Introduction: What This Book Is about and How It Can Be Used

The existence of words is usually taken for granted by the speakers of a language. To speak and understand a language means – among many other things – knowing the words of that language. The average speaker knows thousands of words, and new words enter our minds and our language on a daily basis. This book is about words. More specifically, it deals with the internal structure of complex words, i.e. words that are composed of more than one meaningful element. Take, for example, the word *meaningful*, which could be argued to consist of two elements, *meaning* and *-ful*, or even three, *mean*, *-ing*, and *-ful*. We will address the question of how such words are related to other words and how the language allows speakers to create new words. For example, *meaningful* seems to be clearly related to *colorful*, but perhaps less so to *awful* or *plentiful*. And, given that *meaningful* may be paraphrased as 'having (a definite) meaning,' and *colorful* as 'having (bright or many different) colors,' we could ask whether it is also possible to create the word *coffeeful*, meaning 'having coffee.' Under the assumption that language is a rule-governed system, it should be possible to find meaningful answers to such questions.

This area of study is traditionally referred to as 'word-formation,' and the present book is mainly concerned with word-formation in one particular language, English. As a textbook for an undergraduate readership, it presupposes very little or no prior knowledge of linguistics and introduces and explains linguistic terminology and theoretical apparatus as we go along. Important technical terms appear in bold print when first mentioned. Definitions of terms can be easily located via the Subject Index, in which the respective page numbers are given in bold print.

The purpose of the book is to enable students to engage in (and enjoy!) their own analyses of English (or other languages') complex words. After having worked with the book, the reader should be familiar with the necessary and most recent methodological tools to obtain relevant data, such as introspection, electronic text collections (known as 'corpora'), linguistic databases, various types of dictionaries, basic psycholinguistic experiments, and various other resources available on the internet. Furthermore, the reader should be able to systematically analyze their data and to relate their findings to theoretical problems and debates. The book is not written from the perspective of a particular theoretical framework and draws on insights from various research traditions.

Word-Formation in English can be used as a textbook for a course on word-formation (or for the word-formation parts of morphology courses), as a source book for teachers, as a guidebook for student research projects, as a book for self-study by more advanced students (e.g. for their exam preparation), and as an up-to-date reference concerning selected word-formation processes in English for a more general readership.

For each chapter there are a number of basic and more advanced exercises, which are suitable for in-class work or as students' homework assignments. The more advanced exercises include proper research tasks, which also give the students the opportunity to use the different methodological tools introduced in the text. Students can control their learning success by comparing their results with the answer key provided at the end of the book. The answer key features two kinds of answers. Basic exercises always receive definite answers, while for the more advanced tasks sometimes no 'correct' answers are given. Instead, methodological problems and possible lines of analysis are discussed. Even readers not interested in working on the exercises may find it fruitful to read the answer key texts for the advanced exercises, since they broaden and deepen the discussion of certain questions raised in the pertinent chapters.

Those who consult the book as a general reference on English word-formation may check subject, affix, and author indexes and the list of references in order to quickly find what they need. Each chapter is also followed by a list of recommended further reading.

As every reader knows, English is spoken by hundreds of millions of people, and there exist numerous varieties of English around the world. The variety that has been taken as a reference for this book is General American English. The reason for this choice is purely practical: it is the variety the author knows best. Although there are sometimes differences observable between different varieties (especially concerning aspects of pronunciation), with regard to most of the phenomena discussed in this book, different varieties of English show very similar patterns. Mostly for reasons of space, but also due to the lack of pertinent studies, existing differences will not be discussed. However, the book should hopefully enable the readers to adapt and relate the findings presented with reference to American English to the variety of English they are most familiar with.

The structure of the book is as follows. Chapters 1 through 3 introduce the basic notions needed for the study and description of word-internal structure (Chapter 1), the problems that arise with the implementation of these notions in the actual analysis of complex words in English (Chapter 2), and one of the central problems in word-formation, productivity (Chapter 3). The descriptively oriented Chapters 4 through 6 deal with the different kinds of word-formation processes that can be found in English: Chapter 4 discusses affixation, Chapter 5 non-affixational processes, and Chapter 6 compounding. Chapter 7 is devoted to two theoretical issues, the role of phonology in word-formation and the nature of word-formation rules.

