1
The Nature and Size of Vocabulary

• Are there any patterns among the many thousands of words in a language?
• How should we count vocabulary?
• How much vocabulary do L1 English speakers know? Second language learners?
• How much vocabulary does it take to operate in English?

The White Rabbit put on his spectacles. “Where shall I begin, please your Majesty?” he asked. “Begin at the beginning,” the King said, very gravely, “and go on till you come to the end: then stop.” Lewis Carroll, Alice’s Adventures in Wonderland, p. 106

The advice given in the above quote from Alice in Wonderland seems to be appropriate for an introductory text, so to start at the beginning we must consider what we mean by vocabulary. The first idea which probably springs to mind is words. Every language is made up of words, and they come in an amazing variety. Some are short (a), some are long (antidisestablishmentarianism), some seem old (anon), and some have just entered the language (bling: expensive, ostentatious clothing or jewelry). Some have a single, straightforward meaning (quarantine), some carry several different meanings (bank: financial institution, riverside, a row of things, turning an airplane), and some convey a load of positive/negative connotations (aroma/stench).

Some are common (sky, eat) and some are hardly ever used (punctilious).

This chapter will discuss the nature of vocabulary, and will show that there are useful patterns in the sea of words. If the patterns are understood, it can make teaching and learning vocabulary much more systematic and successful. The chapter will also discuss how much vocabulary there is in language (vocabulary size), and how much must be known in order to be functional in English.

1.1 The Nature of Vocabulary

1.1.1 The Connection between Meaning and Form

At its most basic, vocabulary connects the real world with language. That is, it connects meaning which comes from life experience (the color of leaves, the way
you feel when someone insults you) with linguistic form (i.e., words) which represents those meanings (green, angry). However, it would be a mistake to think that meanings are always connected to just single word forms. To illustrate this, let us consider the following items:

die
expire
pass away
bite the dust
kick the bucket
give up the ghost

The six examples are synonymous, with the meaning “to die.” However, they are made up of from one to four words. Die and expire are single words, with die being by far the more common. Pass away could probably best be described as a phrasal verb, and the last three are idioms. (An idiom is a string of words which taken together has a different meaning than the individual component words. Similarly, a phrasal verb is made up of a verb plus one or more other words, which usually has an idiosyncratic meaning compared to the component words.) Thus, it is clear that there is not necessarily a one-to-one correspondence between a meaning and a single word. Very often, in English at least, meanings are represented by strings of multiple words. Conversely, a single word form often represents several meanings. If the meanings are related, this is called polysemy (e.g., a chip of wood, a computer chip, a gambling chip [all have a small thin shape]). If the meanings are considered unrelated, this is called homonymy (e.g., grave means a place for burial, serious problems or situations, and an accent mark (à)).

Multi-word units are very common. This book will use formulaic language as a cover term for all of the types of multi-word unit, and each individual unit will be called a formulaic sequence. Because the extent of formulaic language only became clear once corpus evidence was available, it will mainly be discussed in Chapter 4 – Corpus Insights: Frequency and Formulaic Language.

If we wish to refer to vocabulary which includes both single words and formulaic sequences, the terms lexeme or lexical item can be used: “an item that functions as a single meaning unit, regardless of the number of words it contains.” However, to enhance the accessibility of this book, we will use the term word unless more precise terminology is required to make a point.

1.1.2 Content Words and Function Words

The meaning–form connection is a key aspect of most vocabulary, whether individual words or formulaic sequences. But not all, as some vocabulary performs grammatical functions. For example, articles (a, an, the) show whether previous information has already been mentioned or not (among other functions), and prepositions show relationships (under, by). These words are called function
words (or grammatical words), and they “knit together” the content words in a sentence. They are necessary regardless of the topic being discussed. The following two sentences concern very different topics, and have a different level of formality, with the first one being written and the second one spoken. Regardless, both include many function words (underlined):

Stephen Hawking described the motion of planets moving around the sun and how gravity exerts force between bodies in space.

I don’t want to put you on the spot, Sam, but is this really a good idea?

Being necessary for all topics, function words are extremely frequent, and make up a large percentage of any spoken or written discourse, as we will see in Chapter 4. (See a webpage by Vivian Cook for one listing of function words: www.viviancook.uk/Words/StructureWordsList.htm.)

