English L2 Vocabulary Learning and Teaching

Concepts, Principles, and Pedagogy

Lawrence J. Zwier
and Frank Boers

ESL & APPLIED LINGUISTICS PROFESSIONAL SERIES

ROUTLEDGE

English L2 Vocabulary Learning and Teaching

Accessible to experts and non-experts alike, this text is a comprehensive entry to teaching and learning vocabulary in ESL and EFL contexts. Firmly grounded in research, it presents frameworks and methods for teaching vocabulary to English L2 speakers. Overviewing key topics as well as providing in-depth research analyses and critiques, Zwier and Boers address all major areas of vocabulary pedagogy and instruction.

Organized in four parts, chapters cover the nature of vocabulary and strands of vocabulary research; curricular approaches; and techniques and activities. Readers are introduced to key topics, including teaching multiword expressions, assessment, discourse, and instruction at different levels. Each chapter includes questions, prompts, and activities to foster discussion. A foundational textbook for courses on L2 instruction and teacher-training courses, it is an essential text for students and scholars in TESOL and Applied Linguistics, and provides the pedagogical grounding future English L2 teachers need to effectively teach vocabulary.

Lawrence J. Zwier is Associate Director for Curriculum and Instruction in the English Language Center at Michigan State University, USA.

Frank Boers is Professor of Applied Linguistics and TESOL at the University of Western Ontario, Canada.

ESL & Applied Linguistics Professional Series

Eli Hinkel, Series Editor

Teaching and Learning Second Language Listening, 2nd Edition
Metacognition in Action
Christine C. M. Goh and Larry Vandergrift

Foundational Principles of Task-Based Language Teaching
Martin East

What English Language Teachers Need to Know Volume III
Designing Curriculum, 2nd Edition
MaryAnn Christison and Denise E. Murray

Doing Reflective Practice in English Language Teaching
120 Activities for Effective Classroom Management, Lesson Planning, and Professional Development
Thomas S. C. Farrell

Creating Classrooms of Peace in English Language Teaching
Edited by Barbara M. Birch

Shaping Learners' Pronunciation
Teaching the Connected Speech of North American English
James Dean Brown and Dustin Crowther

Handbook of Practical Second Language Teaching and Learning
Edited by Eli Hinkel

English L2 Vocabulary Learning and Teaching
Concepts, Principles, and Pedagogy
Lawrence J. Zwier and Frank Boers

For more information about this series, please visit: www.routledge.com/ESL-Applied-Linguistics-Professional-Series/book-series/LEAESLALP

English L2 Vocabulary Learning and Teaching

Concepts, Principles, and Pedagogy

Lawrence J. Zwier and Frank Boers

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Cover image: Getty images

First published 2023
by Routledge
605 Third Avenue, New York, NY 10158

and by Routledge
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
A catalog record for this book has been requested

ISBN: 978-1-032-00156-2 (hbk)
ISBN: 978-0-367-77173-7 (pbk)
ISBN: 978-1-003-17299-4 (ebk)

DOI: 10.4324/9781003172994

Typeset in Goudy
by Apex CoVantage, LLC

Contents

Author Biographies

Lawrence J. Zwier is the Associate Director for Curriculum and Instruction in the English Language Center at Michigan State University in East Lansing, Michigan. His master's degree in Teaching ESL is from the University of Minnesota, and he holds a Certificate in Publishing from the University of Denver Publishing Institute. Specializing mostly in reading and vocabulary, he has written or edited numerous ESL/EFL textbooks, from the beginning level up to advanced. In another idiom, he has produced non-fiction works in history, geography, and light science for middle and high schoolers. He enjoys hiking, exploring in the Great Lakes region, fly-fishing (poorly), crossword puzzles, and songwriting. He and his wife, Jean, have two children and live in Okemos, Michigan.

Frank Boers is currently a Professor in Applied Linguistics and TESOL at the University of Western Ontario, Canada. He grew up in Antwerp, Belgium, where he obtained a master's degree and a PhD degree in Linguistics (with a thesis in the field of lexicology). He also earned a language teaching certificate and taught EFL for about 20 years in various high schools, colleges, and universities in Belgium. Since the early 2000s, Frank has published mostly on issues of instructed second language acquisition. In 2010, he moved to New Zealand, where he taught in the Applied Linguistics and TESOL programs of Victoria University of Wellington for 8 years. He was also co-editor of the journal *Language Teaching Research* for a while in that period. His previous book with Routledge is *Evaluating second language vocabulary and grammar instruction: A synthesis of the research on teaching words, phrases, and patterns* (2021).

Preface

We wanted to write a book on learning and teaching second/foreign language vocabulary that was firmly grounded in research but also easily accessible to readers with limited prior knowledge of this topic. The readers we had in mind were language teachers, course designers, and novice researchers of pedagogy-oriented applied linguistics. We have tried to avoid unnecessary jargon, and, where avoidance was impossible, we have tried to clarify technical terms using lay language. However, the endeavor to write an accessible book does not mean this is an overly simplified account of research findings and their implications. It includes in-depth critiques of the available research and demonstrates the need for further inquiry, especially inquiry with clear practical relevance. In this regard, we believe the book will be useful also for readers who are already quite familiar with vocabulary research and who are looking for ways to make pertinent contributions to it.

The two authors of this book have a lot in common. Both are experienced language teachers (and learners), albeit in different settings, and both are committed to building bridges between the worlds of research and education. At the same time, our different backgrounds and areas of expertise complement each other well. Larry's work has remained mostly within the realm of educational practice (e.g., as an author of ESL textbooks and as a curriculum designer), while Frank has in recent years devoted more and more time to research (and so he now occasionally needs to be reminded what real language courses are like). Larry will often consider what research findings mean for (ESL) curriculum design, while Frank will usually ponder what they may mean for (EFL) lesson design.

We did not write the chapters of this book together, sitting side by side. As is typically the case with co-authored books, each of us wrote the first draft of some chapters, and then the other made suggestions for amendments, ranging from minimal to substantial. Larry wrote first drafts of chapters 1, 5, 7, and 11. Frank took the initiative on the other chapters. However, you will hear both our voices in most chapters (especially chapters 1 and 2, which underwent several rounds of revision). You will probably even be able to discern what passages were written by one or the other because we have very distinct writing styles—including use of vocabulary. Here is a first hint: English is

Larry's mother tongue; Frank learned English as an additional language. Here is a second hint: Larry has used English in a wide variety of written genres; Frank's English writing has been confined largely to academic prose. You will likely notice various additional differences in our writing styles—for example, Larry uses direct quotations far more often than Frank does. We do not consider it problematic if you discern two distinct voices in this book. They tell one story.

Part I

Introduction

1 The Nature of Vocabulary and Vocabulary Items

Introduction

Human reverence for words runs deep. Words have been thought both to summon gods or demons and to drive them away. They can convey heartfelt reflections near the end of life, a parent's joy after a birth, a hopeful foray toward love, or a not-safe-for-work joke. They are shapeshifters, easily slipping the bonds that lawyers, clerics, and prescriptivists of all types try to apply. Societies that treasure holy books or millennium-old oral histories may mold their modern practices to suit the words of ancient times. On the other hand, in nearly any setting, some individuals treat words casually, not paying much attention to them beyond their simple utility. We believe that most readers of this book consider themselves "word people," who want not just to know words but to know *about* them—where they came from, why they are spelled as they are, which other words fit well with them, which slight changes of meaning they can convey in slightly different company.

This chapter aims to help you understand how words are described and categorized by lexicologists and applied linguists. We explain terminology necessary for progressing further through subsequent chapters. While avoiding deep-track jargon, we do present vocabulary and related concepts as an applied linguist might talk about them. Any educator wishing to read research articles about vocabulary or to use important online tools (such as profilers) needs to understand such terms.

A Word About the Lexicon

Linguists usually define a language's *lexicon* or *lexis* as its total stock of available lexical items, including single words (e.g., *rhino* and *when*) and phrases that function as though they were single words (e.g., *beer belly* and *by and large*). The lexicon of a language is thus much larger than the collection of vocabulary items known by individual speakers of the language. Even exceptionally erudite language users do not know all the words that are listed in a dictionary meant to represent a good proportion of a language's lexis. You can therefore praise someone for having a large *vocabulary*, while praising someone

DOI: 10.4324/9781003172994-2

for having a large *lexicon* would sound odd. Put differently, a lexicon could be considered as something a language has, while vocabulary, in a narrow sense, is something that a language user or group of users has. When linguists do use the word *lexicon* to refer to how vocabulary is represented in a person's mind, they typically specify this as this person's *mental lexicon*.

