993). The assumption is that once learners know a meaning of a base word (such as *work*) and have some knowledge of morphology and meanings of the main affixes, they do not need to learn each single word in a language but instead can derive the meanings of word family members (such as *worker*) from the base form. Hence using word families in a way reduces learning burden for L2 learners because they can systematically infer the meaning of the word family members. Especially for the receptive language use (reading or listening) this idea seems to be reasonable. Also, word families seem to have psychological validity and to be represented in the mental lexicon (Nagy, Anderson, Schommer, Scott, & Stallman, 1989). Hence there is an additional argument (apart from lowering the learning load for the learners) to use them as counting units.

However, Bauer and Nation (1993) have claimed that "[a]s a learner's knowledge of affixation develops, the size of the word family increases" (p. 253). While this idea makes sense when considering individual learners, in practice researchers need to pre-set a list of tokens constituting a word family and use it for coverage research, frequency lists, or vocabulary testing. Hence, the common assumption becomes the following: once a base word is known, the whole word family is known. Because of this, the word family approach is problematic. As Nation (2001b) has pointed out, the main problem with using word families is deciding which word forms should be included in a word family. Some affixes, and consequently some word family members, are transparent and thus easy to decipher, but others may not be. Also, what seems transparent for one learner can be beyond the level of comprehension for a different learner (Nation, 2001b). So it is not easy to clearly define a word family that would apply to all language users. Schmitt and Zimmerman (2002) have shown that learners do not necessarily have productive knowledge of all the members of a word family. Kremmel (2016) has also suggested that if a base word is known, it does not mean that all of its word family members are known. Derivation seems not to be easily acquired (at least productively). For example, González Fernández and Schmitt (2019) have asked learners to provide derivational forms of all four word classes (nouns, verbs, adjective, and adverbs) for each of their target words and they have found that learners typically can recognize forms of two to three word classes in a word family, but not all the derivatives. Regarding the ability to produce the word forms, their participants could typically only produce two out of the four forms. Gardner and Davies (2014) have also pointed out that the word family does not take into account part of speech information, and that some members of a word family can be quite distant from each other in terms of their meaning (e.g., *process*, *proceeding*). Overall, it might be that the concept of word family is more problematic for productive rather than for receptive knowledge, but even receptively it may not work reliably. If the learners do not understand/know the members of the word family, then the assumption underlying word families fails.

There are also two more issues with word families. The first one is technical: lists based on word families are more difficult to compute automatically than lists based on lemmas or word forms. The second one is more pedagogical. Teachers and learners (and even researchers) might misinterpret figures based on word families when using research outputs (e.g., lists of words, targets for learning), and simply understand them as "individual words". This could lead to a misleading sense of the vocabulary learning required. Because of all these issues, word families seem to be useful counting units when we are dealing with receptive knowledge and with advanced learners or even native speakers. But they do not reliably work in all the situations with all the different language learners.

Therefore, recently there have been suggestions to move from word families to lemmas as counting units (Kremmel, 2016). Lemmas can be defined as "words with a common stem,

related by inflection only and coming from the same part of speech" (Gardner & Davies, 2014, p. 4). As such, the lemma is a much more straightforward unit: no arbitrary decisions about what to classify as the same lemma need to be made. Because of that, it is much easier to operationalize lemmas computationally as well. Also, it easier and safer to make assumptions about learners' knowledge as inflectional affixes tend to be regular and learners do not need to reach an advanced level to recognize and understand them.

To date, most coverage research has been based on word families. Considering the problems with the word family approach, there has been some discussion of using lemmas instead. However, a lemma-based approach also has limitations. Lemmas might be too restricted for counting, as some derivational affixes are usually transparent (such as *–er* to indicate an agent noun) and do not cause difficulties for learners. On the other hand, lemmas can be used with different inflections, so further research would be needed to establish whether lemmas are not problematic receptively, at least for beginner learners. Finally, it seems that the members of a word family also follow the Zipfian distribution (the most frequent member is much more frequent that the others). So it remains an empirical question if moving from word families to lemmas would actually change much in our understanding of high/mid/low-frequency vocabulary. Overall, it is obvious that both lemmas and word families have limitations, so at the moment there is no way to strongly favor one or the other. Further research on the validity of those two counting units would be very useful to make more informed decisions in the future.

