

# 1 Words

## 1.1 What is a word?

Lexicology is the study of words as linguistic entities. Because we live in a literate society, and have a writing system which sets out words on the page, we generally feel that we know what a word is. This view, unfortunately, confuses cause and effect. In principle, we know how to write English because we know what the words are; we do not know what the words are because we can write English. If this latter view were true, the many languages which have no writing system would not have words, and neither would languages whose writing system does not put spaces between words. On the other hand, if we know where to put the spaces because we know what words are, we might expect that we would be able to reconstruct a way of arriving at that conclusion, but this does not appear to be the case – at least not where English is concerned.

We can certainly see some criteria that provide clues. The clues do not always agree, though. Let us consider a few.

A word is the smallest unit which can stand on its own as an utterance. This is not as simple as it looks. First, we have to distinguish between language **use** and language **mention**. In answer to a question like ‘Is it inadvisable or unadvisable?’, we might answer ‘Un’. But we would not want to say that *un* is therefore a word; the objection would be that we are talking about *un*, not using it in its normal grammatical function. In this book, as in most linguistic texts, words that are mentioned will be put in italics. Even making that distinction, we have problems. In answer to a question such as *What do you hate most in the world?* We might give an answer *Cats*. So *cats* is a word. But it is hard to think of a place where we could have *cat* as an utterance by itself. **Grammatical words**, words which have a function in the grammar, like *of*, *and*, *if* or *the*, cannot be used as utterances by themselves unless they are mentioned. And if we think of something like *whiteboard*, which we write as one word, it

is arguably made up of two units which could stand on their own as utterances.

A word is a unit which, when pronounced in isolation, has a single intonational focus point, or movement of pitch. That point in the word might be perceived as being louder, longer or more important than the surrounding material. In *isolation*, that point would include the sound corresponding to the letter *A*, in *movement* it would include the sound corresponding to the letter *O*. We can call this the **main stress** of the word, and we would probably say that the whole syllable, and not just the vowel, carries the main stress. In *whiteboard* there is just one such point, including the sound corresponding to the letter *I*. Again, there are problems with this criterion. A phrase like *in the beginning* has just one such point, but is not usually thought of as a word. For some speakers a very long word like *antidisestablishmentarianism* might be perceived to have more than one main stress, though this is controversial. And, perhaps most important of all, the position of the main stress seems to be something that fewer and fewer speakers are secure about.

The sound structure of a word is governed by fixed rules which do not apply beyond the word. For example, a word can begin or end with /st/ (as in *stir* and *mist*), but while a word can end /ft/ (as in *soft*), it cannot begin with /ft/. The sequence /sd/ cannot occur either at the beginning or at the end of a word, so if we meet that sequence, it must be over a word-boundary (as in *pats dogs*). This is known as phonotactic structure, or **phonotactics**. Note that on this criterion, *whiteboard* must be a sequence of two words, because /tb/ does not occur at the beginning or end of a word (and occurs in the middle only in just such instances).

A word has a single, unitary meaning. The trouble with this is that we have no definition of a single, unitary meaning. Is *put up* (a guest) a single word because *lodge*, which has the same meaning, is a word? Is *in the end* a single word because it has the same meaning as *finally*, which is a word?

A word is listed in the **dictionary**. Like the spelling convention, this is circular. Things are listed in the dictionary primarily because they are words. In any case, as we shall see below (Section 1.3), there are things that are longer than words which get listed in the dictionary, and some dictionaries also list things that are shorter than words, such as **prefixes** and **suffixes**.

The overall conclusion is that, although these criteria might provide insights into something about the nature of the word, they do not define it neatly. The result is that we are not entirely sure what a word is. For example, we might not know how to write /ɪnsoʊfə:ræz/. Is it one, two, three or four words? People disagree on how to write it. And there is a

large set of words like *coffee-pot* or *rain-forest* that are sometimes written with a hyphen, sometimes as a single word and sometimes as two distinct words. The hyphen is also used where complex items occur in **attributive** position (that is, immediately before the noun it modifies). For example, *His position is well justified* corresponds to *His well-justified position*. This hyphen indicates syntactic constituency, but might also be taken to indicate wordhood.

Finally, we should note that what counts as a word is sometimes unstable, from a historical point of view. *A alright* has only relatively recently been accepted as a spelling of what was earlier *all right*. According to the *Oxford English Dictionary* (the *OED*), *after all* can be spelt *afterall* in the United States. In more recent times, we often see prefixes being written as separate words: things such as *mega deal*.

Despite all of this, there is a large amount of agreement among adult writers as to where words begin and end (children occasionally make what are, by adult standards, errors). There are a few places where there is variation, but people do not write *child ish*, even if they can use *ish* as a word in its own right: *Are you cold? Ish!* Even less do people write things like *difficult*, *sim mer*, *ze bra*, where there is no prefix or suffix involved. Much of this is related to meaning: words are (usually) meaningful units, and things which are not meaningful are not treated as words. This means that even if we must acknowledge areas of insecurity, in most instances speakers seem to be fairly clear about where the words are, and can repeat a sentence one word at a time – for example, if it is required for dictation. At various points in this book, we will need to come back to whether things are or are not words, but in general we will accept the spelling conventions of English as defining words. This might not be terribly scientific, but it has the benefit of being practical.

## 1.2 Are names words?

A word is a linguistic item which helps us discuss the world around us. It is a unit which not only has **form** (a phonological structure or an orthographic one), but also meaning. In this sense, names are words. But names are different from other words in a number of ways.

They have unique reference. *Samantha* refers to one particular person in a given context. *Tree* refers to a set of items, one or more of which is relevant in the given context. *I saw Samantha at the opera* means that I saw a particular individual who we both/all know is called 'Samantha'; *I saw a tree with red flowers* means that I saw a particular tree, but all you know about it is that it is a member of the class of trees with flowers of a particular colour. You might imagine a rhododendron, when I actually

saw a pohutukawa. We will agree that it was a member of the class of trees. One of the results of this is that *Samantha* is not easily made plural, while *tree* is easily pluralisable.

As a result of their unique reference, names are inherently definite. Sometimes names have a *the* in front of them (*The Hague*, *The Gambia*, *The Thames*) but there is rarely a contrast: we can have *Samantha* but not *The Samantha*, we can have *The Hague* but not *Hague*. Neither *Samantha* nor *The Hague* can mean 'any old Samantha/Hague'. Incidentally, this a fact about English: some languages do allow definite articles with names.

Both of those rules can be broken when a name is used as an ordinary word. This happens in sentences like those in (1). Precisely what happens in such cases is a matter of some dispute, but most authorities seem to agree that a proper noun (name) is treated as a common noun, and so takes on the grammar of a common noun.

- (1) This is not the Paris I used to know.

I know the Manchester in New Hampshire, but not the one in England.

We need another Einstein.

Johnny is the new Patricia.

There's a Mr Jones waiting to see you.

Matters are not this simple. There are plenty of unique descriptions which can be used as names: *The Netherlands*, *The Savoy*, *The Himalayas*, *The Duke of Wellington*. There are also unique entities which are named by ordinary nouns, which therefore become rather name-like: *the moon*, *fortune* (for example, in *Fortune favours the brave*), *hell* and so on. We should also note that while English spelling rules treat days of the week and months of the year as names, and use capital letters for them, other languages, such as French, do not.

