2

Words, Phrases and Meanings: Basic Concepts

In chapter 1, I introduced some ideas about the ways in which language is used in different text-types, and gave some initial examples of the importance of phraseology in studying meaning. In this chapter, I will provide a more detailed discussion of the main concepts which are needed for studying phraseology. This involves discussing several concepts which are central to lexical semantics, and which are discussed in many student introductions, including: denotation and connotation; synonymy, antonymy and hyponymy; and lexical fields. However, I will try to show that corpus data can provide a new way of looking at these concepts. In particular, an approach from corpus semantics shows that we have to discuss the relation between words in the lexicon (words in the language system) and words in texts (words in use).

2.1 Terminology

First, we need some essential terms.

Phrase. The unit of meaning in connected language in use is usually not a single word in isolation, but a longer unit of at least a few words in length. Much of this book discusses the nature of these extended lexical units, but when I need a neutral term for a string of words, I will talk of a 'phrase'. For example, the phrases *provide help* and *provide shelter* illustrate frequent uses of the verb PROVIDE.

Collocation. This is a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text. For example, PROVIDE frequently occurs with words which refer to valuable things which people need, such as *help* and *assistance*, *money*, *food* and *shelter*, and *information*. These are some of the frequent collocates of the verb.

Attested language. Almost all the examples which I cite are from real language in use, which was produced for some real communicative purpose. That is, I did not invent the examples just to illustrate a point of argument. I will refer to such data as attested data.

Corpus. Almost all of these examples are taken from corpora (singular *corpus*, plural *corpora*). A corpus is a collection of texts. There are many text collections (such as newspapers published on CD-ROM), which can be useful for some purposes. However, the term 'corpus' is usually used for a text collection which has been designed for linguistic research, in order to represent some aspect of language. It could be a collection from a given text-type (such as casual conversation, scientific research articles or science fiction novels), or it could be designed to sample as wide a range of text-types as possible, including written and spoken, formal and informal, fiction and non-fiction, language produced by or for children and adults, and texts from different historical periods.

2.2 Words: Word-forms and Lemmas

The word 'word' is ambiguous. First, we have to distinguish between 'lemmas' and 'word-forms' as follows. (An alternative term for 'lemma' is 'lexeme'.) I will use upper-case for lemmas and lower-case italics for word-forms. For example, verbs occur in different inflectional forms: the lemma TAKE is realized in text by the word-forms *take*, *takes*, *took*, *taking* and *taken*. Similarly, the lemma of the noun RABBIT is realized by the word-forms *rabbit*, *rabbits*, *rabbit's* and *rabbits'*; and the lemma of the adjective BIG is realized by *big*, *bigger* and *biggest*. Dictionaries of English conventionally use the base form of a verb to represent the lemma (for example, *want* represents WANT), and the singular of a noun (*table* represents TABLE).

Generally, dictionaries group only words from one part of speech under a single lemma, but they are not always consistent on how the grouping is done. For example, CONFUSE would typically include *confuse* and *confused*, but *confusing* might be included under this lemma as part of the verb, or listed separately as an adjective, and the noun *confusion* would typically be listed separately. Lemmatization looks simple, but in fact involves many decisions.

We need to distinguish between word-forms and lemmas, because we need to distinguish between units of texts and units of the vocabulary of a language. Usually the length of an individual text or the size of a corpus is given in statements such as

[1] This novel is 50,000 words long.
[2] This corpus consists of 50 million words.

These statements refer to a sequence of word-forms one after the other. Word-forms are the only lexical units which are directly observable. They are the units which occur in actual texts, and, in a written text, they are strings of letters separated by spaces or punctuation marks. In fact, they provide us with a definition of a text, which consists of a linear string of word-forms. In a written text, they occur one after the other in space; in a spoken text, one after another in time.

In a text or corpus, as in statements [1] and [2], it is likely that the word-form *the* occurs frequently: if it occurs 3,500 times, then I count it 3,500 times. If the forms *want*, *wants*, *wanting* and *wanted* all occur, then I count each occurrence separately. We can count words in a text by counting word-forms, but this is quite different from counting words in the vocabulary of a language. A statement such as

[3] This learner of English has learned 2,000 words

does not refer to the length of a text which someone has produced, but to the size of the vocabulary which they can draw on to produce texts. It means 2,000 different lemmas. The vocabulary of a language can be recorded in a dictionary, either of general English or of a sub-variety (e.g. a dictionary of technical and scientific terms). So we also have statements such as

[4] This dictionary contains 50,000 words.

In a dictionary, I would expect *the* to occur just once as a head-word. Traditionally, dictionaries list the head-word WANT just once, with a note that it occurs in different forms (*want*, *wants*, etc.).

In summary: Word-forms are directly observable units; a text consists of a sequence of word-forms. The sequence is crucial: if we change the sequence, we have changed the text. Lemmas are not directly observable, but abstract classes of word-forms; a vocabulary is usually represented as a list of lemmas. It may be convenient to present the list alphabetically, but this order has nothing to do with the organization of the vocabulary; in fact, it hides many kinds of semantic relations between words. Lemmas could, for example, be grouped according to different semantic areas (this type of book is called a thesaurus).

It is useful to think initially in terms of the correspondences

text	word-forms
vocabulary	lemmas

However, this is a provisional correspondence only. First, we will later need a further distinction between word-tokens and word-types (see chapter 6.6.1). Second, lemmas are not the only lexical units in the vocabulary. The assumption that single lemmas are the main unit of meaning has underlain the construction of English-language dictionaries for hundreds of years. However, corpus work provides a lot of evidence that units of meaning are both smaller and larger than the lemma.

2.2.1 *Example: the lemmas* CONSUME *and* SEEK

The following example shows the importance of the distinction between word-form and lemma. The word-form *consuming* occurs in the phrases *consuming passion* and *time-consuming*. Several other words, such as *costly*, *difficult* and *expensive*, co-occur with this second phrase, in longer phrases, such as

- very expensive and time-consuming; often difficult and time-consuming

The word-forms *consume* and *consumed* do not occur in such phrases at all. However, all three forms share the collocates

- more, quantities, calories, energy, oil

That is, all three forms are used in a literal sense of "consume an amount of fuel", but *consuming* occurs in additional, quite specific phrases. These differences would be missed if the lemma CONSUME was analysed as a whole.

Here is a more complex example. In chapter 1.7.4, I showed that in lonely hearts ads the word-form *seeks* is frequent, as in

- female 31, single, *seeks* well-educated gentleman

In this text-type, it frequently co-occurs with words such as

- attractive, black, caring, female, guy, lady, male, man, professional, similar

However, the word-forms *seek*, *seeking* and *sought* all co-occur with a very different set of words, including

- advice, asylum, help, support

If we looked only at the lemma SEEK, we would miss this striking difference.

In a corpus of 200 million words, I studied the 20 most frequent collocates of the different forms of SEEK: that is, the word-forms which co-occured most frequently with the different forms of the lemma. (The data-base which I used was Cobuild 1995b: this is described in chapter 3.6.) The collocates shared by the word-forms were as follows.