As most of the currently available research is based on word families, most of the calculations in this chapter will be presented in that unit. We will also introduce some suggestions about how the thresholds for high- and low-frequency vocabulary might change if we moved to lemmas instead of word families.

Lexical Coverage

Lexical coverage studies mostly can be divided into two groups:

- Studies focusing on how many words of a text/listening passage one needs to know in order to gain adequate comprehension: coverage as a "percentage of known words in a piece of discourse" (van Zeeland & Schmitt, 2013, p. 457).
- Studies focusing on the frequency profile of certain texts: "Coverage refers to the percentage of tokens in a text which are accounted for (covered by) particular word lists" (Nation, 2004, p. 7).

These two approaches complement each other. For pedagogical purposes, we must first determine the percentage of words in a text a learner needs to know in order to understand it, and then we must establish how many lexical items one needs to learn to reach that percentage.

The lexical thresholds for comprehension have been estimated both for written and for spoken texts. In 1989, Laufer suggested that one needs to comprehend about 95% of a text in order to be able to understand that text. Later she suggested that 3,000 word families constitute a lexical threshold required for reading comprehension (Laufer, 1992). This number has been refined, and now two thresholds for comprehension have been suggested: an optimal one, which is the knowledge of 8,000 word families yielding the coverage of 98% (including proper nouns), and a minimal one of 4,000 to 5,000 word families resulting in the coverage 95% of texts (Laufer & Ravenhorst-Kalovski, 2010). Hsueh-Chao and Nation (2000) have

also suggested that we need to understand 98% of written texts in order to comprehend them, while Schmitt, Jiang, and Grabe (2011) have concluded that there is no clear threshold: comprehension increases almost linearly as coverage increases. However, they suggested that 98% seems to be a reasonable threshold. For listening, on the other hand, it seems that a somewhat lower threshold can work. For example, van Zeeland and Schmitt (2013) have shown that even at 90% coverage levels most of their participants showed adequate comprehension, but at the 95% level, there was less individual variation.

Once the required lexical coverage for comprehension is established, the next question is how many words (lemmas/word families) one needs to acquire to achieve this percentage. This number will depend on what type of texts one wants to read or listen to. For spoken language, corpus research examining the CANCODE corpus seems to show that 2,000 word families are enough for almost 95% coverage (2,000 word families provide 94.76% coverage while 3,000 provide 95.91% coverage) (Adolphs & Schmitt, 2003). For written language, the numbers are higher. For example, Hirsh and Nation (1992) analyzed three novels and calculated that about 5,000 word families are needed to achieve 97% to 98% coverage. For more challenging material, such as academic texts, this number might be even higher.

Nation (2006) summarized the relationship between coverage and frequency bands as illustrated in Figure 6.1. From this figure, it becomes clear that the first thousand words provide by far the highest coverage – about 80% of all texts. But this is partly because of the extremely high frequency of function words, which are almost solely in this frequency band. For example, the first 100 most frequent words in English (virtually all function words) cover about 49% of the running words in texts (Nation, 2001b). The coverage of following frequency bands consistently becomes smaller, and all the words less frequent than the 14K level actually cover only about 1% of English.