Names, because of their meanings, also behave differently in regard to syntax from other words. Names, typically, act like a noun phrase rather than as a noun. *Kim*, in a sentence like *Kim read the entire book*, fills the whole subject function. We cannot add modifiers, defining relative clauses, quantifiers to it – or at least, if we do that, we end up with **constructions** like those in (1) where what was once a name has become a common noun. Even if we have a name that contains *the*, we cannot add adjectives or defining relatives without stopping it being a name: *The beautiful Hague*, *The former Argentine*, *The Gambia that I have come to love*. Most of our syntactic theories treat names as nouns, but we can argue that this is misleading.

All of this indicates that names are unusual words. But like normal words, they provide a phonetic (in sign language, a gestured) or written

representation to a unit to which we assign a semantic interpretation. If that is true of things other than words, it is nevertheless at the heart of wordhood.

### 1.3 Multi-word expressions (MWEs)

So far, the words we have dealt with have mostly been independent of each other: *cat* and *dog*, *cat* and *sat*, *cat* and *independent* have their own meanings, their own usages, their own dictionary entries. But sometimes there are things listed in dictionaries that contain more than one orthographic word. These are called **multi-word expressions** (MWEs for short, although you may also come across the label *MWU* – short for *multi-word unit*).

As a simple example, consider *passion flower* (etymologically, *passion* here has to do with Easter, rather than with strong feelings). Unlike a *magnolia flower*, which could just as easily be a *magnolia blossom*, a *passion flower* is the name of the flower: it is a fixed expression. It happens to be written as two words, unlike *mayflower* and *sunflower* (and you may not even be sure how to write *cornflower*, *wall-flower* or *wild flower*), but the way it is written does not appear to be significant; if we wrote *passion-flower*, nothing else would change. While the spelling of such items is not entirely random in English, it is at least clear that for many such words the spelling is not fixed and does not indicate anything about the status of the construction. So let us say that *passion flower*, *sunflower* and *wall-flower* are items whose names behave just like *cat* and *dog*. But unlike *cat* and *dog*, these words have words as their elements, so they are word-like objects which are made up of (in the examples we have considered) two words. These, then, termed **compounds**, are the most straightforward of MWEs.

Having decided that there are MWEs which have a place in a dictionary, though, we find that there is a plethora of items which might (or might not) need to be treated in the same way.

Another type is provided by what are often called **phrasal verbs**. These are verbs that include a preposition, things like *do up*, *fall out*, *put up with*. Although these are written as two (or more) words in English, and can be divided by other words, as in *do your coat up*, which is a generally accepted sign of separateness for words, their meaning is not always predictable from the meanings of their individual elements. While *fall out* may have a predictable meaning in *My handbag fell out of the car*, it does not in *Kim and Lee fell out over which wine to serve*. Because these verbs have meanings which are often unrelated to the meanings of the verb and the preposition, and because they often have meanings which can

be associated with single words (*He passed out* could be the same as *He fainted*), they seem to act as words, but they contain multiple words, and so are MWEs.

Lack of predictability of meaning is definitional of another class of MWEs, idioms. **Idioms** may be defined as constructions whose meaning is global (that is, the meaning attaches to the whole construction) rather than derived from the meanings of the elements of the construction. Consider the expression *piece of cake*. It does have a literal meaning, equivalent, perhaps, to *slice of gateau*, but in the sentence *The exam was a piece of cake*, the expression *piece of cake* can only be glossed globally as 'very easy'. It has nothing to do with pieces and nothing to do with cake. It is an idiom. As a second example, consider the Australian phrase *big bikkies*. This means 'a lot of money'. It has nothing to do with biscuits (*bikkies*). It can only be glossed globally. It is typical of idioms that they have a fixed form, and cannot be manipulated syntactically. *Kick the bucket* as an idiom means 'to die', but it would be very odd to say *The bucket was kicked by Kim* in this sense, or *What was kicked was the bucket*. *He jumped down my throat when I mentioned the loss* ('he got very angry with me') cannot be changed to *My throat was jumped down when I mentioned the loss*. Having said that, there are many examples which show that the lack of syntactic manipulation is not always absolute. Corresponding to *let the cat out of the bag* ('to divulge a secret unintentionally'), it might be possible to say *The cat has been let out of the bag*.

Some idioms are so obscure that, even if you can guess their meaning in context (such as *Here's mud in your eye* as a toast), you cannot explain how they came to mean what they mean. Others have meanings which can be worked out, with a bit of good will. *Talk the hind leg off a donkey* (to talk excessively) is probably an expression whose meaning can be deduced, given that it is extremely unlikely to be literal. Equally, *When he saw he was caught, the spy came clean* ('told the truth') is probably a meaning that can be deduced. The term 'idiom' is used for the uninterpretable cases and for the cases which depend on some **figurative** use of language, but in terms of understanding, they are not equivalent.

Rhyming slang provides a different set of idioms. Rhyming slang is often attributed to Cockney English, but it is found in Glasgow and in Australia and New Zealand, as well as in London. And, as we shall see in a moment, some rhyming slang has become fairly general English. In rhyming slang, typically, a short phrase replaces the intended word. The last part of the phrase rhymes with the intended word. So, traditionally, instead of *kids* we find *dustbin lids* (occasionally, *teapot lids*) and more recently, instead of *throat*, we get *nanny goat*. Since the meaning of, say, *dustbin lids* is not semantically predictable from either *dustbin* or

*lids*, we have to provide a global gloss, and that indicates an idiom. In a second level of rhyming slang, the rhyming word itself is omitted, so that *kids* are called *dustbins* and a *throat* is called a *nanny*. At this point, even the phonological motivation is removed, and a new **homonym** (a word which sounds the same as another word, but means something unrelated) is created. Because these homonyms cannot easily be reconstituted, it is often the case that we use rhyming slang without being aware of it. Some examples are provided in (2), all of them reasonably widespread.

(2)	<i>Usual expression</i>	<i>Full version</i>	<i>Meaning</i>
	a brass	a brass nail	'a piece of tail, a prostitute'
	blow a raspberry	blow a raspberry tart	'blow a fart'
	bread	bread and honey	'money'
	get down to brass tacks	get down to brass tacks	'get to the facts, get to the practicalities'
	have a butcher's	have a butcher's hook	'have a look'
	my old china	my old china plate	'my old mate'
	on your Tod	on your Tod Sloan	'on your own, alone'
	tell porkies	tell pork pies	'tell lies'
	use your loaf	use your loaf of bread	'use your head'

*Note:* Tod Sloan was a famous jockey at the turn of the twentieth century.

Once these have become established as new homonyms, they are no longer MWEs, but while they have the full version, they are MWEs.

Another place where we find relatively fixed, and not necessarily easily interpretable, usage is with proverbs. Proverbs often have to be interpreted figuratively, and are not always easily manipulable syntactically. *Too many cooks spoil the broth* may be abbreviated to just *too many cooks*, but it would be odd to make it passive (*The broth has been spoilt by too many cooks*) or to extract an element from the sentence (*The broth is what too many cooks have spoilt*). Other examples of proverbs are given in (3).