The lexicon is made up of diverse kinds of lexical items. Many have a lot of "content" and can be described in terms of what they mean—items such as *banana*, *serious*, or *today*. Such items are often labeled *content words*. Other items carry less intrinsic meaning and are notable as *function words*—such items as *the*, *it*, *from*, and *how*. Items replete with content tend to be from very large classes of words (or parts of speech) like nouns, verbs, adjectives, and adverbs; they may be called *open-class* items. The function words tend to be *closed-class* items—prepositions, articles, pronouns, and so on. The open–closed distinction refers to how likely the class is to add new members. The number of open-class words changes constantly, with new items—and new meanings for longstanding items—frequently entering. For example, the word *Internet* is relatively new, and so is the use of *virus* in the sense of computer virus. By contrast, it is very rare for the closed classes to admit new items or shed old ones.

When asked to give an example of a word, people are most likely to think of an open-class item such as a noun (e.g., *table*). While open-class words may intuitively be considered the most typical elements of the lexicon, closed-class words or function words are part of the lexicon as well. This is reflected by the fact that they are included in dictionaries. At the same time, they also commonly feature in books and courses about grammar.

The realm of lexis is often thought of as very distinct from the realm of *syntax*—the latter being described as the regular, systematic "rules" that govern how sentences are constructed. However, lexis shows relatively systematic patterns as well. For example, one lexical pattern in English is that certain verbs may combine with the particle *up* to indicate completion or progress toward completion, as in *eat* vs. *eat up*, *tie* vs. *tie up*, etc. This combination is of course not always possible—we do not say "decrease up" or "endure up"—but the phenomenon is consistent enough to be worth recognizing. What makes it more a lexical phenomenon than syntactic? Its primary effect is on the meaning of words or groups of words, not on the structure and meaning of sentences. Truth be told, the boundaries between lexis and syntax are not clear, and many phenomena are spoken of as lexico-syntactic. Just look at the previous sentence. The multiword lexical item *truth be told* displays a departure from regular syntax. Deletions obscure an implied conditional—*If the truth is to be told.*

Morphemes and Lexemes (or Lemmas)

Another realm of language, besides lexis and syntax, is *morphology*. This is the system of word-forming elements, with the rules for combining them. If lexis and syntax interact, then such interaction is even more obvious with lexis and morphology, as there is a close relationship between the concepts of *lexeme* and *morpheme*. Morphemes are the smallest meaning-bearing elements in language. Words consist of one or more morphemes. For example, *water* cannot be broken up into different meaning-bearing elements, and so *water* consists of a single morpheme. *Disinterested*, by contrast, is a lexeme consisting of three morphemes (*dis, interest*, and *ed*). Each contributes to the word's meaning and function (*dis-* means 'not'; *-ed* may indicate the word functions as an adjective). While lexemes occur as stand-alone units, many morphemes cannot; they need to bind with a "stem" (e.g., *interest* is the stem in *disinterested*).

Some morphological terms to be aware of are displayed in Table 1.1. *Affixes* are identified by their placement before the *stem* (*prefixes*), after the stem (*suffixes*), or within the stem (*infixes*, very rare in English, so not in the table.). They are also identified by their general effect on the word they join. *Inflectional affixes* serve a grammatical purpose, as with the *-s* suffix on a plural noun (*color/colors*) or on a third-person singular present-tense verb (*I eat/He eats*).

Table 1.1 Some Basic Morphological Terms and Examples

Morpheme	an irreducible element of meaning that either is a word or can be used in forming words	water -s (as in *waters*) -ing (as in *watering*) -fer- (as in *transfer*) trans- (as in *transfer*)
Bound Morpheme	a morpheme that is used only in combination with other morphemes to form words	intra- -fer- -dict- -ion
Free Morpheme	a morpheme that may be used on its own as a word, without combining with other morphemes	water map cucumber
Stem	a meaning element onto which affixes are placed.	water map -fer- (as in *transfer*) -dict- (as in *diction*)
Affix	a meaning element that attaches to a stem (or a stem + another affix) to inflect an item or derive a new item	-s -age pre- trans-
Prefix	an affix placed before the stem	pre- (as in *predict*) re- (as in *re-water*)
Suffix	an affix placed after the stem	-s (as in *waters*) -ed (as in *watered*) -ion (as in *diction*)
Compound Word	a word formed by the combination of two free morphemes	waterway railroad skylight

Derivational affixes alter the part of speech of an item (e.g., *-ion* in the noun *decision*, from the verb *decide*) or contribute new meaning in the derived form (e.g., *in-* ["not"] in the noun *indecision*). Some affixes appear in multiple spellings and pronunciations (e.g., the negative prefixes *il-*, *im-*, *in-*, *ir-*); the differences display assimilation with the letter/sound after the prefix (viz., *illegal*, *improper*, *insane*, *irregular*), making the cluster easier to pronounce.

Like lexical items, affixes convey meaning. For example, the prefix *un-* and the suffix *-ion* have clear meanings in novel settings. In fact, some prefixes and suffixes carry enough meaning to evolve from bound morphemes to *free* morphemes, like the noun *ism* ("a system of belief") or the adjective *retro* ("characteristic of an earlier time"). An awareness of how morphemes contribute to lexical meaning is important for language teachers. We can often explain a new vocabulary item in terms of its lexical components (e.g., *impose* means "put onto", *pose-* "put", and *im-* "on/in"; *-lect-* probably has something to do with speaking or reading, as in *dialect* and *lector*; and *frag/frac-* denotes "breaking", as in *fragment* and *fracture*). This technique can also train learners to spot word parts that recur frequently and can thus help them become independent vocabulary learners (see Chapters 3 and 11). Still, reasoning from word parts has its limits. For example, analyzing *remarkable* as *re-* "again" + *mark* "note" + *-able*) will not help learners to infer its meaning ("exceptional").

Let's now move on to the term *lexeme*. In the case of content words (or open-class words), a lexeme very often comprises more than a single word form. For example, *explain*, *explaining*, *explains*, and *explained* are instances of a single lexeme. The non-inflected forms (e.g., *explain*) are used as headwords in dictionary entries, but that does not mean the headword *is* the lexeme. It is simply a convention in lexicography to choose the barest form among the ones that make up a lexeme as shorthand for the whole set. A synonym for lexeme is *lemma*. Like a lexeme, the lexical unit called lemma is defined as a headword (also called base word) plus its inflected forms. It is worth pointing out that a lexeme or lemma does *not* include derivatives, that is, forms connected through derivational instead of inflectional affixation. For example, *explanation* is not part of the lemma comprising *explain*, *explaining*, *explains*, and *explained*. Instead, *explanation* and *explanations* represent their own lemma or lexeme. This is also the way words are categorized in most dictionaries: *explain* (verb) and *explanation* (noun) are usually presented as separate dictionary entries. When you look up a word, it is typically the headword (i.e., the non-inflected form) you are searching for. Dictionaries vary greatly in style, but a common practice is to list a headword, provide its pronunciation, and then list its inflected forms. There are no separate entries for these inflected forms since they are members of the same lexeme (or lemma).

Dictionaries offer a wealth of information, but they provide second language learners with little guidance as to which words merit special attention at the learners' stage of second language development or as to which words will be particularly useful in the learners' specific fields of interest. It is vital to have resources that can guide learners (and their teachers) to prioritize certain sets of lexical items that may serve them best (Nation, 2016). In Chapter 5, we will discuss various word lists that have been developed for such purposes.

Word list developers may follow the dictionary practice of organizing such lists by headwords for lemmas, but they may also wish to work with larger units than what is typically subsumed under a single dictionary entry. For example, they may reason that the verb *educate*, the adjective *educated*, and the noun *education* constitute a single learning target since all three are common and useful words, and it is likely that, if you know one (e.g., the verb *educate*), the others (*educated* and *education*) will also be understood, provided one has basic knowledge of English derivational morphology. We turn to such larger lexical units next.

