M.A.K. Halliday, Wolfgang Teubert, Colin Yallop and Anna Čermáková

Lexicology and Corpus Linguistics

AN INTRODUCTION

Lexicology and Corpus Linguistics

Open Linguistics Series

Series Editor
Robin Fawcett, University of Wales, Cardiff

This series is 'open' in two senses. First, it provides a forum for works associated with any school of linguistics or with none. Most practising linguists have long since outgrown the unhealthy assumption that theorising about language should be left to those working in the generativist–formalist paradigm. Today large and increasing numbers of scholars are seeking to understand the nature of language by exploring one or other of various cognitive models of language, or in terms of the communicative use of language, or both. This series is playing a valuable part in re-establishing the traditional 'openness' of the study of language. The series includes many studies that are in, or on the borders of, various functional theories of language, and especially (because it has been the most widely used of these) Systemic Functional Linguistics. The general trend of the series has been towards a functional view of language, but this simply reflects the works that have been offered to date. The series continues to be open to all approaches, including works in the generativist–formalist tradition.

The second way in which the series is 'open' is that it encourages studies that open out 'core' linguistics in various ways: to encompass discourse and the description of natural texts; to explore the relationships between linguistics and its neighbouring disciplines – psychology, sociology, philosophy, cultural and literary studies – and to apply it in fields such as education, language pathology and law.

Recent titles in this series

Analysing Academic Writing, Louise J. Ravelli and Robert A. Ellis (eds)
Brain, Mind and the Signifying Body, Paul J. Thibault
Classroom Discourse Analysis, Frances Christie
Construing Experience through Meaning: A Language-based Approach to Cognition, M. A. K. Halliday and Christian M. I. M. Matthiessen
Culturally Speaking: Managing Rapport through Talk across Cultures, Helen Spencer-Oatey (ed.)
Development of Language, Geoff Williams and Annabelle Lukin (eds)
Educating Eve: The 'Language Instinct' Debate, Geoffrey Sampson
Empirical Linguistics, Geoffrey Sampson
Genre and Institutions: Social Processes in the Workplace and School, Frances Christie and J. R. Martin (eds)
The Intonation Systems of English, Paul Tench
Language, Education and Discourse, Joseph A. Foley (ed.)
Language Policy in Britain and France: The Processes of Policy, Dennis Ager
Language Relations across Bering Strait: Reappraising the Archaeological and Linguistic Evidence, Michael Fortescue
Learning through Language in Early Childhood, Clare Painter
Multimodal Discourse Analysis, Kay L. O'Halloran (ed.)
Pedagogy and the Shaping of Consciousness: Linguistic and Social Processes, Frances Christie (ed.)
Register Analysis: Theory and Practice, Mohsen Ghadessy (ed.)
Relations and Functions within and around Language, Peter H. Fries, Michael Cummings, David Lockwood and William Spruiell (eds)
Researching Language in Schools and Communities: Functional Linguistic Perspectives, Len Unsworth (ed.)
Summary Justice: Judges Address Juries, Paul Robertshaw
Syntactic Analysis and Description: A Constructional Approach, David G. Lockwood
Thematic Developments in English Texts, Mohsen Ghadessy (ed.)
Ways of Saying: Ways of Meaning. Selected Papers of Ruqaiya Hasan, Carmen Cloran, David Butt and Geoffrey Williams (eds)
Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology, Howard Jackson and Etienne Zé Amvela
Working with Discourse: Meaning beyond the Clause, J. R. Martin and David Rose

Lexicology and Corpus Linguistics

An Introduction

M. A. K. Halliday, Wolfgang Teubert, Colin Yallop
and Anna Čermáková

continuum
LONDON • NEW YORK

Continuum
The Tower Building
11 York Road
London SE1 7NX

15 East 26th Street
New York
NY 10010

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN: 0-8264-4861-5 (hardback)
ISBN: 0-8264-4862-3 (paperback)

Library of Congress Cataloguing-in-Publication Data
A catalogue record for this book is available from the Library of Congress.

Typeset by YHT Ltd, London
Printed and bound in Great Britain by MPG Books Ltd, Bodmin, Cornwall

Contents

This page intentionally left blank

1 Lexicology

M. A. K. Halliday

1.1 What is a word?

To many people the most obvious feature of a language is that it consists of words. If we write English, we recognise words on the page – they have a space on either side; we learn to spell them, play games with them like Scrabble, and look them up in dictionaries. It ought not to be difficult to know what a word is and how to describe it.

Yet when we look a little more closely, a word turns out to be far from the simple and obvious matter we imagine it to be. Even if we are literate English-speaking adults, we are often unsure where a word begins and ends. Is *English-speaking* one word or two? How do we decide about sequences like *lunchtime (lunch-time, lunch time), dinner-time, breakfast time*? How many words in *isn't, pick-me-up, CD*? Children who cannot yet read have little awareness of word boundaries, and often learn about them through word games, like 'I'm thinking of a word that rhymes with . . .'.

Even more problematic is whether two forms are, or are not, instances of the same word. Presumably if they sound alike but are spelled differently, like *horse* and *hoarse*, they are two different words. But how about pairs such as:

like 'similar to'	*like* 'be fond of'
part 'portion'	*part* 'to separate'
shape 'the outline of'	*shape* 'to mould'
content 'happy'	*content* 'that which is contained'

– not to mention *shape* as the old name for a kind of solid custard pudding?

We know that there is no single right answer to these questions, because different dictionaries take different decisions about what to do with them.

Then, what about variants like *take, takes, took, taking, taken*: are these five different words, or is there just one word *take* with many forms? Or

go, goes, went, going, gone? Are *book* and *books*, *friend* and *friendly* one word or two? Are *big, bigger, biggest* three forms of a single word *big*? If so, what about *good, better, best*? Or *four* and *fourth*, *three* and *third*, *two* and *second*?

All these are problems within English, a language where the words are fairly clearly bounded. In Chinese it is much harder, because words are not marked off in writing; Chinese characters stand for **morphemes**, which are components of words. (For example, if English was written with Chinese characters then a word like *freedom* would be written with two characters, one for *free* and one for *dom*.) The Chinese are very conscious of morphemes, even before they are literate, because each one is pronounced as one syllable and hardly ever varies; but they have much less intuition about what a word is. Many other writing systems also, such as Japanese, Thai, Arabic and Hindi, give no very consistent indication of word boundaries. When Ancient Greek was first written down, all the words were joined together without any spaces, and it was a few centuries before the word emerged as a clearly distinct unit.

So writing systems do not always identify words: partly because there are different kinds of writing system, but partly also because the languages themselves are different. There is no universal entity, found in every language, that we can equate with what in English is called a 'word'. And in unwritten languages the 'word' can be a very elusive thing.

Nevertheless there is a general concept underlying all this diversity; that is the **lexical item**. Every language has a **vocabulary**, or 'lexicon', which forms one part of its grammar – or, to use a more accurate term, one part of its **lexicogrammar**. The lexicogrammar of a language consists of a vast network of choices, through which the language construes its meanings: like the choices, in English, between 'positive' and 'negative', or 'singular' and 'plural', or 'past', 'present' and 'future'; or between 'always', 'sometimes' and 'never', or 'on top of' and 'underneath'; or between 'hot' and 'cold', or 'rain', 'snow' and 'hail', or 'walk' and 'run'. Some of these choices are very general, applying to almost everything we say: we always have to choose between positive and negative whenever we make a proposition or a proposal (*it's raining, it isn't raining*; *run! don't run!*). Others are very specific, belonging to just one domain of meaning; these arise only when we are concerned with that particular domain. The choice between rain and snow, for example, arises only if we are talking about the weather. Choices of this second kind are expressed as lexical items: e.g. *hot/cold*; *rain/snow/hail*; *walk/run*.

If we are using the term 'word' to mean a unit of the written lan-

guage, i.e. 'that which (in English) is written between two spaces', then ultimately all these choices are expressed as strings of words, or **wordings**, as in *it always snows on top of the mountain.* But teachers of English have customarily distinguished between **content words**, like *snow* and *mountain*, and **function words**, like *it* and *on* and *of* and *the*; and it is the notion of a content word that corresponds to our lexical item. Lexicology is the study of content words, or lexical items.

The example sentence in the last paragraph shows that the line between content words and function words is not a sharp one: rather, the two form a continuum or cline, and words like *always* and *top* lie somewhere along the middle of the cline. Thus there is no exact point where the lexicologist stops and the grammarian takes over; each one can readily enter into the territory of the other. So dictionaries traditionally deal with words like *the* and *and*, even though there is hardly anything to say about them in strictly lexicological terms, while grammars go on classifying words into smaller and smaller classes as far as they can go – again, with always diminishing returns.

This gives us yet a third sense of the term 'word', namely the element that is assigned to a **word class** ('part of speech') by the grammar. So the reason 'word' turns out to be such a complicated notion, even in English, is that we are trying to define it simultaneously in three different ways. For ordinary everyday discussion this does not matter; the three concepts do not in fact coincide, but they are near enough for most purposes. In studying language systematically, however, we do need to recognise the underlying principles, and keep these three senses apart. The reason our lexicogrammar is divided into 'grammar' and 'lexicology' (as in traditional foreign language textbooks, which had their section of the grammar and then a vocabulary added separately at the end) is because we need different models – different theories and techniques – for investigating these two kinds of phenomena, lexical items on the one hand and grammatical categories on the other. This is why **lexicology** forms a different sub-discipline within linguistics.

1.2 Methods in lexicology: the dictionary

There are two principal methods for describing words (now in our sense of **lexical items**), though the two can also be combined in various ways. One method is by writing a **dictionary**; the other is by writing a **thesaurus.**

The difference between a dictionary and a thesaurus is this. In a thesaurus, words that are similar in meaning are grouped together: so,

for example, all words that are species of fish, or all words for the emotions, or all the words to do with building a house. In a dictionary, on the other hand, words are arranged simply where you can find them (in 'alphabetical order' in English); so the place where a word occurs tells you nothing about what it means. In the dictionary we find a sequence such as *gnome, gnu, go, goad*; and *parrot* is in between *parlour* and *parsley*.

In a dictionary, therefore, each entry stands by itself as an independent piece of work. There may be some cross-referencing to save repetition; but it plays only a relatively small part. Here are some typical entries from a fairly detailed dictionary of English, the two-volume *New Shorter Oxford English Dictionary*, 1993. (The full entries are much longer and omissions are indicated by ... in parentheses; the abridged entries given here serve to show the general structure and to illustrate the kind of detail included.)

bear /bɛː/ *n.* [OE *bera* = MDu. *bere* (Du. *beer*), OHG *bero* (G *Bär*), f. Wgmc: rel. to ON *bjǫrn*.]

1. Any of several large heavily-built mammals constituting the family Ursidae (order Carnivora), with thick fur and a plantigrade gait. OE.

b With specifying wd: an animal resembling or (fancifully) likened to a bear. E17.

2. *Astron. the Bear* (more fully *the Great Bear*) = URSA *Major*; *the Lesser* or *Little Bear* = URSA *Minor*. LME.

3. *fig.* A rough, unmannerly or uncouth person. L16.

(...)

3. LD MACAULAY This great soldier ... was no better than a Low Dutch bear.

(...)

Other phrases: **like a bear with a sore head** *colloq.* angry, ill-tempered.

(...)

bear /bɛː/ *v.* Pa. t. **bore** /bɔː/, (*arch.*) **bare** /bɛː/. Pa.pple & ppl a. **borne** /bɔːn/, BORN. See also YBORN. [OE *beran* = OS, OHG *beran*, ON *bera*, Goth. *bairan* f. Gmc f. IE base also of Skt *bharati*, Armenian *berem*, Gk *pherein*, L *ferre*.]

I *v.t.* Carry, hold, possess.

1 Carry (esp. something weighty), transport, bring or take by carrying; *fig.* have, possess. Now *literary* or *formal*. OE.

(...)

2 Carry about with or upon one, esp. visibly; show, display; be known or recognized by (a name, device, etc.); have (a character, reputation, value, etc.) attached to or associated with one. OE.

(...)

1 CHAUCER On his bak he bar ... Anchises.

R. HOLINSHED This pope Leo ... bare but seauen and thirtie yeeres of age.

SHAKES. *Macb.* I bear a charmed life, which must not yield To one of woman born.

E. WAUGH Music was borne in from the next room.

(...)

2 SHAKES. *Wint. T.* If I Had servants true about me that bare eyes To see alike mine honour as their profits.

STEELE Falshood ... shall hereafter bear a blacker Aspect.

W. H. PRESCOTT Four beautiful girls, bearing the names of the principal goddesses.

A. P. STANLEY The staff like that still borne by Arab chiefs.

(...)

Phrases (...)

bear fruit *fig.* yield results, be productive. (...)

bear in mind not forget, keep in one's thoughts. (...)

cut /kʌt/ *v.* Infl. **-tt-**. Pa. t. & pple **cut.** See also CUT, CUTTED *ppl adjs.* ME [Rel. to Norw. *kutte,* Icel. *kuta* cut with a little knife, *kuti* little blunt knife. Prob. already in OE.]

I *v.t.* Penetrate or wound with a sharp-edged thing; make an incision in. ME.

b *fig.* Wound the feelings of (a person), hurt deeply.

(...)

1 N. MOSLEY The edge of the pipe cut his mouth, which bled. *fig.*: ADDISON Tormenting thought! it cuts into my soul.

b F. BURNEY He says something so painful that it cuts us to the soul.

(...)

Phrases: (...)

cut both ways have a good and bad effect; (of an argument) support both sides.

cut corners *fig.* scamp work, do nothing inessential. (...)

These entries are organised as follows:

1. the headword or **lemma**, often in bold or some other special font;
2. its pronunciation, in some form of alphabetic notation;
3. its word class ('part of speech');
4. its etymology (historical origin and derivation);
5. its definition;
6. citations (examples of its use).

Most dictionaries follow this general structure, but variations are of course found. For example, etymological information may come at the end of the entry rather than near the beginning. Let us look more closely at each item in turn.

1. The **lemma** is the base form under which the word is entered and assigned its place: typically, the 'stem', or simplest form (singular noun, present/infinitive verb, etc.). Other forms may not be entered if they are predictable (such as the plural *bears*, not given here); but the irregular past forms of the verbs are given (irregular in the sense that they do not follow the default pattern of adding *-ed*) and there is also an indication under *cut* that the *t* must be doubled in the spelling of inflected forms like *cutting*. An irregular form may appear as a separate lemma, with cross reference. This dictionary has such an entry for **borne** *v.* pa. pple & ppl a. of BEAR *v.*, indicating that *borne* is the past participle and participial adjective of the verb **bear**. In a language such as Russian, where the stem form of a word typically does not occur alone, a particular variant is chosen as lemma: nominative singular for nouns, infinitive for verbs, etc.

2. In most large and recent dictionaries, the pronunciation is indicated, as here, by the International Phonetic Alphabet in a broad, phonemic transcription. Some older dictionaries use a modified alphabet with a keyword system, e.g. *i* as in 'machine', i as in 'hit', u as in 'hut'; and some dictionaries, especially those intended for use by children, simply use informal respellings, e.g. **emphasis** (EM-fa-sis) or **empirical** (em-PIR-ik-uhl).

3. The word class will be one of the primary word classes (in English, usually verb, noun, adjective, adverb, pronoun, preposition, conjunction, determiner/article). To this class specification may be added some indication of a subclass, for example count or mass noun, intransitive or transitive verb. The senses of the verbs illustrated here, for example, are identified as transitive verbs (*v.t.*). Some dictionaries, especially those compiled for learners of English, give more detailed word class information, showing for example the functional relations into which verbs can enter.

4. The etymology may include, as here, not only the earliest known form and the language in which this occurs (e.g. Old English, OE for short) but also cognate forms in other languages. Some dictionaries may also include a suggested 'proto-' form, a form not found anywhere but reconstructed by the methods of historical linguistics; proto-forms are conventionally marked with an asterisk. The various forms of the noun *bear*, for example, suggest an ancestral form **ber-*, pre-dating the differentiation of languages such as Old English and Old High German. For many words, little or nothing is known of their history, and a common entry is 'origin unknown' (or the more traditional 'etym. dub.'!). This edition of the *Oxford* also indicates the first recorded use against each (sub)definition: OE means the word (or an earlier form of

it) is attested in this sense in Old English texts, E17 means this sense is first recorded in the early seventeenth century, L16 that the sense is first recorded in the late sixteenth century.

5. The definition takes one or both of two forms: description and synonymy. The description may obviously need to include words that are 'harder' (less frequently used) than the lemmatised word. Some dictionaries, such as the *Longman Dictionary of Contemporary English* (first published in 1978), limit the vocabulary that they use in their descriptions. With synonymy, a word or little set of words of similar meaning is brought in, often giving slightly more specific senses. All definition is ultimately circular; but compilers try to avoid very small circles, such as defining *sad* as *sorrowful*, and then *sorrowful* as *sad*.

6. Citations, here grouped together under numbers referring back to definitions or senses, show how the word is used in context. They may illustrate a typical usage, or use in well-known literary texts, or the earliest recorded instances of the word. There may also be various 'fixed expressions' (idioms and cliches) and what the *Oxford* here calls 'phrases', where the expression functions like a single, composite lexical item (e.g. *bear fruit, bear in mind*).

The dictionary will usually use a number of abbreviations to indicate special features or special contexts, for example *fig.* ('figurative'), *Astron.* ('Astronomy') and so on. With a common word such as *bear* or *cut* there are likely to be subdivisions within the entry, corresponding to different meanings of the word.

Compound words, like *cutthroat* (as in *cutthroat competition*), and derivatives, like *cutting* (from a plant) or *uncut*, are often entered under the same lemma; in that case, compounds will appear under the first word (*cutthroat* under *cut*, *haircut* under *hair*), derivatives under the stem (both *cutting* and *uncut* under *cut*). But dictionaries adopt varying practices. In some dictionaries, compounds are given separate lemmata; and sometimes a derivational affix is used as lemma and derivatives grouped under that (for example *antibody, anticlimax, antidote*, etc. all under *anti-*).

1.3 Methods in lexicology: the thesaurus

In a thesaurus, by contrast, there is no separate entry for each word. The word occurs simply as part of a list; and it is the place of a word in the whole construction of the book that tells you what it means.

Thus if we look for *cut* in Roget's *Thesaurus of English Words and Phrases* we will find it (among other places) in the middle of a paragraph as follows:

v. cultivate; till (the soil); farm, garden; sow, plant; reap, mow, cut; manure, dress the ground, dig, delve, dibble, hoe, plough, plow, harrow, rake, weed, lap and top, force, transplant, thin out, bed out, prune, graft.

This may not seem to have very much organisation in it; but it is actually the final layer in a comprehensive **lexical taxonomy.**

A lexical taxonomy is an organisation of words into classes and sub-classes and sub-sub-classes (etc.); not on the basis of form but on the basis of meaning (that is, not grammatical classes but semantic classes). The principal semantic relationship involved is that of **hyponymy** (*x* is a hyponym of *y* means *x* 'is a kind of' *y*, e.g. *melon* is a hyponym of *fruit*). There is also another relationship, that of **meronymy** ('is a part of'), which may be used for classification. Such taxonomies are familiar in the language of everyday life, where they tend to be somewhat irregular and variable according to who is using them. Many of us might organise our shopping around taxonomies such as the one for *fruit* shown in Figure 1, perhaps according to how things are arranged in our local shop or market.

The taxonomies of living things on which biological science was founded in the eighteenth century are systematic variants of the same principle: the five kinds (classes) of *vertebrates* are *fishes, amphibia, reptiles, birds* and *mammals*; the eight kinds (orders) of *mammals* are *pachyderms, carnivores, cetaceans* ... Here each rank in the taxonomy is

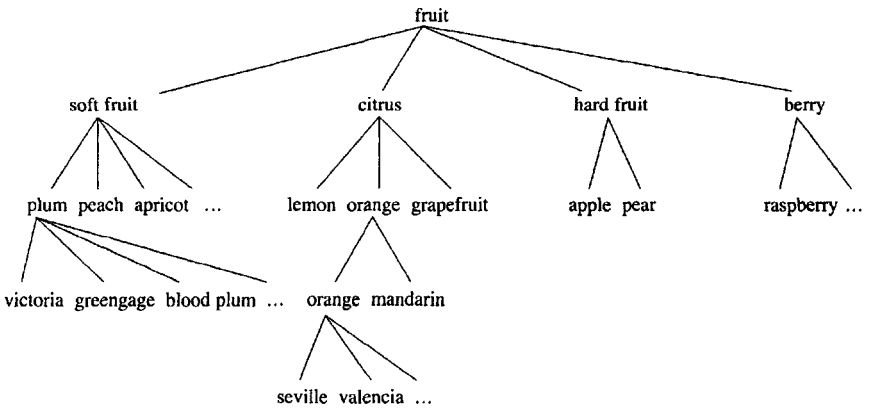

Figure 1 A partial taxonomy for *fruit*

given a special name: *kingdom, phylum, class, order, family, genus, species, variety.*

A thesaurus takes all the lexical items that it contains and arranges them in a single comprehensive taxonomy. Roget's original *Thesaurus*, compiled over four decades from 1810 to 1850, was in fact conceived on the analogy of these scientific taxonomies; in his Introduction, Roget acknowledged his debt to Bishop John Wilkins, whose *Essay towards a Real Character and an Universal Language*, published in 1665, had presented an artificial language for organising the whole of knowledge into an overarching taxonomic framework. Roget's taxonomy started with six primary classes: I, Abstract relations; II, Space; III, Matter; IV, Intellect; V, Volition; VI, Affections. Here is the path leading to one of the entries for the word *cut*. Starting from *Matter*, the path leads to *Organic* Matter, then to *Vitality* and *Special* Vitality (as opposed to Vitality in general); from there to *Agriculture*, then via the verb *cultivate* to the small sub-paragraph consisting of just the three words *reap, mow, cut*, which has no separate heading of its own. Thus there are eight ranks in the taxonomy, the last or **terminal** one being that of the lexical item itself. This path can be traced in the schematic representation shown in Figure 2.