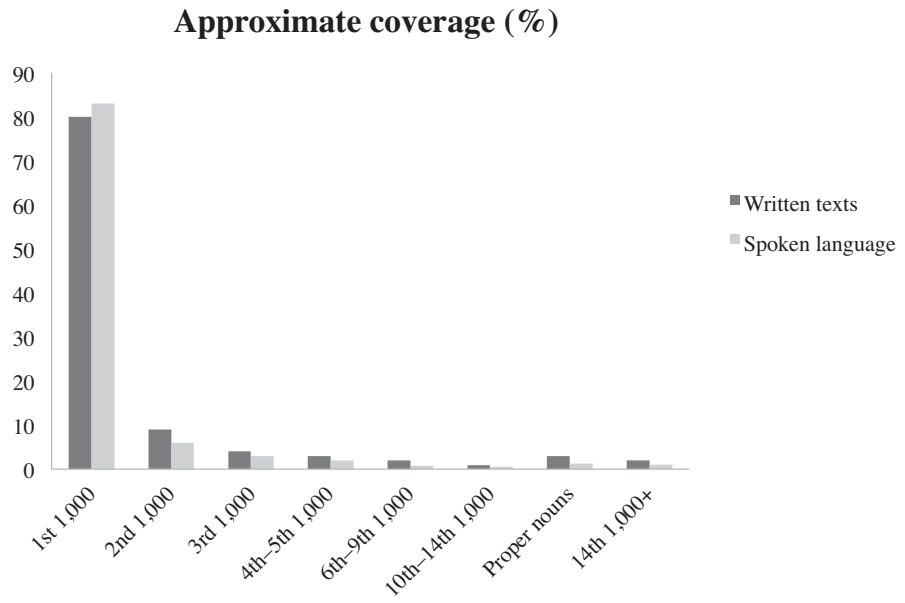

Figure 6.1 Vocabulary size and text coverage

It is important to notice that if one needs, for example, 3,000 word families to reach 95% coverage of a written text, these families are not simply any words in a language, but specifically the most frequent 3,000 ones. This emphasizes the importance of high-frequency words even more. Learners should aim to know enough words to achieve the required coverage of a text/listening passage in order to understand it adequately. Therefore, the thresholds for high-, mid-, and low-frequency words are usually calculated with coverage in mind, so that the resulting figures can indicate what learners knowing this vocabulary can do.

High-, Low-, and Mid-Frequency Words

High-Frequency Words

High-frequency vocabulary consists of words that are the most frequent in language and consequently provide the highest coverage. Therefore, teachers should prioritize high-frequency words because they are the most useful (Nation, 2001a; Read, 2004). The importance of the high-frequency words is not a new idea. As already noted, a very influential list of the most important words (*GSL*) was created in 1950s (West, 1953). While frequency was not the only criterion for creating this list, it was one of the main ones. The need for a list that can be used for pedagogical purposes has not disappeared. New versions of the lists of general vocabulary are continuously being created, such as the *New General Service List* (Brezina & Gablasova, 2015), which lists about 2,000 words occurring across various corpora, and making up core high-frequency vocabulary. While the original GSL is now rather out-dated, in general, the most frequent words tend to remain relatively stable across both time periods and corpora. For example, Nation (2004) compared the GSL with the most frequent word lists compiled based on the BNC and showed that they contain much of the same vocabulary. Hence, it seems the most frequent words in the language are relatively stable, no matter which corpus you use, and do not change quickly over time (Kilgarriff, 2007; Nation, 2004).

There is a general understanding that high-frequency words are important. However, the question of where to draw the line defining high-frequency remains. Different criteria for identifying high-frequency words have been considered, such as relying on coverage research and reaching the 95% lexical threshold, the range of words in different texts and frequency lists, feasibility of teaching these words in a language course, a cost-benefit analysis, etc. (Nation, 2001a). Nation (2001a) suggested that 2,000 most frequent word families should be labeled as high-frequency vocabulary. This figure of 2,000 has been widely cited in teacher guidebooks and research publications (e.g., Nation, 1990; Read, 2000; Schmitt, 2000; Thornbury, 2002). Nation has also set this frequency level for his text coverage analysis tool (*VocabProfiler*, www.lextutor.ca) and his Vocabulary Levels Test (Nation, 1983; Schmitt, Schmitt, & Clapham, 2001), effectively establishing this threshold for high-frequency vocabulary. However, Nation (2001a) himself has clearly stated that this decision is open to debate. Setting the threshold to 3,000 words has also been suggested (e.g., Schmitt & Schmitt, 2014; Waring & Nation, 1997). After Schmitt and Schmitt's (2014) paper, the boundary of 3,000 word families for high-frequency words is becoming more accepted, because the learners who reach this level are able to communicate in a range of situations. Also, the 3,000 most frequent word families often approach the 95% coverage level for many texts (see Figure 6.1).