The author welcomes comments and feedback on all aspects of this book, especially from students. Without students telling their teachers what is good for them (i.e. for the students), teaching cannot become as effective and enjoyable as it should be for both teachers and teachees (oops, was that a possible word of English?).

1 Basic Concepts

Outline

This chapter introduces basic concepts needed for the study and description of morphologically complex words. Since this is a book about the particular branch of morphology called word-formation, we will first take a look at the notion of 'word.' We will then turn to a first analysis of the kinds of phenomena that fall into the domain of word-formation, before we finally discuss how word-formation can be distinguished from the other sub-branch of morphology, inflection.

1.1 What Is a Word?

It has been estimated that average speakers of a language know from 45,000 to 60,000 words. This means that we as speakers must have stored these words somewhere in our heads, in our **mental lexicon**. But what exactly is it that we have stored? What do we mean when we speak of 'words'?

In non-technical everyday talk, we speak about 'words' without ever thinking that this could be a problematic notion. In this section we will see that, perhaps contra our first intuitive feeling, the 'word' as a linguistic unit deserves some attention because it is not as straightforward as one might expect.

If you had to define what a word is, you might first think of the word as a unit in the writing system, the **orthographic word**. You could say, for example, that a word is an uninterrupted string of letters which is preceded by a blank space and followed by either a blank space or a punctuation mark. At first sight, this looks like a good definition that can be easily applied, as we can see in the sentence in example (1):

(1) Linguistics is a fascinating subject.

We count five orthographic words: there are five uninterrupted strings of letters, all of which are preceded by a blank space, four of which are also followed by a blank space, one of which is followed by a period. This count is also in accordance with our intuitive feeling of what a word is. Even without this somewhat formal and technical definition, you might want to argue, you could have told that the sentence in (1) contains five words. However, things are not always that straightforward. Consider the following example, and try to determine how many words there are:

(2) Benjamin's girlfriend lives in a high-rise apartment building.

Your result depends on a number of assumptions. If you consider apostrophes to be punctuation marks, *Benjamin's* constitutes two (orthographic) words. If not, *Benjamin's* is one word. If you consider a hyphen a punctuation mark, *high-rise* is two (orthographic) words, otherwise it's one (orthographic) word. The last two strings, *apartment building*, are easy to classify – they are two (orthographic) words – whereas *girlfriend* must be considered one (orthographic) word. However, there are two basic problems with our orthographic analysis. The first one is that orthography is often variable. Thus, *girlfriend* is also attested with the spellings <girl-friend> and even <girl friend> (fish brackets are used to indicate spellings, i.e. letters). Such variable spellings are quite common (cf. *word-formation, word formation*, and *wordformation*, all of them attested), and even where the spelling is conventionalized, similar words are often spelled differently, as evidenced with *grapefruit* vs. *passion fruit*. For our problem of defining what a word is, such cases are rather annoying. The notion of what a word is, should, after all, not depend on the arbitrariness of the English spelling system or on the fancies of individual writers. The second problem with the orthographically defined word is that it may not always coincide with our intuitions. Thus, most of us would probably agree that *girlfriend* is a word (i.e. one word) which consists of two words (*girl* and *friend*). The technical term for such words is **compound**. If compounds are considered one word, they should be spelled without a blank space separating the elements that together make up the compound. In fact, some spelling systems, for example, German, do treat compounds in this way. But English doesn't. The compound *apartment building*, for example, has a blank space between *apartment* and *building*.

To summarize our discussion of purely orthographic criteria of wordhood, we must say that these criteria are not entirely reliable. Furthermore, a purely orthographic notion of 'word' would have the disadvantage of implying that illiterate speakers would have no idea about what a word might be. This is plainly false.

What, you might ask, is responsible for our intuitions about what a word is, if not the orthography? It has been argued that the word could be defined in four other ways: in terms of sound structure (i.e. phonologically), in terms of its internal integrity, in terms of meaning (i.e. semantically), or in terms of sentence structure (i.e. syntactically). We will discuss each in turn.

You might think that blank spaces in writing reflect pauses in the spoken language, and that perhaps one could define the word as a unit in speech surrounded by pauses. However, if you carefully listen to naturally occurring speech you will realize that speakers do not make pauses before or after each word. Perhaps we could say that words can be surrounded by potential pauses in speech. This criterion works much better, but it runs into problems because speakers can and do make pauses not only between words but also between syllables, for example for emphasis.