- (3) you can take a horse to water but you can't make it drink
- a bird in the hand is worth two in the bush
- make hay while the sun shines
- every cloud has a silver lining
- one man's meat is another man's poison
- you can't have your cake and eat it too

English has a large set of fixed similes, some of which we might expect (*white as snow*), others of which are unpredictable (*cool as a cucumber*, used

only when *cool* means 'unflustered' rather than when it means 'slightly cold'). Some of these are less used today than they were a few decades ago, but many are familiar for at least some speakers. Some are given in (4). Note that other comparisons are possible, and are used; some are even well-established (*as busy as a one-armed paper-hanger*). These are just relatively fixed parts of the English language, and, to the extent that speakers know them, are MWEs.

- (4) as clean as a whistle
- as different as chalk and cheese
- as fit as a fiddle
- as flat as a pancake
- as honest as the day is long
- as long as your arm
- as quick as flash
- as snug as a bug in a rug
- as tough as old boots
- as warm as toast

Finally, we find a whole host of **collocations**, words which habitually occur together, some of which we probably want to classify as MWEs and some of which we probably do not wish to classify in this manner. We can begin with some extreme examples.

If you hear the word *kith*, you probably immediately think of *and kin*. *Kith* occurs, for all practical purposes, only in the phrase *kith and kin*. *Kith* is a **unique morph**, it occurs only in one place in the language. Except when the pattern is disturbed by the name of a cleaning powder, the same is true of *vim*. The phrase *vim and vigour* is the only place *vim* occurs. If *kith* or *vim* occurs without its supporting phrase, it is nearly always a joke. Here one word predicts another with almost 100 per cent accuracy, though the prediction goes only in one direction: *kin* does not predict the presence of *kith*.

At the other end of the scale, consider *in the*. *In the* occurs very frequently, largely because *in* is a preposition, prepositions nearly always occur in prepositional phrases, prepositional phrases are nearly always made up of a preposition and a noun phrase, and noun phrases often begin with *the*. *In* predicts (to some extent) the presence of *the*, but *the* does not predict the preceding *in*. This particular collocation is of value to someone who is programming a phone to support a user to write texts (SMS messages), but is not of great value in determining the meaning of a sentence. It is a collocation, but is probably not an MWE, because it is not lexical but determined by the grammatical system of the language.

In between these two extremes, there is a host of intermediate cases,

sometimes with words that predict each other, sometimes where the prediction is in only one direction.

The class of proverbs, cited above, tails off into various classes of well-known expressions, some of which are quotations, some of which are fixed but not necessarily proverbial. Some examples are given in (5).

- (5) an eye for an eye
- curiouser and curiouser
- diamonds are a girl's best friend
- tomorrow is another day
- a little knowledge is a dangerous thing

There is also a set of binomials, two words linked by a coordinating conjunction (occasionally you find longer strings than this, but the most common are just two items long). Some of these have variable order, but mostly the order is fixed. Some examples are given in (6).

- (6) bed and breakfast
- black and blue
- born and bred
- cut and dried
- done and dusted
- down and out
- fruit and vegetables
- here and there
- life and limb
- loud and clear
- sink or swim

It is noticeable how many of the items in (6) either alliterate or rhyme, and this phonological aspect of the structure seems to be part of the appeal of these expressions. Note also that many **word-classes** are involved, and that the meaning of the whole is not necessarily predictable from the meanings of the parts.

And there are a whole set of different construction types which seem to be fixed expressions. Illustrated below are verbs with direct objects (7), prepositional phrases (8) and adjective-noun combinations (9), but virtually any syntactic structure can have fixed collocations within it.

- (7) bite the bullet
- change (one's) mind
- face the music
- pull (one's) leg
- shake a leg

- smell a rat  
upset the applecart
- (8) at home  
in cold blood  
in the open  
under the weather  
on the blink  
over the moon  
up the duff
- (9) black eye  
commanding officer  
fat cat  
fatty acids  
historical novel  
rhetorical question  
rich pickings

Another class, although one that potentially overlaps with others that have been discussed here, is the class of **formulae**, things which are just the way you say things. Some examples are given in (10).

- (10) all things considered  
as a matter of fact  
excuse me  
good afternoon  
I'm sorry to say  
in other words  
in the long run

A nice example is provided by the way in which we tell the time. To express the time 2:45, there are various possibilities, of which we probably use only one. Consider the versions in (11). Any of these is grammatical in English, but they are not what you say.

- (11) quarter to three (British English)  
quarter of three (American English)  
three o'clock less the quarter (French)  
quarter before three (German)  
three quarters three (German)  
three o'clock less fifteen (Japanese)

These are still fairly constrained, but the Danish for 3:20 translates as 'ten minutes in half four'.

All of this is important because of a general point it makes about linguistic theory. The approach to syntax advocated by Noam Chomsky and his followers assumes that each sentence is freely constructed from the elements that are available and the syntactic rules which allow those elements to be chained together. Examples like the ones that have been provided here suggest that our language is far more constrained than that: many things which appear to be options are not genuine options at all, but limited by the particular words involved.

#### 1.4 Paradigms

In a sentence like *I like watching parrots* we can discern two sets of relationships. The first is the relationship between *I* and *like*, *like* and *watching*, *watching* and *parrots*. That is, there is a relationship between the elements that are strung together. Following Saussure ([1916] 1969), this is termed a **syntagmatic** (/sɪntæg'mætɪk/) relationship. In *I like watching parrots*, the syntagmatic relationship that has been drawn attention to is one between words, but syntagmatic relationships also exist between the elements /w/, /d/ and /ʃ/ in *watch* and the *parrot* and -s elements in *parrots*. A syntagmatic relationship is one between elements which are all present in the stream of speech.

A different kind of relationship is one between elements of which only one is present in the stream of speech. Rather than *I like watching parrots*, we could have had *You like watching parrots*, *We like watching parrots* or *They like watching parrots*; equally we could have had *I hate watching parrots*, *I love watching parrots*, *I abhor watching parrots* and so on. In the first case we have a relationship between *I*, *we*, *you* and *they*, only one of which is present in the original sentence; in the latter case we have a relationship between *like*, *hate*, *love*, *abhor* (and any other words you can think of which would fit into the same slot). Following the Saussurean tradition, these are termed **paradigmatic** (/pærədɪg'mætɪk/) relationships. Paradigmatic relationships are relationships of potential substitutability, and we can also call *I*, *you*, *we*, *they* a **substitution class**. As with the syntagmatic relationships mentioned above, paradigmatic relationships also hold between sounds, for example in the words *bat*, *cat*, *fat*, *bat*, *mat*, *gnat*, *sat*, *vat*, and with parts of words, for example in *employer*, *employee*.