Figure 2 is not how *cut* appears in the thesaurus of course; but we can reconstruct the path from the way the thesaurus is organised into chapters, sections and paragraphs. This particular example relates, obviously, only to one particular meaning of the word *cut*, namely cutting in the context of gardening and farming. But there is no limit on how many times the same word can occur; *cut* will be found in twenty-six different locations, each corresponding to a different context of use. There is an alphabetical index at the end of the book to show where each word can be found.

Thus a thesaurus presents information about words in a very different way from a dictionary. But although it does not give definitions, it provides other evidence for finding out the meaning of an unknown word. Suppose for example that you do not know the meaning of the word *cicuration*. You find that it occurs in a proportional set, as follows:

animal : vegetable
:: zoology : botany
:: cicuration : agriculture

The proportion shows that *cicuration* means 'animal husbandry'.

We cannot always construct such proportionalities. But the fact that a word is entered as one among a small set of related words also tells us a lot about what it means. Such a set of words may be closely

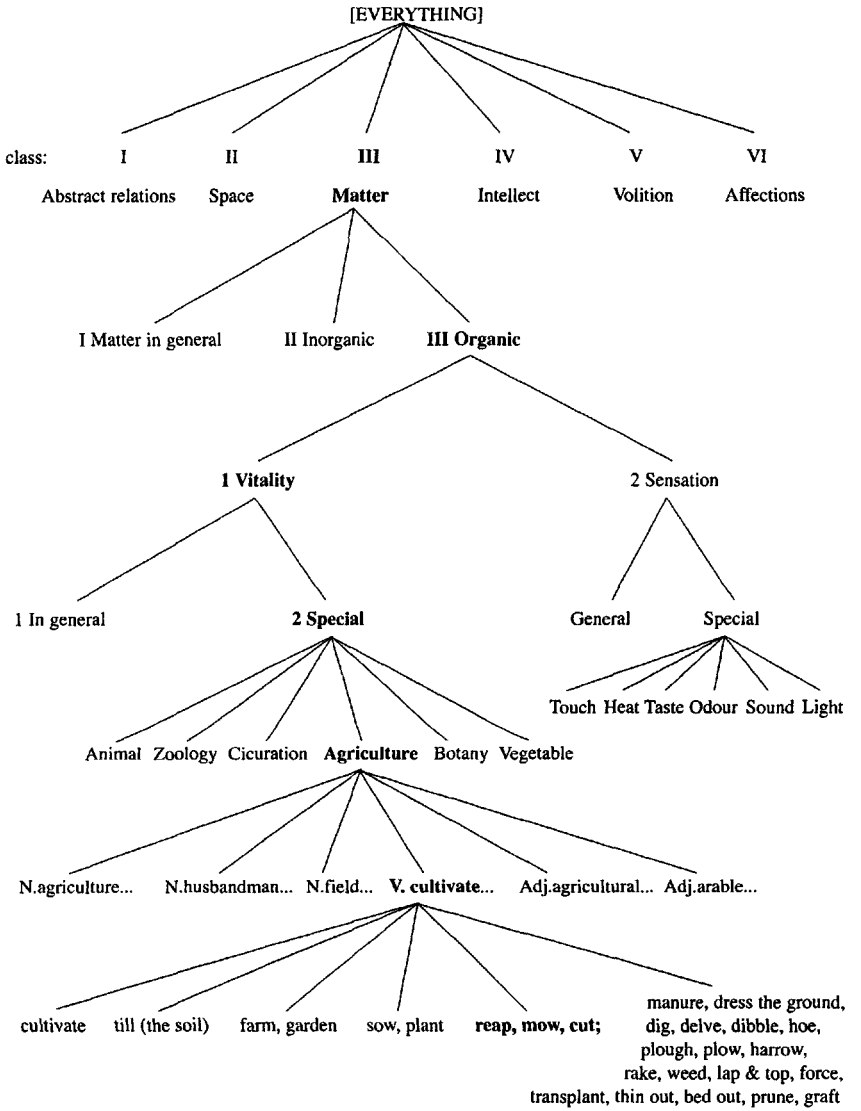

Figure 2 Schematic representation of a thesaurus entry. (Based on Roget's *Thesaurus of English Words and Phrases*, 1936)

synonymous, like *reap, mow, cut* – although not necessarily so; rather, they are **co-hyponyms**, or else **co-meronyms**, of some superordinate term. Thus *reap, mow, cut* (*cut* in this special sense) are co-hyponyms of *cut* in its more general sense; and the items in the next sub-paragraph (*manure, dress the ground ... prune, graft*) all represent stages in the cultivation process – that is, they are co-meronyms of *cultivate*. When we use a thesaurus to search for synonyms, as an aid to writing for example, what we are really looking for are words that share a common privilege of occurrence; they do not ordinarily 'mean the same thing', but they share the same address, as it were, within our overall semantic space.

Another way of thinking about this shared privilege of occurrence that unites the words in one paragraph of the thesaurus is in terms of **collocation**. Collocation is the tendency of words to keep company with each other: like *fork* goes with *knife, lend* goes with *money, theatre* goes with *play*. Of course, if words do regularly collocate in this way, we shall expect to find some semantic relationship among them; but this may be quite complex and indirect. Collocation is a purely lexical relationship; that is, it is an association between one word and another, irrespective of what they mean. It can be defined quantitatively as the degree to which the probability of a word *y* occurring is increased by the presence of another word *x*. If you meet *injure*, you may expect to find *pain* somewhere around: given the presence of the word *injure*, the probability of the word *pain* occurring becomes higher than that determined by its overall frequency in the English language as a whole. The words that are grouped into the same paragraph in a thesaurus are typically words that have a strong collocational bond: either with each other or, more powerfully, each of them with some third party, some common associate that forms a network with them all.

1.4 History of lexicology: India, China, the Islamic world, Europe

When did lexicology begin? Like all systematic study of the **formal** patterns of language, lexicology depends on language being written down. Many oral cultures have developed highly elaborated theories of speech function and rhetoric; but it is only after writing evolves that attention comes to be focused on grammar and vocabulary. This typically began as a way of keeping alive ancient texts whose meanings were beginning to be lost as the language continued to change. In India as early as the third to second century BC, glossaries were drawn up to explain the difficult words in the Vedas, which by that time were already a thousand years old. These glossaries gradually evolved into what we would recognise today as dictionaries. In the seventh century

AD, the scholar Amera Sinha prepared a Sanskrit dictionary, the *Amera Kosha.* More than ten centuries later this was still in use – it was translated into English by Colebrooke, and Colebrooke's translation, published in Serampur in 1808, is acknowledged by Roget as one source of ideas for his *Thesaurus.* Hamacandra's great dictionaries of Sanskrit and of Prakrit, the *Abhidhana Kintamani* and the *Desinamamala,* date from the twelfth century. By this time Indian scholarship in grammar and phonology had reached a high degree of sophistication, and dictionary-making took its place as part of the systematic description of language.

In China the earliest extant lexicological work is in fact a thesaurus, the *Er Ya* 'Treasury of Fine Words'. Compiled in this form in the third century BC, it is a list of about 3,500 words found in ancient texts, arranged under nineteen headings: the first three sections contain words of a general nature – nouns, verbs and figurative expressions; the remaining sixteen being topical groupings, headed Kin, Buildings, Implements, Music, Sky (i.e. calendar and climate), Land, Hills, Mountains, Water (rivers and lakes), Plants, Trees, Insects and Reptiles, Fishes, Birds, Wild Animals and Domestic Animals. Each word is glossed, by a synonym or superordinate term, or else briefly defined. The Chinese paid little attention to grammar: since Chinese words are invariant, the question of why words change in form, which was what led the Indians, Greeks and Arabs to study grammar, simply did not arise. But their study of vocabulary developed in three directions: (1) recording dialect words, as in the *Fang Yan,* by Yang Xiong, in the first century BC; (2) investigating the origin of written characters, in *Shuo Wen Jie Zi,* by Xu Shen, in the first century AD; and (3) describing the sounds of words, classifying them according to rhyme, notably in the *Qie Yun* (AD 600) and *Tang Yun* (AD 750). By the time of the Ming and Qing dynasties, large-scale dictionaries and encyclopaedias were being compiled: notably the *Yongle Encyclopaedia* (1403–9) in 10,000 volumes, few of which however survive; and the *Kangxi Dictionary* (1716), containing some 50,000 characters together with their pronunciation and definition.

Both the Arabic and Hebrew traditions are rich in grammatical scholarship, and the earliest Arab grammarian, al-Khalil ibn Ahmed (died AD 791), is known to have begun work on an Arabic dictionary, using a phonological principle for ordering the words. But the leading lexicographers in the Islamic world were the Persians. The first dictionary of Farsi-dari, the Persian literary language, written by Abu-Hafs Soghdi in the ninth to tenth centuries, is now lost; but the eleventh-century *Lughat-e Fars* (Farsi dictionary), by Asadi Tusi, is extant. Persian

scholars also produced bilingual dictionaries, Persian–Arabic (*Muqaddimat al-adab* 'Literary Expositor', by an eleventh-century scholar from Khwarezm, Abul-Qasim Mohammad al-Zamakhshari) and, from the fifteenth century onwards, Persian–Turkish.

It is known that the Egyptians produced thesaurus-like topically arranged wordlists from as early as 1750 BC, although none has survived. In Greece, as in India, the earliest studies of words were glossaries on the ancient texts – Homeric texts, in the case of Greece. Apollonius, an Alexandrian grammarian of the first century BC, compiled a Homeric lexicon, but both this and the later glossaries by Hesychius are lost. Perhaps the greatest work of the Byzantine period was the *Suda*, a tenth-century etymological and explanatory dictionary of around 30,000 entries from literary works in Ancient, Hellenistic and Byzantine Greek and in Latin.

The development of dictionaries in the modern European context was associated with the spread of education and the promotion of emerging national literary languages. From about 1450 onwards bilingual dictionaries were being produced for use in schools, at first for learning Latin (Latin–German, Latin–English, etc.), but soon afterwards also for the modern languages of Europe. Many of the nation states of southern and eastern Europe then set up national academies, and these were responsible for establishing norms for the definition and usage of words: for example the Italian *Vocabulario degli Academici della Crusca*, 1612; the *Dictionnaire de l'Académie française*, 1694 (the lexicographer Furetière was expelled from the Academy because he published his own dictionary, the *Universal Dictionary Containing All French Words*, in 1690 before the official one had appeared); the dictionary of the Spanish Academy in 1726–39, and that of the Russian Academy in 1789–94. By the nineteenth century the great publishing houses were bringing out extended series of lexicological works: notably in France (Littré, *Dictionnaire de la Langue française*, in four volumes plus supplement, in 1863–78; and Larousse, *Grand Dictionnaire Universel du XIXe siècle*, an encyclopaedic dictionary in 15 volumes, 1865–76) and in Germany (Meyer's *Great Encyclopaedic Lexicon* in 46 plus 6 supplementary volumes, 1840–55). Each of these major works was followed by a large number of 'spinoff' publications of various kinds.

1.5 Evolution of the dictionary and the thesaurus in England

As an illustration of how twentieth-century dictionaries have evolved, we will take the example of English. But it is important to bear in mind that English dictionaries did not evolve in isolation from other

traditions; they were influenced from elsewhere in Europe and even from further afield. Lexicography in England began in the form of glossaries on 'difficult' words in manuscripts of Latin texts: at first these were given in Latin, using simpler words for the purpose, but by the seventh century they were appearing in English (e.g. in the Épinal manuscript preserved at a monastery in France). Next, such glosses were taken out and arranged in a list (a 'glossary'); and then various lists, especially of technical terms, for example in agriculture or in medicine, were collected together into a 'vocabulary'. In the eighth and ninth centuries compilers started arranging the words in alphabetical order. By the thirteenth century the term 'dictionary' had come into use; the collections of words were becoming considerably larger, and English–Latin lists began to be compiled. The *Promptorium Parvulorum sive Clericorum* 'Repository for Children and Clerics', by Geoffrey 'the Grammarian' of Norfolk, dated about 1440, contained some 12,000 words. It was during this century that printing was introduced into Europe; the *Promptorium* was printed in 1499, and from then on the scope and variety of published dictionaries grew rapidly. Sir Thomas Elyot's *Latin–English Dictionary* appeared in 1538; R. Howlet's *Abecedarium Anglico–Latinum* in 1552. Bilingual dictionaries of modern languages began with Palsgrave's English–French dictionary of 1530, and this was soon followed by dictionaries of English–Spanish and English–Italian. The arrangement of words by their strict alphabetical order had now become established practice; and lexicographers began introducing citations from literary works to illustrate usage in the foreign language.

The first monolingual English dictionary was published by Robert Cawdrey in 1604; this was *A Table Alphabeticall of Hard Usuall English Wordes*, which gave the spelling and meaning of about 2,500 terms. In 1616 John Bullokar's *An English Expositor* appeared, and in 1623 Henry Cockeram's *The English Dictionarie*. Cockeram's dictionary contained two parts: one of hard words, one of ordinary words, with words of each group being used to explain those of the other. The first dictionary which set out to include all words, and to define their meanings, was John Kersey's *A New English Dictionary* of 1702; shortly after this, in 1720, Nathan Bailey published his *Universal Etymological English Dictionary*, in which he added a new dimension to lexicography by including the history ('etymology') of each word. This work, along with other publications by Nathan Bailey, was the immediate precursor to Samuel Johnson's *Dictionary of the English Language*, which appeared in 1755. Dr Johnson's dictionary was a landmark not only in setting high professional standards in lexicography but also in establishing the role

of the lexicographer as an authority on the 'correct' spelling, pronunciation and definition of words.

This normative function of a dictionary was a distinctive feature of two major American lexicographers of the first half of the nineteenth century, Noah Webster and Joseph Worcester. Webster in particular, in *An American Dictionary of the English Language* published in 1828, sought to codify American English as a distinct tongue, marked out by its own orthographic conventions; the modifications of spelling which he introduced in his dictionary, while much less radical than his original proposals, became accepted as the American standard.

In nineteenth-century lexicology in England, four achievements stand out.

(1) One was Roget's *Thesaurus*, referred to earlier (1.3). Peter Mark Roget was a doctor who became a leading member of the Royal Society; his work of arranging the words and idiomatic phrases of the English language into one comprehensive semantic taxonomy occupied him for some forty years. As already noted, he was influenced both by his predecessors in the Royal Society of 150 years earlier, in their construction of an artificial language for scientific taxonomy, and by the Indian tradition of lexicology that he knew from Colebrooke's translation of Amera Sinha's seventh-century Sanskrit dictionary.

(2) Another was the *New English Dictionary on Historical Principles*, at first edited by James Murray and published in 12 volumes over the period 1884 to 1928 (by the Oxford University Press; hence its more familiar designation as 'Oxford English Dictionary' or *OED*). This dictionary incorporated both extensive textual citations, a practice established in Charles Richardson's (1837) *New Dictionary of the English Language*, and detailed historical information about each word, following the principle established by Jacob and Wilhelm Grimm in their large-scale historical dictionary of German (begun in 1852, although not finally compiled and published until 1960). The *OED* contains over 400,000 entries and a little under two million citations. Four supplementary volumes appeared between 1933 and 1986, and a revised edition of the entire dictionary was published in 1989 as *The Oxford English Dictionary*, second edition, in 20 volumes. The *Shorter*, *Concise* and *Pocket* Oxford dictionaries are all 'spinoffs' from this venture, and have been through numerous editions since the 1930s (one of which has been used for illustration in 1.2 above).

(3) The third achievement was Joseph Wright's *English Dialect Dictionary*, published in 6 volumes in 1898–1905. This followed the tradition of dialect glossaries that had arisen earlier in various European countries, notably in Germany. Wright assigned each word to the

localities where it was used, county by county; and detailed dialect surveys in the mid-twentieth century confirmed the comprehensiveness and accuracy of his lexicographical work.

(4) Finally, the nineteenth-century dictionaries of the classical languages, Lewis and Short's *Latin–English Dictionary* and Liddell and Scott's *Greek–English Lexicon*, set a new standard that all subsequent bilingual dictionaries, classical or modern, have had to acknowledge.

In English-speaking countries in the twentieth century, dictionaries became a significant proportion of all publishing activity. In general the practices developed in nineteenth-century lexicography continued, but there was further expansion in three main areas: technical dictionaries, both monolingual and bilingual; learners' dictionaries, of English as a foreign or second language; and dictionaries of varieties of English other than those of England and America – principally Scots, Australian, Canadian, New Zealand and South African. In the latter part of the twentieth-century, dictionaries of the so-called 'new varieties of English' also began to appear, for example a *Dictionary of Jamaican English*, first published in 1967 and revised in 1980, and a *Dictionary of Caribbean English Usage*, 1996.

1.6 Recent developments in lexicology

Towards the end of the twentieth century significant changes were taking place in the theory and practice of lexicology, largely brought about by the new technology available for data-processing and text-based research. The two critical resources here are the computer and the corpus. Existing lexicographical techniques have of course been computerised. For example, lexicographers can now check their list of dictionary entries against other lists of words – say a list of words occurring in recent editions of a newspaper – and can run such a check electronically in a fraction of the time that it would take to do this manually. But the computer does much more than speed the processes up – it shifts the boundaries of what is possible. For example, the total content of the 1989 edition of the *OED* is now available on compact disc (CD) to anyone whose computer has a CD drive. It thus becomes a database such that lexical information of all kinds can be retrieved from its half-million entries, with the entire search under any chosen heading usually taking less than one minute.

At the same time, lexical research can now be based on very large corpora of written and spoken language. Corpus work in English originated in the late 1950s, with the Survey of English Usage at the University of London and the Brown University Corpus in Providence,

Rhode Island. The two universities each compiled a corpus of one million words of written text, in selected passages each five thousand words long. By the 1990s lexicographers could draw on massive resources such as the British National Corpus, the International Corpus of English, and the 'Bank of English' at the University of Birmingham in England; and indefinitely large quantities of text, from newspapers to transcripts of enquiries and parliamentary proceedings, began to be accessible in machine-readable form (for further details, see Chapter 3, especially 3.5).

The effect of these resources on dictionary-making is already apparent: the dictionary can now be founded on authentic usage in writing and speech. This means that, in an innovative corpus-based venture such as the Collins COBUILD series of English dictionaries, not only is every citation taken from real-life discourse, but the way the different meanings of a word are described and classified can be worked out afresh from the beginning (instead of relying on previous dictionary practice) by inspecting how the word is actually used – what other words it collocates with, what semantic domains it is associated with, and so on. Here is an example of an entry from the first edition of the *Collins COBUILD English Language Dictionary*. The format of the entry has been changed slightly for presentation here, but the wording and sequence of information are exactly as in the 1987 edition of the dictionary. (A later edition of the dictionary has different wording.)

sturdy /st3:di[1] /, **sturdier, sturdiest.**
Someone who is **sturdy**
1.1 looks strong and is unlikely to be easily tired or injured.
e.g. *He is short and sturdy...*
...Barbara Burke, a sturdy blonde.
sturdily
e.g. *She was sturdily built.*
1.2 is very loyal to their friends, beliefs, and opinions, and is determined to keep to them, although it would sometimes be easier not to do so.
e.g. *With the help of sturdy friends like Robert Benchley he set about rebuilding his life.*
sturdily
e.g. *He replied sturdily that he had only followed her orders.*
2 Something that is **sturdy** looks strong and is unlikely to be easily damaged or knocked over.
e.g. ... *sturdy oak tables...*
...a sturdy branch.

In the *Collins COBUILD English Language Dictionary*, an 'extra column', beside the entry, adds the information that *sturdy* is a qualitative

adjective, in all its senses; and that, in sense 1.2, it is usually used attributively, that is before the noun, as in *sturdy friends.* (This pattern is clearer in an example such as *they are sturdy supporters of the club,* where *sturdy* goes with the verb *support* (= *they support the club sturdily*). If the adjective is used predicatively, that is, after the noun, the sense will typically shift to 1.1: *the club's supporters are sturdy* = 'strong robust people'.)

The extra column also gives, in sense 1.1, the synonym *robust;* in sense 1.2, the synonym *steadfast* and superordinate *dependable;* and in sense 2, the synonym *tough.* This entry may be contrasted with the more traditional entry in another dictionary of approximately the same size, the 1979 *Collins Dictionary of the English Language.* (Again, the presentation here has been slightly changed, with more generous spacing than is normally possible in a large dictionary; and there are later Collins dictionaries than this edition.)

sturdy ('stɜ:dɪ) *adj.* **-di-er, -di-est.**

1. healthy, strong, and vigorous.

2. strongly built; stalwart.

[C13 (in the sense: rash, harsh): from Old French *estordi* dazed, from *estordir* to stun, perhaps ultimately related to Latin *turdus* a thrush (taken as representing drunkenness)]

– **'stur-di-ly** *adv.*

– **'stur-di-ness** *n.*

We said at the beginning that lexicology – the study of words – is one part of the study of the forms of a language, its **lexicogrammar.** Lexicology developed as a distinct sub-discipline because vocabulary and grammar were described by different techniques. Vocabulary, as we have seen, was described by listing words, either topically (as a thesaurus) or indexically (as a dictionary), and adding glosses and definitions. Grammar was described by tabulating the various forms a word could take (as **paradigms,** e.g. the cases of a noun or the tenses of a verb) and then stating how these forms were arranged in sentences (as **constructions,** or **structures** in modern terminology). But vocabulary and grammar are not two separate components of a language. Let us borrow the everyday term **wording,** which includes both vocabulary and grammar in a single unified concept.