But there is another way in which the sound structure can tell us something about the nature of the word as a linguistic unit. Think of stress. In many languages (including English) the word is the unit that is crucial for the occurrence and distribution of stress. Spoken in isolation, every word can have only one **main stress** (also called **primary stress** if there are other stressed syllables in the word), as indicated by the acute accents (ˊ) in the data presented in (3) below (note that we speak of linguistic 'data' when we refer to language examples to be analyzed).

(3)

cárpenter	téxtbook
wáter	análysis
féderal	sýllable
móther	understánd

The main stressed syllable is the syllable which is the most prominent one in a word. Prominence of a syllable is a function of loudness, pitch, and duration. Stressed syllables are pronounced louder, with higher pitch, or with longer duration than the neighboring syllable(s). Longer words often have additional, weaker stresses, commonly named **secondary stresses**, which we ignore here for simplicity's sake. The words in (4) now show that the phonologically defined word is not always identical with the orthographically defined word.

(4)

Bénjamin's
gírlfriend
apártment building

While *apártment building* is two orthographic words, it is only one word in terms of stress behavior. The same holds for other compounds like *trável agency, wéather forecast, spáce shuttle*, etc. We see that in these examples the phonological definition of 'word' comes closer to our intuition of what a word should be.

We have to take into consideration, however, that not all words carry stress. For example, function words like articles or auxiliaries are usually unstressed (*a cár*, *the dóg*, *Máry has a dóg*) or even reduced to a single sound (*Jane's in the garden*, *I'll be there*). Hence, the stress criterion is not readily applicable to function words, and to words that hang on to other words (commonly referred to as clitics, e.g. *'ve*, *'s*, *'ll*).

Let us now consider the integrity criterion, which says that the word is an indivisible unit into which no intervening material may be inserted. If some modificational element is added to a word, it must be done at the edges, but never inside the word. For example, plural endings such as *-s* in *girls*, negative elements such as *un-* in *uncommon,* or endings that create verbs out of adjectives (such as *-ize* in *colonialize*) never occur inside the word they modify but are added either before or after the word; hence, the impossibility of formations such as **gi-s-rl*, **com-un-mon*, **col-ize-onial* (note that the asterisk indicates impossible words, i.e. words that are not formed in accordance with the morphological rules of the language in question).

However, there are some cases in which word integrity is violated. For example, the plural of *son-in-law* is not **son-in-laws* but *sons-in-law*. Under the assumption that *son-in-law* is one word, the plural ending is inserted inside the word and not at the end. Apart from certain compounds, we can find other words that violate the integrity criterion for words. For example, in creations like *abso-bloody-lutely*, the element *bloody* is inserted inside the word, and not, as we would expect, attached at one of the edges. In fact, it is impossible to add *bloody* before or after *absolutely* in order to achieve the same effect. *Absolutely bloody* would mean something completely different, and **bloody absolutely* seems utterly strange and uninterpretable.

We can conclude that there are certain, though marginal, counterexamples to the integrity criterion, but surely these cases should be regarded as the proverbial exceptions that prove the rule.

The semantic definition of 'word' states that a word expresses a unified semantic concept. Although this may be true for most words (even for *son-in-law*, which is ill-behaved with regard to the integrity criterion), it is not sufficient in order to differentiate between words and non-words. The simple reason is that not every unified semantic concept corresponds to one word in a given language. Consider, for example, the smell of fresh rain in a forest in the fall. Certainly a unified concept, but we would not consider *the smell of fresh rain in a forest in the fall* a word. In fact, English simply has no single word for this concept. A similar problem arises with phrases like *the woman who lives next door*. This phrase refers to a particular person and should therefore be considered as something expressing a unified concept. This concept is, however, expressed by more than one word. We learn from these examples that although a word may always express a unified concept, not every unified concept is expressed by one word. Hence, the criterion is not very helpful in distinguishing between words and larger units that are not words. An additional problem arises from the notion of 'unified semantic concept' itself, which seems to be rather vague. For example, does the complicated word *conventionalization* really express a unified concept? If we paraphrase it as 'the act or result of making something conventional,' it is not entirely clear whether this should still be regarded as a 'unified concept.' Before taking the semantic definition of 'word' seriously, it would be necessary to define exactly what 'unified concept' means.