1.1.3 The Meaning Relationships between Words

While function words are a closed set (~150–300 words), the number of content words is huge, with new words and phrases being added all the time. For example, phast (a “phone fast”: a period of time during which someone chooses not to use their smartphone), knitflixing (the activity of knitting and watching a TV program on Netflix at the same time), and jackpotting (the crime of hacking into a cash machine in order to obtain money) were new words noted by the Cambridge Dictionary blog in 2018 (https://dictionaryblog.cambridge.org/category/new-words/). But all of the numerous content words in a language are not completely independent of each other. Rather, many have various semantic (meaning-based: also termed paradigmatic) relationships with each other. The following are among the most common semantic relationships:

**Synonyms** are words which have approximately the same meaning (new/fresh, beautiful/handsome). It is argued that there are no fully interchangeable synonyms, as every pair would have some slight meaning or collocational difference (Paradis, 2012). For example, *new* can be used to describe virtually anything that is innovative or which replaces something (car, idea, year), but *fresh* is largely connected to air or food (bread, vegetables). *Beautiful* and *handsome* both mean “pleasing to the eye” when describing people, but *beautiful* is preferred for women, and *handsome* for men.

**Antonyms** are words which have approximately opposite meanings (hot/cold, expensive/inexpensive). There are two kinds of antonymy: ungraded and graded. Ungraded antonyms are exclusively opposite, either one or the other (dead/alive, pass/fail). Graded antonyms convey oppositeness on a continuum (hot/warm/cool/cold).

**Hyponyms** are sets of words that have a hierarchical relationship from more general to more specific (vehicle/car/Audi), (fruit/apple/crab apple). The more general or
inclusive word is a **superordinate** (*vehicle/car*), the more specific word is a **subordinate** (*vehicle/car*), and words at the same level of generality are **coordinates** (*car/bus/truck*).

**Meronyms** have a whole–part relationship (*bicycle – wheels, handle, seat*).

These categories are important because words do not exist in isolation, and for many, their meanings are defined in relation to other words (Carter, 1998). For example, *dead* can only be defined as “no longer being alive.” Thus, knowledge of these categories can be useful when trying to explain meaning in instructional contexts (Chapters 3, 7, and 8).

### 1.1.4 The Sequential Relationships between Words

Words also relate to each other in sequential ways (sometimes called **syntagmatic** relationships). Although it is possible to combine words in a virtually unlimited range of ways, in day-to-day usage people prefer their language to be more predictable and thus easy to comprehend and produce. Sinclair (1991) was one of the first scholars to discuss these two approaches to language, making the distinction between the **open-choice principle** and the **idiom principle**. The open-choice principle conveys the idea that language is creative, and in most instances, there is a wide variety of possible words which could be put into any grammatical “slot.” For example, if we wished to express the idea of a torrential downpour, the adjective slot in the sentence *The _____ rain caused widespread flooding* could potentially be filled with several synonyms (e.g., *heavy, strong, powerful, forceful, dense*). This is the traditional way of viewing language, and Sinclair stated that “virtually all grammars are constructed on the open-choice principle” (p. 110). However, complementary to this freedom of choice, he observed that language also has a systematicity which constrains vocabulary choice in discourse; constraints which the open-choice principle does not capture. To some extent this systematicity merely reflects real-world phenomena: *fishing* is often done in close physical proximity to a *lake*, so the words expressing these concepts will naturally co-occur as well. But much of the systematicity is strictly linguistic: There is no reason why we do not say *strong rain* or *powerful rain*, but proficient members of the English-speaking speech community know that the appropriate phrase is *heavy rain* (an asterisk before a word indicates that it is inappropriate, ungrammatical, or otherwise nonstandard). The idiom principle highlights the fact that there are patterns in how words co-occur with each other.

In the above discussion of formulaic language, we saw how the sequential patterning resulted in various kinds of formulaic sequences with their own meanings, e.g., idioms and phrasal verbs. But the sequential patterning also extends to word pairs (or sometimes three words) which tend to co-occur in discourse. This “word partnership” is called **collocation**. J. R. Firth first brought this notion to prominence in 1957, and it has become increasingly important since. It refers to
the fact that some pairings are preferred in language use and sound “natural” (cause pain, inflict pain), while other possibilities which would convey the same meaning are not typically used, and just sound wrong (*produce pain, *create pain).