Both syntagmatic and paradigmatic relationships hold between like elements. There is a syntagmatic relationship between *watching* and *parrots* because both *watching* and *parrots* are words. We would not talk about a syntagmatic relationship between /ʃ/ and -ing because the first is an element of sound, the latter is a meaningful word-part; they are not elements from the same level of analysis. Similarly, when we say

there is a paradigmatic relationship between *I* and *we*, it is a relationship between elements which can function as the subjects of the verb *like*, not a relationship between the sound element /aɪ/ and the word *we*, because sound elements and words are not comparable entities.

A set of elements in a syntagmatic relationship is a **syntagm** (/ˈsɪntæm/), and a set of elements in a paradigmatic relationship is a **paradigm** (/ˈpærədaim/). The word *syntagm* is usually avoided in modern usage, in favour of the word *construction*. A construction is a syntagm which has internal structure and coherence (including in its meaning). Very often a construction has a meaning which goes beyond the meanings of the elements within it. For example, in *I like parrots*, there is the relationship between *I* and *like*, which we might gloss as a relationship between the experiencer of the emotion and the emotion, and also a relationship between *like* and *parrots*, which we might gloss as a relationship between the emotion and the object about which the emotion is felt. Similar meaning relationships are found in *We hate cats*, *You admire the Queen* and so on, and belong to the construction rather than to the words in the construction. The result is that although there is a syntagmatic relationship between *man* and *are* in *The sons and daughters of a man are his children* (the two words occur adjacent to each other in the stream of speech), we would not call *man are* a construction (although it may be part of a larger construction).

The term 'paradigm' is in general usage, but its usage is often limited. Although paradigm can justifiably be used of any substitution class it is most often used of substitution classes within the word. Thus the normal use for the term 'paradigm' is the kind of substitution class illustrated in (12).

- (12) walk
- walks
- walked
- walking

This paradigm illustrates two different kinds of word: there is a sense in which all the items in (12) are different words, which we will call **word-forms**, and a contrasting sense in which the paradigm in (12) illustrates different uses of the same word, which we will call a **lexeme**. We standardly write lexemes in small block capitals and word-forms in italics, to distinguish them where the distinction is important. This terminology (and its corresponding notation) allows us to capture the fact that although *walks* and *walked* are different words in the sense that they have different shapes and occur in different syntactic substitution classes (*walked* can occur in *We walked*, but *walks* cannot occur in

\*We walks), there is another sense in which all of these different forms are representations of the same word, so that you would look them up in the dictionary under the heading *walk*, not under *walks* or *walked*, for instance. *Walks* and *walked* (like *walk*) are different word-forms that belong to the lexeme WALK. In English some verbs have just one word-form in their paradigm (for example, *must*), some have three forms in their paradigm in this sense (for example, *bit*, *bits* and *bitting*), some have four (like WALK in (12)), some have five (for example, *throw*, *throws*, *threw*, *thrown*, *throwing*) and one has eight (*be*, *am*, *is*, *are*, *was*, *were*, *been*, *being*). On a world scale, that is fairly modest. Many languages have thousands of word-forms for each verb and Archi, a language of the Caucasus, is reported to have 1.5 million word-forms for every verbal lexeme.

Although the term 'paradigm' is often used for this one very narrow kind of paradigm, that does not prevent there being other kinds of paradigm. One type that is currently controversial is the **derivational paradigm**, where we see different lexemes linked by paradigms of forms within the word. Some examples are given in (13).

- (13) a. deceive  
deception  
deceptive
- b. employ  
employer  
employee
- c. proceed  
process  
processual
- d. theory  
theorist  
theorise

Although the examples in (13) show individual sets of words, these series are often generalised over several **word families**, in the sense that (13d), for example, is just one word family whose overall pattern is also illustrated by *fantasy*, *fantasist*, *fantasise*. And sometimes, the existence of the paradigm seems to provoke its use in new members of the construction, as illustrated in (14).

- (14) a. Like the world is divided into *stompers* and *stompees* and he's a *stompee*. (Stephen Dobyns (1998), *Saratoga Strongbox*, New York: Viking, p. 46; italics in the original)
- b. What it needed was to match *solicitor* with *solicitee*. (Michael Thomas (1983), *Hard Money*, New York: Viking, p. 191)

- c. The bellower was Harmon Crundall – and the bellowee the mysterious Mrs Smith. (Joan Hess (1986), *Murder at the Murder at the Mimosa Inn*, New York: St. Martin's Press, p. 41)
- d. The stabber might want to stay friends but not the stabbee. (Richard Layman (2001), *Night in the Lonesome October*, London: Headline, p. 29; italics in the original)
- e. I found myself more therapist with Julie than therapee. (Robert B. Parker (2000), *Perish Twice*, London: Murray, p. 26)

Examples like those in (14) appear to show the value of the notion of a derivational paradigm: the paradigm supports extensions to the paradigm in places where the relevant words are not (or have not been) established parts of the language. In other words, the paradigm supports the **productivity** of the pattern, its extension to new words. There are other examples which show why some linguists are reluctant to see the derivational paradigm as particularly helpful. Consider the example in (15).

(15)	deceive	receive	conceive	conceive
			(a child)	(an idea)
	deceit	receipt		concept
	deception	reception	conception	conception
	deceptive	receptive		conceptual

In (15) we find that the overall paradigm that we can establish with *deceive* and *receive* does not extend completely to *conceive*, that two meanings of *conceive* seem to be members of different paradigms, and that the meanings of the relevant forms are not always parallel. Not only is the paradigm apparently not reliable, it can be difficult to tell what the boundaries of such a paradigm might be: are *receivable* and *conceivable* part of the same paradigm or not (and if so, why is only one of them used as a noun)?

Another place where we use paradigms and substitution classes is in dividing the words of the language into classes for the purposes of grammatical description. These classes of words are called word-classes or, in a rather older terminology, **parts of speech**, and in principle they are paradigms of words.

Consider, for example, a sentence such as *The parrot is pretty*. In place of *parrot* we can substitute a whole series of words such as *book*, *cat*, *flower*, *girl*, *house*, *ladybird*, *monument*, *statue* and so on. In the place of *pretty* we can substitute *delightful*, *fierce*, *grey*, *incredible*, *large*, *ugly*, *unimaginative* and so on. Each of these substitution classes, or paradigms of words, is a class which can also fit into other slots. For example, *parrot*, *book*, *cat* and so

on could also fit into the sentences *I can see the parrot*, *Why do you want a parrot?*, *A parrot eats insects* and so on. If we generalise over all of these possible environments, we end up with a huge list of possible words, some of which are excluded from individual sentences by virtue of their meaning, but which are nonetheless members of the class because of the way they fit into other comparable slots. So even if it would not make sense to say *I accidentally killed the monument* to parallel *I accidentally killed the parrot*, we know from the examples above that *monument* and *parrot* do belong to the same substitution class in many instances. To make these super-substitution classes easy to refer to, we give them names, usually (in English) names inherited from the Latin grammarians. The class including *parrot* we call the class of nouns (this word comes indirectly from Latin *nomen* 'name', because nouns in many cases are the names of sets of things); the class including *pretty* we call adjectives (because they are so often added to nouns – from Latin *adjectio* 'I add'). The slots that these words fit into include not only syntactic slots such as the ones illustrated above, but slots in words (nouns can typically be made plural in English, can have endings such as *-ish*, *-y* and *-al* added to them; adjectives can often be compared – *prettier*, *prettiest* – or can have *-ise* or *-ify* added to them).