When we speak or write, we produce wordings; and we do this, as we suggested in 1.1 above, by making an ongoing series of choices. Usually, of course, we 'choose' quite unconsciously, although we can also bring conscious planning into our discourse. We also noted that some of these choices are between two or three alternatives of a very

general kind, like positive versus negative (e.g. *it is / it isn't; do it / don't do it*); likewise singular versus plural number, first / second / third person, past / present / future tense, and so on. These 'closed systems' are what we call grammar. Of course, such choices have to be expressed in the wording, and sometimes we have specifically grammatical words to express them ('function words') like *the* and *of* and *if*. But often these general choices are expressed in a number of different ways, some of them quite subtle and indirect; so we tend to label them as **categories** rather than by naming the words or parts of words that express them. For example, we refer to the category 'definite' rather than to the word *the*, because (1) *the* is not in fact always definite, and (2) there are other ways of expressing definiteness besides the word *the*.

Other choices that we make when we use language are choices among more specific items, the 'content words' that we referred to at the beginning. These are not organised in closed systems; they form open sets, and they contrast with each other along different lines. For example, the word *cow* is in contrast (1) with *horse, sheep* and other domestic animals; (2) with *bull*; (3) with *calf* and some more specific terms like *heifer*; (4) with *beef*, and so on. So we refer to it by itself; we talk about 'the word *cow*', and define it in a dictionary or locate it taxonomically in a thesaurus.

We could describe *cow* using the techniques devised for dealing with grammar. We could identify various systems, e.g. 'bovine / equine / ovine', 'female / male', 'mature / immature', 'living organism / carcass', and treat *cow* as the conjunct realisation of 'bovine + female + mature + living'. In this way we would be building the dictionary out of the grammar, so to speak. This may be useful in certain contexts, especially when different languages have to be interfaced as in machine translation – different languages lump different features together, so their words don't exactly correspond. Equally, we could build the grammar out of the dictionary, treating grammatical categories as generalisations about the words that express them: instead of the category of 'definite' we could describe the various meanings and uses of the word *the*. Again there are contexts in which this might be helpful: teaching foreign learners who want only to read English, not to speak or write it, for example.

In general, however, each technique gets less efficient as you approach the other pole: you have to do more and more work and you achieve less and less by doing it (as we put it in our initial summary, there are diminishing returns in both cases). What is important is to gain an overall perspective on lexicogrammar as a unified field – a continuum between two poles requiring different but complementary

strategies for researching and describing the facts. This perspective is essential when we come to deal with the regions of the language that lie around the middle of the continuum, like conjunctions, prepositions and many classes of adverb (temporal, modal, etc.) in English. But it is important also in a more general sense. With our modern resources for investigating language by computer, namely 'natural language processing' (text generation and parsing) and corpus studies, we can construct lexicogrammatical databases which combine the reliability of a large-scale body of authentic text data with the theoretical strengths of both the lexicologist and the grammarian. The user can then explore from a variety of different angles.

One topic that has always been of interest to lexicologists is the recording of neologisms – 'new' words, not known to have occurred before. Earlier dictionary makers depended on written records, which are increasingly patchy as one goes back in time; the first occurrences cited for each word in the *OED* obviously cannot represent the full range of contemporary usage. The huge quantity of text that flows through today's computerised corpora (while still comprising only a fraction of what is being written, and a still smaller fraction of what is being spoken) makes it possible to monitor words occurring for the first time. But the concept of a 'neologism' is itself somewhat misleading, since it suggests that there is something special about a 'new word'. In fact a new word is no more remarkable than a new phrase or a new clause; new words are less common, for obvious reasons, but every language has resources for expanding its lexical stock, no matter how this is organised within the lexicogrammar as a whole. It is a mistake to think of discourse as 'old words in new sentences'. The chance of being 'new' clearly goes up with the size of the unit; but many sentences are repeated time and again, while on the other hand quite a number of the words we meet with every day were used for the first time within the past three generations.

1.7 Sources and resources

The best source of information about lexicology is the dictionary or thesaurus itself. It is important to become familiar with these works, which are now fairly common within the household. (In English-speaking countries at least, most large dictionaries and thesauruses are bought either for family members as Christmas gifts or for the children of the household to help them with their schoolwork.) You can **consult** dictionaries, to find out the meaning and usage of a particular word or phrase; and you can **read** them, dipping in at random or wherever your

fancy takes you. They can be unexpectedly entertaining. Samuel Johnson's 1755 dictionary is famous for several entries that betray a certain personal perspective, such as

> **excise**, a hateful tax levied upon commodities, and adjudged not by the common judges of property, but wretches hired by those to whom excise is paid.

Or you might come across a definition such as the following, from *Chambers Twentieth Century Dictionary*:

> **ranke**, *rangk*, *n.* (Shak., *As You Like It*, III.ii.) app. a jog-trot (perh. a misprint for **rack**(6)): otherwise explained as a repetition of the same rhyme like a file of so many butterwomen.

Nowadays dictionaries and other works of this kind are compiled for a wide range of different purposes. Naturally therefore they vary, both in the information they contain and in the way the information is presented. Consider for example an English–Chinese dictionary, one with English words listed and translated into Chinese. This might be compiled for Chinese students of English; or for English speakers studying Chinese; it might be for use in natural-language processing by computer (e.g. in multilingual text generation), or in the professional work of technical translators. It will be different in all these different cases. It soon becomes apparent that there is no single model that we can set up as the ideal form for a dictionary to take; nor are dictionaries totally distinct from other types of publication such as technical glossaries or travellers' phrasebooks.

This kind of indeterminacy is nothing new in the field. There is no clear line between a dictionary of a regional variety of a language (a dialect dictionary) and a dictionary of a functional variety of a language (a technical dictionary), or of a part of a language, such as a dictionary of slang, or of idioms, or of compounds. Nor is there any clear line between explaining the meaning of a word (dictionary definition) and explaining a literary allusion, or a historical or mythical event. The little dictionaries of hard words for children that used to be produced in various countries of Europe, like the Russian *azbukovniki* ('little alphabets'), included a great deal of useful information besides. In this respect they belong in the same tradition as *Brewer's Dictionary of Phrase and Fable* (first published in 1870, subtitled 'giving the Derivation, Source, or Origin of Common Phrases, Allusions, and Words that have a Tale to Tell') – and are only one or two removes from the great encyclopaedias of China and the encyclopaedic dictionaries of

European countries referred to in 1.4 above. The line between a dictionary and an encyclopaedia has always been uncertain, and has been drawn differently at different times and places throughout the history of scholarship. Equally indeterminate is the line between a dictionary and a scholarly monograph: a dictionary may be conceived of purely as a work of linguistic research, like an etymological dictionary (typified by August Fick's *Comparative Dictionary of the Indo-European Languages* first published in 1868), or dictionaries of the elements that are found in personal or place names.

Finally we might mention the comic dictionaries, like Douglas Adams' *The Meaning of Liff*, which consists of imaginary – and highly imaginative – definitions of place names treated as if they were English words. These in turn are part of the general tradition of lexical humour, which is found in some form or other in every language (the 'play on words' like punning by speakers of English). Related to this are various forms of word games, both traditional and codified: those in English include both competitive card or board games like Lexicon and Scrabble, and individual games such as plain and cryptic crosswords. In quite a few languages people play informal games in which they invert or swap syllables: rather as if in English we were to make *village* into *ageville* or *elbow* into *bowel*. And Indonesians sometimes create an 'explanation' for a word by pretending that its syllables are shortenings of other words; if we tried something comparable in English we might say that an 'expert' is someone who is 'EXpensive' and 'PERTurbing'. These games often fit a particular language – different patterns of phonological word structure lend themselves to different kinds of playful manipulation – but all of them provide insights into the way words work; and the special word games played with children, like 'I'm thinking of a word that rhymes with –', have an important developmental function in giving children a sense of what a word is, and how words are classified and defined.

Standard works written in English on lexicology include Chapman (1948), Hartmann (1983), Hartmann (1986), Householder and Saporta (1962), Landau (1989), McDavid *et al.* (1973) and Zgusta (1971). A more recent general introduction to the field is Jackson and Ze Amvela (1999). Green (1996) is a comprehensive history of lexicography, and Cowie (1990) is also a useful overview, from which much of the information in 1.5 above is drawn.

2 Words and meaning

Colin Yallop

2.1 Words in language

People sometimes play games with words. People may also recite or memorise lists of words, for example when trying to learn the words of another language or to remember technical terms. And they may occasionally leaf through a dictionary looking at words more or less randomly. These are legitimate activities, enjoyable or useful as they may be. But they are not typical uses of words. Typically, human beings use words for their meaning, in context, as part of communicative discourse.

As Halliday has made clear (see especially 1.6 above), vocabulary can be seen as part of lexicogrammar, a lexicogrammar that represents the choices which users of a language make, a lexicogrammar that represents our ability to *mean*. For, ultimately, language is about meaning. The main function of language – and hence of words used in language – is to mean.

This part of the book is particularly concerned with exploring the semantics of words. Section 2.2 offers some comments on meanings as presented in dictionaries. This is followed by brief discussion of potentially misleading notions about 'original meaning' (2.3) and 'correct meaning' (2.4). In 2.5 we try to explain what we mean by a social perspective on language and meaning, followed by some background on the theorising of Saussure and Firth (2.6) and Chomsky and cognitive linguists (2.7). We then look at the implications of our theorising for language and reality (section 2.8) and, to open up a multilingual perspective, we talk about the diversity of languages in the world (section 2.9) and about the process of translating from one language to another (2.10).

2.2 Words and meaning

A dictionary seems the obvious place to find a record of the meanings of words. In many parts of the English-speaking world, dictionaries

have achieved such prestige that people can mention 'the dictionary' as one of their institutional texts, rather in the same way that they might refer to Shakespeare or the Bible. Such status means that a printed dictionary may easily be seen as the model of word-meanings. We may then, uncritically, assume that a dictionary in book form is the appropriate model of words as a component of language or of word-meanings stored as an inventory in the human brain or mind.

In fact a dictionary is a highly abstract construct. To do the job of presenting words more or less individually, in an accessible list, the dictionary takes words away from their common use in their customary settings. While this is in many respects a useful job, the listing of words as a set of isolated items can be highly misleading if used as a basis of theorising about what words and their meanings are.

There is of course no such thing as 'the dictionary'. For a language such as English there are many dictionaries, published in various editions in various countries to suit various markets. The definitions or explanations of meaning in a dictionary have been drawn up by particular lexicographers and editors and are consequently subject to a number of limitations. Even with the benefit of access to corpora, to large quantities of text in electronic form, lexicographers cannot know the full usage of most words across a large community, and may tend to bring individual or even idiosyncratic perspectives to their work.

In the past, dictionaries were quite often obviously stamped by the perspective of an individual. We have already mentioned Samuel Johnson's definition of *excise* as 'a hateful tax' (1.7 above), and, as another example, here is Johnson's definition of *patron*:

> **patron**, one who countenances, supports or protects. Commonly a wretch who supports with insolence and is paid with flattery.

Modern lexicographers generally aim to avoid this kind of tendentiousness. Certainly today's dictionaries tend to be promoted as useful or reliable rather than as personal or provocative. Nevertheless, despite the obvious drawbacks of a dictionary that represents an individual editor's view of the world, it is regrettable that dictionary users are not reminded more often of the extent to which dictionary definitions are distilled from discourse, and often from shifting, contentious discourse. In any event, lexicographers can never claim to give a complete and accurate record of meaning. A team of expert lexicographers may by their very age and experience tend to overlook recent changes in meaning; or they may tend to write definitions which are elegant rather than accurate or simple; or they may follow conventions of definition

which are just that – lexicographical conventions – rather than semantic principles.

Dictionaries often tend to favour certain kinds of technical identification, definitions that describe *dog* as *Canis familiaris*, or *vinegar* as 'dilute and impure acetic acid'. While this kind of information may sometimes be precisely what the dictionary-user is looking for, it is debatable whether it constitutes a realistic account of meaning. Many of us communicate easily and happily about many topics, including domestic animals, food, cooking, and so on, without knowing the zoological classification of animals or the chemical composition of things we keep in the kitchen. Perhaps people *ought* to know information like the technical names of animals or the chemical composition of things they buy and consume, whether as general knowledge or for their health or safety. But it would be a bold move, and a semantic distortion, to claim that people who don't know such information don't know the *meaning* of the words they use.

In general, it is unwise to assume that meaning is captured in dictionary entries, in the definitions or explanations given against the words. Dictionary definitions can and should be informative and helpful, and, when well written, they provide a paraphrase or explanation of meaning. But the meaning is not necessarily fully contained or exhaustively captured within such a definition. This is not to say that meanings are vague or ethereal. Within the conventions of a particular language, meanings contrast with each other in established and often precise ways. Speakers of the same language can convey meanings to each other with considerable precision. Words do not mean whatever we want them to mean, but are governed by social convention. Nonetheless, we cannot assume, without qualification, that the wording of a dictionary definition is an ideal representation of what a word means.

Extending this point, we normally use and respond to meanings in context. As users of language we know that someone's mention of a recent television programme about big cats in Africa implies a different meaning of *cat* from a reference to the number of stray cats in the city of New York. And if someone talks about 'letting the cat out of the bag' or 'setting the cat among the pigeons', we know that the meaning has to be taken from the whole expression, not from a word-by-word reading of *Felis catus* jumping out of a bag or chasing *Columbidae*. Any good dictionary recognises this by such strategies as listing different senses of a word, giving examples of usage, and treating certain combinations of words (such as idioms) as lexical units. But it is important to recognise that this contextualisation of meaning is in the very nature

of language and not some unfortunate deviation from an ideal situation in which every word of the language always makes exactly the same semantic contribution to any utterance or discourse.

For reasons such as these, we should be cautious about the view that words have a basic or core meaning, surrounded by peripheral or subsidiary meaning(s). For example, the very ordering of different definitions or senses in a dictionary may imply that the first sense is the most central or important. In fact there are several reasons for the sequence in which different senses are presented. Some dictionaries, especially modern ones intended for learners of the language, may use a corpus to establish which are the most frequent uses of a word in a large quantity of text, and may list senses of a word in order of frequency. Some lexicographers follow a historical order, giving the oldest recorded senses first (even if these are now obsolete and largely unknown). Or a compiler may order the senses in a way that makes the defining easier and more concise (which is probably of help to the reader, even though it intends no claim about the centrality of the first sense listed).

For instance, the word *season* is commonly used in phrases like *the football season, the rainy season, the tourist season, the silly season, a season ticket, in season, out of season.* These uses taken together probably outnumber what many people may think of as the fundamental meaning of *season* as 'one of the four seasons, spring, summer, autumn and winter'. But the lexicographer may judge it sensible to begin the entry with the 'four seasons of the year' sense, not only because this is perhaps what most readers expect, but also because the subsequent definitions of *season* as 'a period of the year marked by certain conditions' or 'a period of the year when a particular activity takes place', and so on, may seem easier to grasp if preceded by the supposedly basic sense.

To take another example, consider the first four senses listed for the noun *rose* in the *Macquarie Concise Dictionary* (1998). Some of the definitions have been abbreviated for this example:

1. any of the wild or cultivated, usually prickly-stemmed, showy-flowered shrubs constituting the genus *Rosa* ...
2. any of various related or similar plants.
3. the flower of any such shrubs ...
4. an ornament shaped like or suggesting a rose ...

The sequence of these senses is not random and the entry has been written or edited as a whole. The second sense, using the words 'related' and 'similar', assumes the reader has read the first definition;

the third ('any such shrubs') presupposes the first and second; and so on.

The *Macquarie Concise* entry for *rose* also demonstrates that dictionaries are obliged to order items at more than one level. There are of course two quite distinct *roses*, the one we have just been talking about, and the one which is the past tense of *rise*. The *Macquarie* numbers these distinct meanings, as many dictionaries do, with a superscript [1] and [2], giving all the senses of the flower or bush (and the rose-like objects) under the first *rose*, and then simply indicating that the second *rose* is the past tense of *rise*. Probably most dictionary users find this the sensible order. Perhaps nouns seem more important, especially ones which have several different senses. Perhaps the second *rose* seems as though it is here accidentally – it really belongs under *rise*. Evidence from corpora suggests that the verb form *rose* (as in 'the sea level rose by 120 metres' or 'exports rose 2 per cent' or 'the evil genie rose from the jar') is used far more frequently than the noun; but this greater frequency does not seem to give priority to the verb in the minds of dictionary compilers and users.

It sometimes seems to be mere convention to list certain meanings first. Definitions of the word *have* often begin with the sense of 'possess' or 'own', and many people may indeed think of this as the fundamental or ordinary meaning of the word. In fact, corpus evidence indicates that the uses of *have* as an auxiliary verb (as in 'they have shown little interest') and in combinations like *have to* (as in 'we have to do better next time') are more frequent than uses like 'they have two cars' or 'we have a small house'.

Notions of what is a basic or central meaning of a word may thus be encouraged and perpetuated in a variety of ways, including common beliefs about words (which may or may not match actual usage) as well as lexicographical tradition. Sometimes such notions may be given formal recognition. For example, it is common to distinguish denotation from connotation. If taken as a serious semantic or philosophical claim, the distinction tends to separate what a word refers to from the associations that the word conjures up in the mind. More popularly, and sometimes simplistically, the distinction becomes a way of separating a core meaning from peripheral or variable aspects of meaning. But the distinction is by no means straightforward. It is complicated by the fact that what a word refers to in a particular context (as when talking to you I mention 'your cat') is not what is usually intended by *denotation* (which is more like 'any cat' or 'the class of cats'). The notion of denotation also runs the risk of identifying meaning with a class of objects or some idealised version thereof, as if meaning can be

anchored in a world of concrete objects. This is clearly not very helpful in the case of many words, such as abstract nouns in general or verbs like *believe, dream, think, worry* or epithets like *good, kind, mysterious, poor.* And even where a denotation can be satisfactorily identified, it is not self-evident that this is an appropriate way of characterising *meaning.*

The term *connotation* tends to slip awkwardly between something like 'peripheral meaning' and 'emotive meaning' and 'personal associations'. The notion of peripheral meaning simply raises the question of what is central or core meaning and why it should be so. It is clear from examples already given that the most frequently used sense of a word is not always the one that strikes most people as the core meaning. And it is equally clear that the older senses of a word are often neither the most frequent in current usage, nor the most basic by any other conceivable criterion.

Even 'emotive meaning', which might seem a good candidate for the margins of meaning, cannot always be considered peripheral. If I say to you 'Did you hear what happened to poor Sid?', the semantic contribution of *poor* must surely be 'emotive': the word says nothing about Sid's lack of wealth, but seeks to establish and elicit sympathy towards Sid. And this is hardly peripheral, since my question to you is most probably intended to introduce, and engage your interest in, a story of Sid's misfortune. Similar things can be said about the use of adjectives like *lucky* and *unfashionable,* which commonly serve to signal the speaker's attitude, and even about the verb *think* when used in utterances like 'I think the meeting starts at noon' (in which the words 'I think' serve to make the message less authoritative or dogmatic) or 'I think these are your keys' (as a polite way of telling someone they are about to leave their keys behind). Thus what might be termed 'emotive meaning' or 'attitudinal' meaning may sometimes be an integral part of discourse.

On the other hand, if 'associations' really are personal or idiosyncratic, then they hardly qualify as meaning at all, since they cannot contribute to regular meaningful exchanges. Suppose, for example, I have a fondness for a particular kind of flower, say, carnations, perhaps because of some valued childhood memory of them or other such personal experience. This may well have some consequences in my behaviour, including my discourse: I may often buy carnations, whereas you never do, I may mention carnations more than you do, and so on. But does it follow from any of this that you and I have a different meaning of the word *carnation*? Both of us, if we speak English, understand what is meant when someone says 'carnations are beautiful flowers', 'carnations are good value for money' and 'most people like

carnations', whether we agree with the truth of these claims or not. Indeed, to *dis*agree with these statements requires an understanding of what they mean, just as much as agreeing with them does.

Of course to the extent that an association is shared throughout a community, it does contribute to discourse and becomes part of meaning. If a name like *Hitler* or *Stalin* is not only widely known but widely associated with certain kinds of evil behaviour, then it becomes possible for people to say things like 'what a tragedy the country is being run by such a Hitler' or 'the new boss is a real Stalin'. And if people do say things like this, the names are on their way to becoming meaningful words of the language, along a similar path to that followed by words like *boycott* and *sandwich*, which had their origins in names of people associated with particular events or objects. (Note how *boycott* and *sandwich* are now written with initial lower-case letters rather than the capitals which would mark them as names. We might similarly expect to see the forms *hitler* and *stalin* appearing in print, if these names were to become genuine lexical items describing kinds of people.)

There may also be differences of experience and associations within a community which have systematic linguistic consequences. If, for example, some speakers of English love domestic cats while others detest them, this *may* well remain marginal to linguistic systems. But there may be small but regular linguistic differences between the speakers: for example some people may always refer to a cat as 'he' or 'she' while for others a cat is always 'it', and some people may use *cat* as the actor of processes like *tell* and *think* (as in 'my cat tells me when it's time for bed' or 'the cat thinks this is the best room in the flat') whereas others would never use this kind of construction. To that extent we may have (slightly) different linguistic systems, say one in which a cat is quasi-human in contrast to one in which a cat is firmly non-human. In that case, it is legitimate to recognise two somewhat different meanings of *cat* and two minor variants of English lexicogrammar.

For meaning is ultimately shaped and determined by communal usage. A dictionary definition of a word's meaning has authority only in so far as it reflects the way in which those who speak and write the language use that word in genuine communication. In this sense, meaning has a social quality, and while it is sometimes convenient to think of the meaning of a word as a concept, as 'something stored in the human mind', this is legitimate only to the extent that the concept is seen as an abstraction out of observable social behaviour.

An overview of issues to do with word meaning, and references to classic discussions such as Lyons (1977), can be found in the first two

sections of Chapter 3 of Jackson and Ze Amvela (1999). We will return to the issues in the following sections of this chapter, both to elaborate our own views of language as social behaviour and of meaning as a social phenomenon, and to contrast our views with others.