This leaves us with the syntactically oriented criterion of wordhood. Words are usually considered to be syntactic atoms, i.e. the smallest elements syntactic rules can refer to. For example, if we say that auxiliary verbs occupy the initial position in 'yes/no' questions (e.g. *Can you come tomorrow?* vs. *You can come tomorrow*), this is evidence that auxiliary verbs are words. This may not come as a surprise to you, as it is probably this criterion that underlies our intuitions about where to put spaces in writing.

Words as syntactic atoms belong to certain syntactic classes (nouns, verbs, adjectives, prepositions, etc.), which are called **parts of speech**, **word classes**, or **syntactic categories**. The position in which a given word may occur in a

sentence is determined by the syntactic rules of a language. These rules make reference to words and the class they belong to. For example, *the* is said to belong to the class called 'articles,' and there are rules which determine where in a sentence such words, i.e. articles, may occur (usually before nouns and their modifiers, as in *the big house*). We can therefore test whether something is a word by checking whether it belongs to such a word class. If the item in question, for example, follows the rules for nouns, it should be a noun, hence a word. Or consider the fact that only words (and groups of words), but no smaller units, can be moved to a different position in the sentence. For example, in 'yes/no' questions, the auxiliary verb does not occur in its usual position but is moved to the beginning of the sentence (*You **can** read my textbook* vs. ***Can** you read my textbook?*). Hence the auxiliary verb must be a word. Thus syntactic criteria can help to determine the wordhood of a given entity.

To summarize our discussion of the possible definition of 'word,' we can say that, in spite of the intuitive appeal of the notion of 'word,' it is sometimes not easy to decide whether a given string of sounds (or letters) should be regarded as a word or not. In the treatment above, we have concentrated on the discussion of such problematic cases. In most cases, however, the stress criterion, the integrity criterion, and the syntactic criteria lead to sufficiently clear results. The properties of words are summarized in (5):

(5) Properties of words
- words are entities having a part-of-speech specification
- words are syntactic atoms
- words (usually) have one main stress
- words (usually) are indivisible units (no intervening material possible)

Unfortunately, there is yet another problem with the word *word* itself. Thus, even if we have unequivocally decided that a given string is a word, some insecurity remains about what exactly we refer to when we say things like

(6) a. The word *be* occurs twice in the sentence.
b. [ðəwɜːrdbiəkɜːrztwaɪsɪnðəsentəns]

For reasons that will become clear below, the utterance in (6) is given in two forms, in the normal spelling (6a), and in phonetic transcription (6b) (throughout the book I will use phonetic transcriptions as given in the *Longman Dictionary of Contemporary English* (*LDCE*), for the North American English pronunciation).

(6) can be understood in different ways. First, <be> or the sounds [bi] may refer to the letters or the sounds which they stand for. Then sentence (6) would, for example, be true for every written sentence in which the string <BLANK SPACE be BLANK SPACE> occurs twice. Referring to the spoken equivalent of (6a), represented by the phonetic transcription in (6b), (6) would be true for any sentence in which the string of sounds [bi] occurs twice. In this case, [bi] could refer to two different 'words,' e.g. *bee* and *be*. The third possible interpretation is that in (6) we refer to the grammatically specified form *be*, i.e. the infinitive,

imperative, or subjunctive form of the linking verb BE (as in *To be or not to be…*, *Be quiet!*, *Whether they be friend or foe…*, respectively). Such a grammatically specified form is called the **grammatical word** (or **morphosyntactic word**). Under this reading, (6) would be true of any sentence containing two infinitive, two imperative, or two subjunctive forms of *be*, but would not be true of a sentence which contains any of the forms *am*, *is*, *are*, *was*, *were*.

To complicate matters further, even the same form can stand for more than one different grammatical word. Thus, the **word-form** *be* is used for three different grammatical words, expressing subjunctive, infinitive, or imperative, respectively. This brings us to the last possible interpretation, namely that (6) may refer to the linking verb BE in general, as we would find it in a dictionary entry, abstracting away from the different word-forms in which the word BE occurs (*am*, *is*, *are*, *was*, *were*, *be*, *being*, *been*). Under this reading, (6) would be true for any sentence containing any two word-forms of the linking verb, i.e. *am*, *is*, *are*, *was*, *were*, *be*, *being*, and *been*. Under this interpretation, *am*, *is*, *are*, *was*, *were*, *be*, *being*, and *been* are regarded as realizations of an abstract morphological entity. Such abstract entities are called **lexemes**. Coming back to our previous example of *be* and *bee*, we could now say that BE and BEE are two different lexemes that simply sound the same (usually small capitals are used when writing about lexemes). In technical terms, they are **homophonous** words, or simply **homophones**.