There is not a limited and finite set of word-classes. There are always word-classes within word-classes. So, for example, there is a limited number of nouns that can be said to *purr* and that can be substituted for *kitten* in *The kitten purred*. *Kitten* is part of the huge set of nouns, but it also a member of this much smaller word-class which does not have a generally accepted name because we do not need it very often for grammatical description: we could always call it the class of purry nouns. We also find that authorities argue about the legitimacy of some word-classes, and that different word-classes are used at different periods of history as ideas of what is important in defining the word-class varies. It is also the case that the word-classes are not necessarily the same in all languages: in some languages the class of adjectives is very small or subsumed in some larger class.

Nevertheless, there is a set of word-classes which can apply to English and which most authorities agree upon. The classes set out in Table 1.1 are not exhaustive, and the substitution slots illustrated are also examples rather than exhaustive.

Rather than view word-classes as a super-class of substitution classes, we could view them as **canonical** classes. In a canonical class, there are some members which fit all the available criteria for being a member, and some which fit fewer. On this view, *nice* is a more canonical adjective than *afraid*, because *nice* can be used before a noun (attributively)

Table 1.1 Word-classes as substitution classes

as in *a nice person* and **predicatively**, as in *That person is nice* while *afraid* cannot be used attributively: \**the afraid child*. In terms of nouns, *boy* is a better noun than *insect*, is a better noun than *tree*, is a better noun than *freedom* because human beings are more central to human experience than inanimate objects, and because *freedom* isn't even an object that can be touched or moved, and so on. The English language forces us to treat *freedom* just like any other noun, but it is further from the ideal noun than *boy* is.

Probably the least helpful way of defining word-classes is in terms of their meanings. It is traditional to define a noun as 'the name of a person or thing' and to define a verb as 'a doing word', but it is not clear that *freedom*, *insensitivity* or *belief* is the name of a thing, and *seem* is not obviously a 'doing word', even though it is a verb. Having said that, there are those who defend 'notional' definitions of word-classes, especially because they are not subject to language-dependent formal criteria, such as the form of the *-ing* suffix in the word *seeming*.

### 1.5 The dictionary

The term 'dictionary' is usually restricted to real-world dictionaries that appear in print and on-line. Dictionaries provide a list of words of whatever language they deal with – in our case, English – and then give a certain amount of information about each of them. Dictionaries tend to have two functions, which may, on occasions, conflict with each other: to describe the language as it is, and to provide an influence for establishing and maintaining the standard form of the language. Clearly, the second function is not to the fore with dialect dictionaries (except insofar as they usually list words which are excluded from the standard), nor in dictionaries of slang and the like, but is often considered to be a function by readers or users of the dictionary, even if it is not the primary motivation of the makers of the dictionary.

To illustrate this point, consider the reaction to the entry for *ain't* in *Webster's Third New International Dictionary* of 1961. The dictionary said of *ain't*:

Though disapproved by many and more common in less educated speech, used orally in most parts of the U[nited] S[tates] by many cultivated speakers, esp. in the phrase *ain't I*.

In saying this, the dictionary makers believed they were giving a statement of the way in which *ain't* was used in the English of the United States at the period. In other words, they had a **descriptive** goal. And most of the commentators who recognised the descriptive

intent considered that the writers had achieved that descriptive goal. Non-specialist reviewers, however, found this statement offensive. Reviewers wanted the dictionary to tell them that anyone who used *ain't* was, by that very fact, not cultivated ('Cultivated, our foot', scoffed one reviewer). They also objected to a number of other words that were included in the dictionary and not marked as being errors (or just vulgar). These include *boo-boo*, *drain* (in the sense 'to exhaust a person'), *finalise*, *greasy spoon*, *night clothes*, *no-show*, *orientate*. I suspect that none of these would cause a raised eyebrow today. (See Sledd and Ebbitt 1962 for examples and discussion.) People who want dictionaries to tell us how we should use language are said to have a **prescriptive** goal.

The distinction between a descriptive and a prescriptive (or **normative**) approach to language has far wider implications than this single, extreme, example might suggest. Crystal (1984) gives entertaining coverage of a number of linguistic issues that were exercising the British public at the period at which he was writing; most of them are probably still relevant to many people. They involve things such as whether it is 'correct' to stress *kilometre* on the first or on the second syllable, what the difference is between *stationary* and *stationery*, whether it is acceptable to use *they* to refer to a single person, when you should use *shall* and when *will*. These are partly matters of vocabulary, but they also involve pronunciation, spelling and grammar. The descriptive versus prescriptive divide covers all of these. The real difference between the two is whether we think there is a 'right' answer to questions of linguistic usage. If we assume that one of *Klometre* and *kiLOmetre* is right (and the other, therefore, wrong), we assume that there is a unique solution to this question of English usage. We assume that it is a question like *Should you drive on the right or on the left in England?* where there is a single solution laid down by law. We assume it is not like *Should I wear jeans to the party tonight?* where the answer may depend on your age, the dress code for the party, how good you think you look in jeans, whether you prefer to be underdressed or overdressed for any occasion, what else you have to wear, the kind of party that is involved, where the party is being held and so on. Language questions are more often like the jeans question, and less often like the driving-side question. Whatever answer the person you have asked gives you, there will be somebody else, apparently equally authoritative, who will give you a different answer. The moment a dictionary provides only one answer, it becomes de facto an arbiter of usage for anyone who consults it, and thus functions to help define (what its writers consider to be) the standard version of the language.

Linguists often get a bad reputation among the general public for

saying that there is more than one answer to such questions. Linguists are much happier saying that one thing is preferred by the young, another by older speakers, or that one thing is well established but another is starting to be heard, that women tend to prefer one thing and men another, or that North Americans tend to say one thing, Britons another. Often, linguists and lay people might be able to agree that something is not standard, is dialectal usage, is used only in technical language, is poetic and so on. But the underlying distinction between saying what happens and saying what should happen (according to some ill-defined norm) remains. Rather than continue to illustrate this dispute, here we need only acknowledge that dictionaries can have these two incompatible goals, and that sometimes one is more to the fore, sometimes another.

Over time, a set of expectations for a dictionary has grown up, and most dictionaries provide comparable information, within the limits of their format. Here, we will consider some of those pieces of information, and where they might support the notional standard.

First of all, dictionaries provide evidence of the existence of a word. The fact that a word is listed in a dictionary at all is taken to prove that there is such a word. This can be misleading in two ways. From time to time, dictionaries list words erroneously, and the word has no existence outside the dictionary. The *OED* gives a list of such **spurious words** including *banket* ('a term in bricklaying'), *David's staff* ('a navigational instrument') and *sardel*, variously supposed to be a fish (the sardine) or a precious stone. Dictionaries far more often fail to list perfectly good words. This is inevitable. No dictionary can list every word of English, and shorter dictionaries list fewer than larger dictionaries. In the very nature of things, there will be some words whose existence is acknowledged in large dictionaries that do not make it into small dictionaries. One of the problems with on-line dictionaries is that the user has no idea of the extent of the word-list in the dictionary, and so cannot tell whether it is worth looking for the desired word in a different, larger, dictionary, if it is not given in that particular source.