However, it has to be noted that there are also some studies that suggest lowering the threshold for high-frequency words rather than increasing it. They show that students fail to learn 2,000 most frequent words so potentially this goal is too ambitious and a more realistic

goal of 1,000 words should be used for defining high-frequency vocabulary (Dang & Webb, 2017). However, a better way of thinking about this is probably that the 2,000 to 3,000 most frequent words are necessary to engage with English in useful ways, and so they should be considered high-frequency vocabulary (see Figure 6.1). But in terms of *learning goals*, 1,000 words may well be a suitable *initial* goal. Dang and Webb (2017) also note that the cumulative coverage of the frequency bands beyond the first 1,000 drops considerably, suggesting that the first 1,000 most frequent words are clearly the most useful ones. But this disregards the fact that the drop is mainly caused by the function words occurring in the first 1,000 words. If function words are taken out of the frequency profile (as makes pedagogic sense because function words are not typically taught as vocabulary items, but rather as grammar items), then the drop in coverage of *content words* is much more gradual (see Kremmel, 2016 for an illustration of this). Still, we must always take account of how students learn, and so it may often be useful pedagogically to sequence high-frequency items into the essential vocabulary (be it 1,000 words (Dang & Webb, 2017) or 800 lemmas (Dang & Webb, 2016)) to start with at the beginner level, and then move onto the other high-frequency words (up to 3,000 words) required to use English in many contexts.

If we set the threshold for high frequency vocabulary at the 3,000 most frequent word families, we must decide how to best approach this vocabulary in a language classroom. Nation (2001a) has suggested that teachers should directly teach high-frequency words, and students should deliberately learn them using word cards or dictionaries as necessary. Explicit direct teaching seems to be important because even these high-frequency words may not be frequent enough for the learners to get enough exposure to learn them incidentally from reading (Cobb, 2007). Therefore, these words should be the focus of the language syllabus. They can be addressed in various vocabulary exercises, used in graded readers, or even provided as target lists for learning for language learners. Teachers should probably start focusing on the 1,000 most frequent content words first as they are will have the most value for their learners.

However, while high-frequency words are essential, this does not mean that teachers should be completely driven by frequency information. Nation (1990) has noted that many words important for classroom context and classroom management (e.g., *pencil*, *blackboard*) are not necessarily frequent in general English, but will definitely be important in a classroom setting. Also, depending on learners' age, some of the high-frequency words might be not be useful for them. Hence, frequency lists should be seen more as a useful indication rather than a prescription.

The figure of 3,000 cited earlier is based on word families. If the field moved to lemmas as counting units, how would the high-frequency threshold change? It might seem that we need many more lemmas than word families, as a word family on average contains from one and a half to four derivations depending on how inclusive the definition of a family is (Nation, 2001b). Hence, one word family could translate to even more than four lemmas (e.g., nation, nationalize, national, nationally, international, internationalization, and so on). Consequently, the borderlines for high-frequency and mid-frequency vocabulary should increase if we moved from word families to lemmas. However, this increase might be not as large as might be supposed. Some current research shows that around the 3,000 most frequent lemmas are enough to reach 98% coverage in conversations (based on BNC data, Schmitt et al., under review). If this is the case, then we might actually be able to leave the definition of the high-frequency lemmas the same as we had for high-frequency word families – 3,000 – as this amount would still be enough for learners. Furthermore, counting in lemmas would not entail assumptions of knowledge of any lower frequency word family

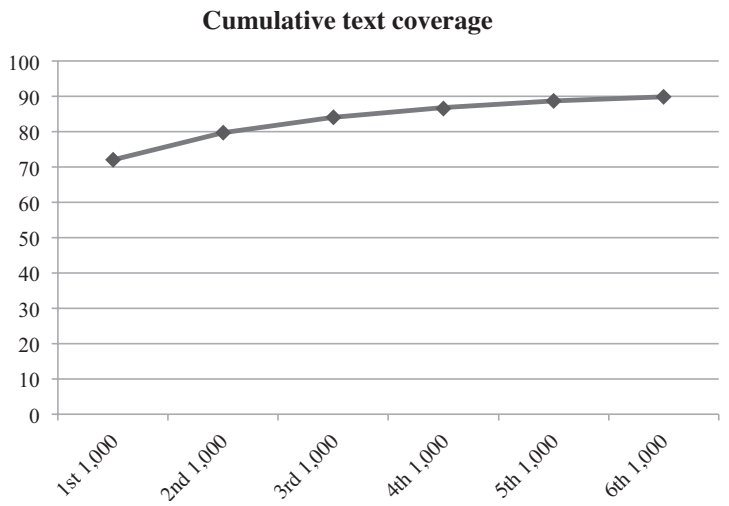

Figure 6.2 Vocabulary size (in lemmas) and cumulative text coverage in the Brown corpus

members. However, regarding reading, a larger threshold might be needed. For example, Schmitt et al. (under review) have calculated that about 6,000 lemmas are needed to reach the 95% threshold for reading adult fiction. Waring and Nation (1997) have also provided figures based on lemmas derived from the Brown corpus (see Figure 6.2 for a summary). In their study, the 3,000 most frequent lemmas give coverage of about 84% of the corpus. It is less than word families would, but not massively so. The actual threshold for high-frequency words based on lemmas is yet to be established, but it does not seem to change dramatically from the one we have for word families.