2.3 Etymology

In this section we look briefly at the relevance of historical development. Changes in language – specifically changes in meaning – are inevitable, but they are sometimes decried, as if language ought to be fixed at some period in time. In fact, attempts to fix meanings or to tie words to their 'original' meanings deny the social reality of linguistic usage. (In the following section, 2.4, we will look more generally at attempts to prescribe and regulate meaning.)

Warburg tells the story of a lawyer who disputed a witness's use of the word *hysterical* (Warburg 1968, pp. 351–2). The witness had described a young man's condition as 'hysterical'. But, the lawyer pointed out, this word was derived from the Greek *hystera*, meaning 'uterus' or 'womb'. The young man didn't have a uterus, so he couldn't possibly be 'hysterical'.

Would a good lawyer really expect to score a point by this kind of appeal to etymology? Few of us are likely to be persuaded to change our view of the current meaning of the word *hysterical*. It is true that the word is based on the Greek for 'uterus' (and the Greek element appears in that sense in English medical terms such as *hysterectomy* and *hysteroscopy*). But it is also true that words may change their meaning and that the modern meaning of *hysterical* has more to do with uncontrolled emotional behaviour, by men or women, than with the uterus as a bodily organ.

Sometimes an older sense of a word survives in limited contexts, while the most frequent meaning has changed. The word *meat*, for example, now has the common meaning of 'animal flesh used as food', but its Old English antecedent was a word that had the more general meaning of 'food'. Traces of the older more general meaning can be seen in phrases and sayings like *meat and drink* (i.e. 'food and drink') and *one man's meat is another man's poison* (i.e. 'one man's food is another man's poison'). The word *sweetmeat* also demonstrates the older sense. Other than in these restricted contexts, the older meaning of the word has become not only obsolete but irrelevant to modern usage. If you ask today whether a certain supermarket sells meat, or talk about the amount of meat consumed in Western Europe, or have an argument about what kind of meat is in a meat pie, no

one who speaks English pauses to wonder whether you really intend *meat* to mean 'food in general' rather than 'animal flesh'.

Indeed, older meanings become lost from view, and phrases and sayings may even be reinterpreted to suit the new meaning. The word *silly* had an older sense of 'happy' (compare German *selig*, 'blessed') but this sense has been ousted by the current meaning of 'foolish' or 'absurd'. A phrase sometimes applied to the county of Suffolk in eastern England, *silly Suffolk*, dates from the days when Suffolk was one of the wealthier counties, and therefore 'happy' or 'fortunate'. But if the saying is quoted at all these days, either it has to be explained, as we have just done here, or it is taken to be an allegation of foolishness or backwardness.

The word *prove* once had the sense of 'try' or 'test' but the most common modern meanings are of course 'show beyond doubt' (as in 'we all suspect him of corruption but no one has been able to prove it') and 'turn out' (as in 'the book proved to have lots of useful information in it'). The saying that *the exception proves the rule* shows the older sense – an exception indeed 'tests' whether a rule is really valid or needs to be reformulated. But the saying is often reinterpreted, with *prove* taken in its modern sense, to mean that an odd exception actually confirms a rule. This is clearly not true – an exception doesn't support a rule, it challenges it – but such is the power of current meaning to efface the old.

There is a long history of interest in etymology, in 'where words have come from', and many large dictionaries of English include etymological information (see McArthur 1992, pp. 384–6, Landau 1989, pp. 98–104, Green 1996, esp. pp. 337–48). Unfortunately, until the development of methodical historical linguistics in the nineteenth century, much etymology was highly speculative and often erroneous. Misguided guesswork about the origins of words can be found in ancient Europe, for example in the work of Varro, a Roman grammarian active in the first century BC (Green 1996, p. 41), and the practice of trying to relate as many words as possible to a relatively small number of allegedly simple or basic words was common until the mid-nineteenth century. Green cites a classic example from the late eighteenth century, in which a whole array of English words were claimed to be derived from or based on the word *bar*: thus a *bar* is a kind of defence or strengthening, and a *barn* is a covered enclosure to protect or defend what is stored in it, a *barge* is a strong boat, the *bark* of a tree is its protection, the *bark* of a dog is its defence, and so on (Green 1996, p. 353). In fact, careful historical research indicates that the word *bar*, as in the bars in a fence or across a window, came into English

from Old French, while *barn* is from an Old English compound meaning 'barley store', *barge* is related to an Old French word for a kind of boat, the *bark* of a tree is a word of Scandinavian origin, and the *bark* of a dog goes back to the Old English verb *beorcan*, 'to bark', which is not related to the other *bark*. These various words are of different origins, there is no evidence that they are all based on *bar*, and the idea that they are all clustered around the notion of defence is pure speculation.

Occasionally, an erroneous origin has become enshrined in the language by a process of 'folk etymology', in which the pronunciation or spelling of a word is modified on a false analogy. The word *bridegroom*, for example, has no historical connection with the *groom* employed to tend horses. The Old English antecedent of *bridegroom* is *brydguma*, where *guma* is a word for 'man'. The word ought to have become *bridegoom* in modern English, but as the word *guma* fell out of use, the form *goom* was popularly reinterpreted (with a change in pronunciation and spelling) as *groom*. A similar process of trying to make the odd seem familiar sometimes applies to words adapted from other languages. The *woodchuck*, or 'ground hog', has a name taken from a North American Algonquian word which, in its nearest anglicised pronunciation, might be something like *otchek* or *odjik*. The word has nothing to do with either *wood* or *chuck*, but was adapted to seem as if it did.

There is nothing wrong with being interested in where a word has come from, and many people who use modern dictionaries expect historical or etymological information to be included. For much of the nineteenth and twentieth centuries, most dictionaries gave considerable prominence to historical information. The first complete edition of what is now commonly referred to as the 'Oxford dictionary' was entitled *A New English Dictionary on Historical Principles*, and it set out to record the history of words, not just their current meanings (see 1.5 above; but not all subsequent Oxford dictionaries, including various abridged editions and dictionaries for learners, have had the same historical priority). It hardly needs to be said that modern professional lexicographers try to avoid speculation and guesswork and to give only information based on good research.

It is indeed often interesting to know something of a word's history and its cognates in other languages, and many (though not all) modern dictionaries still include etymological information. English happens to share with most European languages a reasonably well-documented Indo-European heritage. Languages like Greek, Latin and Sanskrit, as well as a 'proto-Germanic' language ancestral to

modern English, German and other Germanic languages, can be shown to be historically related within an Indo-European 'family' of languages. The entry for *bear* (in the sense of 'carry') in the *New Shorter Oxford*, as cited earlier in 1.2, illustrates the way in which some dictionaries list cognates: the etymology includes not only forms considered to be ancestral to the modern English, in this case Old English *beran*, but also forms from other Germanic languages which are parallel to Old English rather than ancestral to it, such as Old Norse *bera* and Gothic *bairan*. The *Oxford* also lists forms that are parallel to Germanic, including Sanskrit *bharati*, Greek *pherein* and Latin *ferre*. As the *Oxford* entry implies, linguists hypothesise that there was an Indo-European form from which the Sanskrit, Greek, Latin and Proto-Germanic forms were separately derived.

Sometimes there have been intriguing changes of meaning. The word *town*, for example, can be traced back to an Old English form *tun* (with a long vowel, pronounced something like modern English *oo* in *soon*). We can connect this form with related words in other modern Germanic languages, notably *tuin* in Dutch and *Zaun* in German. There are regular patterns of sound change which (partly) explain how the forms have become different: modern English *out, house, mouse*, all pronounced with the same diphthong as in *town*, can be related to Old English *ut, hus, mus* (all with a long *u*) as well as to Dutch *uit, huis, muis* and German *aus, Haus, Maus*. But in the case of the forms related to *town*, Dutch *tuin* means not 'town' but 'garden' and German *Zaun* means neither 'town' nor 'garden' but 'fence'. There was also a similar word in Celtic languages, namely *dun*, meaning something like 'citadel' or 'fortified town'. This element is evident in some Roman place names incorporating Celtic elements, like *Lugdunum*, modern *Lyons*, and in names such as *Dunedin*, an old Celtic name now generally replaced in Scotland by the anglicised form *Edinburgh*, but still the name of a city in New Zealand. Thus the word must once have referred to fortified settlements. By modern times the English word *town* has generalised in meaning to refer to any substantial urban centre (between a village and a city in size and importance) while the Dutch word *tuin* has come to mean 'enclosed cultivated land', that is 'a garden', rather than an enclosed town, and the German *Zaun* has narrowed to the enclosure itself, or 'fence'.

Such information is not only interesting to many readers, it is often valuable as an accompaniment to historical and cultural research. Moreover, modern European languages not only have a certain shared heritage, they have continued to draw on it in various ways. Latin words can still be found in uses as diverse as the English translation of Freud

(the *ego* and the *id*) and the mottoes of army regiments (such as *Ubique* 'everywhere', the motto of the British Royal Artillery). Some Latin phrases are indeed everywhere, even if no longer fully understood. Notable examples are *etc.*, the abbreviated form of *et cetera*, 'and the rest'; *e.g.*, short for *exempli gratia*, 'for (the sake of) example'; and *a.m.* and *p.m.* (*ante meridiem, post meridiem*). Latin has been regularly used in anatomical description (*levator labii superior*, the 'upper lip raiser' muscle, or *corpus callosum*, the 'callous (hard) body' in the brain), and in botany and zoology (*quercus* 'oak' for a genus of trees, or *felis* 'cat' for the genus of animals that includes domestic cats and some closely related species). Latin phrases such as *de facto, in camera, sine die, sub judice* and *ultra vires* are known in legal contexts, and some of them have a wider currency (such as the Australian use, even outside legal contexts, of the phrase 'a de facto' to mean 'a common-law spouse').

Greek and Latin have also provided a rich source of modern coinage. Words like *altimeter, electroencephalogram, hydrophone* and *telespectroscope* are obviously not themselves classical words: they have been built from Latin and Greek elements to deal with relatively recent technological innovation. Indeed, it has become so customary to use such elements as building blocks, that Latin and Greek are often combined in hybrid forms, as in Greek *tele-* with Latin *vision*, or Latin *appendic-* with Greek *-itis*.

But it is by no means just new items of technology, like cardiographs and synthesisers, that attract classical naming. Greek and Latin elements are integral to our standardised systems of calculating and measuring (*centigrade, centimetre, kilogram, millisecond, quadrillion*). Concepts like *social security, multimedia, globalisation* and *privatisation*, though essentially twentieth-century concepts, are conceived in classical forms. A classical heritage similarly underlies terms like *interdisciplinarity* (which I heard used at Macquarie University in discussions about creating links among different academic 'disciplines' or areas of learning) and *interdiscursivity* (which I have seen on a whiteboard in a university lecture theatre but not yet understood). And terms formed with Greek and Latin elements like *intra, non, post, pseudo, ultra* are used as much in administration or business or politics as in science or technology (*intrastate, noncompliance, postdated, pseudo-solution, ultra-conservative*).

Nevertheless, as we have already argued, the history of a word is not the determinant of its current meaning, and the greatest persisting drawback of etymological studies is that they may be misused to support assertions about what words 'ought' to mean. No modern dictionary (including Oxford's *New English Dictionary*) seriously misuses

historical information in this way. And, for the greater part of English vocabulary, no one seriously proposes that an older meaning of a word is the only correct meaning. But where a shift in meaning is relatively recent, and particularly where a newer sense of a word is evidently competing with an older sense, some people may deplore the change and attempt to resist it. Thus in the seventeenth century, the English word *decimate* was used to mean something like 'take or remove one tenth from', as in 'tithing', that is taxing people one-tenth of their income or property, or in the sense of killing one in ten. (Executing one in ten of a group of soldiers was a punishment sometimes used in the ancient Roman empire.) Nowadays the word is most commonly used to mean 'destroy most of', as if the 'decimation' now means reducing to one-tenth, rather than reducing to nine-tenths. Some people, especially those who have had a classical education and are aware of the ancient Roman punishment, condemn the modern usage as loose and unwarranted.

Whatever our feelings about respecting tradition or preserving history, it has to be said that such attempts to resist changes in general usage are rarely if ever successful. What usually happens is that by the time a shift is in progress, a majority accepts or doesn't notice the change, and only a minority condemns or resists the change. At this point, the minority may claim that their usage is educated or correct, and that the majority usage is careless or mistaken. But the minority usage is at risk of seeming unduly conservative and pedantic, and the situation is usually resolved by the disappearance of the minority usage. Over the years, people have deplored the changes in meaning of words like *arrive*, *deprecate* and *obnoxious* and have been able to argue that the older meaning was more faithful to the etymology. Thus *arrive* used to mean 'to reach a shore' rather than to reach anywhere (and the older meaning could be justified by appeal to the French *rive* 'shore, river-bank'); *deprecate* once meant 'to pray against, pray for deliverance from' rather than the modern 'to disapprove of, criticise' (and this too could be justified etymologically, given the Latin *deprecatus* 'prayed against'); and *obnoxious* meant 'liable to criticism or punishment' (Latin *obnoxius* 'exposed to harm') whereas the modern meaning is 'unpleasant, offensive'. Needless to say, the older meanings are now virtually unknown – except to those who find them in dictionaries and other records of the past.

Finally, we should note the need to be cautious about the idea of 'original meaning'. Sometimes we can identify the origin of a word – as for instance with the word *boycott*, which is believed to have come from the name of a land agent in nineteenth-century Ireland, who was

'boycotted' by tenants. But in many cases, there is no justification for calling an earlier meaning 'original'. The most common current meaning of *nice* – pleasant or enjoyable – has probably come from an earlier meaning, something like 'delicate' or 'dainty'. But this meaning can scarcely be called original. It probably came from earlier use of the word to mean 'finely differentiated' or 'requiring care and discrimination' (compare a traditional legal phrase 'a nice point'), which must in turn have come from the Latin *nescius* 'ignorant'. But even the Latin word and its meaning are only original relative to modern English. Latin is also a language with a history. It descended from something spoken previously, just as much as modern Italian came from Latin or modern English from old English. In short, however interesting and instructive the past may be, not all of it is accessible to us and not all of it is relevant. The past is not the present, nor is the history of a word its meaning.

2.4 Prescription

The idea which we have been looking at in the previous section, that a word ought to mean what it used to mean, is just one instance of what can be called a prescriptive approach to language. More generally, there have been many and various attempts to prescribe how language ought to be – prescriptions about pronunciation, for example, or rules about correct grammar, as well as claims about the proper meanings of words. Many of these attempts have been misguided if not perverse, and it became axiomatic in twentieth-century linguistics to reject prescriptivism. A common slogan of linguists was that 'linguistics is descriptive, not prescriptive'.

As a commitment to scientific method and ethical research, the slogan is exemplary. Whether investigating the physiology of speech production, recording what people say to each other in specific situations or examining the frequencies of words in printed texts, linguists, like all scholars and researchers, are under obligation to describe what they find. Even allowing that complete objectivity is unattainable, and that there will always be controversy about what exactly constitutes 'describing what you find', there is an indisputable obligation to aim to describe what is there, rather than to describe what you would like to be there or what you think ought to be there.

The slogan also represents a justifiable reaction to some of the prescriptivism of the past. In seventeenth- and eighteenth-century Europe, for example, some scholars and writers believed that it was necessary to regulate language and to set up academies for this pur-

pose, such as the Académie Française, founded in 1634 and charged with compiling a French dictionary and with ruling on matters of grammar, vocabulary and usage. Though no academy was ever set up in Britain, there were certainly calls to refine and reform the English language. To some extent, these ambitions were motivated by a desire for regularity and consistency. Since it is important both to understand the weakness of prescriptive approaches to language and to recognise the genuine normativity inherent in language, we will consider two examples in some detail, first the history of comparative forms like *(more) bigger*, and second the proposal that prepositions shouldn't end sentences.

In English grammar, by the seventeenth century, the old pattern of forming comparative and superlative adjectives by endings (as in *big, bigger, biggest* or *tall, taller, tallest*) had begun to blend with a newer pattern using the words *more* and *most* (as in *evil, more evil, most evil* or *corrupt, more corrupt, most corrupt*). In Shakespeare's writings, for example, we can find the two patterns combined, as in *more better, more corrupter, most unkindest, most coldest.* But eighteenth-century grammarians began to criticise this practice, apparently on the grounds that only one of the two devices (either the ending or the *more/most*) is logically necessary to convey the meaning. Modern English usage has been partly influenced by these grammatical strictures. People nowadays quite often say things like *more kinder* or *most earliest*, but they tend to avoid them in writing, and editors are likely to delete the *more* or *most.* Written usage is still not exactly regular, however, since the tendency is to use the endings on monosyllabic words (*colder, coldest, higher, highest, later, latest*) and to use *more* and *most* with polysyllabic words (*more difficult, more interesting, most intelligent, most troublesome*). But this is only a generalisation: some monosyllabic words do take *more* (*more tired*, for instance) and for some words of two syllables it seems perfectly acceptable to go either way (*shallower* or *more shallow, commonest* or *most common*). There are also the 'irregular' forms *better, best, worse, worst.* (For an overview of usage see Biber *et al.* 1999, pp. 521–5, and for details of past as well as more modern usage, see Fries 1940, pp. 96–101.)

Despite some variation in usage, forms such as *more bigger* and *most highest* are usually disapproved of by editors and teachers. While there may be a superficial appeal in simplifying such phrases to the single words *bigger* and *highest*, there are two difficulties to be noted. The first is that users of language will rarely if ever be bound by the dictates of individuals and academies, however educated or well informed those authorities may be. Many speakers of English continue to say things

like *more kinder* and *most earliest*, even after they have been told not to. And imagine the reaction (or indifference) of the community at large if linguists or teachers were to recommend that we regularise the language by saying *gooder* and *goodest* rather than *better* or *best*, or *badder* and *baddest* rather than *worse* and *worst*. Whatever arguments might be put forward, that forms like *gooder* are simpler, more regular or more logical than what we actually say, most people would continue to follow their customary practice and would consider the recommendation absurd. With few exceptions, language does not change because of regulation, it changes according to its own communal patterns.

The second problem in making language more logical or regular is that it is not at all self-evident what constitutes logic or regularity in linguistic matters. It is somewhat clearer, and rather more carefully discussed, what logic means in thinking and reasoning, or what regularity means in the study of natural phenomena. But linguistic systems generate their own logics and regularities. Is it really illogical to say *more kinder*? If it is the redundancy that is illogical, then by similar argument, we might claim, for example, that plural forms are redundant and illogical after numerals. A numeral already signals that the noun must be understood as plural, and we could therefore write *five dollar, a hundred student, a thousand spectator*. (And some languages, such as Welsh, do indeed use the singular form of a noun after a numeral.) In fact if we look dispassionately at the patterns of languages, we find a variety of ways of organising the lexicogrammar to express meaning, and it is not at all obvious why any of them should be regarded as more or less logical than others. Is it more logical for adjectives to precede nouns (as they mostly do in English, German or Japanese) or to follow nouns (as they mostly do in French, Italian or Indonesian)? Is there any reason why we should express contrasting verb meanings by suffixes (as English does with, say, *walk, walked, chase, chased*) rather than by auxiliary verbs (as English does with, say, *will walk, might walk, will chase, might chase*)? Is it neater or more regular to signal meanings like 'for', 'in' and 'on' by separate words preceding a noun (as English and most European languages do) or by suffixes on the noun (as languages as diverse as Finnish, Turkish and Australian Aboriginal languages mostly do)? What is logical and regular is the way in which each language underlies the linguistic behaviour of its speakers, the way in which each language builds a system out of its systems. The positioning of adjectives, the mechanics of the verb system, the use of prepositions or noun suffixes are not just trivial and isolated features of a language but are woven together in a complex, coherent and powerful lexicogrammar.

To return to the point about attempts to reform English, our second example is a rule sometimes imposed on English that sentences should not end with prepositions. According to the severest version of this rule, prepositions belong before a noun or pronoun, as in *for Uncle Leo, for me, in Singapore, in the afternoon, on Fridays, on the table.* A sentence in which a preposition appears other than before a noun or pronoun, like 'that's the book which I've been looking for', should be rephrased as 'that's the book for which I've been looking'; and a question like 'what is she looking at?' should be rephrased as 'at what is she looking?' This rule seems to have been invented by Dryden in the seventeenth century (Strang 1970, p. 143) and since then it has been often promoted, possibly beyond Dryden's intentions, and widely ignored or ridiculed.

In modern grammars, a preposition such as the 'for' in 'what are you looking for' is sometimes said to be 'stranded' (see e.g. Biber *et al.* 1999, pp. 105–8). The reasons for wanting to avoid 'stranded' prepositions probably include the fact that prepositions do not occur at the end of sentences in Latin (and Latin has often been held up as a model which other languages should conform to) and the very name *preposition*, which might seem, etymologically, to imply that these words should always be 'pre-posed' before another word.

But Latin grammar is not the same as modern English grammar, and the etymology of the name *preposition* does not impose any requirement on well-established English usage (any more than *premises* must mean '(things) sent beforehand' or *prevent* must mean 'come before'). While many writers, having been schooled in Dryden's rule, may now prefer to avoid sentence-final prepositions in formal English, most of us continue to ask questions like *what were you looking for?* and *who did you give it to?*, and find the rephrased versions awkward or pompous. Indeed, the strength of communal resistance to arbitrary regulation is seen in the way in which the rule is mocked by pronouncements such as 'a preposition is a bad word to end a sentence with' or the witticism ascribed to Winston Churchill 'this is a form of pedantry up with which I will not put'.

While it may sometimes seem desirable to make language more logical or consistent, the fundamental challenge to regulators is that the patterns of language emerge as a matter of social convention. Regularity and consistency are important factors in this process, but not the only ones or the pre-eminent ones. As we have already suggested, the complexity of language and its processes of acquisition and change are such that it is not always clear what exactly logic and consistency mean in linguistic practice. If *most coldest* ought to be simplified or regularised, should it be to *coldest* or to *most cold*? And if this reform is important, why

is it not equally important to get rid of redundant plural forms after numerals or to tidy up the English verb system? Why not get rid of the irregular and redundant word *am*, and simplify *I am* to *I are*, on the analogy of *you are* and *we are*? (We already say *aren't I?* rather than *amn't I?* which takes us some of the way towards this regularisation.) Why not make all verbs regular, replacing *ran* with *runned*, *wrote* with *writed*, and so on? The absurdity of trying to impose some externally conceived general notion of logic and simplicity on language puts a harsh spotlight on the odd details that are on reformist agendas.