Units Larger than Lemmas

The constituents of a lemma are usually uncontroversial because there is general agreement about what is an inflected form. There are marginal instances, such as when a set of words involves inflections for gender (e.g., *actor/actress* or *abbot/abbess*) that are no longer widely productive in English, which some lexicographers might consider to be separate lemmas. Still, the lemma construct is a well-established way of delineating lexical items. As we said previously, however, there are other possibilities. One is the lexical unit labeled *word family* (Bauer & Nation, 1993). A word family consists of a base word (e.g., *understand*), its inflected forms (*understands*, *understanding*, and *understood*), *and* its derived forms (*misunderstanding*, *misunderstood*). It is therefore a larger unit than a lemma. This larger unit was proposed under the assumption that language users (and learners) who understand some members of the family will probably also understand the other members—at least if they are familiar with the affixes used to derive one word form from another. Because it cannot be taken for granted that learners will understand relatively rare affixes or ones that vary in form and in meaning, Bauer and Nation suggested different levels at which words may be considered members of a word family, judged by their similarity to the base word and the frequency of the affixes, in the following descending order:

- Inflected forms (e.g., book*s*; help*ed*; strong*er*)
- Derivatives with frequent affixes that do not change the form of the base word (e.g., drink*able*, home*less*, kind*ness*, self*ish*, ten*th*, *un*usual)
- Derivatives with frequent affixes that change the form of the base word just a little bit (e.g., admir*ation*, apolog*ize*, arma*ment*, dogmat*ism*, fortr*ess*)
- Derivatives with affixes that are less frequent and therefore less likely to be familiar to learners (e.g., clear*ance*, contradict*ory*, revolution*ary*, pictur*esque*, politic*ian*, citizen*ship*, *hyper*active)
- Derivatives where the base word is hard to recognize (e.g., spast*ic*, superst*ition*)

- Derivatives using certain “classical” affixes (e.g., *ab*normal, depart*ure*, *per*spective)

As one moves down these categories, the probability of a learner understanding new words thanks to familiarity with the base word becomes smaller, so the reasoning goes. According to these estimates, *advisable* might be easier to understand than *advisory*, for learners who know the base word *advise*. It is important to note that knowing the derivational affixes at a certain level in Bauer and Nation’s framework does not *guarantee* comprehension of all the words that exhibit these affixes despite familiarity with the base word. A learner might know a highly frequent affix like *-less* and yet fail to understand a word where it appears, like *priceless*.

Nevertheless, the word-family construct has been adopted by numerous vocabulary researchers, and we will be referring to it in several chapters of this book. Importantly, researchers who use the word-family construct typically operationalize it in an inclusive fashion, that is, comprising all the aforementioned types of affixes. It has been argued, however, that it is unrealistic to expect language learners at low proficiency levels to be able to recognize and understand all the members of a given word family thanks to familiarity with one or a few prominent family members. McLean (2018), for example, proposes a smaller lexical unit as a potential alternative, called *flemma*. A flemma includes the inflected form of a base word and additionally members of different word classes provided they look the same. For instance, the flemma for the headword *develop* includes *develops*, *developing*, and *developed*, where *developing* and *developed* can function as verbs (*They developed/were developing a new vaccine*) as well as adjectives (*a developed/developing country*). Thinking of higher-proficiency learners, Cobb and Laufer (2021) proposed a different way of confining the word-family construct. Using corpus data (discussed later in this chapter), they distinguish members of a given word family that occur frequently in discourse and ones that are very rare. Because the latter are very unlikely to cause comprehension problems (since they are unlikely to be met in the first place), they can be deleted from the family-membership list. The result is a word family that is reduced in size—a unit for which Cobb and Laufer have coined the term *nuclear word family*.

Frequencies and Distributions of Words in Discourse

When some words are perceived as more useful and important to know than others, this is often related to their frequency of use, because frequently used words tend to have high utility. One will meet and use the word *light* far more often than *luminescence*, for instance. That said, in the domain of science, the latter word may be used more often than it will be in general conversation, and so it may be slightly more useful for science students than for, say, sociology students. The relative frequency of words in discourse broadly and in specific types of discourse can help to decide which words language learners should prioritize, because knowledge of these words will give the best return on effort. Identifying these useful words then requires tallying their frequency in large samples of discourse (known as *corpora*—discussed in the next section).

It is in this light that the delineation of *word* matters as well. Tallying all instances of a lemma will produce a frequency count that can be quite different from tallying all instances of a word family, especially when the latter includes many more members. For example, the word family represented by the headword *advertise* figures among the 1,000 most frequent word families of English (www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists), but the lemma for the same item does not figure among the 1,000 most frequent lemmas of English (www.wordfrequency.info/samples.asp; www.newgeneralservicelist.org/; http://corpora.lancs.ac.uk/vocab/browse.php). This is because the word family includes several forms beyond those comprised in the lemma, such as *advert*, *adverts*, *advertisement*, *advertisements*, *advertiser*, and *advertisers*. In Chapter 5 we will discuss word lists developed with a view to helping learners (with general or specific language interests) focus on the lexical items likely to be the most relevant for them. We shall see that some such lists—even ones intended to serve the same purpose—are organized by lemma, such as the *Academic Vocabulary List* (Gardner & Davies, 2014), while others are organized by word family, such as the *Academic Word List* (AWL; Coxhead, 2000).

It is not easy to *lemmatize* a text (to sort its word tokens into lemmas), especially a large one. It is especially hard to accurately sort words that have irregular forms or some weakly related/unrelated meanings—workhorse items like *bank*, *bear*, *track*, and *pen*. Software for lemmatizing is designed to look at contexts and distinguish among, say, the varied meanings of *bear* (e.g., "carry," "support," "tolerate," "an animal," "a difficult person or problem," "be relevant to," etc.) and to catch cases in which *bore* is the past form of *bear* and not the verb for "enter by gradually tunneling into" or the verb for "fail to interest an observer." Nonetheless, close calls may simply be beyond the capabilities of the software, and human interpretation is still often required. It goes without saying that sorting word tokens into word families poses even greater challenges. Decisions about what does or does not belong in the same family are often difficult, subjective, and resolved by one list-maker in ways that other list-makers object to. Durrant (2016, p. 51) has observed that the AWL, for instance, includes word families that have members with very diverse meanings, and he gives the example of the headword *constitute*, which subsumes forms such as *constituting*, *constituent*, and *unconstitutional*. For all but etymologically nimble speakers, the relationship between *constitute* and *unconstitutional* is attenuated to the point of obscurity. This illustrates the difficulty of setting the parameters of many word families. Even more than in lemmatization, the judgment of experienced humans is necessary, and even so, many cases are hard to call.

Regardless of the debates around what lexical unit should be used to investigate the frequency and distribution of words in discourse, it is well established that words can differ very markedly in their frequency of use. Function words such as *the* and *of* occur in almost every English sentence; content words such as *debate* and *investigate* occur once in a while. The words that are ranked highest in frequency-based lists make up a disproportionately large amount of the running words (or word tokens) of discourse. The 2,000 most frequent word families of English make up between 86% and 89% of general English discourse (Nation, 2006). Familiarity with these highly frequent words is thus crucial. As we will see in Chapters 3 and 5, however, knowledge of the 2,000 most frequent word families will in many situations be insufficient. Schmitt and Schmitt (2014) have argued that learners should try to master, at a minimum, the 3,000 most frequent word families. Unfortunately, the less frequent a word is, the less likely it is to be met often enough for learners to pick it up "incidentally" (see Chapter 3). This is one of the reasons why it is necessary to examine ways of helping learners expand their vocabulary over and above what they might acquire by chance (see Chapter 4).

Language Corpora

Determining the relative frequency of use of certain words requires very large samples of discourse, now commonly called corpora. *Corpora* is the plural form of *corpus* (Latin for "body"), which is the term for a large collection of language texts that can be studied to discover or confirm lexical/syntactic patterns. By choosing appropriate source texts, some corpus-makers can focus on one arena, such as academic English or medical English. At the other end of the spectrum, an extensive, general-purpose corpus aims to represent the language at large. Even field-specific corpora benefit from size and range to avoid overrepresenting idiosyncratic uses by a few writers or speakers. Corpus research informs word lists, textbooks, dictionaries, and debates about the suitability of teaching targets. Corpora are now so easily accessed and searched that suggestions abound for lessons and exercises that involve corpora (see, e.g., Reppen, 2010).

Prior to the ubiquity of computers and corpus-analysis software, researchers counted items by hand to determine what words were highly frequent in discourse and thus of particular interest to language learners. Michael West's *General Service List* in 1953 referenced a 3 million-word corpus, and Edward L. Thorndike's lists of 10,000 frequent words in 1921 and 1931 were based on sets totaling 4.5 million words (Gilner, 2011). Thanks to technological advances, we can now rely on much larger corpora. The *Corpus of Contemporary American English* (COCA), for instance, stands at 1 billion *running words* as of this writing (Davies, 2008). The figure for running words (or tokens) in a text or a corpus is the total number of words, without any reductions for repetitions, parts of speech, or other characteristics. Digital processing tools are now also advanced enough to allow filtering for part of speech, position in a sentence, word partnerships (or collocations—see Chapter 2), modality (written vs. spoken discourse), setting of use (e.g., an office-hours meeting, a journal article, an instruction manual), the year of each use, and even the demographic characteristics of the speaker/writer. Many such tools have user-friendly interfaces, allowing novices to learn in a few minutes how to use them.

Corpus search tools include frequency displays and tools that compare data across corpora or sub-corpora. These help to determine with far greater precision than in the past what words occur very often in general and what words are characteristic of a specific realm of discourse. Digital tools called *profilers* can be used to analyze a text that the user has on hand—an article, a student paper, or any other text in digital form that can be dropped into a search frame—rendering frequency counts, sorting the words in the text into frequency bands as per reference corpora, identifying phrases, and so forth. For example, the profiler on Tom Cobb's *Compleat Lexical Tutor* (www.lextutor.ca/vp/) generates information that can be very valuable to a teacher trying to decide whether a prospective class reading is appropriate to the students' level.