- *seek*, *seeking* and *sought* have 6 shared collocates: <asylum, court, government, help, political, support>
- *seek* and *seeking* have 10 shared collocates: <advice, also, asylum, court, government, help, new, people, political, support>
- *seek* and *sought* have 9 shared collocates: <advice, also, asylum, court, government, help, political, refuge, support>
- *seeking* and *sought* have 7 shared collocates: <also, asylum, court, government, help, political, support>
- *seeks* and *seek* have only one shared collocate: <professional>
- *seeks* and *sought* have no shared collocates
- *seeks* and *seeking* have no shared collocates

The overlap in their collocates gives us one measure of the semantic distance between the word-forms. We have three word-forms which form a tight cluster, with several overlapping collocates, largely from political and legal contexts, in the semantic field of "help and support", but the word-form *seeks* is only distantly attached to this cluster. (Tuldava, 1998: 142, proposes a simple way of calculating the overlap between two sets of items.)

These findings are not a statement about the whole language, but about the text-types sampled in the corpus which I studied. Obviously, if the corpus had contained no magazines with lonely hearts ads, then I would have found no such examples of *seeks*. Equally obviously, the corpus must have contained enough examples to make collocations such as *seeks–caring* more frequent than other collocations. I therefore checked a separate independent 100-million-word corpus for uses of *seeks* (the British National Corpus). This corpus contained examples from other personal adverts and from newspaper headlines, but it also contained other uses:

- guitarist seeks working band
- Microsoft seeks partners
- where a buyer seeks to reject goods supplied under a sale contract
- in his Symphonic Etudes, he consciously seeks an orchestral sonority

Adverts and headlines share the need to use short words. Other uses tend to be from formal, frequently legal, texts.

This example illustrates important principles. First, an exclusive concentration on only the most frequent collocations may hide variation in the language. Second, collocations may differ quite sharply in different text-types. Many text-types are specialized in their uses of language, and no corpus can fairly represent every one of them.

2.3 Collocation

The CONSUME and SEEK examples introduce the concept of collocation: the co-occurrence of words. We can talk of a node-word co-occurring with collocates in a span of words to left and right:

collocates . . . node . . . collocates
——— span ———

A 'node' is the word-form or lemma being investigated. A 'collocate' is a word-form or lemma which co-occurs with a node in a corpus. Usually it is frequent co-occurrences which are of interest, and corpus linguistics is based on the assumption that events which are frequent are significant. My definition is therefore a statistical one: 'collocation' is frequent co-occurrence.

What is node and what is collocate depends on the focus of study, and relations are rarely symmetrical. In a phrase such as *bonsai tree*, there is a much stronger prediction from left to right than from right to left, and such asymmetry is much more general. For example, the word *cushy* is quite rare. When it occurs, there is a high probability (about one chance in seven) that it will occur in the phrase *cushy job*. Other recurring collocates of *cushy* include *up-bringing*, and general nouns such as *number* and *situation*. However, the word *job* is much more frequent, co-occurs with a wide range of other words, and has only a low probability (about one in 5,000) of co-occurring with *cushy*.

One further term allows us to state collocations succinctly. A 'span' is the number of word-forms, before and/or after the node (e.g. 4:4, 0:3), within which collocates are studied. Position in the span can be given as N – 1 (one word to the left of the node), N + 3 (three words to the right), and so on. There is some consensus, but no total agreement, that significant collocates are usually found within a span of 4:4 (Jones and Sinclair 1974). There is a problem here, to which there is currently no solution. Lexical units may consist of collocations, and be larger than individual word-forms. Yet I am using word-forms – whose orthographic representation is often arbitrary: e.g. *already* but *all right* – to measure span (Mason 1999).

We now have a convenient notation for presenting information on collocations. A statement such as

- node <. . . list of collocates . . . >

says that the collocates listed are those that typically co-occur within a given span of the node, usually 3:3 or 4:4. Here is a real example:

- seeking 11,735 <asylum, help, advice, support, information>9%

This says: in a corpus, the word-form *seeking* occurred 11,735 times; in 9 per cent of cases it occurred with one of these five collocates. (The data here are from Cobuild (1995b), which uses evidence from a 200-million-word corpus to calculate the most frequent collocates of word-forms in a span of 4:4. In chapters 3 and 4, I use this data-base for an extended case study.)

Collocation is a relation between words in a linear string: a node predicts that a preceding or following word also occurs. Linear co-occurrence is traditionally referred to as a 'syntagmatic' relation. The prediction will only rarely be 100 per cent: there is usually choice, and *seeking* can obviously co-occur with different words. The term for this relationship of choice is 'paradigmatic'. However, this choice is not entirely free either, and often it is surprisingly restricted. With *seeking* there is almost a one-in-ten chance that it co-occurs with one of only five semantically related words.

These syntagmatic co-occurrence relations often cross-cut the way in which dictionaries have traditionally represented head-words. Sometimes different forms of a lemma behave differently (the SEEK example), but sometimes forms which are usually regarded as separate lemmas behave similarly. One such case is the collocational relation between the lemmas ARGUE and HEAT. One finds

- argue heatedly; heated argument; in the heat of the argument

These phrases cross-cut the traditional parts of speech, since the collocations are, respectively, between verb (ARGUE) and adverb (HEAT), adjective (HEAT) and noun (ARGUE), and noun (HEAT) and noun (ARGUE). In this case, the collocation is between semantic units, irrespective of grammatical category; but there is still a restriction on word-form, since the form *heat* has to occur: *heated argument*, but not **hot argument*.

2.4 Words and Units of Meaning

Dictionaries are mainly organized around individual words (lemmas), which are listed alphabetically with their meanings, but they do also usually list other longer phrases, where the meaning may not be predictable from the individual word-forms. The term 'lexical item' is therefore used to cover a range of individual words and phrases such as

- near, near-sighted, Near East
- nurse, nursery, nursery rhyme, nursery school
- nuclear family, nuclear winter

However, dictionaries differ greatly on how many such larger units are identified (these examples are from Cobuild 1995a).

