as absolute figures, as normalised figures and as percentages. The notion of representativity was also discussed. In the final section we saw how corpora can be tagged for part of speech, parsed for grammatical structure, or annotated for semantic, pragmatic/discourse or prosodic features.

Study questions

1. What is the difference between qualitative and quantitative methods? How can quantitative methods be combined with qualitative analysis?
2. What is lemmatising? How does it influence frequency counts?
3. Why is it important to normalise frequencies when you compare results from different corpora?
4. What arguments are relevant when we discuss whether a corpus is representative or not?
5. What are the main types of corpus annotation? Why is it important to know the principles behind the tagging of a corpus that you are using?

Corpus exercises

Hands-on exercises related to this chapter are available at the following web address:
http://www.euppublishing.com/series/ETOTELAdvanced/Lindquist.

Further reading

It is hard to find easy-to-digest information on statistics. Even Oakes (1998), which is specifically for corpus linguistics, is rather tough going, and the chapters on statistics in introductory books on corpus linguistics such as McEnery and Wilson [1996] (2001), Biber et al. (1998) and Meyer (2002) are quite brief. A clear but somewhat technical introduction is given in Baroni and Evert (2009). There are also tutorials on the web.

Aston and Burnard (1998) deal with some of the questions raised here, including representativity. The classic paper on representativity is otherwise Biber (1993). The best introduction to corpus annotation is still Garside et al. (1997), especially the first two chapters. More detailed and up-to-date treatments of questions related to annotation can be found in Lüdeling and Kytö (2008: 484–705). Treebanks are treated in Wallis (2008). Nelson et al. (2002) describe the philosophy behind the parsed version of ICE-GB and give many examples of studies that can be made by means of the special search program ICE-CUP.

3 Looking for lexis

3.1 The role of the lexicon in language

The traditional distinction between the grammar and the vocabulary of a language is reflected in the fact that we have separate grammar books and dictionaries. Generative linguistics in the second half of the twentieth century emphasised this separation and strove to describe grammatical structures in which lexical items were inserted only at a late stage in the derivation (construction) of a sentence, when the grammatical structure had already been decided. The lexicon in such a model has a peripheral role as a storehouse of words, idioms and oddities that cannot be described by rules.

In many alternative linguistic models, however, the lexicon is given a more central role. In these models, lexical items or small classes of lexical items are believed to have not only their own meaning but also their own "local" grammars, which can be discovered by the close study of language in use. The ideal method to carry out such studies is to use corpora. It is therefore reasonable that in this book we devote several chapters to the lexicon. We will begin by saying a few words about using corpora in the production of dictionaries.

3.2 How lexicographers use corpora

In the old days, lexicographers collected slips of paper with text excerpts for all the words they were going to include in their dictionary, just as Jespersen did for his grammar. The editors of the *OED*, for instance, had hundreds of correspondents who collected and sent in slips with citations of interesting new words or new uses of old words. In the end there were hundreds of thousands of handwritten slips in the purpose-built pigeonholes in the garden shed of the editor, James Murray, in Oxford (where a large number of helpers were busy sorting them for him). This was very cumbersome work, so computers have been a great invention

for dictionary-makers. As mentioned in Chapter 1, the Cobuild corpus was originally compiled for lexicographical purposes, and today all major British dictionary publishers have their own corpora (many of which unfortunately are not accessible to outside researchers). The editors use concordances to find out the typical meanings and constructions in which each word is used, and try to evaluate which of these are worth mentioning in the dictionary. Many dictionaries also quote authentic examples from corpora, either verbatim or in a slightly doctored form. It is also possible for modern lexicographers to get a good grip on differences between genres and registers by studying specialised corpora or subcorpora. Finally, corpora give rich information on common words like *take, go* and *time*, which often have very many meanings and uses which tend to be overlooked in introspection and by human citation collectors. In the next section we will look at a dictionary entry and compare it with findings from a corpus which is different from the one that the dictionary is based on.

3.3 The meaning of words

As mentioned in the discussion about tagging, the easiest way into a corpus is to search for lexical items. But even that is not always totally straightforward. First we must decide whether we are interested in a word form or a lemma. If we are interested in the verb *squeeze*, we are probably interested in all the forms of that verb: *squeeze, squeezes, squeezed* and *squeezing*. These forms together constitute the verb lemma SQUEEZE. We might also decide to study the noun lemma SQUEEZE as in *one squeeze* and *several squeezes*. Remember that lemmas are often written in small caps. Before we go on, try think of all the meanings of the verb *to squeeze* that you know. You probably have a general idea of the meaning, and some contexts probably crop up in your mind, maybe something like *She squeezed my arm* or *He squeezed out the last of the toothpaste.*

Let us now see what the *Longman Dictionary of Contemporary English*, a learner's dictionary, says about the verb *squeeze* (Figure 3.1).