Another basic tool is a sorting application called a *concordancer*. The user enters a search "string" (a sequence of alphabetic characters or symbols), which could be a single word or more. The output is a display of lines in which the search term occurs in the corpus, with a certain number of words to the left and right. With most concordancers, users can specify the part of speech they want—for example, does one want to see instances of *deposit* as a noun, a verb, or both? The searcher can apply *wild cards* that allow inflected forms to display along with base forms. Concordance lines provide a moderate amount of context for the search term (and, for many corpora, a longer context displays if you click on a line in the concordance).

Homonymy and Polysemy

We hinted earlier at another issue that can make delineating lexical items challenging: A single word form may have more than a single meaning. In some cases, the different meanings are unrelated, and it is mere coincidence that the same word form denotes the distinct meanings. If so, the words are considered *homonyms* (derived from Greek *homo* = the same; *nym* = name) For example, *row* meaning "line" (e.g., a row of seats in a theater) and *row* in the sense of using oars to make a boat move through the water are homonyms. Because they are spelled and pronounced the same way, they are also examples of the more specific notions *homograph* (*graph* derived from Greek "write") and *homophone* (*phone* derived from "sound"). Some such pairs of words are homonyms only in their written form. For instance, *row* meaning "line" and *row* meaning "dispute" are identical in writing, but they are pronounced differently. Others may be pronounced identically, but differ in spelling (e.g., *rain* and *reign*). Most dictionaries will have separate entries for homonyms because these are semantically (i.e., meaning-wise) distinct lexical items.

Far more common than homonymy is the phenomenon called *polysemy*. Polysemy (*poly* = many; *seme* = meaning) refers to words that have more than one meaning, but these meanings (or "senses") are related in one way or another. For example, *row* is not only used in the sense of a physical line of objects (e.g., a row of desks in the classroom), but also in phrases such as *several times in a row*, where it refers to an uninterrupted sequence of a type of event. It is not hard to see how this second sense of the word is related to (and derived from) the first. We often think of time in spatial terms (e.g., the days of the week laid out in a row on a calendar). *Row* in its physical sense denotes a line of objects without other objects in between, and therefore it denotes a line that is "not interrupted" by other objects. It is this implication that is preserved in the more abstract, temporal use of, for instance, *they won six times in a row*.

Polysemy abounds in language. Rather than coining new words to express something for which no words exist yet, it is much more economical to extend the meanings of already existing words. Over time, a language's lexicon not only changes by adding (and shedding) words, but also by adding (and shedding) particular uses of words. Highly frequent words are especially likely to have various uses or senses (e.g., Skoufaki & Petrić, 2021). Most dictionaries list the various senses of a polyseme (or polysemous word) within a single dictionary entry (i.e., under the same headword). As a result, dictionary entries for highly polysemous words are generally longer than ones for only moderately polysemous ones, and these long entries almost invariably concern highly frequent words, such as prepositions (e.g., *over*) and common verbs (e.g., *run*). In fact, their high frequency is partly attributable to their polysemy—the more functions a word can serve, the more frequently it will be used.

In lexicology, highly polysemous words are often described as a network of different senses, with a basic or "prototypical" sense as the hub, and the other senses either directly or indirectly connected to it. Because the central sense in the network is usually the most common one, learners in a general L2 proficiency course will often become familiar with this sense first. From the perspective of the language learner, then, building a vocabulary is not only a matter of adding new words to one's repertoire, but also a matter of extending one's knowledge of already familiar-looking words. A language learner may be familiar with the most common sense of a word but may still be unfamiliar with additional senses.

At least two mechanisms drive the development of additional uses or senses of an existing word. One is simply to broaden the use of a word through semantic inferencing. If you can feel *confident* about a project, then you can also feel *confident* about your own abilities, and hence the sense of *self*-confidence. If you *examine* something, you will be able to assess its qualities, and hence the sense of assessment (as in the noun *examiner*). You can use your *elbows* to push people aside when you want to move through a crowd, and hence the use of *elbowing* as a verb. Note that these meaning extensions often cross word classes (e.g., deriving verbs from nouns, or vice versa: *an advocate—to advocate*). The other major driver of meaning extensions is *metaphorization*. This occurs when words with a literal meaning develop figurative uses to refer to abstract things. This is a very common phenomenon. You can *wield an axe* and you can also *wield power*; you can *embrace a person* and you can also *embrace an ideology*; a gardener may *prune branches of a tree* and an entrepreneur may *prune branches of a company*; food can be *bitter* and so can feelings; a *landslide* can change a physical landscape while a *landslide victory* in elections can change the *political landscape*; one can *break a stick* and one can also *break the silence*; flowers *blossom* and so may a romantic relationship; and so on.

So Much to Learn!

Building a vocabulary in a second or additional language is clearly a daunting task. There are a great many words to be learned, regardless of the unit one uses to count "words." Even with a modest aim of mastering, say, the 3,000 most frequent word families of English (Schmitt & Schmitt, 2014), this amounts to learning a quantity of lemmas that is considerably greater than 3,000 because a word family usually comprises several lemmas through derivational affixation, and it of course amounts to an even greater number of individual word forms because a lemma usually comprises different inflected forms.

Moreover, there is a lot to be learned about a word. Nation (2013, p. 49) distinguishes aspects of word knowledge regarding form (spelling and pronunciation), regarding meaning, and regarding use (when and how to use the word). Various factors can make each of these aspects challenging. Learning the spelling of a word may be especially challenging if it includes letters that do not correspond to its pronunciation (e.g., *doubt*, *knight*, *phlegm*) and if there is a homophone that is written differently (e.g., *principle* and *principal*). Conversely, guessing the pronunciation of words based on their written form is challenging in a language such as English, where sound-spelling correspondences have become unreliable (e.g., the pronunciation of *au* is very different in *launch* and *gauge* and the pronunciation of *ei* is very different in *seize* and *reign*). Confusion regarding spelling and pronunciation may also be caused by interference from known words in the learner's first language (or other languages that the person is familiar with) that bear incomplete resemblance to the to-be-learned word (e.g., French *exercice* vs. English *exercise*; French *développement* vs. English *development*).

Interference from the mother tongue may hinder the learning of a word's meaning as well. For example, a Dutch-speaking learner of English may misinterpret *eventually* as "possibly" because this is the meaning of the Dutch word *eventueel*. A new word may be mistaken for one that has already been learned in the target language if the two look alike and a minimal difference in form goes unnoticed (e.g., *adapt* and *adopt*; *terrible* and *terrific*). Furthermore, many words are polysemous, and so it is not just one form-meaning correspondence that needs to be established and remembered (e.g., *intelligence* as cognitive ability and as information of military value; *run* as in running a race, running a bath, running an errand, running a business, and running out of money).

When it comes to knowing how to use a word, learners need to be aware of usage restrictions (as in the case of words that are rude or offensive). They also need to be aware of the grammatical features of the word, such as whether the noun exists in a plural form (e.g., *advice* and *stuff* do not have a plural *-s* form, i.e., *advices* or *stuffs*), whether a verb has an irregular past form (e.g., the past form of *drive* is *drove*, not *drived*), and how the word influences the features of other words in its vicinity (e.g., *decide to go* is conventional but not *decide going*; *consider going* but not *consider to go*). This illustrates that lexis and morpho-syntax are impossible to separate once you take words out of the dictionary and release them into their natural habitat of discourse.

Knowing how to use a word also involves familiarity with its typical companions (or collocations). In English, one *makes an effort*, while the counterpart in Dutch is "do an effort". *Meeting the deadline* sounds right, but *hitting the deadline* sounds odd (while *hit the headlines* does sound conventional). Similarly, we can go somewhere *on foot* but not *on* (or *in*) *car, train, bike,* and so on. Knowing a word (e.g., *corner*; *tongue*) very well thus also includes knowledge of the fixed expressions in which it occurs (e.g., *cut corners*; *tongue in cheek*). We mentioned at the beginning of his chapter that a language's lexicon includes numerous items consisting of more than a single word. We have said very little about such items in the present chapter for purely practical reasons—they are the focus of Chapter 2.

Loanwords and Coinages

Finally, let us consider the dynamic nature of the lexicon, illustrated more specifically by examining where vocabulary items in English come from, including some that might be placed at the margins of English lexis. Can foreign words and phrases (e.g., *déjà vu*, *kibbitz*) be considered English? How about abbreviations like *lol* and *bff*?