Indeed, many people have tried to reform or regularise a language or to stop it from changing, but few have had much success. In general, languages change as societies and cultures do: as we differ from our grandparents, whether radically or not, in our beliefs, our perspectives, our social behaviour, our hobbies, our dress, so we differ from them, significantly or trivially, in our accent, in our idiom, in the words we use and the meanings we exploit. Changes in language do not happen uniformly across the world, and perhaps not even at a constant rate – there may be periods of rapid change and periods of relative stability. But change is observable, everywhere where the history of languages can be studied.

We should nevertheless be clear that an argument against regulation and prescription is not an argument against normativity in principle. The social nature of language brings a normativity of its own. As children we learn our linguistic patterns in the community in which we function, from our peers and from the adults with whom we interact. We learn the conventions of the written language which our community has inherited. And the patterns and conventions that underlie linguistic behaviour around us exert a strong pressure to conform: as human beings we are powerfully motivated, not only to understand and be understood, but to belong.

As we enter places of formal education and employment, we may be subject to specific linguistic norms, the kinds of norms that govern the writing of university essays or press releases or product information or government reports. Here we may well be in relatively circumscribed domains, where norms may be imposed more directly and more authoritatively. Thus a commercial company may have rules about the structure and wording of the memorandums written by its employees, a journal may have requirements about the style and presentation of papers which it is prepared to publish, a government department may follow conventional guidelines about the format and style of its documentation, and so on. (For more discussion of 'controlled' language, especially nomenclatures, see 2.8 below.)

It is in such domains that arbitrary prescriptions of the kind that tell us to write *shallower*, not *more shallow*, or to avoid ending sentences with prepositions, may have some measure of success. To some extent, arbitrary rulings in well-defined contexts are necessary, simply to yield consistency in, for example, the way in which dates are written or bibliographies compiled or reports presented. Hopefully the focus of those who write the relevant style guides or otherwise determine conventions in such settings is on clarity and consistency and efficiency, and on meaningful rather than empty traditions.

Moreover, even in society at large, it is important, even essentially human, to bring moral perspectives to bear on social and cultural changes. Social and cultural changes can, and should be, evaluated for their effects on human wellbeing, on the distribution of resources, on fairness and justice, difficult and contentious though the processes and criteria of evaluation may be. And to the extent that language reflects and supports behaviour and social structures, it is open to moral evaluation. Without such evaluation there would be no debate about sexism and racism in language, no possibility for argument about clarity and truth in language. Thus most of us do accept style guides that promote inclusive or egalitarian language, guidelines that provide for a certain degree of consistency of format in journals and bibliographies, courses that teach report writing, and so on.

The argument against prescription is not an argument against normativity in principle. But linguistic norms must be founded in social agreement on issues that matter to people – in a recognition by most people that we ought to eliminate racist words from the language, or that it is worth some effort to make instruction manuals as clear as possible, or that bibliographies are much easier to use if they follow standard conventions. This kind of commitment does not constitute justification for prescriptions about whether you can end a sentence with a preposition, and it gives no support to rulings based on individual interpretations of what might make language more regular, nor to arguments that language should be fixed once and for all in some supposedly golden age.

2.5 A social view of language and meaning

In this book we take the view that language is social behaviour and meaning a social phenomenon. By this we mean that language is more than an individual possession or ability, that language 'exists' because of its life in social interaction, that meaning is shaped and negotiated

in social interaction and that meaning must be studied with due recognition of its social setting.

The concept of meaning itself is difficult to define and it is no exaggeration to say that modern linguistics has failed to formulate a widely agreed theory of meaning. But the fact that there is something elusive and mysterious about meaning need not embarrass us, any more than humans should be embarrassed by the difficulty of understanding and defining exactly what we mean by time, number, life and other fundamental concepts of our existence. Most of us readily acknowledge that we cannot give a snappy definition of what time is, but we are still conscious of what we call the passing of time, we know the difference between yesterday and tomorrow, we even make it possible for ourselves to measure and quantify time by counting the alternations of daylight and darkness, constructing a twenty-four-hour day, and so on. Similarly, it is hard to give a technical definition of life. Dictionaries resort to phrases like 'the state of being alive' or to descriptions of what distinguishes living beings from dead ones or living beings from inanimate objects. In so doing they demonstrate both the difficulty of what they are trying to do and the good sense of drawing on our experience: we know that some things (people, animals, plants) live, that other things do not, that living beings sooner or later die. We will try to take a similar approach to meaning: it may be hard to define, but we all experience it; we negotiate meanings in our daily life; we (mostly) know what we mean and what others mean.

In societies with well-developed literacy and a tradition of publishing and using dictionaries and other reference books, there is always a danger that a language will be equated with some written account of the language. We have already referred to the dangers of assuming that a dictionary of English *is* the vocabulary of English (2.2 above), and a book describing the grammar of English may likewise seem to *be* the grammar of English. But dictionaries and grammar books are only representations of the language (and limited representations of certain aspects of the language). If they have value, it is because they represent, in some generalising abstract way, what people do linguistically. The meanings of words or the rules of grammar have not been laid down by some expert or authoritative decree at some point in the past and then enshrined in print. Dictionaries and grammar books are not legislation enacted by a linguistic parliament, nor are they the official manuals issued by people who created the language. If dictionaries and grammar books have authority, it is because they reflect general usage. Thus a language exists or lives not because it is described or recorded but because it is in use among people who know the language.

We say that people 'know' a language. And this, perhaps as well as images of language as recorded rules and inventories, may imply that language exists in the human mind. While it is obviously true that adult speakers of a language have large resources of knowledge – including for example knowledge of words and meanings and experience of using and understanding them – it would be misleading to suggest that an individual's linguistic knowledge is a complete and adequate version of 'the language'. For an individual, taken in isolation, is just that, an isolated individual. We cannot really speak of a language unless individual human beings are communicating with each other, bringing the language to life. Our individual knowledge of language comes from interaction with others, at first particularly with parents and family, later also with other children with whom we spend time, with schoolteachers, and so on. Some bases of our linguistic behaviour seem to be established relatively early and firmly. Most people acquire their accent or patterns of pronunciation fairly early and seem to change very little, even if they move to an area where people speak differently (although some people do make substantial changes in their pronunciation, for example at secondary school or at university). People similarly tend to maintain basic vocabulary and idioms that they have used frequently in their early years, although again they may yield to strong pressures to change, for example if they realise there are substantial social and economic advantages in making changes, or if they move to an area where some different words and idioms are customary. But even those whose language seems to change little during their lifetime are still using and experiencing language. For most of us, in most parts of the world, language is realised – actualised, made real – in a wide range of settings, such as homes and schools and workplaces and shops among many others. Our sense of what is normal usage, of what words mean, is constantly shaped by such experience.

Consider for example the word *stakeholder*. Until the latter part of the twentieth century, the meaning of the word was something like 'the person who holds the stakes in a bet'. English-language dictionaries published before the 1980s record only that sense. By the end of the 1980s, however, the word was being used in a commercial sense, as in an Australian newspaper's reference to 'the best interests of the company taking into account the stakeholders'. From this kind of use in commercial and financial contexts, the word extended into other institutional uses, so that we find, during the 1990s, a university talking about its 'accountability and information provision to external stakeholders' and a water supply authority talking about workshops attended by 'stakeholders, managers and scientists'. A website relevant to

the construction industry speaks of the importance of the 'collaborative efforts of all stakeholders' and then helpfully specifies stakeholders as designers, engineers, property consultants, technologists and clients 'among many others'. From uses such as these it is clear that *stakeholder* can no longer be taken in the sense of someone who is holding or directly investing money.

While it would be unwarranted to attach too much significance to a single word, the shift and extension of *stakeholder* not only illustrates how words and our understanding of them can change, but also how changes in words reflect social movements, in this case the widening scope of *stakeholder* going hand in hand with an increasingly commercialised perspective on services such as education and health through the 1990s and the extension of many commercial or financial terms into general administrative discourse.

The word *gender* has also shifted in recent years, again reflecting social changes. Until quite recently English-language dictionaries gave as the main use of *gender* its meaning in grammar, as in talking about the two genders (masculine and feminine) of nouns in French or Spanish, or the three genders (masculine, feminine and neuter) of nouns in Latin or German. Some dictionaries also recorded a technical biological use of the word, as in talking about gender differentiation within a species, and an informal, possibly jocular or euphemistic use, as in talking about people 'of the opposite gender'.

By the end of the 1980s, dictionaries are recording *gender* as having a significant and formal use for something like 'the fact of being male or female'. The word has largely replaced *sex* in this sense, for *sex* has increasingly been used as shorthand for 'sexual intercourse'. At the same time the word *gender* has increasingly appeared in various kinds of official and academic discourse. A corpus search suggests that in formal written discourse in the 1990s, references to grammatical gender were now vastly outnumbered by the use of the word in phrases like 'redefining gender roles' or 'gender balance (in the workforce)' or 'gender and sexuality'. Thus demographers can refer to the 'age/gender profiles' of population groups and a trade union can raise the question of 'gender inequities in the existing staff structure', while universities offer courses with titles such as 'Gender and Policy' and the 'Politics of Culture and Gender'. Readers may like to ask themselves what they would take to be the current difference in meaning between 'the politics of gender' and 'the politics of sex'.

There is a sense in which the meaning of (most) words is constantly being negotiated. Our notion of what words like *stakeholder, gender* and *sex* mean is dependent on our discourse, on our experience of these

words, on our experience of how others use these words in real situations. Older readers may remember uses that are now archaic or obsolete, like 'the gentle sex' and 'the second sex'. Even phrases that are current may reveal a certain competition between different senses: note for instance how we understand the word *sex* in 'sex discrimination' compared with 'safe sex', or 'sex stereotyping of women' compared with 'gratuitous sex scenes'. (Compare examples given earlier of meanings which may be associated with particular contexts, or of meanings which may disappear other than in a few phrases, such as *meat* in the sense of food in general, 2.3 above.)

The word *patron* comes from a Latin word that meant something like 'protector' or 'guardian'. In English, the word has had a similar meaning, still evident in the phrase 'patron saint' for example. When we read about the eighteenth-century lexicographer Samuel Johnson and his need for patrons (and see his biting definition of *patron*, 2.2 above), we also understand the word against a background of benefactors and their dependants. Current corpus evidence shows continuing use of *patron* in this kind of meaning ('galleries which were trustees of public art, with local government as their major patrons') but also shows the word with a meaning that is closer to *customer* or *client*, especially a customer in a hotel or restaurant ('most diners want privacy ... some patrons, however, do not mind being observed'). Meanwhile the French word *patron* has come to be used in the sense of 'manager'. Thus in a restaurant in France, someone who asks for *le patron* is looking for 'the boss', not any of the customers. That two words of one origin can end up with contrasting, almost opposite meanings demonstrates again that meanings are negotiable and negotiated.

In the following section, we will further develop this perspective by looking briefly at the contribution to linguistic theory of the Swiss linguist Ferdinand de Saussure and the British linguist J. R. Firth. Saussure is widely considered to be the founder of modern structural linguistics and Firth a leading figure in mid-twentieth-century British linguistics. While these are by no means the only two linguists whose ideas we respect and draw on, they are both influential and explicit theoreticians who have shaped the way many linguists talk about meaning.

2.6 Saussure and Firth

Saussure

Ferdinand de Saussure was a francophone Swiss, born in Geneva in 1857. He seems to have had a great talent for languages and at the age of 15 was said to be already competent in Latin, Greek, German and English (as well as French, his mother tongue, of course). He came from a family with a tradition of scientific achievement – his father was a well-known naturalist, for example – and he entered the University of Geneva as a student of physics and chemistry in 1875. But his talents and enthusiasm were focused on language, and after a year of studying science in Geneva, he persuaded his parents to send him to Germany to study Indo-European languages.

Saussure studied in Germany for four years, mixing with learned and creative scholars, acquiring extremely useful experience in the research methodology of the times. He then taught for ten years in Paris, where he seems to have been highly regarded and influential, before returning, in 1891, to a professorship in Geneva. He taught mostly the linguistics of the time – Sanskrit, comparative and historical linguistics – but there is some evidence from his correspondence that he was dissatisfied with general linguistic thinking, that he thought there was need to reform the jargon and terminology of the day, and that he thought linguists needed to think more about what they were doing.

In 1906, the University of Geneva asked him to take over the responsibility for teaching general linguistics, and from then until 1911 he gave a series of lectures in alternate years. In 1912 he fell ill and he died in 1913. (For a concise account of Saussure's life and work, see Culler 1976.)

He had written a substantial amount about Indo-European languages and historical reconstruction, by which he had maintained his high reputation, but he had written nothing about his ideas on language in general. His colleagues and his students were so impressed by what they had heard from him that they thought they should try to preserve the lectures from the last years of his life. Two of his students put together what they could, from Saussure's own lecture notes and their and other students' notes, and created a book now known as Saussure's *Cours de Linguistique Générale* or *Course in General Linguistics.* The *Cours* was first published in Paris in 1916 and has been through several editions since then. A critical edition of the French text, prepared by Tullio de Mauro, was published in 1972 (Saussure 1972) and includes copious background and notes on the text. An English

translation (translated by Wade Baskin) was published in 1960 and another (translated and annotated by Roy Harris) in 1983. Harris has also written a critical commentary on the text (Harris 1987).

Saussure is now famous for various points which are developed in the *Cours*. He made a clear distinction, for example, between describing the history of a language and describing how it is at any particular point in its history, a distinction between a historical (or diachronic) perspective on language and a current (or synchronic) perspective. If that distinction seems self-evident to us nowadays, that is partly because Saussure firmly established it.

Saussure devotes considerable attention to the nature of the linguistic sign, which he describes as an inseparable combination of a *signified*, a concept or meaning, and a *signifier*, the spoken or written form which conveys or represents that meaning. This view contrasts with a long and continuing tradition in philosophy and linguistics in which it is assumed or claimed that you can separate form and meaning. This difference of theoretical stance has many consequences – for example for one's view of what translation is (see 2.10 below). We will therefore be returning to this point, but for the moment we note that Saussure says you can no more separate the signifier from the signified than you can separate the front and back of a sheet of paper.

Saussure's *Cours* also emphasises the point that linguistic signs are arbitrary (although he elaborates and qualifies the point in ways that make a simple summary difficult). Arbitrariness is not just a matter of the lack of logical or natural connection – in most instances – between the meaning of a word or phrase and the spoken sounds or written form which represent that meaning. Arbitrariness is also evident when we compare languages and find that their signs and meanings do not neatly match each other. The Dutch *slak* could be either 'snail' or 'slug' when we translate it into English. English *blue* is two different colours in Russian. And in some Australian Aboriginal languages, what looks like the word for 'father' is a term referring not just to an individual but to a range of male persons, not only one's biological father but also to brothers of one's father, parallel cousins of one's father and even certain great-grandsons.

Thus to speak of arbitrariness in language is not only to say that one concept in one language can become two in another, or that two can be collapsed into one. More than that, languages often see the world very differently. They divide reality up differently, they focus on different criteria, they structure experience in different ways. In the case of kinship terms like 'father' and 'mother', English highlights biological relationships, whereas Australian Aboriginal languages focus on

social structure in such a way that a word which English speakers might expect to refer to a unique individual refers rather to a group of people who share a similar place or role in the system.

In the kind of linguistics promoted by Saussure, it is important to do justice to the structures and systems which language itself generates or embodies. If you want to understand the kinship terms of an Australian Aboriginal language, don't try to set up some universal transcendental framework, try to get inside the language itself. If there's a word that looks as though it means 'father' but evidently does not correspond with English *father*, the questions to ask are: what are the other kinship words in this language? How do they contrast in meaning with each other? How do they appear in discourse? What kind of systems and structures do they form or enter into?

These meanings may be arbitrary in the sense that there is no predetermined framework that says all languages must make this or that distinction, but they are certainly not arbitrary in the sense that individuals can play freely and randomly with the language. While there is of course scope for creative excursions, whether in the strikingly unusual turn of phrase of a poet or in the entertaining zaniness of a comedian, what holds a language together, what makes it work as a language, is the social convention or agreement that undergirds it. A word means what it means because that is what people here and now in this community take it to mean. At its heart, language rests on social convention.

For reasons such as these, Saussure is considered a modernist and sometimes compared with figures like Freud (born the year before Saussure) and Durkheim (the 'founder' of modern sociology, born the year after Saussure). The three of them, among others, were leaders in a powerful movement that brought into the twentieth century new kinds of science and scholarship, behavioural and social sciences with their own thinking and methods.

Despite the fact that the Saussurean approach is not universally approved (see the following section for some brief comments on Chomsky's criticism of Saussure), it has shown its strength in its continuing appeal to substantial numbers of linguists and social scientists.

Firth

John Rupert Firth was born in England in 1890 and taught at the University of the Punjab from 1919 until 1928. Returning to England, he held posts in London, first at University College, then at the School of Oriental and African Studies, where he was the Professor of General

Linguistics from 1944 to 1956. Much of Firth's work was in phonology, a field in which he was descriptively and theoretically innovative. (For introductory overviews of this work, see for example Robins 1979, pp. 214–21, or Sampson 1980, pp. 215–23.) But Firth wrote also about meaning and about language in general. Unlike many of his European contemporaries, Firth had extensive experience outside Europe. (In phonology, for example, he was alert to the dangers of assuming that a European alphabetic writing system was a good model of the organisation of spoken language: while it is possible to draw an analogy between the letters of an alphabet and the phonemes or sounds of spoken language, there are significant differences as well as similarities.) Firth also read the work of anthropologists like Malinowski, whose charmingly entitled *Coral Gardens and their Magic* (1935) gave an account of the culture of the people of the Trobriand Islands, in what is now Papua New Guinea. Malinowksi stressed the importance of understanding language in its context and spoke of language as activity, explicitly rejecting the notion that language was a means of transferring thoughts or ideas from one person's head to another's.

For Firth, meaning is function in context, and, consistently with this broad claim, not only words but also grammatical structures and even the sounds of language have meaning. At times Firth seems to equate meaning with use (a word, for example, is meaningful because it serves some purpose in genuine contexts) or with context itself (a word's meaning is the range of contexts in which it occurs). While this has struck – and still strikes – many people as an unusual if not perverse extension of the notion of meaning, what is significant here is Firth's attention to what could be observed, and to genuine usage. Firth takes a theoretical stand not only against the kind of linguistic description which deals with invented examples considered outside any real context, but also against the kind of theoretical mentalism which presents speculations about the contents and workings of the human mind as if they were scientific observations.

The influence of Firth's views is evident in much of British linguistics: he was a major influence on Halliday, and hence in the development of modern systemic functional linguistics (see for example Sampson 1980, pp. 227ff., Martin 1992, p. 4, Eggins 1994, pp. 51–2), and on Sinclair and the development of corpus linguistics (to be explored in detail in Chapters 3 and 4). The development of corpora – the large electronically accessible collections of textual material – has made Firth's seemingly bizarre statements about meaning as use and meaning as context far more believable. Now that it has become possible to track thousands of occurrences of words and phrases, in their

real settings, linguists have begun to see just how informative a record of use in context can be – and how wrong our intuitions sometimes are. As we will see in Chapters 3 and 4, modern corpus linguistics brings a new seriousness to observation of actual usage.

2.7 Cognitive linguistics

In contrast to Saussure and Firth, many linguists writing in the latter part of the twentieth century have been avowedly 'mentalist' or 'cognitivist'. The most famous of these is Noam Chomsky.

Chomsky was born in Philadelphia in 1928. He studied linguistics, mathematics and philosophy and qualified for his doctorate at the University of Pennsylvania, before taking up an academic post at the Massachusetts Institute of Technology, where he became famous not only as a theoretical linguist but also as an outspoken critic of the war waged by the USA in Vietnam in the 1960s and 1970s, and as a writer and speaker on US foreign policy, politics and the mass media. Encyclopedias and dictionaries describe him variously as 'a linguist, writer, and political activist', 'a political observer and critic' and 'one of the leading critics of American foreign policy [since 1965]'. His published books include not only widely read works on linguistics but also political works such as *Manufacturing Consent: the Political Economy of the Mass Media* (with Edward S. Herman, 1988) and *Rethinking Camelot: JFK, the Vietnam War, and US Political Culture* (1993). The titles of these works already give some idea of Chomsky's stance: *American Power and the New Mandarins* was dedicated to 'the brave young men who refuse to serve in a criminal war'; and the phrase 'manufacturing consent' is often quoted by critics of the modern 'free enterprise' mass media.

As with Saussure and Firth, it will be impossible to do full justice here to an influential and widely discussed scholar. (A brief but useful evaluation of the earlier years of Chomsky's contribution to linguistics, psychology and philosophy can be found in Lyons 1970; a later and more critical account is Chapter 6 of Sampson 1980; and Chomsky's more recent views can be found in Chomsky 2000.) Our concern here is with approaches to meaning, and in particular with twentieth-century mentalism and cognitivism, rather than with an overall assessment of Chomsky's work. And it is Chomsky's *Cartesian Linguistics* (1966) which offers us a classic defence of mentalism: the book is significantly subtitled 'a chapter in the history of rationalist thought' and it seeks to draw on and continue the work of the seventeenth-century philosopher Descartes.

In this view, there is a 'fundamental distinction between body and

mind' (Chomsky 1966, p. 32) and the mind and its structure and processes are deemed to be a proper object of study. It is assumed 'that linguistic and mental processes are virtually identical, language providing the primary means for free expression of thought and feeling, as well as for the functioning of the creative imagination' (Chomsky 1966, p. 31). Thus the human mind has a certain structure and certain ways of operating, which in some sense determine – or even *are* – the structures and processes of language itself.