Such phrases are often seen as an exception to the usual word–meaning correspondence, however there is often a lack of correspondence between words and units of meaning. Sometimes this is evident in arbitrary word-divisions and spelling conventions. We write *already*, and *almighty*, but *all right* (though many people write *alright*). We write *another* as one word, but *of course* as two. We write both *maybe she'll go away* and *she may be going away*. And it is largely a matter of personal choice whether we write *match box*, *match-box* or *matchbox*. Forms such as *I'll* and *it's* are written as one word-form (at least without a space in the middle), but are easily interpretable as two: *I will* and *it is*. (Though many people confuse *it's* and *its*.) Sometimes the apostrophe-*s* represents a word (as in *she's*), but sometimes it represents the possessive-*s* (*the man's hat*). We would usually think of the apostrophe-*s* as being attached to individual words (*Susan's bicycle*), but in *the king of England's hat* and *the boy across the road's bicycle*, the possessive-*s* is attached to a larger unit (see Bloomfield 1933: 178–9):

- the [king of England]'s hat
- the [boy across the road]'s bicycle

There are also cases where conventional word boundaries (spaces in writing) cross-cut the semantic units. A *small farmer* is not a farmer who is only five feet tall, but a farmer with a *small farm* or *smallholding*. A *heavy smoker* is not a smoker who weighs twenty stone, but someone who smokes heavily. Palmer (1971: 45) gives several examples, such as:

- a [small farm] -er, a [heavy smok] -er, a [criminal law] -yer, an [artificial flor] -ist

Some words have no independent existence at all, but occur only in one combination, for example *dint*, *kith* and *spick*, as in *by dint of*, *kith and kin* and *spick and span*. And *span* here is arguably not the same word as in *a span of six years* or *attention span*. In a few such often-cited cases, given one word, a hearer can predict with almost 100 per cent certainty what the following

one or two words will be. Words such as *kith* are certainly very restricted in their occurrence, though even apparently fixed phrases can be manipulated, as in the attested example:

- no more expensive to call your kith in Sydney than your kin in Southampton

Similarly, the word *amok* is almost always immediately preceded by RUN, but I have two examples of *an era gone amok* and *journalism gone amok*.

2.5 Delexicalization

The following examples also show that individual words are not always the unit of meaning. Some verbs, which are written as separate words, seem to carry little meaning. Quirk and Stein (1991) discuss examples of some common verbs in V–NP constructions:

- take a decision; take a look; take a shower; take a sip
- have a chat; have a drink; have a look; have a shower; have a swim; have a try
- give a scream; give a shout; give a speech
- make a mistake; make a note; make a suggestion

In these cases, almost all the meaning seems to be in the noun, and some of the phrases mean almost the same as corresponding verbs: for example, *to take a look = to look*, *to have a wash = to wash*. In such cases the verb is said to be delexicalized (although desemanticized would be a more logical term). By far the most frequent use of TAKE and MAKE is in phrases such as

- take place, take part, take care, make sure, make sense, make clear

I searched for the lemma pair *TAKE a* in a corpus of over two million words. There were over 400 examples, but in only about 10 per cent of these did TAKE have a literal meaning of "grasp with the hand" or "transport". The most common use by far is in combinations such as

- take a close look at; took an interest in; take a deep breath; takes a photograph; take a decision

where TAKE is delexicalized, and where almost all the meaning is carried by the noun.

The phenomenon of delexicalization is much more common than even these examples might suggest. Adjectives are usually thought of as narrowing the meaning of a noun. Thus *a red house* is more specific than *a house*, and the class of *dangerous dogs* is smaller than the class of *dogs*. Sinclair (1992) calls this use 'selective': the adjective selects a smaller set from the larger set. However, he distinguishes this from a 'focusing' use, and argues that:

> The meaning of words chosen together is different from their independent meanings. They are at least partly delexicalized. This is the necessary correlate of co-selection. . . . [T]here is a strong tendency to delexicalize in the normal phraseology of modern English.

He gives examples of adjective-noun pairs where the adjective is co-selected with the noun and shares part of the meaning. If the noun occurs on its own, little meaning would be lost:

- physical attack, physical damage, physical proximity
- scientific analysis, scientific experiment, scientific study
- general drift, general opinion, general public, general trend

Selective and focusing adjectives can be distinguished as follows.

selective	focusing
outward-looking	inward-looking
independent	dependent
separate choice	co-selected with noun
adds separate meaning	repeats part of meaning of noun
narrows meaning of noun	intensifies meaning of noun

Lorenz (1999) discusses similar examples of focusing adverbs:

- diametrically opposed, firmly entrenched, heavily loaded, instantly recognizable, irretrievably lost, readily available, ruthlessly exploited

In these cases, the adverb contributes little to the propositional meaning, but it emphasizes what the speaker regards as important. Similarly, *distinctly* is frequently used either in phrases where it adds little propositional meaning, or where it emphasizes the speaker's disapproval of something:

- distinctly different, distinctly audible, distinctly visible

- distinctly alarming, distinctly dated, distinctly inferior, distinctly nervous, distinctly odd, distinctly peculiar, distinctly queasy, distinctly uncomfortable, distinctly uneasy, distinctly unimpressed

Positive phrases do occur, but even phrases such as *distinctly better* imply that, up till now, things have been pretty bad.

We now have several cases where units of meaning do not coincide with individual words. Taken separately, they look like minor exceptions to the idea that individual words have individual meanings, but taken together, they start to throw considerable doubt on the status of words as the normal units of meaning.

2.6 Denotation and Connotation

So far, all my examples have been of relations between words and words (e.g. collocation), but words are used to talk about things in the world, and we therefore also need concepts to talk about relations between words and the world: reference and denotation.

Reference is the relation in a particular instance of use. If I say *Look at that huge dog over there*, then I have made an act of reference. It is not individual nouns which refer, but noun phrases: in this case *that huge dog*. Denotation means the appropriate range of reference of a word: for example, the word *dog* can appropriately be used to refer not only to small spaniels but also to large Saint Bernards. Reference and denotation are most obviously relevant to noun phrases and nouns, but they also apply to verbs and adjectives: for example, we might debate whether we could agree on the exact denotational boundaries between WALK, STROLL and HIKE, or between RED and PINK.

In summary: Reference is a speech act which picks out a referent in a concrete situation. Reference concerns language use. Denotation is a relation between a term in the language and a range of potential referents in the world. Denotation concerns the language system.

Different terms are used in this area. Denotation is also referred to as cognitive, conceptual, logical, ideational and propositional meaning. An everyday term is the 'literal meaning' of a word. This is often contrasted with connotation, which is also called affective, associative, attitudinal and emotive meaning.

Words can have the same denotation but different connotations. For example, *die* is a neutral word, but *pass away* attempts to express the speaker's sympathy, and *snuff it* expresses no sympathy at all. Such alternative words often exist in taboo areas, such as death: a *coffin* is neutral, but a *casket* sounds more dignified. Some words express little denotational meaning. I gave the example in chapter 1, that nothing is inherently *super*: the word expresses much about the opinion of the speaker, but little if anything factual about the world. Connotation is sometimes thought of as personal or emotional associations, conveying the attitude of an individual speaker, and if such meanings were purely personal and subjective, then they would be of limited interest. However, connotations are also widely shared within a speech community (see chapters 7, 8 and 9).

Denotation is usually taken to be a stylistically neutral and objective relation between a word and the world. It is often thought of as the most important part of the meaning: the basic or core meaning, which is not deniable. Connotations are often thought of as subjective, second-order or peripheral meanings, which depend on a relation between the word and the speaker/hearer. However, what is primary or secondary depends on one's point of view, and the expression of attitude may be the main function of the utterance. The distinction between denotation and connotation is usually clear, although the boundary can be hazy. It is often not easy to decide what is the primary denotation and what is the secondary connotation, and different dictionaries can differ considerably in what they present as part of the inalienable, undeniable denotational meaning, and what is merely implied or connoted.