There are two factors that are key to the notion of collocation. The first is that words co-occur together and the second is that these relationships have varying degrees of exclusivity. A commonly given example of collocation involves the word blonde. Blonde occurs almost exclusively with the word hair and a few other animate nouns like woman or lady. But it never occurs with words like paint or wallpaper, even though there is no reason semantically why they should not fit together. Because blonde has such an exclusive relationship with hair, they are said to collocate strongly. Most words do not collocate this strongly, however. Sometimes the collocation can be much weaker, as in the case of the word nice. This commonly occurs with almost any noun which one would want to associate with pleasantness, such as a nice view, nice car, or nice salary. These combinations could be said to collocate weakly. Some words combine so indiscriminately that there is not enough exclusivity to warrant the notion of collocation. An example is the word the, which co-occurs with virtually every non-proper noun. So to be considered a collocation, words must co-occur in discourse, but there must also be an element of exclusiveness.

Most authors who discuss collocation agree that there are two basic kinds of collocations: grammatical/syntactic collocations and semantic/lexical collocations (e.g., Benson, 1985; Biskup, 1992; Bahns, 1993). Grammatical collocations are the type in which a dominant word “fits together” with a function word, typically a noun, verb, or adjective followed by a preposition. Examples are abide by, access to, and acquainted with. Lexical collocations, on the other hand, normally consist of combinations of two basically “equal” words such as noun + verb (ball bounces), verb + noun (spend money), and adjective + noun (chilly night), in which both words contribute to the meaning (Benson, 1985). In addition to these two basic collocational categories, Allerton (1984) proposes a third, consisting of collocations that are not based on grammatical or semantic patterning. The relatively arbitrary prepositions attached to time fit in this category, since there does not seem to be any logical reason why we should say at six o’clock but on Monday.

1.1.5 Grammatical and Morphological Relationships between Words

Words are also related through word forms which reflect grammatical and morphological relationships. Walk, walked, walking, and walks are closely related, consisting of the simplest verb form walk (base, root, or stem form), and its grammatical inflections walked, walking, and walks (base form + grammatical suffixes). Similarly, the noun base walk has the plural inflection walks. Inflected forms are created according to regular and transparent grammatical rules, which do
not change a word’s meaning. If we package a base word and its inflections together, we call this unit a **lemma**.

Often, we also want to use a word in contexts that require a different word class, for example, **persist**:

Noun: The judge changed his mind because of the lawyer’s **persistence**.
Verb: The lawyer **persisted** until the judge changed his mind.
Adjective: The **persistent** lawyer persuaded the judge to change his mind.
Adverb: The lawyer argued **persistently**.

In these cases, the meaning does not change, but the word class (part of speech) does. These word-class variations are called **derivations**. If we package the base word, its inflections, and its derivatives together, the unit is called a **word family**. There is clearly a relationship between all members of lemmas and word families, but the choice of which package to use will depend on pedagogical purpose and the proficiency of the learner.

### 1.1.6 Frequency: How Commonly Is a Word Used?

One does not need to know much about a language before realizing that some words occur more often than others. For instance, almost everyone with much exposure to English would have the intuition that **weak** is much more common than **puny**. **Weak** occurs frequently in many different contexts (weak economy, weak knees, weak arguments), and thus is much more frequent than **puny**, which is used mainly to refer to people and body parts (puny arms, puny muscles). So, some words are more frequent than others. But vocabulary frequency does not follow a simple, linearly decreasing curve as frequency steadily tapers off. Rather, it follows what is called **Zipf’s Law**, a pattern where a relatively small number of high-frequency items are extremely frequent, but then frequency drops off exponentially, with the vast majority of items becoming relatively rare quite quickly.

This is good news for language learners, because it means that a large percentage of language is made up of a relatively small set of words. In English, the top 10 words make up about 20 percent, the top 50 words 35 percent, and the top 100 words 41 percent of all words in a typical written text. The top 2,000 words usually make up around 80 percent of typical English texts (Nation & Waring, 1997). Considering that the vocabulary of the English language is very large (see below), we find that a relative handful of words contribute the bulk of the vocabulary that a person will come across when reading and listening, while the others occur rather infrequently.