The *Longman Dictionary of Contemporary English* is a corpus-based dictionary, so the definitions given here are already based on corpus findings. As you can see, the dictionary lists no fewer than six meanings plus a number of phrases like *squeeze sth out* and *squeeze up*. Clearly, describing the meaning of even a relatively simple verb like *squeeze* is far from easy, and it would be extremely difficult to arrive at all these meanings by introspection only.

To get a feeling of what kind of material lexicographers work with, we will now compare the description given in the dictionary with what can find in a corpus. In Figure 3.2 the first twenty concordance lines

squeeze[1] /skwiz/ *v* **1 PRESS** [T] to press something firmly together with your fingers or hand: *She smiled as he squeezed her hand.* | *He squeezed the trigger, but nothing happened.*
2 PRESS OUT LIQUID [T] to get liquid from something by pressing it: *Squeeze the oranges.* | **squeeze sth out** Try to squeeze a bit more out. | **squeeze sth on/onto sth** *Squeeze a bit of lemon juice onto the fish.*
3 SMALL SPACE [I, T always + adv/prep] to try to make something fit into a space that is too small, or to try to get into such a space (cf. **squash**): [+**into**] *Five of us squeezed into the back seat.* | [+**through/past**] *He had squeezed through a gap in the fence.* | **squeeze sb/sth in** *We could probably squeeze in a few more people.*
4 squeeze your eyes shut to close your eyes very tightly
5 JUST SUCCEED [I always + adv/prep] to succeed, win, or pass a test by a very small amount so that you only just avoid failure: *Greece just squeezed through to the next round.*
6 LIMIT MONEY [T] To strictly limit the amount of money that is available to a company or organization: *The government is squeezing the railways' investment budget.*
squeeze sb/sth **in <-> also squeeze sth into sth** *phr v*
to manage to do something although you are very busy: *How do you manage to squeeze so much into one day?* | *I can squeeze you in at four o'clock.*
squeeze sth **<-> out** *phr v*
1 to do something so that someone or something is no longer included or able to continue: *If budgets are cut, vital research may be squeezed out.*
2 to squeeze something wet in order to remove the liquid from it: *Squeeze the cloth out first.*
3 to squeeze sth out of sb to force someone to tell you something: *See if you can squeeze more information out of them.*
squeeze up *phr v BrE*
To move close to the person next to you to make place for someone else

Figure 3.1 Dictionary entry for the verb *squeeze*.
Source: Longman Dictionary of Contemporary English [1978] (2003: 1606–1607)

containing the infinitive form of the verb *squeeze* found in the 1990s section in the Time Corpus are given. They have been ordered according to the five different meanings given in the dictionary.

Meaning 1: PRESS

The two examples given in the dictionary entry are both about concrete things being pressed together (a hand, the trigger of a gun), but among our twenty Time Corpus tokens only one, example (1), is of this kind. Examples (2)–(5) are about (political) forces applying pressure on a country and its ruler. This metaphorical meaning is not mentioned in the dictionary (for a definition of metaphor, see Chapter 6). Example (6) is about a period of time which is made shorter by compression, so to speak. Examples (2)–(6) illustrate the fact that many words and expressions which are originally based on physical phenomena (and which we think of as physical) are more often used metaphorically about abstract phenomena. Dictionaries tend to begin with the physical explanation, and sometimes leave it at that.

Meaning 2: PRESS OUT LIQUID

There were no examples concerning liquids among our twenty concordance lines, but instead a number of metaphorical uses that seem to fit best under this heading: (7)–(10), where the stuff that is pressed