2.7 Relational Lexical Semantics

The vocabulary of a language is not an unstructured list of words. In addition to the syntagmatic and paradigmatic relations between words, which I have started to illustrate, there are other relations which are repeated across many pairs and sets of words, and which make broad cuts across the vocabulary:

semantic fields
content and function words
core and non-core vocabulary

2.7.1 Semantic fields

The vocabulary of a language is internally structured by many clusters of words, which stand in different relations to each other, sometimes logical relations of sameness, difference and entailment, and sometimes vaguer relations within a topic area or semantic field. For example, there is an elaborate vocabulary for talking about horses in English. The following

'horsy' words all occurred as collocates (in a span of about 10:10) of 230 examples of *horse* in a 2-million-word corpus. They include words for types and colours of horse, movements that horses make, equipment used with horses, people who deal with horses, along with phrases and idioms which contain the word *horse*:

- bay, mare, pony, racehorse, roan, thoroughbred
- bolt, canter, gallop, rear, trot
- flank, hock, hooves
- mount, ride, on horseback
- harness, horseshoe, reins, saddle
- blacksmith, cowboy, jockey
- rocking-horse, runaway horse
- horse box, horse trough, stable
- don't look a gift horse in the mouth; don't get on your high horse; you can take a horse to water, but you can't make it drink; you're a dark horse; straight from the horse's mouth

Often words cluster because things in the world cluster; such as *horse*, *saddle* and *ride*, but there are always also conventional and recurrent ways of phrasing things.

2.7.2 *Synonyms, antonyms and hyponyms*

Semantic fields are not merely lists of words related by topic: they are also organized by relations amongst these words. Although words are inherently fuzzy in meaning, the vocabulary is structured. A *bush* is smaller than a *tree*, even if this is not a logical distinction and the boundary is unclear, and even if some large bushes are larger than some small trees. To *stroll* is to move more slowly than to *walk*, which is to move more slowly than to *run*: even if some people walk very fast.

Synonyms are words which mean the same. It is often said that it is difficult to find examples which are entirely convincing. After all, there would seem to be no reason why a language should have words which mean exactly the same. Certainly, it is rare to find words which are equivalent in both denotation and connotation. Candidates are *couch*, *settee* and *sofa*, which have the same denotation, though they occur in different collocations:

- casting couch, couch-potato, psychiatrist's couch, sofa-bed

Other candidates are pairs of words such as *glasses* and *spectacles*, where the second is stylistically more formal; and *car* and *automobile*, which are used in different national varieties. In some taboo areas of the vocabulary there are many synonyms which are close in both meaning and use. For example, there are many informal and pejorative words meaning "mad", such as

- bananas, barmy, bonkers, crackers, crazy, cuckoo, dotty, loony, loopy, nuts, potty, unhinged

Death is another taboo area where there are many approximate synonyms, such as the many words and phrases for "die":

- expire, give up the ghost, pass away, perish, shuffle off this mortal coil, snuff it

and many more. There are also several words and expressions for the dead human body, which illustrate relations between denotation, connotation and text-type. As usual all examples below are attested. *Body* is a neutral term, which is used in a wide range of contexts. *The deceased* denotes someone who has recently died, it connotes respect, and it is often used in legal contexts. A *corpse* connotes unpleasantness and often occurs in reports of a crime. A *stiff* is a slang term for a corpse, certainly disrespectful, possibly slightly old-fashioned, and possibly largely restricted to American detective fiction. A *cadaver* is a technical term, often used in medical contexts, especially with reference to study by medical students.

- Lenin's body lay in state
- a body was washed up on the beach
- determine the identities of the deceased
- the family of the deceased
- the corpse was barely recognizable
- the corpse was found floating in the river
- they found a stiff in the river
- anatomical investigations of a human cadaver

A *carcass* is used of larger animals, especially if they can be useful as meat, for either humans or animals. If it is used of humans, it is rude (*move your carcass over here*). *Carrion* is used of the decayed bodies of animals which are food for scavenging animals and birds.

- the carcass of a dead buffalo
- vultures picking at a lion's carcass
- meat left on the chicken carcass
- crows feeding on carrion

The approximate synonyms are distinguished partly by their denotations, but also by their connotations, and by the text-types they typically occur in. The contemporary variation in the lexis is due to historical changes in English. It was the influence of French and Latin after the Norman invasion of 1066 which contributed greatly to the expansion of the English vocabulary, via the semantic fields which were developed within social institutions such as the law and medicine and their associated bodies of knowledge and text-types. The core word is the Germanic *body*. Others are of Romance origin: *corpse* (compare French *corps*, Latin *corpus*), *cadaver* (compare French *cadavre*, Latin *cadaver*), *deceased* (compare French *décès*, Latin *decessus*).

Such sets of words also provide an insight into the way in which English categorizes a small part of the social world. There are several terms for dead humans. There are terms for dead animals, if they are useful as a food source. There are terms for large dead trees (*log*, *lumber*, *timber*), especially if they are useful and/or cultivated. But there are no terms for dead insects or smaller dead plants. The vocabulary embodies a hierarchy of importance and gives decreasing attention to humans, animals and plants.

Antonyms are words which are opposite in meaning. Speakers can often give immediate clang responses when asked for the opposite of a word (for example, *wet–dry*, *up–down*, *hot–cold*), but this provides another example of the limited relevance of asking for the meaning of isolated words. Does it really make sense to ask for the meaning of *dry*? Or to ask for its opposite? The word has a core meaning and a prototypical antonym. If you ask for a clang response in isolation, people will probably say *wet*. However, this only works in some cases. Compare

dry socks	wet socks
dry season	wet season *or* rainy season
dry wine	sweet wine
dry skin	moist skin?
dry humour	unsubtle humour?
a dry area	an area which has pubs
a dry run	the real thing?
dry land	sea?
dry-cleaning	washing?

Dry means different things in these different phrases, not to mention *high and dry* and *there were few dry eyes in the house*. Conversely, several uses of *wet* do not have an obvious opposite at all: *wet blanket*, *wet nurse*, *feel like a wet rag*, *a Tory wet* (a British English term for a Conservative politician, especially in Margaret Thatcher's government, who holds moderate views).

There are many similar examples. A clang response to the opposite of *white* would probably be *black*, but compare:

white wine	red wine
white collar	blue collar
white coffee	black coffee

(and white coffee is, well, coffee-coloured, that is, light brown).

Antonymy has traditionally been regarded as a paradigmatic opposition permanently available in the lexicon of the language. However, it is better seen in addition as a syntagmatic relation, which is realized in co-text. For example, the commonest collocate of *bride* is its antonym *groom*: that is, the words often co-occur (usually in the phrase *bride and groom*), rather than being available in paradigmatic opposition to each other (and often they could *not* substitute for each other). Out of context, the antonym of *conventional* might be *modern*, but in a text about weapons, the antonym might be *chemical* or *nuclear*.