The programme of cognitive linguistics initiated by Chomsky and his colleagues in the 1950s and 1960s proposed a distinction between 'deep' and 'surface' structure in language. At least in the early stages of this programme, deep structure was assumed to have a mental reality closely related to meaning: 'It is the deep structure underlying the actual utterance, a structure that is purely mental, that conveys the semantic content of the sentence' (Chomsky 1966, p. 35). It was also suggested that this deep structure might be universal: 'The deep structure that expresses the meaning is common to all languages, so it is claimed, being a simple reflection of the forms of thought' (Chomsky 1966, p. 35). Those who followed Descartes 'characteristically assumed that mental processes are common to all normal humans and that languages may therefore differ in the manner of expression but not in the thoughts expressed' (Chomsky 1966, p. 96). This universalism is itself tied to the mentalism: 'The discovery of universal principles would provide a partial explanation for the facts of particular languages, in so far as these could be shown to be simply specific instances of the general features of language structure ... Beyond this, the universal features themselves might be explained on the basis of general assumptions about human mental processes or the contingencies of language use ...' (Chomsky 1966, p. 54).

As Chomsky himself sees it, his late-twentieth-century mentalist linguistics thus revives the concerns and perspectives of the rationalists of the seventeenth and eighteenth centuries and links them with modern psychology: 'it seems that after a long interruption, linguistics and cognitive psychology are now turning their attention to approaches to the study of language structure and mental processes which in part originated and in part were revitalized in the "century of genius" and which were fruitfully developed until well into the nineteenth century' (Chomsky 1966, p. 72).

Judged in this cognitivist light, the kind of linguistics which builds on the work of Saussure and Firth (2.6 above) is too sceptical about the mind and mental processes, and too oriented to what is observable 'on the surface'. In Chomsky's own words:

> From the standpoint of modern linguistic theory, this attempt to discover and characterize deep structure and to study the transformational rules that relate it to surface form ... indicates lack of respect for the 'real language' ... and lack of concern for 'linguistic fact'. Such criticism is based on a restriction of the domain of 'linguistic fact' to physically identifiable subparts of actual utterances and their formally marked relations. Restricted in this way, linguistics studies the use of language for the expression of thought only incidentally, to the quite limited extent to which deep and surface structure coincide; in particular, it studies 'sound–meaning correspondences' only in so far as they are representable in terms of surface structure. From this limitation follows the general disparagement of Cartesian and earlier linguistics, which attempted to give a full account of deep structure even where it is not correlated in strict point-by-point fashion to observable features of speech.
>
> (Chomsky 1966, p. 51)

This focus on mind and thought, backed by a dualistic perspective on mind and body, tends to assume that meanings are mental concepts which have real existence in the mind (as opposed to being convenient or theoretical abstractions or constructs). Previous sections of this chapter have already indicated that our view is somewhat different. Like the linguists whom Chomsky criticises, we take it that the distinction of mind and body is an assumption, not a proven fact, and we are indeed sceptical about how much can be discerned within the mind. In fact the mind–body dichotomy represents a particular conception of humanity, a conception that is by no means self-evident and universal.

Firth was clear on this point: 'As we know so little about mind and as our study is essentially social I shall cease to respect the duality of mind and body, thought and word ...' (Firth 1957, p. 19). For Firth and many other linguists of the twentieth century (see Hasan 1987, esp. pp. 117ff., Halliday 1994b), the postulation of mental entities is not well justified and too easily takes linguistics away from its proper concerns with the physical, biological, social and semiotic character of language.

This section has given no more than a thumbnail sketch of some of the theorising of Chomsky and cognitive linguists, and it is certainly not intended as a thorough review of this theorising. Nevertheless it serves no good purpose to avoid or disguise serious differences in theoretical stance which affect modern linguistics. We hope that some indication of the differences between Saussurean and cognitivist linguistics helps to clarify our approach as well as to remind readers that in linguistics, as in most human enquiry, there is no one theoretical position which is taken for granted by everyone. Chapters 3 and 4 will expand and illustrate further the theoretical stance of this book.

2.8 Language and reality

It seems an obvious and necessary truth that language connects with reality, that language is in some sense grounded in reality. Words seem to refer to things that have an existence independent of human language, discourse somehow relates to actions and situations, language at large must be grounded in a world at large.

The fact that it seems self-evident to talk about a 'real world' to which language refers or relates actually has more to do with traditions and habits of talking and thinking than it does with objective necessity. It is customary to talk about words referring to things and about language connecting with reality; this does not mean that this is necessarily the best way of thinking about language and reality. We have already mentioned (2.2 above) the awkwardness of treating meaning as reference, of assuming that all words refer to things. For some words, it does seem quite reasonable to make a connection with a reality that is 'external' to language. But for many others, such a connection is speculative.

Part of being human is to try to make sense of the world and our place in it, and part of this endeavour is ordering and classifying the world, as we perceive and experience it. To a large extent, our language does the job for us. As children learn their first language, they learn categories and classes, usually without being at all conscious of it. We learn words for objects which we see and talk about, and these words imply categorisation: a stick is different from a stone, a hill different from a mountain, a flower different from a fruit, a sheep different from a goat, a pen different from a pencil, a book different from a magazine, and so on. We learn words for colours, which give us a division of the colour spectrum, we learn words for human relationships, such as *aunt* and *cousin*, which bring with them ways of structuring our kinship, we learn verbs like *say, speak, stand, stay, steal, stumble,* among many others, which imply all kinds of distinctions and judgements relevant to human actions and behaviour.

It may be convenient for us to assume that this categorisation is natural and universal. But this assumption will be constantly disturbed, as our experience becomes wide enough to realise that not all human beings live in the same environments, that there is more than one way of defining what flowers and fruits are, that some languages don't have a simple lexical distinction between hills and mountains or between sheep and goats, that some books look more like magazines and some magazines more like books, that communities have different ways of describing kinship, and so on.

Indeed, the more we widen our experience – for example by learning new languages or by empirical scientific investigation of the nature of reality – the more we are forced to recognise that what we call 'reality' or 'the real world' is by no means as natural and self-explanatory as we sometimes like to believe. Consider, for example, the scientific discovery that colour is a spectrum, not a set of discrete colours, combined with the observation that different languages divide the spectrum differently. Descriptions like 'green' or 'blue' and properties like 'greenness' and 'blueness' cannot be considered part of an objective reality: they are at least as much due to the English language as they are to the 'physical' world. Or consider an example already mentioned in 2.6, the difference between the English word *father* and what looks like the equivalent word in some Australian Aboriginal languages: the Aboriginal word refers not just to the person we call *father*, but also to brothers of one's father, and even to male parallel cousins of one's father. There are many other related differences between the English and Aboriginal ways of seeing kinship. In general, the English terms highlight genetic relationships, while the Aboriginal terms focus on social structure. From the English-speaking point of view, my father and mother are individuals who are biologically or genetically related to me. From the Aboriginal point of view, my fathers and mothers are groups of people who are related to me communally or socially, by a structure of obligations and responsibilities.

At least as far back as Aristotle, human beings have also tried to describe their world more deliberately and self-consciously, in ways that might transcend or improve upon 'ordinary' language or 'naïve' thinking. Attempts like these underlie much of what we now call a scientific description of the world. We now have, for example, elaborate classifications of plants and animals that extend – and in some respects clash with – our everyday vocabulary. Thus most Australian speakers of English have a notion of what a 'pine' tree is, based largely on the nature of the foliage (evergreen needle-shaped leaves) and the overall appearance of the tree (with a relatively straight trunk and long branches bending out from it) and perhaps also on its smell and its sticky resin. The word *pine* is part of an informal classification of trees implied by the (Australian) English lexicon: pine trees are different from gum trees, wattle trees, palm trees, and so on. But in modern discourse we also have access to a far more elaborate classification of plants, the naming system sometimes called botanical nomenclature or the Linnean system (after the Swedish botanist usually credited with introducing the system in the 1750s, Carl von

Linné, or in the Latinised version of his name, Carolus Linnaeus). In the Linnean system, pine trees belong to a genus known as *Pinus*, and particular kinds or 'species' of pine are identified in a standard way, by putting the name of the species after the genus, as in *Pinus radiata* (radiata pine) or *Pinus palustris* (longleaf pine).

Now the 'scientific' way of naming plants is not simply a refinement of 'ordinary' vocabulary. For a start, the Linnean classification is based largely on observation of the stamens and pistils of plants, features which are significant in plant reproduction but not nearly as relevant in 'ordinary' discourse as the overall shape and appearance of a plant or its usefulness to humans. Partly for that very reason, there are trees which are not scientifically classified as *Pinus* species but which are nevertheless popularly known as pines – for example the Huon pine (scientific name *Dacrydium franklinii*) and the Norfolk Island pine (scientific name *Araucaria heterophylla*). Similarly, there are 'gum' trees which do not belong to the *Eucalyptus* genus (such as the Sydney red gum, *Angophora costata*) and lilies which do not belong to the *Lilium* genus (such as the belladonna lily, *Amaryllis belladonna*).

Since the eighteenth century there has been an enormous expansion of taxonomies. The nomenclature of plants and animals are just two of the most widely known examples. Other fields in which classificatory naming systems have been developed include geology and mineralogy, anatomy (names of muscles, nerves and so on), medicine (names of diseases, surgical procedures, and so on) and chemistry (names of chemical compounds). Indeed, many large industries have created their own nomenclature, such as an organised set of names for tools and procedures, or a systematic classification of products, components and spare parts.

Many of these taxonomies are supervised and regulated, by a company or an industry or by some international body like the International Union for Pure and Applied Chemistry, in ways that are unthinkable for everyday discourse. (Compare our earlier remarks on prescription and regulation in 2.4 above.) In the twentieth century, terminography or terminology processing (see e.g. Sager 1990, Pavel and Nolet 2002) became a field in which people could train and work. Terminologists may collect information on specialist terms, may provide information, whether in published glossaries or terminological databases or through an advisory service, and may provide advice and recommendations on terms and their use. They may be employed by companies and industries who maintain databanks of technical terms, or by publishers, or by bodies such as the European Union or the government of Canada who maintain large terminological resources

particularly to support translation work. (If we include the many people working in non-English-speaking countries in agencies that coin and promote indigenous terminology, there must be far more people now employed in terminological work than in conventional lexicography.)

The classification enshrined in a taxonomy is (in theory at least) rigorous, and the naming conventions are precise and strict. For example, any species of plant can be placed within the 'Plant Kingdom' which is in turn divided into phyla, classes, orders, families, genera and species. The example below shows the classification of one species of pine tree mentioned earlier. The use of Latinised forms ('Plantae', not 'plants', 'Coniferales', not 'conifers') is conventional and highlights the distinction between scientific description and everyday language. Note also the conventions governing the mention of a species: both genus name and species name are written in italics, the species name follows the genus, and the genus name takes an initial capital, while the species name is always given a lower-case initial letter.

Kingdom	Plantae (plants)
Phylum	Tracheophyta (plants with a vascular system)
Class	Pteropsida (plants with leaves with branched venation)
Order	Coniferales (trees and shrubs producing bare seeds, usually on cones)
Family	Pinaceae (trees with needle-shaped leaves, including firs, larches and spruces, as well as pines)
Genus	*Pinus* (pine trees, comprising about a hundred species)
Species	*Pinus radiata* (radiata pine, also known as insignis pine or Monterey pine)

Here are two more examples, first another plant, the musk rose (*Rosa moschata*) and then, from the animal kingdom, the silver gull, the common seagull of Australia (*Larus novaehollandiae*).

Kingdom	Plantae
Phylum	Tracheophyta
Class	Angiospermae (plants with their seeds enclosed in ovaries; flowering plants)
Order	Rosales (families of flowering plants incl. cherry, plum, strawberry, as well as roses)
Family	Rosaceae (flowering plants with typically five-petalled flowers)
Genus	*Rosa* (roses)

Species	*Rosa moschata* (musk rose)
Kingdom	Animalia (animals)
Phylum	Chordata (animals with vertebrae or a notochord)
Class	Aves (birds)
Order	Charadriiformes (families of gulls, puffins and waders such as curlews and plovers)
Family	Laridae (gulls and terns)
Genus	*Larus* (gulls)
Species	*Larus novaehollandiae* (silver gull, in Australia usually referred to as gull or seagull)

Conventions such as we have just mentioned – the use of italics and so on – are by no means obvious. They can be enforced reasonably successfully, however, precisely because the nomenclature is used mostly in professional writing, subject to careful editing, as in scientific journals, technical reports and textbooks.

The discrepancies between such taxonomies and everyday language may be considerable. We have already mentioned pine trees which are not species of *Pinus*, gum trees which are not eucalypts and lilies which are not *Lilium*. In general, taxonomies serve to identify and classify large numbers of items: many of these items may be rarely if ever talked about by most people and the criteria by which they are classified in the taxonomy may also be marginal in daily discourse. Thus roses belong botanically in the genus *Rosa*, within the family *Rosaceae*. This family happens also to include blackberry and strawberry plants as well as the (often decorative and ornamental) herbs and shrubs of the genus *Spiraea*. But this scientifically established family of plants does not have any relevance in everyday discourse. Indeed, most people find it surprising that such a diverse group of plants should form one family. Similarly, it goes against habitual discourse to say that, botanically, a tomato is a fruit rather than a vegetable, or indeed that nuts are fruits.

This brings us back to the question of an objective description of reality. It is clear that nomenclatures of the kind developed for describing and classifying animals and plants and chemicals serve an important purpose: they are generally more comprehensive than everyday language, they are based on careful and often highly detailed observation, and they may bring with them valuable insights from empirical research. To that extent, a scientifically validated taxonomy may be closer to reality, or more revealing of reality, than everyday language.

Nevertheless, this does not justify the further step of claiming that everyday language is defective, misleading or in need of reform. In

daily life, the categories of everyday language are likely to be more useful than a scientific nomenclature. The everyday English distinction between fruit and vegetables may not be entirely scientifically 'correct', but it is highly relevant to our eating habits and shopping practices. If I am planning meals and making up a shopping list, thinking perhaps about salads as light meals, or about cooked vegetables to accompany other food, or about desserts of fresh fruit, then it makes sense to think, as speakers of English habitually do, in terms of everyday categories. For my purposes, fruits do not include tomatoes or nuts, and it would be foolish and inefficient to suppose that they ought to. If I am asking a friend about fruit currently available at the market, or looking for fruit in a greengrocer's shop, or offering my guests a choice of fresh fruit to eat, none of us should feel any need to defer to a botanical classification based on careful investigation of plant reproductive systems.

Moreover, it should not be assumed that scientific taxonomies, once developed, reveal objective truth once and for all. The botanical and zoological nomenclatures, for example, are always open to revision and some areas of the taxonomies remain controversial. Sometimes a simple renaming has proved necessary: when the Australian platypus was first described scientifically, in 1799, it was given the species name *Platypus anatinus*; but it turned out that the term *Platypus* was already in use for a group of beetles, and a new genus name *Ornithorhynchus* was devised, so that the platypus is now described as *Ornithorhynchus anatinus*. Sometimes the taxonomy itself has had to be extended. Linnaeus and his contemporaries in the eighteenth century probably believed that species of plants were invariant and invariable; subsequent research, including the development of evolutionary theory and empirical studies of diverse environments around the world, has led to a more flexible view. The plant taxonomy now includes subcategories (such as subspecies) as well as varieties within species. And sometimes, as a result of further research, a particular plant is relocated in the system, say from variety to subspecies or from subspecies to species. (The example given above, of the place of the silver gull in the animal kingdom, should actually include a suborder Lari, below the order Charadriiformes and above the family Laridae, and a subfamily Larinae, below the family Laridae and above the genus *Larus*. For further discussion of the provisional nature of scientific taxonomies, see 3.4.)

The terms of a scientific taxonomy are in some ways more like a naming system than a vocabulary. In the Linnean plant nomenclature, for example, it is normal to refer to genus and plant 'names', and the typical genus species name, say *Pinus radiata*, is sometimes likened to a

surname plus given name. Nomenclatures also tend to be recorded and explained in encyclopaedias and technical publications rather than in general-purpose dictionaries. Tendencies such as these inspire a tradition of distinguishing between encyclopaedic knowledge and linguistic knowledge, between 'knowledge of the world' and 'knowledge of language'. Thus, it may be argued, knowing the names of individual people, knowing historical facts and knowing about particular objects are all part of knowing about our world, and not part of our language. And it has to be said that there are things we know which are, on the face of it, quite outside language: telephone numbers, addresses, names of people and places, historical dates, and so on. Obviously, it is possible to be a fluent and competent speaker of English without knowing who the premier of Tasmania is, which is the largest city in California or when the kingdoms of England and Scotland began to be ruled by one and the same monarch.

But the line between factual knowledge and linguistic knowledge cannot be drawn sharply. We have referred earlier to the way in which names can become words (e.g. *boycott*, *sandwich*, 2.2 above). Some names of people and places – and 'facts' about them – are so well known in a community that users of the language do assume that everyone knows them. An old Australian idiom, *to do a Melba*, 'to keep saying goodbye, to make repeated farewells', drew on common knowledge of the singer Dame Nellie Melba and her several 'farewell' appearances. Legendary figures may figure in discourse as if they were common nouns, like King Canute, who is supposed to have commanded the tide to turn, unsuccessfully of course, but deliberately so, in order to demonstrate to his followers that there were limits to human power, even the power of a king. Thus a fiction writer says of a character that he was 'Canute controlling the waves' and assumes that readers will know the story of Canute so that they grasp the ironic meaning. In fact, the meaning of 'Canute' may have generalised to anyone who resists or denies evidence – or even to the act of resistance itself, as in the phrase 'doing a Canute'. On 24 July 2002, the Melbourne *Age* had a headline in its business section 'Bush does a Canute with falling US stockmarkets'. The article reported President George W. Bush's claim that the future was 'going to be bright', despite, in the words of the article, 'much evidence to the contrary'.

It may not be essential to one's ability to speak English to know who the first president of the USA was or who the prime minister of England was in 1945. But discourse does sometimes assume such knowledge in its meaningful progress. Some historical figures do carry meaning. An American writer refers to 'George Washington's cherry

tree': according to the story, the young George chopped down a cherry tree and when questioned by his father, confessed to the misdeed, saying that he was unable to lie. The writer assumes that most or all readers will know the background story. Or, to take the example of the British wartime prime minister, a search of a few corpora for references to Churchill naturally produces many references to the man – in historical accounts, political discussions, and so on – but also yields some uses where the name is used descriptively, again presupposing that author and audience have some shared understanding or image of the man. For example, someone is described as 'of Churchillian mien'; a politician is recorded as having told reporters that a recent 'stirring' speech was 'his Churchill speech'.

In fact there is no way of drawing a principled distinction between knowledge of the language – the lexicogrammar – and extra-linguistic knowledge. Not long ago I was walking out of a particularly complicated car park in Canberra when a car pulled up beside me. The driver asked me if I could point him towards the exit – *any* exit – and added that he'd been driving round the car park for some time and had 'done more miles than Burke and Wills'. Now I'm not sure whether I have ever heard that phrase before, and I don't recognise this as a familiar Australian idiom; but I do know (as probably most Australians do without having to look them up) that Burke and Wills were explorers who undertook an ambitious journey across Australia from south to north and then back again, but died of starvation before completing their expedition. Presumably the man assumed I knew that much, to be able to share in his self-deprecating joke about arduous and fruitless travels across a car park. (The Bank of English corpus records a couple of idiomatic uses: 'She's seen more Australia than Burke and Wills' is similar to the phrase I heard, while 'Waugh and Healy [Australian cricketers] are as much an Aussie institution as Burke and Wills' at least implies that Burke and Wills are well known in Australia.)

An example like this illustrates the uncertain edges of social discourse. Perhaps the man who spoke to me came from an area of Australia where his turn of phrase was a familiar idiom to most people. I might have simply been ignorant of his usage, just as any of us can easily find ourselves out of our depth when we move into a community where we are not accustomed to local usage. Perhaps he was simply an individual with a liking for a certain kind of Aussie imagery, and I will never hear the phrase again. Perhaps the phrase is in fact more widely used than I realise, and it's just that I have failed to come across it. Perhaps even my mention of it in this book might cause it to be quoted

more often. Whatever the possibilities might be, the eventual status and meaning of the wording will depend on further usage, on uses which bring the phrase into play as an increasingly well-known idiom, or on absence of use which will ensure that the phrase does not enter a pool of linguistic resources nor find its way into dictionaries and phrase books.

For words are first and foremost elements of text, elements occurring in actual discourse, not isolated items listed in a dictionary (2.2 above). Traditional lexicographers have separated linguistic knowledge from encyclopaedic knowledge by a process of decontextualisation, trying to describe the meaning of words in isolation from their contexts. In this view, if we could detach from a word all its links to relevant contexts, we should be left with the isolated unadulterated meaning. But access to modern corpora has made it possible to study texts far more intensively, and corpus linguists are now able to show the semantic cohesion of textual segments. If we are no longer limited to single words detached from their contexts, if we do away with decontextualisation, we need not insist on the distinction between linguistic and encyclopaedic knowledge.

What we normally call encyclopaedic knowledge is in fact almost always discourse knowledge. For most of us nowadays, everything we know and are able to know about King Canute, George Washington, the explorers Burke and Wills, and Winston Churchill, is based on texts. Even photos and film and video mean relatively little without accompanying text. If we consider how much our encyclopaedic knowledge owes to our discourse knowledge, the distinction virtually disappears. This too is a topic we will revisit in Chapters 3 and 4.