Dictionaries list the spelling of a word. Most dictionaries list a single spelling, occasionally two (for example, for *judgement/judgment*). The *OED* lists many spellings, representing spellings at various stages of the development of English, but uses the expected modern spelling for the **headword** (the heading under which the word is discussed). The spelling is important for print dictionaries, because the headwords are ordered according to the spelling. Having said that, 'alphabetical order' is not necessarily obvious. It can be enlightening to put the sets of words like the following in what you consider to be alphabetical order, and

then check the order in which they occur in a number of dictionaries, *batch, batchery, batchback, batchbet, batching, batbox, batter, bat trick*.

Many dictionaries provide a pronunciation. Some of them have abbreviatory systems which allow them to give alternative pronunciations in a single transcription (for example, pronunciations with or without the /r/ sound at the end of *prefer*). Dictionaries of pronunciation (for example, Upton et al. 2001; Wells 2008) typically give rather more options, though *Webster's Third New International Dictionary* is said to give twenty-five pronunciations for *berserk* (which is more than either Upton et al. 2001 or Wells 2008). While more and more dictionaries use a transcription system based on the International Phonetic Alphabet, many use their own respelling system. In either case, you will need to be able to interpret the system to use this information in the dictionary.

Dictionaries also provide a meaning, or several meanings, for each word. We can term each of the meanings under a single heading a **polyseme** (see further Section 4.7.1). The number of polysemes provided for a given word is largely a function of the size of the dictionary. Some of the polysemes provided will be largely predictable: for example, the use of *head* to denote part of the human body or part of an animal's body may not need to be given as a separate polyseme. Others are unpredictable, such as the *head* in a tape-recorder being the magnetic part that converts analogue sound to a record on the tape or vice versa. The style of the definition is one of the key decisions facing lexicographers. Should the definition always reflect the word-class of the headword (nouns being defined by other nouns, adjectives by other adjectives and so on), should the definition be a single word or phrase, or is a whole sentence required, is the definition of a plant species by its Latin name sufficient, should all the definitions use a restricted vocabulary, or can they use any words (even ones not in the dictionary) and so on? Very often the style of the definition in a dictionary will be a reason for the reader to like or dislike the particular dictionary, depending on the taste of the reader or the purpose for which the reader requires the dictionary (solvers of cryptic crosswords and learners of English as a foreign language are likely to have different demands, for instance).

Dictionaries usually provide a certain amount of grammatical information. This may be fairly minimal: that *cross* can be a noun or a verb, for example, or that *walk* can be a transitive or an intransitive verb. Other dictionaries, especially dictionaries for foreign learners, give much more grammatical information. The *Longman Dictionary of Contemporary English* (2009), for instance, gives typical subjects and/or direct objects of the verb *cross* with different meanings, and tells you which prepositions to use with the different meanings.

Most dictionaries provide some examples of usage. Sometimes these are just MWEs of various kinds (see Section 1.3), sometimes these are sentences to illustrate grammatical patterns (either from real usage or made up by the lexicographer), sometimes they are examples to illustrate the overall way (including style) in which the word is used. The use of citations from literature to illustrate good usage was made popular by Dr Johnson in the eighteenth century; these days, illustrations from less literary sources are rather more common.

Many dictionaries provide a brief section on the *etymology* of the word. The etymology is the history of the word, where it was borrowed from, when it was first used, what words from other languages derive historically from the same source and so on. Although the word *etymology* comes from a Greek root meaning 'true', and at one time some scholars thought that in providing an etymology they were providing the 'true meaning' of the word, the etymology should never be considered to represent the meaning of the word. For example, the word *innocent* derives etymologically from Latin elements meaning 'not harmful', but *innocent* today does not mean 'harmless' but 'having little experience, not guilty'.

## 1.6 The lexicon

The **lexicon** is what linguists call the dictionary that is assumed to be in people's heads, or the linguist's best approximation to that. That is, it is fundamentally a psychological entity, and correspondingly, its contents cannot be observed directly, but must be deduced from speakers' and hearers' behaviour. There are two fundamental approaches to the lexicon. The first is that, in the slogan, the lexicon is the home of the lawless (Di Sciullo and Williams 1987): that is, the lexicon contains whatever cannot be predicted by general rule. The second, wider, approach sees the lexicon as not only containing the lawless, but also containing anything to do with the structure of words, whether it is lawless or not.

Among the most fundamentally lawless things is that *cat* means 'cat' (and so on, for thousands of fundamental elements). There is nothing in the form *cat* that indicates that it must mean 'cat', so that is an unpredictable fact. It is so unpredictable that we are surprised when it turns out that the word for 'dog' in the Australian aboriginal language Mbabaram is actually *dog* (and that the word is not borrowed from English). We do not expect the same meaning to be carried by the same form in widely different languages. While not all dictionaries list prefixes and suffixes separately from the words in which they occur (though some do), it seems likely that the lexicon must list not only the fundamental words,

like *cat*, that do not contain any prefixes or suffixes, but also the prefixes and suffixes, or at least those which can be used in the creation of new words (that is, those **affixes** which are productive). The usual statement on this is that the lexicon must contain a list of the **morphemes** of the relevant language (where a morpheme is the smallest meaningful unit in the language – including affixes). At the present time, the notion of morpheme is theoretically disputed, which makes it difficult to be sure just what must be included here. Nevertheless, to the extent that we recognise recurrent meaningful chunks within words, we assume them to be listed in the lexicon.

Sometimes, though, words must be in the lexicon even if they contain meaningful elements within them. There are various pieces of evidence that point to this. Derwing (1973: 124) cites the evidence of his three-year-old daughter being delighted when she discovered that *orange juice* had something to do with *oranges*: she had managed to use both words without linking them. A psycholinguistic experiment by Wheeler and Schumsky (1980) found that some speakers did not relate *citizenship* to *citizen* or *baker* to *bake*. Most speakers do not see *foul* as related to *filth* in the same way that *long* is related to *length* (though, in this case, the *-th* suffix is not productive). And in general terms, where a derivative is more frequent than its base it is not clear that speakers break it up into smaller bits: so we probably do not automatically relate *government* to *govern*, even if we relate *confinement* to *confine*, because *government* is heard much more often than *govern* (Hay 2003). This means that for some words, we have two routes to understanding them: we can 'look them up' in our lexicons, or we can work out what they mean on the basis of the elements that make them up. Models of the lexicon which permit such a procedure are called **dual-route models**.

We must assume, then, that we have in our lexicons a list of all the fundamental, unanalysable, words, other meaningful elements (morphemes, if we allow for the presence of such a unit), and also a number of words which, etymologically, are made up of several meaningful elements. However meaningful the elements of *blackmail* were to the person who first used the word, they are not helpful to people today in understanding the word.

This raises the question of how these words are listed in the lexicon. Alphabetical order cannot be a relevant factor – there are far too many people, even today, who speak English but cannot read or write English for this to be the answer. That does not mean that literate speakers of English are not also influenced by spellings, just that spelling is not the fundamental organisational principle of the lexicon. Rather, it seems that there are multiple links between items in the lexicon, some of them

dependent upon the form of the word (the sounds it is made up of) and some of them dependent upon meaning.