Hyponymy is the logical relation of class inclusion. A *buttercup* is a kind of *flower*, which is a kind of *plant*; a *spaniel* is a kind of *dog*, which is a kind of *animal*. There is a large number of approximate synonyms, and more and less specialized hyponyms, for groups of people: this is not surprising, since the different ways in which people can be grouped is of inherent social interest. *Group* is a neutral superordinate word for a collection of things, animals or people. One hyponym is *crowd*: a "very large group of people". In turn, a hyponym of *crowd* is *mob*: an "unruly crowd".

2.8 Frequent and Less Frequent Words

Words in texts are distributed very unevenly: a few words are very frequent, some are fairly frequent, and most are very rare. These facts are due to two distinctions which provide ways of talking about the vocabulary of a language and about the distribution of the vocabulary in texts: function and content words, and core and non-core words.

2.8.1 Content and function words: lexical density

In English and in many other languages, there is a distinction which divides the whole vocabulary into two major categories: content words tell us what a text is about, and function words relate content words to each other. The distinction is made in most grammars of English, but since many linguists make essentially the same distinction, there are several terms in use. Content

words are also referred to as major, full and lexical words. They carry most of the lexical content, in the sense of being able to make reference outside language. Function words are also referred to as minor, empty, form, structural and grammatical words. They are essential to the grammatical structure of sentences. Their function is internal to the language, for example in making explicit the relation of lexical words to each other. The distinction is made by Henry Sweet in his famous grammar of 1891:

> In a sentence such as *The earth is round*, we have no difficulty in recognizing *earth* and *round* as ultimate independent sense-units. . . . Such words as *the* and *is*, on the other hand, though independent in form, are not independent in meaning: *the* and *is* by themselves do not convey any ideas, as *earth* and *round* do. We call such words as *the* and *is* form-words, because they are words in form only. When a form-word is entirely devoid of meaning, we may call it an empty word, as opposed to full words such as *earth* and *round*. (Sweet 1891: 22)

It is possible to conceive of a communicative system which has only content words, but not of a system which has only function words. For example, in a telegram one can omit function words and still have a comprehensible message:

- Please meet Harry airport six Saturday evening [I]

These two semantic categories divide the traditional parts of speech into two broad sets:

content words: noun, adjective, adverb, main verb
function words: auxiliary verb, modal verb, pronoun, preposition, determiner, conjunction

The boundary between the two word classes is not perfectly clear-cut. For example, modal verbs (*must*, *can*, *should*, etc.) express obligation, permission and ability, and therefore convey content; and pronouns can have extra-linguistic reference. However, as well as the rough semantic distinction, content and function words have strikingly different formal characteristics. Briefly: content classes have many members (there are tens of thousands of nouns, but only a couple of dozen pronouns), and are open to new words (for example, new nouns and verbs are being constantly invented; it is very rare for new pronouns to enter the language). And only content words take inflections (such as plural inflections on nouns, person endings on verbs).

This distinction between two classes in the vocabulary is relevant to text structure, because different types of texts have predictably different proportions of content and function words. Certain restricted text-types, mainly lists of different kinds, consist entirely of content words. However, usually the difference between text-types is one of proportion. On average, written texts have a higher proportion of content words than spoken texts, because written texts can be more tightly packed with information.

The lexical density of a text is the proportion of lexical words expressed as a percentage. If N is the number of running word-forms in text, and L is the number of lexical word-forms, then

$$\text{lexical density} = 100 \times L/N$$

Ure (1971) studied corpora of 42,000 words of spoken and written texts, and showed a strong tendency for written texts to have a lexical density of over 40 per cent (range 36 to 57) and for spoken texts to be under 40 per cent (range 24 to 43). There are functional interpretations for these findings. On average, a written text is shorter and has fewer repetitions than a comparable spoken text. It is permanent, highly edited and redrafted, rather than being unplanned and spontaneous, as casual conversation is. A written text is relatively context-free, though never entirely so, whereas a spoken text can rely to a large extent on the immediate physical context. We would therefore expect the information load to be higher in a written text: since it is permanent, readers can reread obscure sections. Spoken texts must, on the other hand, be understood while they are being produced: they must be more predictable.

So, on average, written texts are less predictable, and spoken texts are more predictable. In turn, content words are less predictable: there are thousands of them. Function words are more predictable: there are small numbers of them. For example, there are only half a dozen frequent conjunctions. We would expect, therefore, that written texts have a higher proportion of unpredictable content words, and that spoken texts have a higher proportion of more predictable function words. More recent studies with larger corpora confirm Ure's findings (Stubbs 1996: 71–6).

2.8.2 *Core vocabulary*

Another way of comparing texts is to calculate what percentage of words from the core vocabulary they contain. By definition, the core vocabulary is known to all native speakers of the language. It is that portion of the vocabulary which speakers could simply not do without.

Suppose we have sets of words which are related by approximate synonymy and hyponymy:

- break, burst, chip, crack, shatter, smash, snap

- gaze, glance, glimpse, look, peer, watch
- quake, quiver, shake, shudder, tremble
- display, exhibit, expose, flaunt, show
- drudgery, labour, toil, work
- dirty, filthy, grimy, grubby, soiled, unclean

I think there would be widespread agreement that one word in each list is somehow more basic than the others:

- break, look, shake, show, work, dirty

Such intuitions are partly based on frequency, but also on functional criteria, such as which words would be most easily understood by children or non-native speakers, or which words it would be most useful to introduce in the early stages of teaching English as a foreign language.

The core vocabulary will certainly contain the most frequent words in the language. The 100 most frequent word-forms from a large general corpus will be mainly function words such as

- the, of, and, to

plus a few content words such as

- think, know, time, people, two, see, way, first, new, say, man, little, good

And the 2,000 or 3,000 most frequent word-forms will include words which are indispensable for discussion of a wide range of topics. However, beyond the top few hundred words in different general corpora, word frequency varies greatly, and merely reflects the content of the texts in the corpora. Therefore raw frequency lists often have odd gaps, because, for example, the word *Sunday* is twice as frequent as *Tuesday* (see chapter 1.7.3). However, the vocabulary is a structured whole, not an unordered list of words. Therefore, the core vocabulary contains common closed sets of words, such as days of the week, months and seasons, numbers, and sets with a few frequent members such as colours, major family members, parts of the body, and common professions.