Low-Frequency Words

Let's now look at the other extreme of the frequency continuum – low-frequency words. What to consider low-frequency is not so well-established. Nation (2001b), for example, has divided vocabulary into high-frequency vocabulary, academic words, technical words, and low-frequency vocabulary, but does not clearly indicate where the low-frequency words begin. Schmitt and Schmitt (2014) have suggested that 9,000+ word families should be labeled as low-frequency words. However, they admitted that a clear boundary is difficult to establish because each different 1000-word level from around 9,000 words does not add much to the coverage. Hence, they have based their threshold on the fact that the 9,000 most frequent word families in the Nation's BNC frequency lists cover 95.5% of the Corpus of Contemporary American English (COCA). As COCA is a huge collection of language, they concluded that reaching an adequate coverage of such a vast variety of texts is a good criterion to define everything beyond that as low-frequency vocabulary.

Nation (2001b) has noted that low-frequency vocabulary consists of some words that are moderately frequent but do not make it to high-frequency, and also includes proper names that are usually quite easy to identify and understand. There are also some words that in general language are infrequent, but actually belong to a specific field. For some people these words will be important and widely used, while for others they will be low-frequency vocabulary. After that, there are a lot of words that are simply very rare in a language. This

distinction is important, as specialist vocabulary will definitely be useful for people learning English for specific purposes, while words that are simply very infrequent for everybody do not require much pedagogical attention. For example, *scalpel* is low frequency in general, but important specialist vocabulary for surgeons. Conversely, *umbrage* is a low-frequency word that is unlikely to be particularly useful for anyone. As the benefits of learning low-frequency words in terms of added coverage are rather limited, and there are so many them, it is not very useful to dedicate a lot of classroom attention to low-frequency words. Rather, Nation (2001b) advocates focusing on learning strategies, such as guessing their meaning, drawing on word part knowledge, or using dictionaries to deal with these words.

If we move towards lemmas as counting units, where should we draw the line for low-frequency vocabulary? We could follow the criterion of Schmitt and Schmitt (2014) and ask how much vocabulary we need to reach the 95% to 98% coverage of a broad range of texts, as illustrated in a large corpus. Based on the recent corpus study (Schmitt et al., under review) a not overly large number of lemmas are needed to reach this coverage: e.g., around 9,000 lemmas are enough for understanding academic writing, and 11,000 for comprehension of adult fiction. Spoken language requires much less: about 3,000 lemmas to reach 95% coverage of conversations, and around 6,000 lemmas to reach this level of coverage of TV shows. Magazines seem to need the largest number of lemmas – about 14,000 lemmas. So, if we took the largest figure as a threshold between mid-frequency and low-frequency vocabulary, it would be around 14,000 lemmas. If, on the other hand, we sought some kind of average of words needed for understanding various kind of texts, it would seem to be around 10,000 to 11,000 lemmas. This figure is larger than the before-mentioned 9,000 word families, but using lemmas as counting units would not push it unrealistically too far. While more research is needed to be able to reliably draw the threshold for low-frequency words in lemmas, this number seems to be a couple of thousands rather than several times larger than for word families. This is probably the case because most of the coverage of the typical word family is usually provided by its most frequent member(s).

Mid-Frequency Words

Schmitt and Schmitt (2014) labeled the vocabulary between high-frequency words (3,000) and low-frequency words (9,000+) as *mid-frequency vocabulary*. Mid-frequency vocabulary is a relatively new term. This frequency range includes words that were traditionally labeled as academic vocabulary and technical vocabulary, but also some of the other words that are more frequent than in the 9,000 frequency band. Nation (2001b) noted that beyond the high-frequency level, one's vocabulary grows depending on one's interests, jobs, professions, and therefore is more idiosyncratic. Gardner (2013) has also noted that beyond the very highest frequency bands, words start becoming increasingly more domain-dependent and frequency lists drawn from different corpora could differ considerably at each thousand frequency band. Because of that, it is not that easy to provide mid-frequency lists useful for the majority of learners. Therefore, after the high-frequency words are acquired, the words to teach depend on the specific interests and needs of the learners (Read, 2004).