It used to be thought that readers translated what they saw on the page into a phonological representation in order to link it to a meaning. Now it is far less clear that things are as simple as this. For instance, when we look at pathological cases, we find that some patients are able to understand a written text but not name the words they are reading, or understand a spoken word without being able to link the meaning to a spelling. This suggests that although there is a link between pronunciation and spelling, those two are linked independently to representations of meaning in the brain (Allport and Funnell 1981). We must conclude that spelling is important for literate people as well as pronunciation, and to some extent independent of it.

When we recognise spoken words, we obviously use the pronunciation, and we use the elements within the word in the order in which we hear them. If we hear /gʌ/ we seem to get our brains fired up ready to hear *gubbins* or *guppy* or *gust* or any other relevant word. If we next hear /t/ we do not know whether the word will be *gut* or *gutta-percha* or *gutter* or *guttural* and all these options remain open. If we next hear /ə/ we can rule out *gut* (assuming the /ə/ is part of the same word), and if that is the end of the word we have *gutter*. When we get to the point where no other word will fit what we have heard, we reach the **recognition point**, and we do not need to bother with any further information, because it is redundant. This means that it is easier to recognise *redundant* from /ri'dʌ/ than from /dənt/ (Marslen-Wilson and Zwitserlood 1989). This is how we can tell so quickly that *cown* is not a word of (our) English. If we know some 60,000 lexemes (which is the kind of figure that is often provided; see, for example, Aitchison 2003: 8), some of which have several word-forms associated with them, and we could check them off at, say, 60 per second, it would still take us over 1,000 seconds to make sure we have no word in *cown* in our vocabularies. But if we have been checking each sound-element as we hear it, and cutting down the number of options at each step, we already know by the end of the word that we have nothing matching that string in our lexicons.

When we produce words, as opposed to recognising them, we cannot start with the sounds, we must start with the meanings (unless somebody tells you to list as many words as you can starting with /b/). It seems we work in much the same way: we activate a whole lot of words in the relevant part of our vocabulary, and then make a choice between the activated words. For example, if we want to say something about a cow, we may activate a whole lot of words for farm animals (*donkey*, *sheep*, *horse*, *goat* and so on) and from that list choose the word *cow*. At

some stage, and it is not entirely clear what that stage is, we assign a pronunciation to the word we have chosen. Sometimes, though, we make a mistake, and we might say *We have to milk the bulls*, because both *cow* and *bull* have been activated in our brains, and we have grabbed the phonological representation for the wrong one. If you talk to someone about electricity, and then make them read something that includes the phrase *sham dock*, they might very well say *damn shock*, because of the prior activation of words associated with electricity (Aitchison 2003: 218); and the phenomenon of the Freudian slip depends upon some subject area being readily activated in the brain, and brought out inadvertently, whether because of a word related in meaning or a word related in pronunciation.

There are multiple ways in which words can be linked semantically. As was just suggested, words for farm animals may be linked by our experience of the entities they denote being linked. Words can be linked by meaning (at least in some contexts) more or less the same thing: *remedy*, *rectify* and *restore*, for instance. They may be linked by being opposites: *happy* and *sad*, *black* and *white*. They can be linked because they often occur together: *keep* and *calm*, *round* and *table*. Words can be linked by being part of the same classification: *tulip* and *rose* (both flowers), *Tyne* and *Severn* (both rivers), *cod* and *whiting* (both fish) and so on. Words can be linked by having the same form but different meanings: *cricket* (the game) and *cricket* (the insect), *guy* (a rope) and *guy* (a person), *potter* (a maker of pots) and *potter* (to undertake small tasks with no haste). We shall deal with such relationships as these in a later section (see Section 4.7.1). For the present, note that these various types of relationship are not necessarily mutually exclusive: *ewe* and *ram* are both farm animals, both types of sheep and opposites in terms of adult sex.

As well as providing some information on meaning, the lexicon must provide some grammatical information. Just how much such information must be there and how it is stored are controversial topics, but some of the information is not predictable from general principles, and is probably therefore listed in the lexicon because it is lawless. Word-class is one such piece of information. There are those who argue that word-class is determined by the syntactic tree into which the word is placed, but somehow the speaker needs to know that *dog* and *cat* will both fit into trees as nouns, but while *dog* will fit into a tree as a verb (*He seems to dog my footsteps everywhere I go*), *cat* – for most speakers of modern English – will not. Equally, while *up* can be a verb (*The doctor upped the dose*) or possibly an adjective (*the up train*), *through* cannot be a verb, seem that it is lexical information. More generally, the lexicon must

Table 1.2 Grammatical patterns with some verbs

	<i>Sbe_d</i> <i>that I should leave</i>	<i>Sbe_d</i> <i>me to leave</i>	<i>Sbe_d</i> <i>my departure</i>	<i>Sbe_d</i> <i>my leaving</i>	<i>Sbe_d</i> <i>leaving</i>	<i>Sbe_d</i> <i>me that I should leave</i>
advise	✓	✓	✓	?	✓	✓
counsel	X	✓	X	X	X	?
exhort	X	✓	X	X	X	X
propose	✓	X	✓	?✓	✗	X
recommend	✓	X	✓	✓	✓	X
suggest	✓	X	?	?✓	✓	X
tell	X	✓	X	X	X	✓
urge	✓	✓	✓	X	✓	X

provide information on the constructions the word can occur in. This will include such usage as transitive and intransitive in verbs (if you can *shoulder a burden*, why can't you *back a coat* or *finger a ring* – especially as you can *back a horse* and *finger a perpetrator of a crime*), but also much more complex constructions. Consider the set of words and the patterns in which they can occur freely shown in Table 1.2, where the verbs all have to do with communicating something. To the extent that such information is not predictable, it must be in the lexicon.

Some words mean different things when they appear in different contexts. The obvious example is *dry*, which can be used to modify *politician*, *throat*, *towel*, *weather*, *wine* or *wit*, with a different meaning in each case. Even if some of these meanings arise from perfectly normal extensions of an original meaning, they have become established and, to a large extent, unpredictable. In other cases, even the fact that particular words tend to go together is unpredictable. It does not seem to be predictable that we say *warm welcome* but *friendly wave*, that we have a *vernal equinox* or a *spring equinox* but only *spring flowers*, that you *hide your eyes* but *conceal your thoughts*. When such collocations are unpredictable and linked to the individual words involved, they must be in the lexicon.

The various kinds of MWE discussed in Section 1.3 must also be part of the lexicon. Since such things are either known or not known, when they are known, they must be in the lexicon.