The main defining criterion of core vocabulary is that of maximum usefulness. This criterion can be operationalized in two main ways. We can discover which words are widely and relatively evenly distributed in texts of different kinds, and we can discover which words can be used for defining other words:

1 *Distribution in texts.* The core vocabulary includes words which occur not only frequently, but with a relatively even distribution across a wide variety of texts and text-types. For example, *doctor* will occur in texts of many kinds, both everyday and specialist, whereas *paediatrician* may be common in a few texts, but only on restricted specialist subjects. Core vocabulary is not restricted to specialist fields or genres: for example *children* (versus *offspring* or *progeny*), *brothers* and *sisters* (versus *siblings*), and *stomach* (versus *abdomen*). And core vocabulary is neutral stylistically, neither markedly casual nor formal: for example, *child* (versus *kid* or *kiddy*), *drunk* (versus *pissed* or *inebriated*), and *give* (versus *award* or *donate*).
2 *Semantic usefulness.* Core words are often useful for defining other words: that is, they are not hyponyms with a narrow denotation. For example, the core words *laugh* and *softly* can be used to define non-core *chuckle*. Similarly, *clumsy* and *walk* can be used to define *waddle*.

Sometimes, the two criteria coincide: for example, *paediatrician* is a hyponym of *doctor*, and *award* and *donate* are hyponyms of *give*.

2.9 Two Examples

Here are two small case studies which use corpus data to document the main principle of this chapter: that observable corpus data can provide evidence of both denotational and connotational meaning.

2.9.1 Example 1: Bloomfield's analysis of SALT

In what was for many years the main student textbook in American structuralist linguistics, semantics was regarded as 'the weak point in language study', since a study of meaning would require human knowledge to advance 'very far beyond its present state' (Bloomfield 1933: 140). Bloomfield put forward a general argument that meanings were simply too complex to analyse systematically:

> The situations which prompt people to utter speech, include every object and happening in their universe. In order to give a scientifically accurate definition for every form of a language, we should have to have a scientifically accurate knowledge of everything in the speaker's world. (p. 139)

He concluded that meanings 'could be analysed or systematically listed only by a well-nigh omniscient observer' (p. 162). In addition, he

attributed a special status to a particular form of 'scientifically accurate knowledge':

> We can define the names of minerals, for example, in terms of chemistry and mineralogy, as when we say that the ordinary meaning of the English word *salt* is "sodium chloride (NaCl)" . . . but we have no way of defining words like *love* or *hate*, which concern situations that have not been accurately classified – and these latter are in the great majority. (p. 139)

Both of these arguments are usually regarded today as faulty. First, 'everything in the speaker's world' is not an unorganized flux, but categorized by social cognition into lexical fields. Second, it is odd to argue that the 'ordinary meaning' of *salt* is NaCl. It is not necessary to know this meaning at all in order to use the word appropriately in most everyday situations. Most native speakers of English probably do not know the chemical formula for table salt, and the "NaCl" meaning is often quite irrelevant to the use of the word.

Third, some of Bloomfield's arguments seem very strange indeed:

> We have defined the *meaning* of a linguistic form as the situation in which the speaker utters it and the response which it calls forth. (p. 139)

This statement is understandable in view of Bloomfield's behaviourist assumptions: situations provide stimuli which evoke responses in speakers. It is nevertheless odd to say that the meaning *is* the situation. More reasonable might be an argument which runs: meanings are essentially mental or psychological events, they take place inside people's brains or minds, and we have no idea how this works. The process is unobservable and we may as well give up trying to study it. Bloomfield was driven to this position (which he stated very clearly, even if he did not always follow it himself) because of his view of scientific methodology that linguistics must be based on observable facts.

An answer to Bloomfield's pessimistic view is to look at some data on usage. We can then agree with Bloomfield's own argument about the necessity for observable facts, but use it against his position. There are many observable patterns, and some aspects of meaning are observable: meaning is use. So, to answer the question 'what does *salt* mean?', we will observe how it is used in attested data. In a corpus of over two million words, the most frequent combinations were *salt and pepper* and *salt water*. The first of these occurs frequently in recipes, often in longer phrases such as

- season with salt and pepper; a good sprinkling of salt and pepper; salt, pepper and mustard

In cooking contexts, the antonym pairs are *salt–sugar*, or *salty–sweet*, but the antonym of *salt water* is *fresh water*. However, in German, fresh water is *Süsswasser* (= "sweet water"). In other words, these are linguistic facts, not facts which relate directly to the world. The plural *salts* occurs with quite different collocates:

- copper salts, iron salts, mineral salts, vegetable salts

In these cases *salt* does not mean "NaCl". We see here again the principle that different forms of a lemma may not have the same meaning. In addition, there are several idioms where the "NaCl" meaning may be remote to modern speakers:

- rubbed salt into the wounds; has to be taken with a pinch of salt; the salt of the earth; if he is worth his salt

The literal denotation of "NaCl" can explain the history of the phrases (for example, salt used to be a very valuable commodity, used for preserving food), but now this is largely lost in extended metaphorical meanings, which may rely in turn on intertextual Biblical allusions.

In conclusion: (1) The lemma SALT does not always mean "NaCl". (2) Admittedly, we cannot directly observe the meanings of SALT, but corpus data provide much evidence for these meanings. (3) These meanings depend on relations with other words in the co-text, or with other words in other texts.

2.9.2 *Example 2:* CAUSE problems *and* CAUSE amusement

Here is a second small case study which further illustrates three principles. (1) Words should be studied, not in isolation, but in collocations. (2) Findings from one corpus should be checked against an independent corpus. (3) Potential counter-examples should be carefully checked.

The lemma CAUSE almost always co-occurs with unpleasant collocates. Evidence of this can be seen in concordance 2.1, which presents some raw data on the verb lemma from a spoken corpus. A concordance is the main tool of corpus linguistics. The computer is programmed to search for all examples of a node word in a corpus and to print them out in the centre of the page or screen within a given context, of a few words to left and right.

In a detailed study (Stubbs 1995a), I looked at the 38,000 occurrences of the lemma CAUSE (verb and noun) in a corpus of 120 million words of general English. Amongst its 50 most frequent collocates, within a span of 3:3, there were only words (most frequently abstract nouns) with unpleasant connotations. The most frequent were