Mid-frequency vocabulary is essentially a gap learners need to fill in order to move on to reading authentic texts, especially in academic contexts. For instance, Nation's (2006) and Schmitt et al.'s (under review) research has indicated that learners need to know many thousands of items beyond high-frequency vocabulary to read virtually any kind of authentic text in English. So these mid-frequency words might not be relevant for all the L2 learners: for those who have less ambitious goals in a second language (e.g., just casual conversation),

high-frequency words might suffice. But for those that have more ambitious goals in learning an L2, such as studying or working in an L2 environment, these words will be very important.

Schmitt and Schmitt (2014) reviewed a number of studies and have shown that mid-frequency vocabulary is not addressed well in classrooms at the moment: textbooks do not typically cover it systematically, and teachers do not focus on it or use it enough in their classroom talk. We do not have conclusive research on this band of vocabulary to give very specific advice on how to deal with it and more studies are needed.

Mid-frequency words can be acquired incidentally, but the research suggests leaving them all to be picked up in this manner is problematic. To start with, studies on incidental word acquisition from reading (e.g., Day et al., 1991; Horst, 2005; Horst, Cobb, & Meara, 1998; Pellicer-Sánchez & Schmitt, 2010; Pigada & Schmitt, 2006) show that incidental word acquisition is possible, but rather slow and a number of encounters is needed for the acquisition to occur. This does not mean that incidental acquisition is not useful: different aspects of word knowledge seem to be enhanced and the depth of vocabulary knowledge is increased. However, for acquiring a form-meaning link of new words, incidental learning is not very effective and its gains are modest (Read, 2004). This is partially the case because words beyond high-frequency level are simply not encountered frequently enough (Cobb, 2007).

One way of addressing mid-frequency words is focusing on lists of specific words for specific purposes, such as academic vocabulary (Coxhead, 2000) or Engineering English (Hsu, 2014). Another approach is to promote increased exposure, e.g., through extensive reading and graded readers. There are now suggestions to write mid-frequency readers that focus on recycling mid-frequency vocabulary (Nation & Anthony, 2013; see also Paul Nation's website[3] for a number of these readers free of charge). Computerized programs for learning vocabulary can also be a way to give students exposure and enough recycling of new vocabulary (Cobb, 2010). This seems to be a viable potential way forward, as nowadays computers/personal notebooks/smartphones are owned by most schools and learners.

Final Remarks

If one wants to learn a second language for general purposes, frequency seems to be an objective and useful criterion to prioritize words to focus on. It does not require any subjective judgment and frequency lists can be easily obtained for many languages. For English, there seem to be around 3,000 most frequent word families that deserve explicit direct attention, in order to give learners the best chances to be able to comprehend the majority of the words in a wide range of texts. If learners want to be able to use English in a variety of contexts, engaging in conversations and reaching an optimal understanding of written texts, knowledge of the 9,000 most frequent word families (i.e., mid-frequency vocabulary) is necessary. Above this threshold, low-frequency vocabulary becomes relatively rare, and it is of little pedagogical relevance unless the words are domain-specific. Dividing vocabulary based on frequency is not the only way to choose what to teach/learn first. Especially after high-frequency words are mastered, the needs of individual learners might differ considerably. However, at least the high-frequency end of the continuum is definitely worth attention in any classroom, as it provides a platform for all language use.

So far we have only focused on high-, mid-, and low-frequency words for choosing how many words and which words to teach. However, it is worth briefly mentioning that frequency bands are important for other applications as well. One of them can be language assessment. For example, Laufer and Nation (1995) suggested calculating learners' lexical

profiles in order to measure their lexical knowledge in L2 writing. According to these scholars, the amount of high-frequency words one uses in their writing is an indication of their overall L2 command, and discriminates between learners of different proficiencies. Most vocabulary size tests are also based on the idea of frequency as a sampling rationale, e.g., Meara (1992), Nation and Beglar (2007), Schmitt et al. (2001), and Webb, Sasao, and Ballance (2017) sample words from frequency bands in order to estimate vocabulary size.