2.9 Language and languages

The diversity of human languages is an inescapable truth. Some languages, such as those of Western Europe or the group of languages sometimes called the 'dialects' of Chinese, do show similarities, because of common ancestry or a history of contact, but many languages are strikingly different from each other. Even where languages have much in common – as English and German do, two languages which are historically related and which show many cultural similarities, including a long tradition of being influenced by Latin and French – differences are still of some consequence. Modern English and German are not mutually intelligible and it takes considerable time and effort for adult speakers of the one language to learn to function reasonably well in the other.

Taking a wider sweep across the world, languages differ more radically than English and German do. Phonetically, some languages have sounds and patterns of pronunciation which seem quite impossible to speakers of other languages. The click sounds of some languages of southern Africa seem odd and difficult to those who have not grown up speaking such a language; needless to say, there is nothing difficult or bizarre about these sounds to those who do habitually use them. The dental fricative consonant at the beginning of English words like *thin* and *thorn* is a constant challenge to those whose mother tongues do not have the consonant, while the various uvular and glottal consonants of Arabic strike a speaker of English as impossible to pronounce.

Grammatically, the patterns of one's own language become so habitual that alternatives seem perverse and sometimes beyond learning. Hence we hear people who have learned English as a second language saying things like 'you like coffee, isn't it?' (instead of 'you like coffee, don't you?') or 'I'm working here since 1995' (instead of 'I've been working here since 1995'). In so doing, they are simply following the patterns of another language and failing to follow those of English. And of course speakers of English learning other languages make other – but comparable – errors. The patterns of one's own language are 'natural', ingrained enough to interfere systematically with the learning of different patterns.

What is true of pronunciation and grammar is also true of meaning. Even related words which look or sound similar often differ in meaning. An example is a word already referred to more than once in this chapter (2.2 and 2.5), namely *patron.* Commonly used in English to refer to the customers in a hotel or restaurant, the seemingly equivalent word in French means 'boss' rather than 'customer'. Other deceptive differences between French and English include French *large,* which corresponds to English 'broad' or 'wide' rather than to 'large', and French *sensible,* which is closer to the meaning of English 'sensitive' than to 'sensible'. In French, 'sensitive skin' is *peau sensible,* and a sensitive or tender spot might be described as *l'endroit sensible.* But note how the words and meanings of different languages do not line up as perfect equivalents across languages: when the French *endroit sensible* is used metaphorically it is probably better translated into English as 'sore point' rather than 'sensitive spot'.

To take an example from Dutch, the word *serieus* looks and sounds to an English speaker as though it ought to correspond to English 'serious'. And in a sense it does, in some contexts, particularly where a contrast is implied with humorousness or lightheartedness, as in a

person looking a bit serious or a happy occasion turning out to be too serious. But this word is not used of, for example, a 'serious problem' or 'serious illness'. Here the relevant Dutch word is *ernstig*. You might shrug off a minor injury as *niet ernstig*, 'not serious', or you might be accused of *(iets) niet ernstig nemen*, 'not taking (something) seriously'.

More seriously, whole areas of meaning are differentiated and elaborated in some languages but seemingly unimportant in others. Some languages, like Dutch and Italian, have morphological devices for expressing diminutives which are used to signal not just smaller size of an object but also (sometimes) endearment and informality. Compare Dutch *kast* 'cupboard, wardrobe', *kastje* 'little cupboard, locker', *kop* 'mug', *kopje* 'cup', *hand* 'hand', *handje* 'little hand'. But these so-called diminutive forms may be used in various ways: for example *handje* may be used in talking about a young child's hands but it is also the appropriate form in the metaphorical 'lend a hand' with a job. The informal or casual effect of diminutives is also evident in a request like *mag ik een sigaretje van je?* 'may I (get) a cigarette from you?', where the diminutive form *sigaretje* of course does not indicate that the speaker is asking for a small cigarette but is rather a device to downplay the request (somewhat as an English speaker might ask, strictly inaccurately, to 'borrow' a cigarette, or might add the word 'just', as in 'could I just ask you ...'). Some languages have similarly extensive use of diminutives – Czech and Italian, for example – but while English does have some comparable morphology, as shown by *book* and *booklet* or *dog* and *doggie*, it is not nearly as widely used, nor used with the same elaboration of interpersonal meaning.

A language like English has an infinitely expandable set of numerals and considerable resources for talking mathematically – ways of talking about addition and multiplication and solving equations and so on. By contrast Australian Aboriginal languages have relatively few terms for numerals and little comparable resources (although with the arrival of a more technologically-oriented culture in Australia they have started to acquire such resources). And so one could go on, comparing the more elaborate semantics of Australian Aboriginal kinship and clan structure with the simpler resources of English, among many other possible examples.

Languages do influence each other semantically, and this is an important observation for two reasons. First, it underlines the point that languages differ from each other, for if they were not significantly different, there would be nothing significant for other languages to imitate or acquire. Second, it is a reminder that while differences are real enough, languages are not always separated by impenetrable

boundaries or yawning chasms. Just as individuals can learn foreign languages, so cultures can acquire the characteristics of other cultures – although it must be said that they never seem to end up identical.

In Australian Aboriginal languages there is usually a verb which refers to hitting or striking with an implement, potentially hurting or even killing, as in clubbing or spearing an animal. (A different verb is used of hitting someone or something with a missile such as a stone.) In Aboriginal English, the word *kill* is now used regularly not with the sense of causing to die or ending life, but with the sense of attacking or hitting or beating up. The history of languages is full of such semantic readjustments, often in conjunction with major cultural changes. When Christianity came to England in the seventh century, not only did Old English adopt Latin words already in Christian use (such as *maesse* 'mass' from Latin *missa*, and *scrin* 'shrine' from Latin *scrinium*) but Old English words took on new meanings. The Old English word for 'build' started to be used to mean 'edify', on the analogy of Latin *aedificare*, which already had the sense of 'build up' or 'edify' as well as 'build' in a more material sense. The Old English *halig* 'holy' was probably derived from a word to do with health or wellbeing (compare Modern English words like *hale* and *whole*) but it came to be used in a specifically Christian way. In fact in the Old English period, the plural of the word was used to translate the Biblical 'saints', i.e. 'the holy ones'. This usage survives in certain names such as 'Allhallows' (All Saints) and most notably 'Halloween' (Allhallows Eve), but, in another semantic adjustment, the word 'saint' (Old English *sanct*, from Latin *sanctus*) has now taken on the Christian sense of 'a holy one'.

Just as Latin has influenced English, so elsewhere languages which were in one way or another dominant or prestigious, like Arabic as the language of Islam, or English as the language of the British Empire, have left their mark on many other languages. Thus Arabic has influenced Malay (now Indonesian and Malaysian) and Urdu, and English has influenced many languages of sub-Saharan Africa.

When the Netherlands ruled what is now Indonesia as the Dutch East Indies, the Malay that was widely used in the area took over many words from Dutch, many of them still evident in modern Indonesian, from *rem* for the brakes of a vehicle to *bank* for the financial institution, from *dokter* for a medical doctor to *gang* for a lane or passageway. As English words extended their meaning in the Christianisation of England, so Indonesian words acquired wider uses in the period of Dutch colonial rule, as illustrated by the word *pusat* which refers to the navel or to the centre of a (more or less) circular pattern like a thumbprint, but now also has a far wider range of uses for abstract and

institutional 'centres' such as 'centre of gravity' or 'language centre'. As always, the semantic patterns of language shift and adjust. To take another example, the Indonesian word *rumah* 'house' now enters into a series of specialised combinations such as *rumah penatu* 'laundry' and *rumah sakit* 'hospital' (compare Dutch *washuis* 'laundry', *ziekenhuis* 'hospital', based on the Dutch *huis* 'house').

Given the evident diversity of human languages and cultures, and the ways in which they interact, often influencing each other and copying from each other, but never quite ending up the same, it makes sense to say that languages have their own semantic strengths, their own areas of richness and elaboration. It is this that often makes learning another language a rewarding experience, an experience which changes one's horizon and opens up new views of the world. And this may make it seem all the more surprising that anyone has ever entertained the notion of universal grammar or universal semantics. In fact there have been a number of attempts to generalise across languages, to find a kind of ideal model or to find something that could be said to underlie all human languages. An arrogant but not unknown way of denying or minimising language differences is to focus on one or a few languages and to regard any language that is not similar to them as deviant or degraded. European respect for Latin has sometimes led to this kind of view, especially when accompanied by an imperialistic willingness to dismiss many non-European languages as not really fully-fledged languages. But there have also been more thoughtful and more scholarly attempts to define some kind of universal grammar or universal semantics. We have referred earlier (2.7) to Chomsky's postulation in the 1960s of a 'deep structure' that might be common to all languages. Chomsky looked back to those who had thought along similar lines – for example the grammarians working at the convent of Port Royal in France in the seventeenth century, who theorised that the categories and structures of grammar could be related to universal logic or universal thinking.

Universalism, as a theoretical position on language, usually rests on one of two strategies. One is to postulate something which is actually not observable, like a set of 'universal concepts' or Chomsky's 'deep structure'. Universal concepts, for example, could exist only in human minds, or perhaps in some common human consciousness, if there is such a thing. We cannot observe and record what is in the human mind in the same way that we can observe and record human behaviour, in particular what people say or write. This is in itself no objection to universalism as a belief, since most of us have beliefs of one kind or another, whether belief in God or in fellow humans or in ghosts or in

good or bad luck, or beliefs about the future or about what is valuable and significant in human living. But it is important to recognise the role and nature of belief here. Those who do believe in universal concepts, underlying the semantics of all languages, will argue that one can only put forward theoretical postulates and then check their explanatory power or test them against the evidence, for example by looking for their consequences in observable behaviour. It then becomes necessary to face questions about what exactly constitutes a valid check or test of one's theoretical position, and not simply to begin to take theoretical hypotheses as probable or self-evident. Of course one can live by faith – as we all do to a greater or lesser extent – but faith needs to be acknowledged as faith, not presented as indisputable scientific finding.

The other strategy found in universalism is, in one way or another, to set up a supposedly universal framework or inventory from which all languages make some kind of selection. Thus one might claim that there is a vast inventory of universal concepts or components of meaning, including presumably very general ones like 'human' and 'animate' and 'concrete' (which might be semantic components of many words in many languages) as well as much more specific ones that would differentiate (semantically) a snail from a slug, a mountain from a hill, saying from telling, hitting with an implement from hitting with a missile, and so on. The fact that languages differ from each other semantically – for example Dutch makes no lexical distinction between 'snail' and 'slug', just as English does not have separate lexical items for 'hit with an implement' and 'hit with a missile' – is then allowed for by saying that each language makes its own selection from the universal inventory. This is an interesting ploy. On the one hand it recognises the difficulty of the universalist position, for the 'universal' inventory is no longer genuinely common to all languages. On the other hand it raises the question of what kind of existential status this inventory has. Since the inventory is by definition larger or more comprehensive than the semantics of any one language, it must exist beyond or above specific languages. If it resides in human minds, then part of it is redundant or irrelevant to the language(s) known to any individual mind, which must surely put that part of it well beyond any kind of empirical verification. And if it is not confined within individual minds, where is it to be found and how can we access and study it?

Much has been written about languages and their differences and similarities. What we have said here goes only some way towards justifying our reluctance to postulate universal grammar and universal concepts and our preference for a more cautiously descriptive

approach to linguistic behaviour. We emphasise again that we are not suggesting that languages are so different from each other that they constitute totally different worlds, cut off from each other. We do acknowledge that languages show similarities. But except where languages happen to be quite closely related, their similarities cannot be grounded in a core vocabulary or an underlying and invariant set of concepts or anything as temptingly concrete or specific as that. Rather, the similarities are better understood in terms of functions and general design rather than in terms of inventories of items or components or rules.

The analytical and theoretical problem here is not unique to linguistics or semantics, for it affects most of our study and understanding of humans and their behaviour and institutions. It is rather as if we set out to see what was common to wedding ceremonies around the world; or what was universal about food and eating; or what was common to all the world's practices of religious worship. We might try to find the objects common to weddings (such as rings or flowers or special clothing) or we might look for a universal underlying structure (for example with people arriving, participating and departing in a certain typical sequence). But if we really pursued such a project along these lines, we would soon find it futile. Rings and bouquets and wedding cakes are indeed part of many weddings in many countries but they are not universal. They were certainly not part of most marriage ceremonies in Australia or Papua New Guinea or the Amazon Basin before the arrival of white colonists and their culture. In fact, the very notion of 'wedding ceremony' already suggests a European perspective on the event. If we wanted to assess universality in a more open-minded and realistic way, we would do better to step back from our immediate experience of weddings and to start to think in a more broadly functional way: how human beings form alliances or partnerships for sexual intercourse and parenting, how these partnerships are integrated into wider social structures, whether and how these partnerships need to be endorsed or recognised by other members of the larger society, and how these partnerships are entered into and characterised, in theory or in practice, by commitment and loyalty. Even here, we are still talking in English, using modern English words like *parenting* and *partnership*, which already project a certain light on what we think we are looking for and talking about. But at least at this point we have lifted our sights above a mere search for shared objects and entities, a search which is bound to fail, and we have started to think in a more general and productive way about what it is that characterises people and their social behaviour as human. The wording used here may not satisfy

everyone – I can think of several lines of objection to the phrase 'partnerships for sexual intercourse and parenting' – but if it is hard even to frame what we are studying, that is precisely because we are facing the genuinely rich complexity and diversity of humankind.

Much the same could be said about food and eating, or about religious worship. There are few if any foodstuffs which are truly universal. Even if certain items such as sandwiches and hamburgers are now obtainable in some kinds of hotels and restaurants around the world, they are definitely not consumed by everyone everywhere. Even items that are very widespread – say bread – take different forms and shapes and are eaten in different ways. (Indian bread typically has a different appearance and function from French bread, for example.) What might be universal is rather the human need to eat, the need for substances such as starch and sugar, human enjoyment of eating, and so on. Likewise with the practice of worship in settings as diverse as the mosque, the synagogue, the temple, the church and the chapel: universals are found not in the objects and components that are present in worship but in the ways in which humans function as worshipping beings.

So also with language. If there are universals of language, they are best approached from the perspective of how language functions in human life and how it serves human purposes. All languages seem to be systems for making meanings, meanings encoded in wording which is expressed in spoken form (or, in the case of many languages, spoken and written form). All languages seem to provide ways of talking about things or entities and, by contrast, ways of talking about events or processes or relationships. (This distinction is often related to the grammatical distinction between nouns and verbs, but the relationship is by no means a direct and simple one.) All languages seem to project both experiential or representational meanings (relating to what can be said about the world and facts and events and so on) and what can be called interpersonal meanings (relating to how speakers or writers are interacting with hearers or readers). This is a quite different approach to universals from one which seeks to find a common core vocabulary or a universal set of concepts. (For more detailed exposition of this kind of functional perspective on language, see Eggins 1994, esp. Chapter 1, or Halliday 1994a, esp. pp. xvii–xx, xxvi–xxxv.)

2.10 Translation

Translation from one language to another is sometimes described as if it were a process of rewording the same meaning, a process of finding

new words to express the same meaning. While this may sometimes be a convenient way of describing the process, and good translators do have a commitment to what we might call loyalty to the original, there are several objections to conceptualising translation as if it were a process of taking meaning out of the words of one language and re-expressing it, unchanged, in the words of another language.

In the first place, most translators know from experience the rashness of claiming that they are preserving meaning unchanged. As we have seen in the previous section of this chapter, meaning is not isomorphic across languages. To take a simple example, if you translate the English word *sister* into the Australian Aboriginal language Pitjantjatjara, you have to choose between a word meaning 'older sister' and one meaning 'younger sibling'. (There is of course another Pitjantjatjara word meaning 'older brother', but there is no lexical distinction between 'younger sister' and 'younger brother'.) You cannot simply transfer 'the same meaning'. Information about the relative age of the sister may be implicit in the English text or may be entirely unmentioned and irretrievable. And even if you can establish that the sister is in fact a younger sister, you still won't be expressing exactly the same meaning in the relevant Pitjantjatjara word, since the sex of the sibling will now become as invisible as relative age is in English. Of course you can make a special effort to bring information to the fore, in both English and Pitjantjatjara: for example in English it is perfectly possible to use expressions like 'older sister' or 'younger sibling', as we have just done above; but the words are still not exactly equivalent. English *sibling* is not a word which is normal in the English-speaking world in the same way as the Pitjantjatjara words in the Pitjantjatjara community. It belongs to anthropological or sociological discourse (or to discussions of translation!) rather than to talk of family and friends. I sometimes heard my father talk about his brother and sister, but never about his 'two siblings'; and I have sometimes heard my wife refer to her sister and (two) brothers but never to her 'three siblings'. In fact, even at this point, we have not exhausted the problem of translation, since the Pitjantjatjara words actually refer not only to brothers and sisters but also to parallel cousins (children of mother's sisters and children of father's brothers). But enough has been said to indicate that even apparently simple words cannot be assumed to match each other across languages.

This example has been a little too abstract. In real translation work, one has a context and purpose (say translating a service manual or interpreting in a court of law or assisting in a land claim) and problems have to be solved in their context. Let's take another example and

place it in context. Suppose I want to send a letter to a number of people around the world. Let's say it is a letter inviting them to contribute a paper to a journal. As I draft this letter in English I will have to make a decision on how to begin it. There are quite a few options. If I know all the names and can adapt each letter, I might begin each letter with a personal address, choosing among options like 'Dear Professor Jones' or 'Dear Susan' or 'Dear Sue'. If I am unable or unwilling to make each letter specific in that way, and am prepared to be rather formal, I can choose among options like 'Dear Colleague' and 'Dear Sir or Madam'. I can even take the option of omitting such an opening entirely. Without going through all the reasons why some people dislike letters beginning 'Dear Sir or Madam' and some dislike letters without any salutation at all, let us say that I opt to begin my letter 'Dear Colleague'.

Now I want to translate my letter, and I want it to be 'the same letter' in several languages. If I translate the letter into Dutch, I now have options which were not available in English. At the point where 'Dear' occurs in the English there are two possibilities in Dutch: *Beste*, which is appropriate for friends, and *Geachte* which is typical of official or business correspondence. (There is actually a third option, *Lieve*, but this is familiar and affectionate and not an option to consider in this context.) Thus there is no simple way to match the generality of English 'Dear ...', which can be used quite intimately ('Dear Susie') as well as very formally ('Dear Madam'). The Dutch version of the letter forces a choice between a more familiar option and a more formal one. Even in this small detail, we cannot claim that the Dutch letter will have exactly the same meaning as the English one.

In the second place, it is not at all clear that we have any way of separating meaning from wording. To hark back to Saussure's classic metaphor, a linguistic sign is like a sheet of paper, with 'thought' (or a concept or meaning) on one side and its expression (the form or actual word) on the other (2.6 above). One cannot isolate either side from the other (Saussure 1972, p. 157). What translators actually do when 'discovering' or 'analysing' the meaning of a text involves paraphrasing within the relevant languages rather than thinking in any genuine sense 'outside' the languages. Thus, when translators ponder what the text really means or search for the right words in the translation, they range over words of similar or contrasting meaning, over phrases that might expand the meaning or words that might condense the meaning, both in the language of the text in front of them and in the language into which they are translating. What they do not do, as far as we can understand the process, is to engage in some kind of

abstract thinking that is independent of both languages. Consider the example we have just been through, of translating 'Dear Colleague' into Dutch. The translator, aware of the context, runs through options in both languages and thinks about what sort of equivalence might be achieved. It seems highly unlikely that translators engage in any sort of higher level abstraction in which they categorise kinds of 'dearness' (whatever that might be) independently of both Dutch and English.

Third, suppose that we could somehow separate meaning from wording. How could we then express meaning, other than through language itself? The suggestion that we can extract meaning from the words of one language and then put it into the words of another, poses the question of where this meaning is and how it is represented when it is, so to speak, in between the two languages. In some cases, depending on the kind of text they are translating and its meaning, translators may be able to visualise objects and situations that are referred to, but even here it is doubtful whether they do this in a way that is independent of language. Is it really desirable, let alone possible, for a translator to imagine an agricultural tractor or a fluorescent lamp or a voicemail system without thinking of descriptions of it in language?

The examples that we have considered should make it clear that scepticism about metaphors of 'extracting' and 'transferring' or 'rewording' meaning is not the same as saying that translation is impossible. Experienced translators work quickly and skilfully with their linguistic material but they do not deceive themselves that they handle meaning detached from texts, nor do they claim to translate in such a way that their output is a perfect semantic match of the original text.

As Haas puts it

> The translator ... constructs freely. [A translator] is not changing vehicles or clothing. [A translator] is not transferring wine from one bottle to another. Language is no receptacle, and there is nothing to transfer. To produce a likeness is to follow a model's lines. The language [the translator] works in is the translator's clay.
>
> (1962, p. 228).

This page intentionally left blank

3 Language and corpus linguistics

Wolfgang Teubert

3.1 Are all languages the same?

'According to Chomsky, a visiting Martian scientist would surely conclude that aside from their mutually unintelligible vocabularies, Earthlings speak a single language' (Pinker, 1994, p. 232). Indeed, if we discount the meaning of words, sentences and texts, our natural languages share many characteristics. They are linear. Utterances have a beginning and end, and between beginning and end we find a string of sounds or of characters, perhaps ideographic as in Chinese, or alphabetical as in most European languages. This is, of course, also the case for sign languages. An utterance in a sign language is again a string, in this case of signs such as hand and finger movements and facial expressions.

Utterances differ from pictures. Utterances are one-dimensional, pictures are two-dimensional. Even if we try to describe a picture, the description will be inherently one-dimensional. Linearity would also be a characteristic of the language of the visiting Martian scientist. All languages are systems for signifying content. Each utterance has a content. But the content is not the utterance. The utterance is a sequence of signs which represent the content, which stand in place of the content. The utterance 'a Martian scientist visits Earthlings' can be said to represent an image, a photograph or a mental image which is two- or even three-dimensional. But the utterance is always a one-dimensional string of signs. John Sinclair, one of the pioneers of corpus linguistics, is fond of repeating what he believes to be a quote of the grammarian E. O. Winter that 'grammar is needed because you cannot say everything at the same time'. This is certainly the reason why all natural languages need grammar, and perhaps also why these various grammars can be described if not in identical, then in very similar terms.