1.	ody's land as long as you don't	cause	a criminal offence then you've g
2.	erm bankrupt some firms and so	cause	a lot of social disruption
3.	t you get a pay rise that would	cause	a public outcry?" And the Guardi
4.	at to say the wrong thing would	cause	a row er Joanna said er don't
5.	here. Erm originally it used to	cause	problems between the children an
6.	But it's not the sutures that	cause	the wound to heal [FOX] it's
7.	t make weapons of war you would	cause	unemployment but there's no reas
8.	ly go and do anything they want	cause	whatever misery they want cause
9.	blizzards for fifty years have	caused	a state of emergency in souther
10.	that's another area that that's	caused	antagonism between us is the fa
11.	erm has MX's behaviours ever	caused	argument or conflict between yo
12.	illiam Hague in the by-election	caused	by the erm er move er [ZF1] of
13.	on are are of are generated and	caused	by the Holy Spirit Himself. You
14.	nine per cent of all illness is	caused	directly or indirectly by a bas
15.	now it sort of [pause] If I say	caused	problems I don't mean it full-
16.	ay it was total negligence that	caused	this and I don't feel that thes
17.	events that were happening that	caused	us to go downhill effectively e
18.	nd the harm if you like that is	caused	you if you can't have children.
19.	any issues which have caud you	caused	you particular stress or distre
20.	ed to any school so that always	causes	a bit of erm er er er confusion
21.	t so many kilograms per hectare	causes	a loss of something or [F01] Mm
22.	ir own crowd so to speak and it	causes	a major disruption not so much
23.	[M02] right. Oh uh the air u	causes	a vacuum and that's why it stic
24.	d a bit of a smokescreen. If he	causes	chaos in class then the teacher
25.	and that the sheer trauma of it	causes	him a heart attack? [M01] Mm.
26.	Y the lack of air on the inside	causes	it to stay down. [M02] Pulls it
27.	r than to look what cau at what	causes	it which would mean you'd have
28.	there are many theories on what	causes	its stages. Too much dairy in t
29.	ies away from home. This always	causes	pressures doesn't it. [F02] I t
30.	rea is a horrific disease which	causes	severe dementia in middle age.
31.	so that the depression is what	causes	the spending. It's not like the
32.	asked them that there that it	causes	this many problems with that ma
33.	here's more than one thing that	causes	warming and cooling you know.
34.	No. [M01] Would you say that it	causes	you inconvenience [F01] Mm. [M0
35.	how much extra work it [pause]	causes.	[M0X] Well what they've done i
36.	ll see people getting drunk and	causing	a fuss and running amok. [M21]
37.	hree years [M01] Yeah [F01] Erm	causing	a great deal of furore in the
38.	to erm to that behaviour that's	causing	a problem for us we can call
39.	break off at speed and that is	causing	City all sorts of problems.
40.	round on the street of 'er kids	causing	criminal damage to property
41.	n remember [laughs] practically	causing	G B H on my best friend [laugh
42.	ildren by going out to work and	causing	infant mortality because a chi
43.	it was extra-terrestrial beings	causing	it or maybe some kind of pract
44.	esday afternoons. This last was	causing	some concern to students becau
45.	minister to those whose past is	causing	the future to look very bleak
46.	tion you cause an ice age or by	causing	the ice age you shut down the
47.	st to find out what is actually	causing	the problem [F02] Mhm [F01] An
48.	if there's a if a group's been	causing	trouble we'll try and get them
49.	ery much. [F01] Okay. [F0X] I'm	causing	you problems aren't I? [F01] N
50.	what er the the thing that was	causing	you the upset was it the becau

Concordance 2.1 Fifty random examples of CAUSE (verb)
Notes: The data are from the spoken language sub-corpus of the Bank of English. [FOX] etc. are speaker identification codes.

- CAUSE <problem(s) 1806, damage 1519, death(s) 1109, disease 591, concern 598, cancer 572, pain 514, trouble 471>

The lemma often occurs in longer combinations of verb plus adjective plus noun, such as

- cause considerable damage; cause great problems; cause major disruption; cause severe pain

Not all widely used dictionaries explicitly draw attention to these negative uses. Some do, but others give a neutral definition such as "a cause is something which produces an effect". However, the examples in corpus-based dictionaries (such as CIDE 1995; Cobuild 1995a; LDOCE 1995; OALD 1995) include:

- heavy traffic is causing long delays; the cold weather caused the plants to die; it was a genuine mistake but it did cause me some worry; the cause of the fire was carelessness; causes of war; cause for anxiety; cause of the accident; cause of the crime problem; her rudeness was a cause for complaint

A minority of examples in these dictionaries are neutral or positive, such as *every cause for confidence*, though there is no indication of how much less likely positive examples are.

Corpus data allow collocates to be extensively documented. In my data, collocates which occurred as subject or object of the verb CAUSE or as prepositional object of the noun CAUSE include:

- abandonment, accident, alarm, anger, annoyance, antagonism, anxiety, apathy, apprehension, breakage, burning, catastrophe, chaos, clash, commotion, complaint, concern, confusion, consternation, corrosion, crisis, crowding, damage, danger, death, deficiency, delay, despondency, destruction, deterioration, difficulty, disaster, disease, disorganization, disruption, disturbance, disunity, doubt, errors, frustration, harm, hostility, hurt, inconvenience, interference, injury, interruption, mistake, nuisance, pain, pandemonium, quarrel, rejection, ruckus, rupture, sorrows, split, suffering, suspicion, trouble, uneasiness, upset

I give quite a long list because it shows very clearly that there is a very simple semantic pattern ("bad things get caused"), which is realized by considerable lexical variation.

(A much less frequent sense of CAUSE as "aim or principle" is signalled by different collocations, including: *devoted service to this cause*; *conviction that your cause is right*; *plead a cause*; *take up causes*. Causes in this sense are *good*, *glorious*, *just* and *worthy*, but also *lost* and *foolish*.)

If different samples of data gave different results, then these unpleasant associations might be a feature of the corpora, not a collocational property of the word. I had no reason to suspect that my corpus was biased by containing lots of texts about gloomy things, but I carried out the same analysis on other independent corpora. For example, a corpus of 425,000 words, comprising texts about environmental issues, contained three or more examples of the collocates

- blindness, cancer, concern, damage, depletion, harm, loss, ozone, problems, radiation, warming

These collocates reflect the environmental topics (in phrases such as *cause global warming*), but the same simple semantic pattern holds. (See Gerbig 1996 for a detailed analysis.)

However, the question now arises as to whether there are counter-examples to the generalization. A possible collocate of the verb CAUSE is *amusement* as in *to cause someone amusement*. Are such occurrences genuine counter-examples to the expected negative uses of CAUSE? First, I studied the 100 occurrences of the word-form *amusement* itself in over six million words of running text. I ignored a few phrases such as *amusement arcade* and *amusement park*. In some cases, negative connotations are signalled by an adjective at N−1; many of the remaining examples implied a degree of *schadenfreude* towards the butt of the amusement, and thereby disapproval of those who are amused:

- derived malicious amusement; with wicked amusement; for his own twisted amusement; with a little sardonic amusement; silly boyish amusement; a look of contemptuous amusement
- she was listening with more amusement than respect
- suppressed amusement at his outrageous manners
- vulgar aspects seem to have been a source of some amusement
- landing flat on my back, much to the amusement of the lads

(The last example is from Cobuild 1995a.) Certainly, not all examples are disapproving, though even more positive cases may imply a condescending, patronizing attitude:

- Lovat listened with affectionate amusement
- Victor stood watching her in fond amusement

However, the crucial question is whether collocations of CAUSE-*amusement* are disapproving. I found eight examples in nearly 60 million words of data. In six of these cases it is evident, even from a small context, that disapproval is being expressed of the person who is amused, or the amusement is at someone's expense:

- it caused a certain amount of amusement in the lab
- the affair with Kim caused her a great deal of amusement
- the veiled hints caused us plenty of amusement
- the amusement caused by my looking so hot
- the unexpected reference caused titters of amusement
- are also cause for sardonic amusement

In summary: The collocation CAUSE-*amusement* does not provide counter-examples to the generalization that CAUSE has overwhelmingly unpleasant connotations.