English is a voracious language, hoovering up words from other tongues and adopting terms from various disciplines like there's no tomorrow. The previous sentence has been packed, only a little artificially, with lexical items of various backgrounds, including:

- West Germanic: *English*, *word*, *tongue*, *is*, *up*, *other*, *a*
- North Germanic: *from*
- French: *language*, *term*, *adopt*
- Latin: *voracious*, *various*, *discipline*
- Commerce: *hoover*
- Idiomatic (of unknown source): *like there's no tomorrow*

English is, taxonomically, a West Germanic language (closest modern relative: Frisian), an attribute that is apparent lexically in many function words and workaday items like *word* and *tongue*. Yet it is also a language that cohabited with Danish for more than a century (and with Scots, Welsh, and Irish longer than that); retreated to largely proletarian status under Norman French; was injected with Latin, Greek, and Arabic during periods of scholastic and scientific advance; went to church amid ecclesiastical Latin; had daily commerce with Malay, Hindi, Māori, Zulu, and Virginia Algonquian; picked up some Wolof and Swahili under unspeakable circumstances; and has served thought leaders and thoughtless ones worldwide in countless pursuits. The result is a high proportion of *loanwords* in even the most common texts of English. Durkin (2015, p. 3) defines a loanword as a case "where both the form and (at least some aspect of) the meaning of a word from another language have been borrowed into English." He also points out that borrowing is perhaps not the right metaphor for the process, since the acquiring language does not simply use the words for a while and then return them. After examining the 1,000 most frequent English words in the *British National Corpus*, Durkin identified 52% of them as loanwords.

Borrowing words from another language is only one way of extending a lexicon. Another way is to create or *coin* new words or phrases. New words deliberately invented are sometimes called *neologisms*, a very general term, often with the negative implication that the word is fake. That sort of debate keeps usage gurus employed, but it does not have much bearing on descriptive analyses of the lexicon. Table 1.2 lists some types of origin or invention of new lexical items.

Although many prescriptive forces have operated on English over the centuries—from stylebooks to religious leaders to at least one president—its lexis has faced no serious vetting from any standardization body. There is no "academy" (à la French or Italian) in any of the large hubs of varietal English (the UK, the USA/Canada, Australia, New Zealand, South Africa, etc.) to say whether a word is really a word, so the lexicon has grown in fascinating ways, taking on additions from a potpourri of cultural, technological, scientific, and entertainment sources. Lexical items rise, survive, fall, or disappear in response to the needs of a great mass of English speakers worldwide, not by fiat.

Dictionaries have an aura of authority in the realm of words. They influence public attitudes toward the legitimacy of an item as a word, with many English speakers reluctant to accept a lexical item, at least explicitly, until a well-known dictionary has accepted it first. The expressions *according to Webster* and *Webster says* invoke the proper name *Webster* to mean "dictionaries" (after Noah Webster [1758–1843], whose *An American Dictionary of the English Language* was first published in 1828). High-prestige dictionaries have become de facto gatekeepers, whether or not they want the role.

Lexicographers are unlikely to discourage their reputation as authorities, but the best of them usually strive to describe English, not prescribe it (c*f.* Merriam-Webster, 2021a). Inevitably, any description will be out of sync with actual use, to some degree, by the time it is published, since lexical innovation moves faster than the dictionary-adoption process, which has some

Table 1.2 Some Types of English Lexical Items, by Means of Entry or Invention

Loanwords	Words entering English from another language	*déjà vu*, *amok*, *berserk*, *tsunami*, *salsa*, *pied à terre*, *mesa*, *hygge* (from Danish, added in 2021 to Merriam-Webster online), *barista*, *gumbo*
Abbreviations (and Acronyms)	An *abbreviation* is pronounced as letters.	*DNA*, *H_2O*, *LED*, *FBI*, *SOS*, *BTW*, *TGIF*, *EMT*
	An *acronym* is pronounced as if it were a spelled word.	*HVAC* (pron: "aitch-vak") *FEMA*, *NASA*, *scuba*, *radar*, *laser*
	Some abbreviations and acronyms, esp. in social media, appear as unabbreviated phrases only rarely.	*lol* (either as an abbreviation or acronym), *jk*, *omg*, *bff*, *tl;dr*
Scientific & Technical Terms in Common Use	Often from Greco-Latin word parts. Unlike loanwords, they did not have currency in another language before entering English.	*oxygen*, *polyurethane*, *carbon dioxide*, *methamphetamine*, *solar plexus*, *schizophrenic*
Clipping	Shortening an existing word or compound	*bro*, *nuke*, *trans*, *Ivy* (for Ivy League College), *meds*, *gym*
Blending	Combining parts of two or more words (or part of one with the whole of another)	*motel*, *smog*, *staycation*, *Europop*, Mc*Mansion*, *Brexit*, *the Potter-verse*
Compounding	Combining two or more whole words (spelling may be fused, hyphenated, or with word space)	*houseboat*, *loanword*, *lockdown*, *stir-fry*, *four-wheeler*, *camper van*, *pork chop*
Affixation	Affixes added by invention; often involves trade names/personal names	*gentrification*, McCarthy*ism*, *Dylan-esque*, *de-Baathification*, *unmute*
Rhyming, Assonance, Alliteration	Combining words (or non-words) into pairs that either rhyme or have similar sounds	*shop 'til you drop*, *rom com*, *artsy-fartsy*, *willy-nilly*, *splish-splash*, *ding-dong*
Symmetrical Pairs (often opposites)/ Correlatives	Related terms (often opposites) pair in a somewhat balanced phrase, often with *and/or*	*in and out*, *here and there*, *life and times*, *one or the other*, *no work no pay*, *from top to bottom*
Genericized Commercial Terms	Though businesses mark their trade names, English uses them anyway as non-proper nouns or verbs (Capital letters used here; not always so in common use.)	*Zoom*, *Escalator*, *Google*, *Kleenex*, *Zipper/Zip*, *Photoshop* (v)
Quotations and Catchphrases	In earlier years, drawn largely from literature or oratory; presently from multiple forms of entertainment and the speech of celebrities	*deplorables* (H. Clinton), *doh!* (Homer Simpson), *They're ba-ack* (*Poltergeist II*), *the Force be with you* (*Star Wars*), *axis of evil* (G.W. Bush)

Proverbs and Sayings	Of longer standing than most quotations, often untraceable to a specific source; part of a saying may be used to imply the whole.	*Reap what you sow* *A stitch in time saves nine* *Cold hands, warm heart* *People who live in glass houses . . .* *Live by the sword, . . .*

brakes built into it. Not every usage in a given year merits a place in a broad-spectrum dictionary. As Stamper says of the process at Merriam-Webster, to even be considered for inclusion, a word must not only be used widely and have meaning but also demonstrate "shelf life" (Stamper, 2017). That is, the Merriam-Webster staff need evidence that the word has remained in use for a considerable time (although Stamper does not specify a length). Green (2016) interviewed Merriam-Webster editor at large Peter Sokolowski and reports,

> In its entries for this year's new gender-related words, Merriam-Webster dates many to the early '90s; Sokolowski said it's common for words to be around for decades before they make their way into the dictionary. "We're a lag indicator," he said. "We're not trying to be avant-garde."

At one time all dictionaries were print vessels, and updates were slow. If a new edition appeared every ten years or so (as the *Merriam-Webster New Collegiate Dictionary* has done since its 7th edition in 1963), that was fast. Now, some dictionaries release occasional updates online—lists of new words being added to the dictionary's Web services and for printing the next time the dictionary comes out in a full new edition. The *Oxford English Dictionary* releases updates four times a year; since the *OED* is a historical dictionary, many of its new entries are not new words at all but older forms recently found to be important enough for inclusion. The *Merriam-Webster Collegiate Dictionary* puts out an annual update of between approximately 450 and 850 new items, an event that usually generates news stories. In their update in April 2021, Merriam-Webster included 520 new words, including *cancel culture*, *hard pass*, *hygge*, and *pod* in a new meaning, "a usually small group of people (such as family members, friends, coworkers, or classmates) who regularly interact closely with one another but with few or no others in order to minimize exposure and reduce the transmission of infection during an outbreak of a contagious disease" (Merriam-Webster, 2021b).

Summary

The principal purpose of this chapter was is to make you familiar with (or remind you of) key terms, many of which will be recurring throughout the book. These include *lexis*, *lexico-syntax*, *content words* and *function words*, *morpheme*, *lexeme* or *lemma*, *word family*, *inflectional* and *derivational affixes*, *homonymy*, and *polysemy*. We briefly mentioned the role of the frequency of lexical items in discourse for deciding what items should be prioritized considering the learners' proficiency level and the purpose for which they are learning their target language (more about which will be said in Chapter 5). In this context we also briefly reviewed corpora and the information they can yield about the frequency of words in discourse broadly or in specific types of discourse.