The word *turquoise* is probably used more by women than by men (though *cyan* may be used more by men); the word *doggy* is probably used more by children or to young children than among adults; the term of address *bro* used to be used (and possibly still is) more among some ethnic groups than others; the word *discombobulate* is not entirely serious; the word *ardour* is formal or literary; the word *micturition* is used

scientifically or medically. Some of these things are, to a certain extent at least, predictable: words with the *-y* suffix tend to be used more by and for children, but not all of them – *Chevy* (from *Chevrolet*) is not a word children use; *discombobulate* shows its made-up nature by containing unrecognisable elements. Nevertheless, just when a particular word is used is not something that can be determined from basic principles. It can even change over time. The expression *more than somewhat* started life as a joke, but has become serious. Although it is not entirely clear just what information people gather about such words, or what categories they work with, some information on usage must be listed in the lexicon.

All of the material that has been discussed above is the 'lawless' material: things that cannot be predicted by general principles or deduced from other aspects of the words involved. But in the more inclusive view of the lexicon, everything that has to do with the construction of words is there, whether it is predictable or not. According to this view, the terms 'lexical' ('to do with the structure of the lexicon') and 'lexical' ('to do with the structure of words') can be seen as synonymous – or at least, a lot closer to the synonymous than would be the case under the alternative view.

Part of the justification for this view is that the way in which words are made up is not equivalent to the way in which sentences are made up. In a sentence like (16), for instance, the meaning of the sentence can be deduced from the meanings of the words in the sentence and the syntactic structure of the sentence.

- (16) The dog barked loudly at the postman.

If we know what *dog*, *bark*, *loudly* and *postman* mean, and we know that *the dog* is the subject of the verb *bark*, we can work out the meaning of the sentence from basic principles. In such cases, the semantics of the sentence is said to be compositional. But if we look at words like *marriage* and *carriage*, we cannot tell from the elements that make up the words and their structure that *marriage* means a legal relationship while *carriage* means a vehicle (in both cases there are other meanings, but these are the most likely ones). Neither can we predict from general principles of English that the noun from *marry* will be *marriage* rather than *marriment* (compare *embodiment*), *marriance* (compare *dalliance*, *defiance*), *marrial* (compare *burial*) or *marry* (compare *worry*). Furthermore, we cannot predict that if we have *marriage* and *carriage* we will not have *tarriage* from *tarry* or *worriage* from *worry*. Unpredictable meanings and very different from what goes on in word-formation.

Words like *marriage* and *carriage* exemplify **derivational morphology**, the creation of new lexemes. **Inflectional morphology**, the creation of new word-forms of known lexemes, is rather different. Given a verb such as *sing*, it is entirely predictable what *sings* means and where it will occur. Moreover, every non-modal verb has a form with a final -s, and it always means precisely the same thing. It would be possible to treat this final -s form as being a part of syntax without any problem. The difficulty here lies in how far we can push this. *Sing*, like all other non-modal verbs, also has an -ing form, *singing*, but *sing* does not have an -ed form, *singed*: rather we find *sang* fulfilling the same role. Still, there is a form in the appropriate slot in the paradigm, even if the form itself is not entirely predictable: why is it *sing* and *sang* but *swing* and *swung*? What about adding -ly to adjectives to give things like *commonly*? Here there are some adjectives which do not follow the expected pattern: the adverb from *fast* is *fast*, for instance, and the adverb from *hard* is not *hardly*, which means something else. With plurals there are some nouns which do not have a plural (*knowledge*, for example), some which do not have a singular (*mumps*, for example) and some which have an unpredictable plural (*genera* as the plural of *genus*, for example).

All theories of the lexicon, whether they wish to include word-building in the lexicon or not, whether they want to distinguish between inflection and derivation or not, and whatever the structure of the theory, have to deal with questions like these. How the problem is dealt with may change, but the general nature of the problem does not.

## 1.7 Summary

In this chapter, we have seen that the notion of word is difficult to define, and that there are elements that are longer than words, which nevertheless share many of the features of words. We have also looked at traditional dictionaries, and talked about the kind of content that is likely to be found there, and then contrasted that with the lexicon, the dictionary that belongs in people's heads, showing how many of the things in the dictionary are also in the lexicon, but that the two do not always match each other.

### Exercises

1. Choose any category of MWEs and find another five examples which fit in the category.
2. Find any printed dictionary (preferably one with which you are familiar) and work out what the structure of the dictionary entry

- is. What information does it give, what does it not give? How much grammatical information does it give? How does it mark different parts of the entry? How does it deal with MWEs – does it deal with all of them in the same way, or does it distinguish between types?
3. Find five words which you would consider literary or poetic, and five which you would consider colloquial or familiar. If you have access to any large corpora, check the environments in which your words occur. Is your feeling supported by the evidence?
  4. Can you think of any prescriptive rules that you have been taught at home or in school? Why do you think you were taught these rules? Did the overt teaching change your behaviour, and if so, how?
  5. Can you find, or do you know, any words which have more than one correct spelling? How many can you list (or better, how many can a group of you list)? Are any of these spellings indicative of geographical location? What does all this tell you about the degree to which spelling is fixed by prescription in English?
  6. Re-read the section on dictionaries above (Section 1.5) and then, without consulting a dictionary, write dictionary entries for the words *chair*, *drive* and *late*. Explain what decisions you have taken in drafting your entries, and why you took them. Finally, compare your entries with ones from published dictionaries. Are the differences due to ignorance on your part, or to something else? If you were to write a dictionary, what would you learn from this experience?
  7. Do you agree with all the ticks and crosses in Table 1.1? Can you improve upon any of the environments proposed in the table?
  8. Do you agree with the data presented in Table 1.2? Are there other relevant environments? Are there generalisations which make some of the information predictable?

### Recommendations for reading

On the notion of word from a cross-linguistic perspective, see Bauer (2000), Dixon and Aikhenvald (2002) and Hippisley (2015). On MWEs, see Moon (2015). On dictionaries in general, see Mugglestone (2011). On formulaic language in general, see Wray (2012).

# 2 Making new words

## 2.1 Introduction

We can divide words into two major categories. On the one hand, we have the words whose meaning we cannot predict at all, we just have to know what they mean. The example of *cat* was given above. There is nothing in *cat* to tell us that it means 'feline quadruped' rather than, say, 'coniferous tree'. We have to know what *cat* means. On the other hand, a word like *friendliness* provides us with some clues as to why it should mean what it means. We still have to know that *friend* means 'friend', but the *-li* and the *-ness*, although they have fairly abstract meanings, tell us that we are dealing with an abstract noun derived from an adjective. The meaning of a word like *cat* is **arbitrary**; the meaning of a word like *friendliness* is partly **motivated**.

In this chapter we are concerned with partly motivated words, and how they are, can be, or have been created in English. In order to create a word, it is typically the case that some procedure is carried out on some other word. In *friendliness*, that procedure is the addition of suffixes. We will see other procedures later in the chapter. We call the original word on which the procedure is carried out the **base** of the new word. Some words like *friendly* have one base, while others like *clocktower* have more than one. *Friend* is the base for the creation of *friendly*; *friendly* is the base for the creation of *friendliness*. Any process which can be used to create new words is said to be productive (as we have already seen in Chapter 1).

## 2.2 Blends

Consider the English word *hangry*. It is used to describe a fractious child (or indeed, adult) who is simultaneously hungry and angry. It is formed by taking the words *hungry* and *angry* and telescoping them together, omitting some of the phonological or orthographical structure, to give *hangry*. We could write this as in (1).