1. gun to your head and say, "I'm gonna **squeeze** it five times, and if there's not a bullet
2. confronts Saddam Hussein, encircles Iraq and Kuwait, and begins to **squeeze**.
3. As an international embargo begins to **squeeze**, Saddam adds American diplomats to his collection of Western hostages
4. the view that the economic embargo, if it could ever **squeeze** Saddam sufficiently to cause his unilateral withdrawal from Kuwait, would
5. their best tactic is to **squeeze** Saddam between rebellious Kurds to the north and hostile Shi'ites
6. Part of the solution: **squeeze** the interval between final editing and distribution of the magazine
7. From so little they glean so much: **squeeze** the last ounce of joy from a flower with no petals
8. He did manage to **squeeze** out of Israel an agreement that might finesse the problem of
9. the contender who reminds no one of a President, might **squeeze** victory out of the state's mercurial mood.
10. The struggle to **squeeze** more aid dollars out of a finite pool brings with it
11. (RUC) and British intelligence forces had too often managed to **squeeze** information out of its members.
12. second performance for all the healers and spiritualists who were unable to **squeeze** into the auditorium for the first one.
13. The time traveler would have to survive the crushing pressure inside a black hole and somehow **squeeze** through an opening smaller than a single atom.
14. They have found a way to **squeeze** up to 45 billion bits of data onto a square inch
15. Budget Director Richard Darman was able to **squeeze** under the Gramm-Rudman-Hollings deficit target of $64 billion but only by
16. How many hits can a movie mogul **squeeze** into a box office?
17. He managed to **squeeze** in concern for the middle class about as often as Bob
18. He will also try to **squeeze** in a drama workshop at Los Angeles' Mark Taper Forum
19. much time for relaxation, but he claims to be able to **squeeze** in trips to the opera and other cultural events.
20. Saddam's dictatorship can and will **squeeze** the civilian economy as hard as may be necessary to maintain

Figure 3.2 The first twenty concordance lines for *squeeze* (verb) in the Time Corpus, 1990s, ordered according to meanings given in the *Longman Dictionary of Contemporary English* [1978] (2003).

out is *joy, an agreement, victory* and *aid dollars*. Example (11) is related to these, although "to force someone to tell you something" has been given a place of its own under the phrase **squeeze sth out 3** in the dictionary (see below).

Meaning 3: SMALL SPACE

The examples in the dictionary are all about physical space, like our examples (12)–(14). In addition, however, we have examples (15)–(17), which are about various metaphorical spaces, and (18)–(19), which are

squeeze[2] *n* |C| **1 a (tight) squeeze** a situation in which there is only just enough room for things or people to fit somewhere: *It will be a squeeze with six people in the car.* **2** an act of pressing something firmly with your fingers: *Marty gave her hand a little squeeze.* **3 squeeze of lemon/lime etc.** a small amount of juice obtained by squeezing a fruit **4** a situation in which wages, prices, borrowing money etc. are strictly controlled or reduced: |**+on**| *cuts due to the squeeze on public sector spending | a credit squeeze | All manufacturers are feeling the squeeze (=noticing the effects of a difficult financial situation.* **5 put the squeeze on sb** *informal* to try to persuade someone to do something **6 sb's (main) squeeze** *especially AmE* someone's BOYFRIEND or GIRLFRIEND

Figure 3.3 Dictionary entry for the noun *squeeze*.
Source: Longman Dictionary of Contemporary English [1978] (2003: 1607)

about fitting something into a tight schedule. The schedule meaning is given separately in the dictionary under the phrase **squeeze sb/sth in**.

The dictionary categories 4 and 5 do not occur in our small sample, but example (20) belongs to Meaning 6: LIMIT MONEY. Note that this mini-study was made on just the infinitive form; similar studies of the other verb forms indicated that different meanings dominate different tenses, so that there is a difference between the progressive *squeezing* and the simple past *squeezed*. For lack of space we will not go into that here. Instead we will turn to the meanings given for the noun *squeeze* in the dictionary, which are in part parallel to the verb meanings. The dictionary entry is given in Figure 3.3.

The twenty first concordance lines from the corpus have been ordered according to these meanings in Figure 3.4.

The results in Figure 3.4 are a bit surprising at first. There were no examples of tight physical squeezes with just enough room, and no physical squeezing of things (but in example (21), not on the list, there was one, referring to the handling of a gun: *those rapid-fire guns that require a single* ***squeeze*** *of the trigger for every round discharged*). Examples (1)–(3), however, might be seen as metaphorical extensions of physical squeezing. There was no squeezing of fruit either, and no instances of the phrases **put the squeeze on sb** or **sb's main squeeze**. Instead, all the remaining seventeen examples fit more or less well under the dictionary's Meaning 4, a situation in which wages, prices, borrowing money etc. are strictly controlled or reduced. These examples indicate that the noun *squeeze* has become part of a number of fixed compounds, e.g. *money squeeze, budget squeeze, financial squeeze* and *energy squeeze*. In the next chapter we will return to the phenomenon of words tending to co-occur or collocate frequently with certain other words.