We also highlighted the multifarious nature of vocabulary knowledge. Not only do learners face the formidable challenge of mastering a large quantity of words, but they also need to learn diverse things about individual words, including aspects of their form(s), meaning(s), and use(s). We have given examples of factors that can make learning one or more of these facets difficult.

Finally, we illustrated the dynamic nature of the (English) lexicon by examining its history and how dictionaries have attempted to keep up with changes in lexis. In so doing, we also looked at some categories for classifying lexical items according to the means by which they entered English.

Activities and Discussion

New Kids on the Block?

With one or two partners, draw up a list of 10 very new words or phrases in English that, in your opinion, should be included in any new edition of a major all-purpose dictionary. This may require you to observe for several days the English usages you encounter (both spoken and written). Then choose a dictionary and check (probably online) to see if your candidates are listed in it.

More Patterns

This statement was made earlier in this chapter: "One lexical pattern in English is that certain verbs may combine with the particle *up* to indicate completion or progress toward completion, as in *eat* vs. *eat up*, *tie* vs. *tie up*, etc." Other lexical patterns can be recognized in English. By yourself or with one or two partners, consider each set of words below. Articulate any pattern(s) you see in a set.

Set 1: snow cover, dust storm, word list, treasure trove, waterfall
Set 2: overseer, afterparty, outlier, upgrade, offputting
Set 3: the Netherlands, Belgium, the United States, China, the Democratic Republic of Congo
Set 4: Traffic was flowing faster in the northbound lane.
The criminals diverted their cash into several secret channels.
The river of time runs in only one direction.

Most flashlights draw about 3 amps of current from their batteries.

I can't watch TV while I work because it interrupts my flow of thought.

Tightly Closed or Slightly Ajar?

Although change is much more common in the open classes of lexical items, sometimes new members or new meanings may enter so-called closed classes as well. Pronouns may be undergoing some change, with more effort being put into finding non–gender-specific forms, such as *they/them/their* for singular reference, as in the sentence *Each student should do* ***their*** *own work*. In fact, in 2020, the American Dialect Society declared singular *they* the "Word of the Decade." With one or two partners, discuss whether it is necessary to have a non–gender-specific personal pronoun. If so, is singular *they* the best choice?

There are also advocates for the use of *new* gender-neutral pronouns, such as *ze* (instead of *he/she*), *zem* (instead of *him/her*), and *zir* (instead of *his/her*) to respect people who do not identify with a binary gender. Do you know of similar trends in other languages than English?

References

American Dialect Society. (2020). *2019 word of the year is "(my) pronouns"; Word of the decade is singular "they."* www.americandialect.org/2019-word-of-the-year-is-my-pronouns-word-of-the-decade-is-singular-they

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

Cobb, T. (n.d.). *Compleat Lexical tutor* (version 8.3). www.lextutor.ca/

Cobb, T., & Laufer, B. (2021). The nuclear word family list: A list of the most frequent family members, including base and affixed words. *Language Learning*, *71*(3), 834–871. https://doi.org/10.1111/lang.12452

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

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. www.english-corpora.org/coca/

Durkin, P. (2015). *Borrowed words: A history of loanwords in English*. Oxford University Press.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes*, *43*(1), 49–61. https://doi.org/10.1016/j.esp.2016.01.004

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

Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language*, *23*(1), 65–83.

Green, E. (2016). Does adding "genderqueer" to the dictionary make it "real"? *The Atlantic*, April 25. www.theatlantic.com/entertainment/archive/2016/04/genderqueer-cisgender-transphobia-merriam-webster/479406/

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823–845. https://doi.org/10.1093/applin/amw050

Merriam-Webster. (2021a). We added new words to the dictionary for 2021. *Merriam-Webster.com*. www.merriam-webster.com/words-at-play/new-words-in-the-dictionary

Merriam-Webster. (2021b). *A word on descriptive and prescriptive defining*. www.merriam-webster.com/words-at-play/descriptive-vs-prescriptive-defining-lexicography

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

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P. (Ed.). (2016). *Making and using word lists for language learning and testing*. John Benjamins.

Oxford English Dictionary online. (2020). Oxford University Press. www.oed.com/

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge University Press.

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/10.1017/S0261444812000018

Skoufaki, S., & Petrić, B. (2021). Exploring polysemy in the Academic Vocabulary List: A lexicographic approach. *Journal of English for Academic Purposes*, *54*. Online First. https://doi.org/10.1016/j.jeap.2021.101038

Stamper, K. (2017). How a dictionary writer defines English. (video) *Vox*. www.youtube.com/watch?v=uLgn3geod9Q

2 Multiword Expressions

Introduction

When we hear or read the term *vocabulary*, we tend to think of words first and foremost. However, language abounds with lexical items made up of more than a single word. The clearest cases are fixed expressions whose meaning does not follow from adding up the meanings of the constituent words, such as *by and large*, *put up with*, *through thick and thin*, *fit the bill*, and *out on a limb*. These types of expressions are usually included in dictionaries, typically as sub-entries under one of the constituent headwords. They are also found in inventories of conventionalized phrases of a language, such as dictionaries of idioms. Idioms are fixed or semi-fixed expressions that are "non-compositional" in meaning, in the sense that disassembling them and examining the separate parts is not much use for working out their meanings. Beyond this long-acknowledged class of idioms, as we shall see, vast numbers of other phrases and recurring word combinations can be considered lexical items as well and also merit attention from language learners and teachers.

The principal aim of this chapter is to illustrate the diversity of multiword lexis, and to clarify terms and concepts that will help you read the subsequent chapters of the book, where we discuss research regarding the importance of this dimension of vocabulary (Chapter 3) and research regarding ways of teaching multiword expressions (Chapter 4).

A Large Collection

We will be using *multiword expression* (MWE) as an umbrella term for a wide variety of items consisting of more than a single word, which may include (for starters):

- compounds (e.g., *mobile phone*, *role model*, *ballot box*, *love handles*)
- figurative idioms (e.g., *a wet blanket*, *jump the gun*, *follow suit*, *a loose cannon*, *across the board*, *beyond the pale*)
- proverbs (e.g., *kill two birds with one stone*, *better safe than sorry*)

DOI: 10.4324/9781003172994-3

- phrasal verbs (e.g., *give up*, *find out*, *back down*, *hang out*, *come over*, *get along with*)
- prepositional phrases (e.g., *at school*, *on foot*, *over time*, *in love*, *below standard*)
- complex prepositions (e.g., *in front of*, *next to*)
- standardized similes (e.g., *cold as ice*, *good as gold*)
- standardized "A and B" phrases (e.g., *loud and clear*, *toss and turn*)
- standardized "A of B" phrases (e.g., *a bunch of grapes*, *a pack of wolves*).
- conversational formulas for smooth interaction (e.g., *How's it going? Have a great day! I'm so sorry*, *Thank you*, *No problem*, *See you later*).
- highly frequent word strings that function as discourse connectors or transitions (e.g., *for example*, *such as*, *and so on*) or simply serve as "fillers" (e.g., *sort of*, *you know*).
- common phrases to introduce a point (e.g., *It's well known that* . . . , *There's no doubt that* . . . , *It's safe to say that* . . .)
- structural frames with internal slots left to be completed (e.g., *not only* . . . , *but* . . . , *as far as* . . . *is concerned*).

In addition, there are also countless recurring combinations of words, or "word partnerships." For example, the verb *conduct* often occurs with the noun *research*, and the adjective *contagious* often occurs with the noun *disease*. Such word partnerships are called *collocations* in the literature. Many scholars distinguish collocations from idioms, because understanding the constituent words of a collocation is thought to be sufficient to understand the whole expression.

Unfortunately, the use of MWE terminology is not always consistent in the literature. John Sinclair, one of the founding fathers of modern corpus linguistics, considered all conventionalized word combinations a manifestation of "idiomaticity" (what he called "the idiom principle"; Sinclair, 1991). The term idiomaticity can thus be understood in a broad sense (referring to a language's conventional word combinations in general) as Sinclair did, or in a narrower sense, where it refers only to non-compositional idioms (Grant & Bauer, 2004).

Dictionaries of English idioms (which usually include some proverbs and similes as well) typically list up to 10,000 expressions (e.g., Ayto, 2020). Adopting the broader definition of idiomaticity entails recognizing a substantially larger collection of MWEs. The *Oxford Collocations Dictionary for Learners of English* (2nd ed., McIntosh, 2009), for example, includes more than 250,000 word combinations. Because it is not easy to delineate the broad class of MWEs, it is difficult to determine what proportion of discourse is made up of MWEs. Some researchers have claimed it may be more than half (e.g., Erman & Warren, 2000). Clearly, then, L2 vocabulary programs cannot neglect this phraseological dimension of lexis. But where to begin if the language's stock of MWEs is so large?