In everyday speech, these rather subtle ambiguities in our use of the term 'word' are easily tolerated and are often not even noticed, but when discussing linguistics, it is sometimes necessary to be more explicit about what exactly one talks about. Having discussed what we can mean when we speak of 'words,' we may now turn to the question of what exactly we are dealing with in the study of word-formation.

1.2 Studying Word-Formation

As the term 'word-formation' suggests, we are dealing with the formation of words, but what does that mean? Let us look at a number of words that fall into the domain of word-formation and a number of words that do not:

(7)

a.	b.	c.
employee	apartment building	chair
inventor	greenhouse	neighbor
inability	team manager	matter
meaningless	truck driver	brow
suddenness	blackboard	great
unhappy	son-in-law	promise
decolonialization	pickpocket	discuss

In columns (7a) and (7b), we find words that are obviously composed by putting together smaller elements to form larger words with more complex meanings. We can say that we are dealing with morphologically **complex words**. For

example, *employee* can be analyzed as being composed of the verb *employ* and the ending *-ee*, and the adjective *unhappy* can be analyzed as being derived from the adjective *happy* by the attachment of the element *un-*. And the word *decolonialization* can be segmented into the smallest parts *de-*, *colony*, *-al*, *-ize*, and *-ation*. We can thus decompose complex words into their smallest meaningful units. These units are called **morphemes**.

In contrast to those in (7a) and (7b), the words in (7c) cannot be decomposed into smaller meaningful units, they consist of only one morpheme, they are called 'monomorphemic' or 'simplex.' *Neighbor*, for example, is not composed of *neighb-* and *-or*, although the word looks rather similar to a word such as *inventor*. *Inventor* ('someone who invents (something)') is decomposable into two morphemes, because both *invent-* and *-or* are meaningful elements, whereas neither *neighb-* nor *-or* carry any meaning in *neighbor* (a neighbor is not someone who neighbs, whatever that may be . . .).

As we can see from the complex words in (7a), some morphemes can occur only if attached to some other morpheme(s). Such morphemes are called **bound morphemes**, in contrast to **free morphemes**, which do occur on their own. Some bound morphemes, for example *un-*, must always be attached before the central meaningful element of the word, i.e. the **root**, **stem**, or **base**, whereas other bound morphemes, such as *-ity, -ness*, or *-less*, must follow the base. Using Latin-influenced terminology, *un-* is called a **prefix**, *-ity* a **suffix**, with **affix** being the cover term for all bound morphemes that attach to roots. Note that there are also **bound roots**, i.e. roots that occur only in combination with some other bound morpheme. Examples of bound roots are often of Latin origin, e.g. *circul-* (as in *circulate*, *circulation*, *circulatory*, *circular*), *approb-* (as in *approbate*, *approbation*, *approbatory*, *approbator*), or *simul-* (as in *simulant*, *simulate*, *simulation*), but occasional native bound roots can also be found (e.g. *hap-*, as in *hapless*).

Before we turn to the application of the terms introduced in this section, we should perhaps clarify the distinction between 'root,' 'stem,' and 'base,' because these terms are not always clearly defined in the morphological literature and are therefore a potential source of confusion. One reason for this lamentable lack of clarity is that languages differ remarkably in their morphological make-up, so that different terminologies reflect different organizational principles in the different languages. The part of a word which an affix is attached to is called **base**. We will use the term **root** to refer to bases that cannot be analyzed further into morphemes. The term 'stem' is usually used for bases of inflections, and occasionally also for bases of derivational affixes. To avoid terminological confusion, we will avoid the use of the term 'stem' altogether and speak of 'roots' and 'bases' only.

The term 'root' is used when we want to explicitly refer to the indivisible central part of a complex word. In all other cases, where the status of a form as indivisible or not is not at issue, we can just speak of **bases** (or, if the base is a word, of base words). The derived word is often referred to as a **derivative**. The base of the suffix *-al* in the derivative *colonial* is *colony*, the base of the