While people have intuitions about the frequency of vocabulary, they are best at differentiating between very frequent and very infrequent words (Schmitt & Dunham, 1999; Alderson, 2007; McCrostie, 2007). We can get much more robust and fine-grained results about frequency from counting how many times words occur in large **corpora** (language databases). For instance, **weak** occurs 19,839 times in...
1.1.7 General and Specialized Vocabulary

Another way of finding patterning in the many words in a language is to consider how they are used and what topics they cover. In an average day, a person will encounter vocabulary across a range of domains (e.g., various topics when reading the newspaper, the vocabulary for their particular occupation, business vocabulary when doing taxes or paying bills, and fairy tales when reading to their children at night). The categories that have been found useful in describing this diverse range of vocabulary include “general vocabulary,” “academic vocabulary,” and “technical vocabulary.” General vocabulary is the term that is used to describe vocabulary that is useful across a wide range of topics and contexts, in both speech and writing. This consists of high-frequency vocabulary, which is very frequent precisely because it occurs regularly across a wide range of contexts. It is impossible to say exactly how much general vocabulary there is, as words gradually become less common and occur in fewer topics as frequency decreases. That is, there is no obvious boundary where general vocabulary stops and thereafter all words are more specialized. For a long time, the definition of general vocabulary has been synonymous with the General Service List (GSL), a list developed by West (1953) (see Chapter 2) which contains about 2,000 headwords. More recently, Gardner (2013) offered lists of core vocabulary totalling 2,857 words and Davies and Gardner (2010) produced a dictionary of the 5,000 most frequently used words in American English. The 5,000 figure chimes with our experience that at around the 5,000 frequency level, vocabulary use is better indicated by topic or domain than by frequency.

Academic vocabulary is vocabulary which is particularly useful for engaging with academic contexts, particularly reading academic texts. It is not topic-specific but rather serves to provide a greater level of precision which contributes to the academic tone of academic texts and speech. Compare the following two sentences:

A. The company changed its marketing ideas to try to make more money.
B. The company modified its marketing strategy to try to increase revenue.

Sentence A uses only high-frequency words and is easy to understand, but words like change, ideas, and money have many possible meanings, so the sentence lacks the precision required in academic texts. The meaning of Sentence B is nearly the same, but the underlined words are more precise. For example, modify means to make small changes to something to improve it, strategy means a planned series of
actions, and revenue is not just any money, but money which is earned from doing business or from collecting taxes.

Academic vocabulary also represents academic activities (to define, investigate, hypothesize) and signals rhetorical functions (e.g. furthermore, nevertheless, whereas) in texts.

**Technical vocabulary** is the jargon that is specific to particular domains (e.g., business, medicine, chemistry) and that represents the concepts and ideas specific to those domains (ledger, scalpel, catalyst). Technical vocabulary is crucial to understanding particular domains, because many of the key concepts are represented by this vocabulary. In this book, some of the technical vocabulary you have been exposed to include lexeme, collocation, and meronym, which are unlikely to occur very prominently in other domains like chemistry or politics.

General vocabulary is important for all language use and so will need to be prioritized as part of teaching. Specialized vocabulary, whether academic or technical, may be important for learners, depending on which purposes they wish to use language for. There will be more discussion on identifying and teaching these categories of vocabulary in Chapter 5: Categories of Vocabulary and Word Lists.

### 1.2 Vocabulary Size

Languages contain a lot of words. They have enough words to represent all of the things and concepts that a culture wants to talk about in the world, ranging from remembering last weekend’s family get-together to musing about the origin of the universe. Most languages have vocabularies reaching into the hundreds of thousands. English is commonly believed to have the largest of all, largely because it has freely borrowed from the many other languages it came into contact with during the years of the far-flung British empire, and then later American influence.

#### 1.2.1 How Many Words Are There in English?

Reports of the size of the English language in the popular press range widely: from 400,000 to 600,000 words (Claiborne, 1983, p. 5), from ½ million to over 2 million (Crystal, 1988, p. 32), about 1 million (Nurnberg & Rosenblum, 1977, p. 11), and 200,000 words in common use, although adding technical and scientific terms would stretch the total into the millions (Bryson, 1990). The largest English dictionary, the *Oxford English Dictionary*, claims to include more than 600,000 words, although many of these are historical and no longer in general usage. The discrepancy in size estimates is due largely to differing definitions of a word, and so a study attempted to produce a more reliable estimate by using word families instead of words as the unit of counting. Goulden, Nation, and Read (1990) counted the number of word families in *Webster’s Third International Dictionary* (1963), which was one of the largest non-historical dictionaries of English at the
1.2 Vocabulary Size

It may seem that counting words is easy, but we have seen that estimates of vocabulary size depend on how a “word” is defined, and what counting unit is used. Ideally, we would count meanings rather than word forms (i.e., words, lemmas, word families), but this would entail a manual analysis, as computer software is currently limited in how well it can discern meaning. Due to the large number of words in a language, we usually need computers to do the counting automatically for us. This limits us to counting word forms, which computers are very good at. To illustrate this, let us take the example of persist, persisted, persisting, persists, persistence, persistent, and persistently. We could count these as seven different individual words. However, we could also use a lemma counting unit, where the base word persist and its inflections persisted, persisting, and persists are packaged together as a single unit. The other derivative forms would count as their own lemmas. But there is also a close semantic relationship between all of the words, so it may make sense to package them together (base + inflections + derivatives) and count them as one word family (Table 1.1).