Perspectives on Identifying and Prioritizing Multiword Expressions

Focus on Non-Transparent Expressions

We made a distinction earlier between idioms and collocations, where idioms were characterized as semantically non-compositional (e.g., Cacciari & Glucksberg, 1991). For example, the meaning of *a close call* ('very nearly escaping disaster') does not follow straightforwardly from considering the individual meanings of *close* and *call*. We characterized standardized word combinations that *are* semantically compositional as collocations. If learners understand the meaning of the constituent words, then they will understand the expression, so the reasoning goes. For example, if a learner understands both *conduct* and *research*, then the phrase *conduct research* will be semantically transparent and therefore, in this scheme, will be called a collocation rather than an idiom.

The notion of compositionality (and by extension transparency) is not so clear-cut, however. A collocation that is transparent to one person (e.g., an L1 user) may not be to another (e.g., an L2 learner) even if the constituent words look familiar (Boers & Webb, 2015). This is because words that make up a collocation may not be used in their primary or "basic" sense (see Chapter 1). For example, *pay* in *pay a visit* is not used in its sense of financial transaction, *run* in *run risks* is not used in its sense of physical motion, and *make the bed* does not mean creating or assembling a bed. The question whether a given MWE is transparent should therefore be specified as "transparent to whom?"

Within the class of idioms, not all expressions are equally opaque either (Moon, 1998, p. 8; Titone & Connine, 1999). Some have a figurative meaning that follows rather smoothly from a literal reading of the phrase. For example, if one is familiar with traffic lights, then the idiom *get the green light* (receive permission to proceed with a certain plan) is likely to be transparent if it is encountered in context. In comparison, the idiom *face the music* (accept criticism or punishment for something you have done wrong) is probably not as easy to interpret, because one does not usually associate music with a negative experience. We will say more about figurative idioms and ways of helping learners deal with them in Chapters 4 and 9. Of course, some phrases listed in some idiom dictionaries (e.g., *by and large*) are very hard to trace back to a literal use (Grant & Bauer, 2004).

Idioms and collocations each pose challenges for learners. In the case of collocations, learners need to remember the appropriate word combinations (e.g., *conduct research*, not *perform research*, or *make research*) if they want their productive use of L2 to sound "right." In comparison, idioms pose a double challenge—learning their meaning as well as learning their lexical makeup. From this perspective, it could be argued that it is idioms (e.g., *by a long shot*; *sit on the fence*; *the worm turns*) that language learners need most assistance with.

The likely non-transparency of MWEs can signal potential comprehension problems for language learners, but, as we have illustrated, the signals are not always clear in actual practice. Transparency varies among language users and is not an either–or phenomenon. It is nonetheless a factor considered by many researchers interested in singling out MWEs for learning and teaching. For example, Martinez and Schmitt (2012) compiled a useful list of "phrasal expressions" (downloadable from www.norbertschmitt.co.uk/vocabulary-resources) for this purpose by considering not only their high frequency but also their likely non-transparency.

Focus on Frequent Expressions

Another approach is to simply prioritize MWEs that are used very frequently and that must for that reason alone be useful to learn. The frequency-oriented approach to identifying and selecting MWEs is heavily based on the use of corpora (see Chapter 1). The most straightforward method is to use software to extract word strings that occur a predetermined number of times per, say, one million words of text—usually in a certain genre, such as academic discourse (e.g., Biber et al., 2004; Liu, 2012). These highly frequent word strings (or *n*-grams) are sometimes called *lexical bundles* in the literature. Depending on the frequency threshold set by the researchers, they may include word sequences such as *a lot of*, *as well as*, *in order to*, *at the same time*, and *on the basis of*, which many teachers of (pre-)intermediate students will indeed recognize as phrases worth directing their students' attention to. However, the software will also pick up highly frequent strings such as *and of the* and *or in a*, which do not correspond to what language users may consider a lexical unit.

Going by frequency alone may therefore not suffice to single out meaningful MWEs for learning and teaching. Asking language users which of a collection of highly frequent word strings they consider to be MWEs may then help. This was the approach taken by Simpson-Vlach and Ellis (2010), for example, when they set out to compile a list of expressions that are valuable for EAP (English for academic purposes). They first extracted highly frequent word sequences from a corpus of academic discourse and then consulted experienced teachers about whether these were indeed "phrases" and worth teaching. Their resulting *Academic Formulas List* excludes strings like *and of the*, which the teachers did not consider to be real phrases, but includes strings like *as a consequence*, which did correspond to the teachers' intuitions.

Focus on Prefabs

Many MWEs are used by L1 users as prefabricated units, functioning as though they were single words. Such MWEs seem to be stored in memory as ready-made chunks that can be retrieved without any need to assemble them word by word. Wray (2002, p. 9) calls such MWEs *formulaic sequences*. She defines formulaic sequence as "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar." This definition highlights that a formulaic sequence travels as a unit. That is how a speaker learns it, stores it mentally, and retrieves it from memory. Someone using *as a matter of fact* probably pulls from memory the whole sequence, unified, rather than in pieces (that is, *as* + *a* + *matter* + *of* + *fact*).

It is important to note that Wray (2002) had L1 users in mind first and foremost when she described formulaic sequences. In fact, she argued that post-childhood *L2* learners are much less likely than L1 speakers to have at their disposal a large repertoire of MWEs that are stored in memory as unanalyzed prefabs. When you start learning an additional language after childhood, you will have started considering words as the building blocks of discourse. This is in part a side effect of learning to read and write—words are presented as distinct elements in written texts (i.e., with blanks between them). Unlike the way children segment speech into semantic units that are very often larger than single words (Peters, 1983), the tendency among post-childhood L2 learners is to approach MWEs as combinations of words instead of prefabricated "chunks," so the theory goes.

If formulaic sequences are retrieved from memory as prefabricated chunks, then they should be expected to benefit fluency (smooth comprehension and production of discourse despite time pressure). They will aid receptive fluency because on hearing the beginning of a formulaic sequence (e.g., *as a matter* . . .), you will be able to predict what follows (. . . of fact), and so you do not need to carefully process every single word. Formulaic sequences will aid productive fluency because they provide stepping-stones (or "islands of reliability"; Dechert, 1983), so to speak, between other, more "creative," segments of language use. Research has indeed shown that speakers deploy their arsenal of formulaic sequences especially when they need to communicate under time pressure and so there is a high demand for fluency. Kuiper (1995), for instance, found a much greater abundance of formulaic sequences in reporters' live commentaries of fast-paced sports, such as horse racing, than in slow-paced ones, such as cricket. Research with L2 learners has also found a significant association between speaking fluency and the use of MWEs, provided the learners know the MWEs well enough to produce them without hesitation (e.g., Tavakoli & Uchihara, 2020; see Chapter 3 for a review of this strand of research).

Whether all MWEs used by an L1 speaker should be considered formulaic sequences, as defined by Wray (2002), is difficult to determine. Conversely, we cannot rule out the possibility that a fair number of MWEs are acquired holistically by L2 learners as well. Besides, an MWE that is initially learned as a combination of words may eventually be represented in the learner's mental lexicon as a unified whole, thanks to frequent exposure and use. Still, it is reasonable to assume that L2 learners are more inclined than L1 users to experience MWEs as combinations of words instead of prefabs. That having been said, when Wray (2002) characterizes formulaic sequences as being stored and processed holistically, this does not preclude the possibility of language users, including L1 users, adopting an analytic approach to an expression they acquired as an unanalyzed prefab. You may have acquired a sequence like *as a matter of fact* as a single chunk, but later you can take a step back and see how the meanings of the constituent words add up (e.g., *of fact* is an adjectival prepositional phrase modifying *matter*). This expression is clearly compositional and analyzable. This is not always the case. Many formulaic expressions—like *for crying out loud* or *kangaroo court*—have such obscure origins that they resist analysis except by unusually knowledgeable speakers. Nonetheless, language users *can* adopt a meta-linguistic stance and attempt to analyze a formulaic sequence, but—if the sequence is well known—they will under normal conditions retrieve it from memory as a prefabricated chunk, similarly to how they retrieve well-known single words. It stands to reason that it is the latter, "natural" use of formulaic sequences that is associated with fluency.

When it comes to identifying formulaic sequences in discourse, it can be challenging to agree on what does and what does not constitute a unitary chunk. Researchers who have tried this typically rely on several coders and their intercoder agreement to identify such units (e.g., Eyckmans et al., 2007; Stengers et al., 2011). Majorana and Zwier (unpublished) administered an informal survey to 39 directors of Intensive English Programs in North America in order to discover their opinions about how many "vocabulary items" were contained in certain common sequences of words. The majority considered the sequences *pick up* (as in *pick up a phone*) and *in other words* single units, but *pay attention to* was considered a single unit by only half of them. The majority also identified the idiom *cat got your tongue* as a single vocabulary item. The latter suggests that the length of the sequence does not determine whether it is perceived to be a formulaic unit. For example, most respondents considered the short sequence *my opinion* as a sequence of two items, *my* and *opinion*, rather than a unit. Although *my* and *opinion* frequently co-occur, they lack the "bondedness" that you find in, for example, *in other words* and *in my opinion*.