Future Directions

We would like to propose three potential directions for future research: incorporating formulaic sequences in our understanding of high-frequency vocabulary, triangulating research findings, and reconsidering the idea of word family. We will discuss these in turn.

The first suggestion is moving away from only individual words, and incorporating formulaic sequences into our vocabulary lists, coverage studies, and subsequently into our understanding of what high-, mid-, and low-frequency vocabulary consists of. Formulaic language is made up of "combinations of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization" (Erman & Warren, 2000, p. 31). There are a number of reasons why taking formulaic language into consideration is important. First, it seems to make up from one-third to one-half of discourse (Conklin & Schmitt, 2012) which means it is very widespread in language use. Second, if individual words are known, but the meaning of a figurative sequence is not, it hinders understanding (Martinez & Murphy, 2011). Hence, we can calculate a lexical threshold needed for an adequate comprehension of a text, but if it contains idiomatic formulaic sequences, the vocabulary figures will tell us little of the actual comprehension a learner will achieve.

Ideally, we would like to have a frequency list that incorporates both words and formulaic sequences; then we could recalculate the coverage figures and divide vocabulary based on frequency of lexical items instead of the frequency of single words. The idea is not new. Waring and Nation (1997) also mentioned that some idioms and expressions behave like high-frequency words, while Schmitt and Schmitt (2014) have discussed the limitation of looking only at individual words. However, so far, we have no list that includes both words and longer lexical items.

Second, replication of reported studies and triangulation of the methods employed would be extremely useful. So far the research on high-, mid-, and low-frequency vocabulary is based mostly on corpus data or on studies of lexical coverage. Coverage figures are mostly drawn from studies where participants read texts with certain percentages of words missing. However, more empirical evidence from quasi-experimental classroom studies would be useful to see what learners can actually do in real-world teaching contexts. Questions like the following would be interesting to ask:

- What can learners actually do if they know only high-frequency words?
- Do typical learners actually learn mid-frequency vocabulary, or can they survive perfectly well without it in various contexts?
- What strategies do learners use to cope with texts if they do not know low-frequency words? Or do they merely skip over them without problems?

We believe integrating corpus results with empirical research on actual learner performance would go some way in determining the amount of vocabulary learners need to get things done in language.

Finally, we feel the idea of word family has to be critically reconsidered. The assumption underlying word family is that learners can recognize the members of a word family, so it is a meaningful unit to use for calculations of receptive knowledge. This assumption has been challenged: learners typically recognize some but not all members of a word family and perform even worse on a productive level. This suggests moving to lemmas as a more pedagogically sound counting unit. But this might not be the optimal solution either. We have to admit that learners can typically use some very frequent and consistent affixes, e.g., *–er* (*learn/learner*). Therefore, the lemma might actually be too small a counting unit. Ideally then, we could like to use lemma +, that is a lemma with a few transparent affixes that learners can consistently and reliably comprehend/produce. The problem at the moment is that we do not know what these affixes are and it remains an empirical question.

So for the moment, it is probably best to use the lemma until the research has been carried out to establish the best pedagogical unit that learners can actually use. This means we would need to reconsider the thresholds for high- and mid-frequency vocabulary with lemmas in mind. Some research is now being done to establish the size targets necessary to use English in lemma terms (e.g., Schmitt et al., under review), and some of the suggested numbers are already discussed in this chapter. More research is needed until these numbers can be reliably reported. However, initial evidence suggests that moving from word families to lemmas would not change the thresholds for high-, mid-, and low-frequency vocabulary drastically. At the same time, they could be based on fewer assumptions and may be easier to interpret.

Further Reading

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/org/10.1017/S0261444812000018

This is the original article that has suggested moving the threshold of high-frequency words to 3,000 word families and lowering the boundary of low frequency words to 9,000 word families. It also labeled the vocabulary in between as "mid-frequency" vocabulary. The paper gives research-based reasons for these boundaries and addresses the pedagogical challenges with mid-frequency vocabulary.

Nation, I. S. P. (2001). How many high frequency words are there in English. *Language, Learning and Literature: Studies Presented to Hakan Ringbom.* Abo Akademi University, Abo: English Department Publications, *4*, 167–181.