2.10 Summary and Implications

In this chapter I have introduced the main terms and concepts which I need in the subsequent chapters:

word-form and lemma
collocation (node, span and collocates)
denotation and connotation
semantic (or lexical) fields
content and function words
core and non-core vocabulary
I have used two main arguments:

1 Individual words often do not correspond to units of meaning. Individual forms of a lemma may have quite different uses, and often the unit of meaning is a longer phrase or collocation.
2 There are many structural relations within the vocabulary of a language, including logical relations between words, such as synonymy, antonymy and hyponymy. These relations hold between words in the vocabulary, but also between word-forms in texts where they contribute to text cohesion (see chapters 5 and 6). In addition, words can be divided into broad classes, such as content and function words, core and non-core

vocabulary. These distinctions also concern how words are used in texts: for example, the density of information in a text, or how specialized the text-type is.

Two brief case studies showed that corpus data can provide evidence of both denotation (the SALT example) and connotation (the CAUSE example). The principle 'meaning is use' leads to observational methods of corpus semantics. The main tool of corpus semantics is the concordance, which allows words and their characteristic collocates to be studied in detail.

I have also now given initial examples of the main empirical methods which underlie corpus semantics. The primary data are texts. Linguistics studies human language: but this is not directly observable. Even individual languages (such as English, German or Swahili) are highly abstract objects, and also not directly observable. However, languages are realized in texts, and these texts are observable. They exist independently of the observer, and provide publicly accessible, objective data. The patterns in large collections of texts are also not directly accessible to the individual human observer, but if texts are stored as a corpus, in computer-readable form, then computer-assisted methods can be used to discover their structure and regularities.

The interpretation of such data involves constant subjective decisions. However, these decisions are testable and can be checked by independent observers. Data and methods therefore make possible the replicable and empirical analysis of meaning. Chapter 3 will describe in more detail how patterns can be discovered in corpus data.

2.11 Background and Further Reading

Amongst the most influential early discussions of the main concepts of lexical semantics (including semantic fields, synonymy, antonymy and hyponymy, and semantic features) were two textbooks by Lyons (1968, 1977). These discussions are, however, not based on textual or corpus data: indeed, in the two volumes and over 800 pages of Lyons (1977), there is not a single example of a naturally occurring text. Cruse (1986) provides a widely used textbook on lexical semantics, and Aitchison (1987) provides a very readable student introduction to many aspects of meaning and to the organization of the 'mental lexicon'. These are only four books out of very many, and lexical semantics is discussed in most introductions to linguistics.

For more detailed discussions of core vocabulary and lexical density, see Stubbs (1986) and Stubbs (1996: 71ff), respectively, and further references there. There are many further references to work on collocations and phraseology in general in chapter 3.11, below.

2.12 Topics for Further Study

(1) Use corpus data to state the collocates of BOGGLE. (Does it always collocate with *mind*?) And *blithering*. (Are there any other forms of a lemma BLITHER?) And GRUFF. (The phrase *gruff voice* is common, but there are also other collocates. Do the collocates share a semantic feature?)

(2) Some words and phrases usually occur in the negative (Buyssens 1959; Laduslaw 1996: 328; Sinclair 1998):

- not bad-looking; wouldn't budge; didn't cut much ice; didn't drink a drop; wouldn't lift a finger to help me; not so much as a red cent; I've never set eyes on her

Are these phrases always negative? Or are there exceptions?

(3) Study examples of adverb–adjective phrases such as

- absolutely certain, potentially dangerous, singularly stupid, specially designed, totally different, understandably reluctant, virtually impossible

Is it possible to make generalizations about the adjectives which typically follow these adverbs? A useful article is by Louw (1993), who analyses the negative implications of *utterly*, as in *utterly confused* and *utterly ridiculous*.

(4) Study the sets of words which make up semantic fields, such as cooking and furniture. Some examples include

- boil, cook, fry, grill, roast, sauté
- garlic, herbs, parsley, spices
- bookcase, chair, cupboard, desk, furniture, table
- chair, chaise longue, couch, sofa, stool

State the semantic relations which hold between the words in such sets, such as hyponymy and approximate synonymy. And list some of the conventional phrases in which the words occur.

(5) Compare two or three different dictionaries to see whether they agree on the central and more peripheral meanings of the following words:

- acquisitive, affectation, aloof, antiquated, avaricious

- gaggle, garish, gimmicky, glamorize, glib, grandiose, grovel

For example, the word *gaggle* denotes a kind of "group", but also expresses criticism or disapproval. Is the denotation more basic? Or are denotation and connotation equally central in a phrase such as *a gaggle of teenagers*? Consider whether the words express disapproval as an inherent part of their denotation, or whether the disapproval is a deniable connotation.

(6) There are many hyponyms for *group*, and denotations and connotations differ in different phrases. There are specialized words for groups of animals which collocate only with certain animal names, and therefore have a more restricted denotation than the superordinate:

- flock, gaggle, herd, pride, shoal

Some of these words can be used for people, often, though not always, with pejorative connotations:

- following the herd, a clergyman's flock

Use corpus data to study the collocates and meanings of these words, and others such as

- band, bunch, clan, crew, crowd, gang, horde, mob, rabble, team, tribe

Why should there be so many different words for talking about groups of people?

(7) Study the concordance lines for the verb CAUSE in section 2.9.2, and identify all the object noun phrases. What are the most frequent collocates? Are there any counter-examples to the generalization that what is caused is something bad? If yes, what explains the counter-examples?

(8) Set up a list of core vocabulary for English or some other language, as the basis for vocabulary teaching in the initial stages of learning a foreign language. This would involve quite a substantial project, rather than just a study question: perhaps a project for a final-year undergraduate dissertation. The project would involve at least the following steps.

Use frequency lists (from a large raw corpus or from a corpus-based dictionary) to establish a starting list. For English, the Cobuild Dictionary (1995a) gives lemmas in frequency bands up to 30,000, and the Cobuild Collocations Dictionary on CD-ROM (1995b) gives the 10,000 most frequent word-forms (see chapter 3, below). Extract the top 2,000 or 3,000 lemmas. Develop this list by completing sets of words which are incompletely represented. Test the list in various ways. Check that the candidate words are evenly distributed across many texts and not clustered in just one or two texts. Check that the list does not include hyponyms which are too specialized. Check what text coverage the list provides: a reasonable aim might be 95 per cent in texts from a general corpus. Compare the list against well-known published lists (such as those at the end of LDOCE, 1995, or OALD, 1995).

Any list is only as good as the corpus or data-base on which it is based. The original corpus must at least contain a wide selection of text-types: spoken and written, formal and informal, fiction and non-fiction, intended for children and adults, and must sample widely used genres.