Proficient speakers of English sense that the complex preposition *in front of* is more tightly bonded than the prepositional phrase *inside a box*. Not that there is anything odd about *inside a box*. Intuition tells us that the combination of *inside* with *box* is common in contexts of all sorts. Nonetheless, something special is going on with *in front of*, something that is not operating with *inside a box*. That special thing is hard to quantify, but it has a psycholinguistic reality. Incidentally, we chose *inside* ***a*** *box* as our loosely bonded example—not *inside* ***the*** *box*—because *inside the box* could be considered tightly bonded when it means "within conventional limits," as in *We'll never solve this problem if we keep thinking inside the box*.

Focus on Strong Collocations

The prioritization of highly frequent word sequences in L2 learning and teaching makes good sense, but it stands to reason that, if the aim is to progress beyond intermediate proficiency, less frequent expressions merit attention as well. The most frequent MWEs consist of highly frequent words, and highly frequent words in turn feature in numerous MWEs. For example, the high-frequency verb *make* is part of *make a suggestion*, *make no mistake*, *make a presentation*, *make an effort*, *make progress*, *make a comment*, *make a mess of something*, *make a fool of yourself*, *make friends*, *make money*, *make amends*, *make yourself understood*, *make up*, *make up for*, *make out*, and countless other phrases. Less frequent words, by comparison, usually prefer the company of fewer other words. For instance, *perform* also features in various expressions (e.g., *perform a play*, *perform a piece of music*, *perform a task*, *perform a function*, and *perform a miracle*), but this is a smaller set than in the case of *make*. At the lower end of the frequency continuum, such sets of typical combinations (or collocations) get smaller still. For example, the verb *wreak* occurs in *wreak havoc*, *wreak destruction*, *wreak vengeance*, and just a few other common combinations.

The greater-than-chance likelihood of a word co-occurring with a certain other word is the criterion used by corpus linguists to call a word combination a *collocation*. Recall that in another perspective that we discussed earlier, collocation referred narrowly to MWEs which, unlike idioms, are considered compositional and thus semantically transparent (provided one understands the constituent words). In the perspective discussed here, however, collocation refers to the above-chance likelihood of co-occurrence of words—according to corpus counts—regardless of the semantic characteristics of the phrase. You will thus occasionally see the term *collocation* used in the literature with reference to expressions that would traditionally be considered idioms. For example, *pull strings* is an idiom, but at the same time it manifests an above-chance co-occurrence of *pull* and *strings*, and so it manifests the phenomenon of collocation in this broader sense.

A well-established statistical measurement of association, or strength of partnership, is mutual information (MI), a tool used in information science. In very simple terms, it measures how often a combination occurs as compared to how often it would occur by chance given all the other possible combinations that are theoretically available. For example, the noun *book* will frequently co-occur with a variety of verbs, including *read*, *write*, *buy*, *publish*, *recommend*, and *finish*, whereas the noun *suicide* will occur almost exclusively with *commit*. The more exclusive the partnership, the higher the MI score will be.

Importantly, MI scores are different from raw frequency counts of word combinations. A combination of very common words (e.g., *make something*) may occur very often in a corpus, but because each of the words also occurs in the company of countless other words, the MI score will be relatively low. The combination *excruciating pain*, by comparison, will occur far less often in the corpus, but the MI score will be high because *excruciating* occurs in the company of only a small number of nouns other than *pain* (e.g., *excruciating agony*, *excruciating headache*, *excruciating ordeal*). It follows that very high MI scores are found more often for combinations of (non-frequent) content words (e.g., *commit suicide*) than for combinations involving function words, such as prepositions (e.g., *on the phone*). This is because the latter co-occur with so many other words. Word-partnerships of content words are sometimes called lexical collocations, while ones involving function words are called grammatical collocations.

A list of close to 2,500 lexical collocations that are used frequently in English academic discourse was compiled by Ackermann and Chen (2013) and can be found at www.eapfoundation.com/vocab/academic/acl/. Like Simpson-Vlach and Ellis (2010), these researchers did not just rely on corpus data but also consulted teachers about which MWEs to include. Ackermann and Chen's *Academic Collocations List* gives precedence to strongly associated content words (e.g., *adverse effect*; *adopt an approach*) rather than highly frequent word strings and is therefore more useful for learners at higher proficiency levels than the other MWE lists mentioned this far.

In Chapter 1, we introduced collocation as one of the many things that needs to be learned about a word (along with a word's pronunciation, spelling, meanings, grammatical patterning, and register). So, you may wonder if collocation should not be thought of as an aspect of word knowledge instead of as a kind of lexical unit. The answer is not clear-cut. Some collocations "feel" more like a single lexical item than others. Some collocations are very strong (e.g., *hard-boiled egg*), some moderately strong (*hard-boiled detective*), and some relatively weak (*hard-boiled attitude*). At the strong end of the continuum, a word sequence is more likely to be perceived as a *unit*, while at the weaker end it is probably perceived as a word *combination*. For those reasons, we will in this book consider collocation knowledge from both angles—as an aspect of word knowledge and as knowledge of multiword expressions.

Like many other aspects of lexis, the collocational strength of word combinations is susceptible to change over time, and what may start out as a weak collocation (e.g., *do an experiment*) may through repeated use gain in strength and begin to "compete" with a long-established stronger collocation (e.g., *conduct an experiment*). Many disagreements about what should be regarded as "proper" language use have centered on whether a sequence is collocationally appropriate. The verb *center* itself offers a prime example. Should it be followed by *on* or by *around*? Usage authorities have long opposed *center around* because they see it as illogical: The center of a circle does not go around anything but rather locates at one point. Yet common usage has perpetuated *center around*. Now, with corpus tools, you can check for yourself to what degree *center* collocates with *around* or with *on*. A quick search through the Corpus of Contemporary American English (COCA; Davies, 2008) portrays it as collocating with both, though much more frequently with *on* than with *around*. Still, *center* and *around* do collocate with moderate strength, no matter what prescriptions there may be against their liaison. With the advantage of such research, modern usage guides have started relaxing their strictures against it.

Collocational patterns are far more than a sport for mavens. Without knowledge of at least the most common collocational possibilities for a vocabulary item, a learner lacks productive control over the item and may have difficulty interpreting it in a written or spoken context. As obvious as this is, collocations received relatively little emphasis in vocabulary teaching (and were underplayed in some dictionaries) until the last years of the twentieth century. Only in the early 1990s were collocations and their importance brought truly front and center by lexis-focused volumes like John Sinclair's *Corpus, Concordance, Collocation* (1991), Nattinger and DeCarrico's *Lexical Phrases and Language Teaching* (1992), and Michael Lewis's *The Lexical Approach* (1993). The influence of these relatively accessible works gave working teachers and curriculum developers a window on patterns that corpus researchers had been explicating in more arcane form for decades.

The general view among applied linguists is that collocation is arbitrary, meaning there is no compelling explanation for why two words form a partnership. Why does *make* collocate with *call*, *dinner*, and *bed*, whereas *do* collocates with *homework*, *puzzle*, and *the dishes*? Why is *woefully* about three times as likely to occur before *inadequate* as before any other adjective? Why does *utterly* have strong associations with *ridiculous*, *different*, and *useless*? To some extent, these collocations are indeed matters of convention, and it is often said that learners cannot reason them out and predict them. Walker (2008, p. 291) offers a less absolute characterization, noting that "collocation need not be arbitrary at all . . . [V]arious nouns and verbs invite distinct, 'characteristic collocates' that fit their semantic make-up, their etymology, their prototypical literal sense, and their semantic prosody." Liu (2010) shares this view and illustrates how knowledge of the basic meaning of, for example, *tall* can help to explain the different collocations (e.g., *tall building*, not *high building*). Others (e.g., Boers & Lindstromberg, 2009, pp. 106–119) have pointed out that some collocations are more likely than others, owing to catchy phonological repetition such as rhyme (e.g., *deep sleep*) and alliteration (e.g., *fast food*).

Summary

The perspectives on MWEs outlined in this chapter do not contradict one another. Each highlights a dimension of phraseology that another is less interested in. When considering students' needs as well as their challenges in acquiring MWEs, course designers and teachers may want to give special attention to MWEs that (a) are highly useful, going by their frequency of use; (b) show a high degree of bondedness, going by how exclusive their word partnership is; and (c) are not easy to understand even if the constituent words look familiar. Criteria (a) and (b) are especially pertinent if the aim is for learners to be able to incorporate MWEs fluently and "accurately" in their own L2 discourse, while criterion (c) is especially important to assist them with comprehending L2 discourse.