The fact that some of the meanings from the dictionary did not turn up in our sample, and that some other meanings were heavily over-represented, can be explained by two circumstances. First of all, the

1. To overcome Shamir's qualms, Bush and Gorbachev staged a diplomatic **squeeze** play.
2. Arafat is also caught in a political **squeeze**.
3. of the railroads, such schemes were unlikely to break Moscow's **squeeze**.
4. Or if he can, you get caught in a "short **squeeze**," in which the stock gets bid up to even
5. but they did just that when a money **squeeze** threatened to shut down twelve of the city's 25 branches
6. other regimes also confront a money **squeeze** as Soviet funds dry up.
7. A cash **squeeze** was in fact one element in the pressure that Washington put
8. Post-secondary institutions are feeling both an economic and a demographic **squeeze**.
9. While the **squeeze** has so far been greatest in New England and neighboring states
10. tighter capital is beginning to put the **squeeze** even on healthy industries!
11. This **squeeze** on families bodes ill for children.
12. is battling its most awesome and implacable enemy: the defense budget **squeeze**.
13. But the budget **squeeze** has sparked a debate about whether the U.S. can afford three
14. Democrats feared that the budget **squeeze** on other domestic programs, already harsh, would be still
15. but a NASA budget **squeeze** killed the project.
16. Crimes Statistics Act is already caught up in the Government's financial **squeeze**.
17. In the U.S. the scientific community is beset by a budget **squeeze** and bureaucratic demands, internal squabbling, harassment by activists
18. devastated by corruption and the financial **squeeze** applied by the U.S. during the final two years of Noriega
19. While the energy **squeeze** is far less severe than the shocks of the '70s
20. Symptoms of an energy **squeeze** are breaking out all over.

Figure 3.4 The first twenty concordance lines for *squeeze* (noun) in the Time Corpus, 1990s, ordered according to meanings given in the *Longman Dictionary of Contemporary English* [1978] (2003).

Time Corpus is biased towards some registers, such as politics and arts, but contains less of others, such as recipes (squeezing of lemons), and virtually no authentic spoken everyday conversations. This shows that a multipurpose learner's dictionary has to be based on a very varied corpus. Second, twenty concordance lines are far too few. How many concordance lines a lexicographer needs to write an entry for a word depends on how many different meanings the word in question has, but twenty is seldom enough.

In searches for lexical items quite often some of the hits consist of names of people, organisations etc. Such hits should normally be discounted, so that if you intend to investigate 100 examples of a particular word and it turns out that 5 are proper nouns (i.e. names), these 5 tokens should be excluded and another 5 relevant ones added (using the same sampling method) so that the total number of relevant tokens is 100. The easiest way to avoid extra work is to make a quick pilot search first and check the rough proportion of names in your data, and then make your search big enough so that you will have enough examples even after the names have been deleted. Out of the first 100 hits in a search

for the colour noun *green* in the BNC, for instance, 11 referred to people by the name of *Green* and another 4 referred to places (*Blade Bone Green*, *Dock Green*, *Abbey Green* and *Juniper Green*). To be certain to retrieve the right number of colour nouns, therefore, about 25% should be added to the figure in this case.

However, sometimes it can be interesting to see how a particular word has been used to name companies, rock bands and so on. Such an example is (1) from the *squeeze* search.

(1) [. . .] by presenting such offbeat performers as Sinead O'Connor, Neil Young and **Squeeze**. (Time Corpus, 1990s)

Squeeze here refers to a 1970s British New Wave band. In all, the *squeeze* exercise has shown that the meaning of a word can only be ascertained by looking at the contexts in which it occurs.

3.4 Semantic preference, semantic prosody and evaluation

We saw in the last section that the meaning of a word depends on its context, so that the meaning of a verb like *squeeze* depends on both its subject and its object. It may also be the case that words acquire a sort of hidden meaning, which was not there from the beginning, from the words they frequently occur together with. This has been called 'semantic prosody'. Semantic prosody is part of a system of lexical relations suggested by Sinclair (1998) and explained by Stubbs (2009: 22) in the following way (the wording has been slightly simplified):

> COLLOCATION is the relation between a word and individual word-forms which co-occur frequently with it.
> COLLIGATION is the relation between a word and grammatical categories which co-occur frequently with it.
> SEMANTIC PREFERENCE is the relation between a word and semantically related words in a lexical field.
> SEMANTIC PROSODY is the discourse function of the word: it describes the speaker's communicative purpose.

We will return to colligation and collocation in the next chapter and will not discuss semantic preference further, since it is a concept that has not been widely used in corpus linguistics. Semantic prosody, however, is frequently referred to. A famous example used by Sinclair is the phrasal verb *set in*. Basically, it means just 'begin', but Sinclair claims that looking in a corpus one will find that most of the things that *set in* are negative. Figure 3.5 shows the first ten concordance lines with the relevant sense of *set in* in COCA for the 2000s.