<table>
<thead>
<tr>
<th>TABLE 1.1 COUNTING UNITS FOR PERSIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting unit</td>
</tr>
<tr>
<td>Individual word</td>
</tr>
<tr>
<td>Words</td>
</tr>
<tr>
<td>• persist • persisted</td>
</tr>
<tr>
<td>• persisting • persists</td>
</tr>
<tr>
<td>• persistence • persistent</td>
</tr>
<tr>
<td>• persistently</td>
</tr>
<tr>
<td>• persistently</td>
</tr>
<tr>
<td>Number of units</td>
</tr>
</tbody>
</table>

time. After excluding entries like proper names and alternative spellings, Goulden et al. found that the dictionary contained about 54,000 word families. Dictionaries obviously cannot contain every current word family, but they were the best resource available before the widespread use of large corpora, and therefore early estimates of the number of words in a language were usually based on them. Brysbaert et al. (2016) describe a more recent statistical analysis of a trillion-word English corpus. They calculated that it contained about 10 million different individual words. Ultimately, it is impossible to say precisely how many words there are in English, as new words are always being added and old words are falling out of usage. However, what we can take from both dictionary and corpus approaches is that English clearly has a huge vocabulary (remembering that each word family can contain numerous individual words, as in the persist example below).
So how do we decide which counting unit to use? The most principled way of
deciding is based on how well people can perceive the relationships between the
various word forms, but this will differ depending on the language user in question.
Nation (2016) argues that word families make sense for L1 English speakers, as they
are likely to have a reasonable grasp of morphology and therefore will recognize the
various members of a family as related. He also believes that word families might be
appropriate for second language learners for the receptive skills of reading and
listening, arguing that if learners know one family member (persist), they should be
able to recognize or work out unfamiliar derivatives as semantically related words
when they are encountered in a meaningful context (e.g., persistence, persistent).
However, some evidence shows that this may be overly optimistic (especially for
beginners), as recognizing derivative word forms is often more difficult than Nation
has proposed (McLean, 2018). In terms of productive ability, almost all scholars
agree that even at advanced levels, learners do not reliably know all of the various
word-family members and so will not necessarily be able to use the appropriate
derivative forms when required in speaking and writing (e.g., Schmitt & Zimmer-
man, 2002). However, learners appear to be better able to produce the inflected forms
(McLean, 2018), presumably because they are based on rules which generally operate
in a regular and consistent manner, e.g., the progressive form of a verb involves -ing.
The upshot is that there is probably no single counting unit that is best in all
circumstances. The choice will depend on the proficiency of the learner and whether
the focus of use/learning is receptive or productive ability. Few researchers use the
“individual word” unit, as learners typically exhibit some knowledge of relationships
between the words. A number of researchers support the use of the lemma
counting unit (e.g., Gardner, 2013; Kremmel, 2016; Schmitt, 2010). The argument is
that this more conservative unit is less likely to overestimate learners’ knowledge.
This reflects the evidence that learners are more likely to know and use the inflec-
tional morphemes that make up the lemma than they are to know the derivational
morphology needed to make up a word family. An increasing amount of research
uses lemmas when dealing with second language learners (e.g., Brezina & Gablasova,
2015; Gardner & Davies, 2014; Kremmel, 2016). Despite the apparent movement of
the field toward using lemmas as the most common counting unit, most vocabulary
size research to date has been denominated in word families (e.g., Nation, 2006).
Therefore, the vocabulary size literature at the time of writing this book is denom-
ninated in several different units, and so it is important to carefully note the counting
unit when considering research which reports vocabulary size, as the reported figures
will always vary to some extent depending on the counting unit chosen.

1.2.3 How Many Words Do L1 English Speakers Know?

Estimates of the vocabularies of L1 English speakers provide an idea of what the
vocabulary size of lexically proficient users might be. Most studies have estimated
that L1 English speakers (usually university students) have a vocabulary size of
between 10,000 and 20,000 word families. Goulden et al. (1990) and D’Anna,