This paper explains why the distinction between high-frequency and low-frequency words is critical and gives pedagogical advice on how to treat those words differently.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

This article reports on coverage research that sets out to answer a question how many word families one needs to read a novel, a newspaper, a graded reader, to watch a children's movie, or to understand unscripted spoken English. This paper suggests that 8,000 to 9,000 words families are needed for dealing with written texts and 6,000 to 7,000 word families for dealing with spoken discourse.

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

The paper looks at lexical coverage and tries to estimate more accurately how many words one needs to understand in a text in order to achieve comprehension. The authors suggest 98% for optimal and 95% coverage for adequate comprehension.

Related Topics

Academic vocabulary, technical vocabulary, L1 and L2 vocabulary size and growth, word lists

Notes

1 *Range* program online: www.victoria.ac.nz/lals/about/staff/paul-nation#publications
2 Although the idea of frequency as a pedagogical tool presumably is useful for all languages, the vast majority of frequency-based research has been done on the English language. Thus, the figures reported in this chapter are for English, and may differ for other languages.
3 Freely available graded readers: www.victoria.ac.nz/lals/about/staff/paul-nation#free-graded-readers

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

Barlow, M. (2011). Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics*, *16*(1), 3–44. https://doi.org/10.1075/ijcl.16.1.02bar

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*(1), 1–22. https://doi.org/10.1093/applin/amt018

Bybee, J. (1998). The emergent lexicon. *Chicago Linguistic Society*, *34*(2), 421–435.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. London: Greenwood Publishing Group.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38–63.

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language*, *22*(1), 181–200.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation (Ed.), *Making and using word lists for language learning and testing* (pp. 153–167). Amsterdam: John Benjamins.

Dang, T. N. Y., & Webb, S. (2017). Evaluating lists of high-frequency words. *ITL-International Journal of Applied Linguistics*, *167*(2), 132–158. https://doi.org/10.1075/itl.167.2.02dan

Day, R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, *7*(2), 541–551.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text – Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29–62.

Gardner, D. (2013). *Exploring vocabulary: Language in action*. London, UK: Routledge.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327. https://doi.org/10.1093/applin/amt015

González Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, amy057, https://doi.org/10.1093/applin/amy057

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*, 689–689.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *Canadian Modern Language Review*, *61*(3), 355–382. https://doi.org/10.3138/cmlr.61.3.355

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*(2), 207–223.

Howatt, A. P. R. (1984). *A history of English language teaching*. Oxford, UK: Oxford University Press.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. *English for Specific Purposes*, *33*, 54–65.

Hsueh-Chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Hunston, S. (2012). Corpus linguistics: Historical development. In *The encyclopaedia of applied linguistics*. John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0257/abstract

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, *33*(1), 147–151. https://doi.org/10.1162/coli.2007.33.1.147

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976–987. https://doi.org/10.1002/tesq.329

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension. In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Philadelphia: Multilingual Maters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & P. Arnaud (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). London: Macmillan.

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

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

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*(2), 267–290. https://doi.org/10.5054/tq.2011.247708

Meara, P. (1992). *EFL vocabulary tests*. ERIC Clearinghouse. Retrieved from www.lognostics.co.uk/vlibrary/meara1992z.pdf

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, 262–282.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Heinle & Heinle Publishers.

Nation, I. S. P. (2001a). How many high frequency words are there in English. *Language, Learning and Literature: Studies Presented to Hakan Ringbom. Abo Akademi University, Abo: English Department Publications*, *4*, 167–181.

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

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. *Vocabulary in a Second Language: Selection, Acquisition, and Testing*, 3–13.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading*, *1*, 5–16.

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

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, *22*(1), 31–55.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1–28.
Read, J. (2000). *Assessing vocabulary*. Cambridge and New York, NY: Cambridge University Press.
Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics*, *24*, 146–161. https://doi.org/10.1017/S0267190504000078
Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.
Schmitt, N. et al. (under review). How much vocabulary is required for listening and reading in English?
Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x
Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/org/10.1017/S0261444812000018
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103
Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145–171. https://doi.org/10.2307/3588328
Thornbury, S. (2002). *How to teach vocabulary*. Harlow: Pearson Education.
van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521
van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. https://doi.org/10.1093/applin/ams074
Waring, R., & Nation, I. S. P. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge, UK: Cambridge University Press.
Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL – International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web
West, M. (1953). *A general service list of English words*. London, UK: Longman, Green and Co.