Is this what Noam Chomsky meant (Chomsky 1957)? Not quite. Chomsky argues that all humans share the same language faculty, an

innate faculty that regulates the ways signs are to be organised so that they become utterances. This is what is called grammar. In Chomsky's view, the innate language faculty shapes the grammar. This is not to say that all languages share the same grammar, not even on a deeper level. Today, in his minimalist programme, Chomsky sees the language organ as an apparatus that gives limited options. Adjectives, for example, can precede the noun they modify, or they can follow it. But all languages have adjectives and nouns and several other parts of speech. They are universal, they are shared by all human languages. So, and this is the important point, the language faculty is contingent, i.e. it happens to be the way it is, but it could have been different (and the language faculty of Martians might be different). The philosophical problem connected with this stance is that its credibility depends on conceiving of a convincing language, a language that could exist but does not exist – a language that does not comply with the settings of the language organ but is otherwise, in functional terms, equivalent to existing natural languages.

Chomsky's views on universal grammar (in a more recent version than referred to earlier in 2.7) are found in his book *New Horizons in the Study of Language and Mind* (Chomsky 2000, pp. 7–15). Whether he has succeeded in presenting his case convincingly is a matter of contention. Geoffrey Sampson (1997) in *Educating Eve: The 'Language Instinct' Debate* shows that there is evidence to the contrary in respect of many of the language features that Chomsky and Pinker claim as universals.

Traditional linguistics has been good at describing how syntax, morphology and inflection work. There is a set of basic assumptions, most of which have been around since classical times and which are used for describing any language that linguists stumble across. These assumptions include the facts that there is an entity we call a sentence, another entity we call a clause, that there are subjects, objects and predicates, and that there are words. There are different kinds of words, so-called parts of speech (from Latin *partes orationis*), featuring prominently among them: nouns, adjectives, verbs and adverbs (the big four), and less prominently others, such as pronouns, determiners, prepositions, and depending on the language or the particular grammatical theory, a few more or many more. In a language such as English, a word can come in different forms. The noun *table*, for example, can be a singular form (*table*) or a plural (*tables*). In many European languages a finite verb can be characterised by the properties person and number (e.g. first-person singular as in English 'I laughed', or first-person plural 'we laughed'), tense (e.g. past tense, present tense), mood (e.g. indicative, subjunctive) and voice (active,

passive). Words can be combined to form larger units such as noun phrases, or verb phrases, or other kinds of phrases, and several phrases can be put together to form a clause, or even a sentence.

There are of course differences in the details of grammatical description and theory, and all these entities form sets with fuzzy edges. For instance, some English *-ing* forms are usually described as verb forms ('she was laughing'), others as nouns ('laughing uses quite a few muscles'). Different linguistic schools tend to define these entities in different ways, and they give them different names. For example, in the sentence 'I enjoyed the concert', many linguists would call 'the concert' the object (of the verb 'enjoy'); but a more general term such as 'complement' may also be used, while some linguists would differentiate various kinds of 'objects', distinguishing for example between the material goal of verbs like 'hit' and 'break' and the object of behavioural or attitudinal verbs like 'enjoy' and 'dislike'.

The basic entities and categories of grammar are nevertheless common ground for many linguists. Whatever a specific school of linguists may call them, they are to a large extent translatable into each other. Noam Chomsky also subscribes to them. They are used to describe not just English, or other Indo-European languages, but, in principle, all languages. Some languages may display features that others do not have: for example, many Australian Aboriginal languages have a dual category in contrast to the singular and the plural, to indicate that there are exactly two, or a pair of entities; compare Pitjantjatjara *ngayulu* 'I', *ngali* 'we two', *nganana* 'we three or more'. Some languages, like Indonesian, do not have categories of the verb such as tense and mood. But principally it is the same finite set of entities and properties that we use to describe any of the Earthlings' languages, and it wouldn't be surprising if we used them also for all the Martian dialects once we come across them.

Smaller entities can be combined to form larger entities. Syntactic rules tell us which combinations are grammatical, and which are not. For many linguists, the smallest syntactic entities are words. For some, the morpheme is the smallest unit. Morphemes are parts of words, the smallest linguistic elements to which we can assign a meaning or a function. The word form *singing* consists of two morphemes: *sing* and *-ing*. The morpheme *-ing* can occur in most other verbs, as well; we find it in certain syntactic constructions, e.g. after a certain set of verbs like *help, see, hear*: 'he heard her singing in the rain'. Because its occurrence may be said to be caused by syntax, some linguists take morphosyntax to be part of syntax, and for them, morphemes are part of syntax. But generally, if syntax is held to be something different from the rest of

the lexicogrammatical systems, it is understood to describe how words can be assembled to form a grammatical sentence.

Seen in this light, words are the basic tissue of syntax. They make up the vocabulary, the lexicon of a language. Linguists, including Chomsky, agree that the lexicon is a more or less finite list of lexical entries. Each lexical entry consists of the word, an indication of the part of speech it belongs to, and the syntactic and semantic properties it has. The entry for *boy* would tell us that it is a noun, that it is countable (hence there is a plural *boys*), and that it fits, according to specifiable rules and constraints, into a slot (i.e. a terminal element of the syntactic structure of a given sentence), which asks for a word denoting a human being (such as the subject and the object position of the verb *love*). The sentence 'Big boys love intelligent girls' could be described as having the structure: adjective + noun + (transitive) verb + adjective + noun. Each noun and the verb exemplifies a slot into which we can insert a suitable lexical element taken from the lexicon.

Entities, properties and rules: this is the stuff that, according to Chomsky, constitutes each language. Therefore Chomsky's claim about the similarity of languages is not totally implausible. Languages resemble each other because their phonology, syntax, and morphology can be described in the same – or at least similar – terms. For mainstream linguists, languages are all more or less the same. They may follow different rules, but they are made up of the same entities and share many properties.

But does this mean that entities, property types and rule types are language universals? This is not a question to which there is an easy answer. When we describe language, what kind of a reality are we describing? There are sound sequences, or chains of alphabetic characters (or other kinds of characters in languages that have non-alphabetic writing systems), which we are accustomed to interpret (successfully) as language. Linguists cut these strings into little bits and pieces and assign various functions to them. Certain bits (say in English those that can be preceded by a determiner, that can serve as heads of noun phrases or prepositional phrases, and that can be modified by an adjective phrase or a prepositional phrase) we call nouns. But does that mean that nouns are more than bundles of properties that we construe in our theory? In the sentence 'This is a fake diamond', is *fake* a noun or an adjective? Obviously it is modifying the indisputable noun *diamond*. In this sense, it shares the properties of adjectives. But usually adjectives are gradable (*big, bigger, biggest; short, shorter, shortest*), whereas *fake* is not. And usually adjectives can be used predicatively, as in 'the house is big, but the garden is small', or 'isn't his hair short!' The word

fake can occur predicatively ('this diamond is fake') but many people might prefer to say 'this diamond is a fake'. Grammatical description would seem to require that we say that *fake* is an adjective in 'this diamond is fake', but a noun in 'this diamond is a fake'. So it may be up to the linguist or the lexicographer to decide whether they describe *fake* as a noun that can be used as an adjective, or as an adjective that can be used as a noun. Observations like these should throw some doubt on the widespread belief that entities or categories such as nouns exist independently of their description, in the way that apples and pears would still exist, even as something categorically different, if there was no one trying to categorise them.

Linguistics, Chomsky tells us, should describe the human faculty of generating an unlimited number of different grammatical sentences. This is why he and many of his followers are opposed to an empirical study of language (where empirical means the analysis of existing texts). No amount of text, Chomsky claims, can account for the competence to distinguish non-grammatical structures from grammatical structures. If we accept the premise that we can always utter a (grammatical) sentence that has never been uttered before, then the criterion of grammaticality is not something that can be found in texts. Rather, it is a feature of our language faculty. It is the application of the rules that can generate endlessly new, never heard before, sentences, all of which are grammatical, because they comply with the rules. This competence to produce new grammatical sentences is something (ideal) native speakers have.

The language faculty is therefore a feature of the mind. If we want to find out how language works, we have to look at the mind, and not at texts. Let us, for a moment, return to the sentence 'Big boys love intelligent girls'. This sentence structure can demonstrate the generative power of the language faculty. We can say that this sentence structure consists of two parts, the noun phrase *big boys* and the verb phrase *love intelligent girls*. This verb phrase consists of a transitive verb (*love*) and another noun phrase (*intelligent girls*). Noun phrases must have a head, usually a noun (such as *boy* or *girl*), either in the singular or in the plural, which can be preceded by a determiner (*a* or *the*), and modified by an adjective (such as *big* or *intelligent*). Now, this structure can easily yield a seemingly endless amount of different sentences, by the insertion of other nouns and verbs into the respective slots ('little girls hate spiteful boys', 'intelligent women admire intelligent men', and so on). Some verbs may not go well with some nouns as in: 'Fake diamonds hate eternity'. It seems we must therefore apply other rules as well that make sure that only those nouns are selected which go together with a parti-

cular verb. (For Chomsky, those so-called sub-categorisation rules are part of syntax, not of semantics, a position that is arguable.)

Chomsky's revolution in linguistics is about the generative power of rules. Rules, he says, do not describe what is there but what is possible. This focus on the generative aspect of language has changed the agenda of linguistics. The role of linguistics is no longer to interpret what we find in existing texts, but to describe the language faculty, or, in abstract terms, the competence of a speaker to produce new grammatical sentences. While rules were once formulated by language experts in order to facilitate the understanding of existing texts, or to help us to learn a foreign language, the task for a Chomskyan linguist is to discover the rules we follow as native speakers without even being aware of them, i.e. the rules which constitute the language faculty of human beings. In traditional linguistics, entities or categories like nouns, or tense, or person, were useful constructs in the framework of a theory. Rules were expressions of the linguist's ingenuity to make sense of the language evidence. Under the new agenda, language is like a game of chess. We are born with the capability to follow the rules without ever having to learn them. Chomskyan linguistics thus changes the status of linguistic rules. Rather than being tools for language analysis, they now become the metaphysically real essence of language.

Pre-modern linguistics in Europe was not concerned with the productivity of language. From the Middle Ages well into the nineteenth century, linguists were philologists, which was, at the time, more or less synonymous with classicists. Their research was on 'dead' languages: Latin, Greek and Hebrew. Their aim was not to produce new texts in these languages; they wanted to understand the texts we had inherited from ancient times. The rules they came up with were rules to help us make sense of the sentences. The rules were meant to describe what we were confronted with in the texts; they were not designed to empower us to become competent speakers of ancient Greek. The grammatical rules philologists were interested in were those that explained the specificity of Greek as compared to other languages, those that helped to understand their texts. Philologists were not interested in what was universal. Their rules were descriptive; they had to facilitate the analysis of textual evidence.

The philologists may not have had a scientific method. And yet we inherited from them the academic editions of classical and oriental texts we are still using today, together with comprehensive dictionaries, or rather glossaries, citing each noteworthy occurrence of any word embedded in its contexts and still providing an irreplaceable aid in understanding these texts.

Hermeneutics was the philosophical basis not of linguistics as we know it today but of philology. Hermeneutics is the art (or craft) of interpretation. In the early Middle Ages, this meant interpretation particularly of the Bible, but later also of the other classical texts. The goal of hermeneutics is to find out what a text means. What, indeed, does a text mean? Do we have to find out what the authors *thought* was the meaning of their texts? The authors might not tell us that explicitly, or they might tell us but be deceiving us in one way or another. Whatever they say, it is not the meaning of their texts. Or is the meaning of a text what the text means to me? Then meaning is something subjective, individual, something that cannot be validated by other readers. Meaning must be something else. When we encounter the word *love* in a medieval text, can we find out what the word meant then? Is there a methodology to answer this question? Is there a possibility of coming to an understanding that is shared by our fellow linguists? This is the key question hermeneutics is concerned with.

Particularly in the English-speaking countries, hermeneutics and philology have lost much of their earlier appeal. Since the first years of the twentieth century, British empiricism has given way to the new paradigm of analytic philosophy. This brand of philosophy, dating back both to Cambridge and connected with names such as Bertrand Russell, and equally to the Vienna circle and connected with names such as Mach, Carnap and (the young) Wittgenstein, is concerned with truth and reality. The question that is at the core of the current mainstream paradigm of the philosophy of language is not what a text, a sentence, a word means but how we can know whether it is true, whether it truly reflects the discourse-external reality or not. This is not a question hermeneutics, or philology, is concerned with. Philologists do not want to know under which conditions the sentence 'Mary, the mother of Jesus, was a virgin' is true; they content themselves with the exploration of the meanings of words, for example with questions such as whether the English word *virgin*, Latin *virgo*, Greek *parthenos* are appropriate translations of Hebrew *almah*, a word which usually just means 'young woman'.

Today, hermeneutics and philology are often considered dull, continental and old-fashioned. Edward Said, the famous Lebanese-American orientalist, is a noble exception. For him, philology is 'the extraordinarily rich and celebrated cultural position' that (not only) gave classics and orientalism their methodological basis. The philologist is the interpreter of bygone texts on the horizon of our own modernity. The philologist makes us understand cultural and

intellectual history. This act of understanding is two-directional. Our understanding of these texts always also presents a challenge to the way in which we understand ourselves. Thus, 'philology problematises – itself, its practitioner, the present'. Said quotes Ernest Renan, a nineteenth-century orientalist: '"The founders of the modern mind are philologists". And what is the modern mind ... if not "rationalism, criticism, liberalism [all of which] were founded on the same day as philology"' (Said 1995, p. 132). What has made philology so unattractive in the twentieth century? Perhaps it is the sense of arbitrariness, of subjectivity, the lack of a truly scientific method. Interpreting a text is always an act, as opposed to a process that follows clearcut rules. The art of hermeneutics, the craft of philology always involves making decisions. It means choosing between alternatives, without unambiguous instructions on how to select one of the options.

In the nineteenth century we find a novel interest in languages, different from traditional philology. It was the century when the enlightenment finally bore fruit and nature began to be understood in terms of the laws of nature. The main foundations of the sciences as we know them today were laid. All the academic glamour now rested with the sciences; and the liberal arts, including the humanities, were relegated to backstage. The hermeneutical approach to language was not interested in immutable, eternal laws or rules. But that did not necessarily mean that there weren't any. The first domain of this new 'scientific' approach to language was the study of relationships among languages. That became the starting point of modern linguistics. It seemed that many languages spoken in Europe, in the near East and even as far away as India, were somehow related to each other, some closer, like Gaelic and Breton as Celtic languages, Lithuanian, Latvian and Old Prussian as Baltic languages, or Czech, Polish and Russian as Slavonic languages. There was Sanskrit, there were the Romance languages, there were the Germanic languages and many more, dead or alive. They all seemed to descend from one single language, Indo-European, and in the course of history they seemed to have become more and more separated from each other. Over the course of their existence, all these languages underwent change. What was *patēr* in Greek and *pater* in Latin became *padre* in Italian, *père* in French, *Vater* in German and *vader* in Dutch. English *father* developed from Old English *fæder*. All of them share, ultimately, the same ancestor. Similarly we can work out that the English word *rich* is related to the German *reich*, that early Germanic took the ancestral form of these words from Celtic, and that they are also related to the Latin *rex*, 'king'; or that the English word *glamour* is borrowed from Scots, while, in turn, the Scots word is

derived from English *grammar*, which is, in turn, taken from Latin *(ars) grammatica*. (For more examples of historical changes, see 2.3.)

To the linguists of the nineteenth century who studied these phenomena, it seemed that the phonetic changes these words underwent in the course of history were governed by laws. The new linguists were less concerned with interpreting the meanings of texts, sentences, or words; they wanted to discover the laws of phonetic change. They were so confident in their scientific powers that they did not shy away from reconstructing ancestral languages, like Indo-European, even though no texts had survived. For the first time, it had become possible to describe language in terms of rules; rules that did not involve any decision-making on the part of the linguists, rules that produced results that had to be objectively correct once you accepted the premises. And if there were laws in phonetic change, there must also be laws for grammar. Therefore we can find, from the middle of the nineteenth century, a surge of literature on grammar, coinciding with a relegation of linguistic literature dealing with the vocabulary and the meaning of words to a less prominent position. This is still the situation in which we find ourselves today.

The modern linguists who succeeded the philologists saw themselves as scientists. However, from Ferdinand de Saussure (2.6 above) and the structuralists of the Prague school, to Louis Hjelmslev and Roman Jacobson, these linguists were not interested in the mental processes linked to language. They wanted to investigate the structure of language, based on analyses of texts, in order to understand the language system behind it, what Saussure called *la parole*. They wanted to describe a system of rules and means that existed independently of its individual speakers and its historical development (language synchrony) – although this system could also be studied from the historical point of view as a system gradually undergoing change according to language laws (language diachrony).

Thus the preoccupation with rules and laws characterises both non-Chomskyan modern linguistics (henceforth: standard linguistics, preoccupied with the idea of the system) and the Chomskyan variety of language studies (less interested in the system). Both varieties look at language as a system, which can be described in terms of rules, entities, categories and properties. From the structural point of view, these laws, entities and properties are, on a general level, more or less identical for all human languages, though rather, and at times profoundly, different in particulars.

Yet while Chomsky insists on the fundamental sameness of all languages (on a biological level), he also points out something very

important: the vocabularies of all these languages across the world are (mostly) mutually unintelligible. People do speak different languages, and we do not understand each other. Doesn't this contradict the claim of sameness? In general, Chomsky's interest in the lexicon is, contrary to structuralists, only marginal. But, how important is the lexicon? How important is it to find out about the meanings of words?

3.2 Standard linguistics and word meaning

Even if Chomsky is technically wrong in positing an innate mechanism that determines, by a minimum of external input, the grammar of the language we grow up with, it still remains a fact that we seem to have much less difficulty in learning the syntax of a foreign language than its vocabulary. It is not always too difficult to construe grammatically correct sentences in a second language. But unless we are acquainted with it very thoroughly, we will make mistakes when we try to put our thoughts into words or to translate a text from our native language. We can follow rules easily. But how can we do the right thing if it seems all but impossible to teach us what is the right thing? This is indeed the impression if we attempt to let ourselves be guided by bilingual dictionaries. They offer many choices but few instructions.

The difference between grammar and vocabulary is largely a matter of perspective or method (1.6). For vocabulary, at least at first sight, there seem to be few rules which we can follow. Rules we can learn, and instructions we can follow. But no bilingual dictionary seems to be big enough to tell us how to translate an apparently quite simple word, like *grief*, into French. There are, according to the *Collins–Robert French Dictionary* (1998, repr. 2001), two main options: *chagrin* and *peine*. We are, however, not clearly told which of the alternatives to choose when. In the absence of clear instructions, even the most comprehensive bilingual dictionaries let us down when we want to translate a text into a non-native language.

The same dictionary gives us, as the equivalents for *sorrow*, the same two words it has given us for grief, *peine* and *chagrin*, plus another word, *douleur*, which is preceded by the ominous comment: '(stronger)'. It is not quite clear what this means: is this the word to use if your grief is stronger than average grief, or is *douleur* a stronger word than *peine* or *chagrin*? From the French perspective the two equivalents *sorrow* and *grief* appear to be synonyms. However, most native speakers of English agree that in these two sentences 'Grief gave way to a guilt that gnawed at him' and 'A magic harp music made its listeners forget sorrow', *grief* cannot be replaced by *sorrow*, and vice versa, so that, at least from the

monolingual English perspective, they cannot be regarded as synonyms. Things get even more confused when we look up, in our bilingual dictionary, the English equivalents for the French word *chagrin.* For *chagrin* we find: '(= affliction) grief, sorrow', and thus we become curious what French *affliction* means in English. The only English equivalent we are offered, though, is *affliction.* The French equivalents of English affliction are *affliction* and *détresse,* while *détresse* is, we are told, *distress* in English. As the English equivalents of *peine* we find *sorrow* and *sadness,* but not *grief.* Our analysis thus reveals a distressing absence of systematicity, and we are left wondering whether this is due to the languages as they are or due to our inability to describe them properly. (And it has to be said that the *Collins–Robert* is not just any French–English dictionary. Together with the *Oxford–Hachette French Dictionary,* it represents the apogee of modern bilingual lexicography.)

The meaning of words, as compared with the regularities of phonetic change and sentence construction, is generally fuzzy and vague, not only when we compare one language with another, but also from a monolingual perspective. Words, single words, may be the ideal core units when it comes to describing the working of grammar. But they are much less the appropriate core units when we are interested in meaning. Single words are commonly ambiguous. Dictionaries capture this ambiguity by assigning two or more word senses to a word. As shown above, we are confronted with the ambiguity of single words whenever we want to translate into a foreign language. Then we have to choose between several options, only one of which is acceptable. But when we read a sentence or text we are not fooled, under normal conditions, by any ambiguity. Usually we have no problem understanding what a sentence means. This is because we do not look at the words in isolation, but embedded in a context. We read a word together with the words to its left and to its right; we have no problem in knowing what a word means. Ambiguity is a consequence of our misguided belief that the single word is the unit of meaning. Units of meaning are, by definition, unambiguous; they have only one meaning. While some words are units of meaning, many are not.

This enquiry into meaning makes the case that meaning is an aspect of language and cannot be found outside of it. It is entirely within the confines of the discourse that we can find the answer to what a unit of meaning means, be it a single word or, more commonly, a collocation, i.e. the co-occurrence of two or more words. A unit of meaning is a word (often called the node or keyword) plus all those words within its textual context that are needed to disambiguate this word, to make it monosemous. As most of the more frequent words are indeed

polysemous, they do not, as single words, constitute units of meaning. As any larger dictionary tells us, for example, the word *fire* is ambiguous. It is therefore not a unit of meaning. In combination with the noun *enemy* it becomes a part of the collocation *enemy fire*, meaning 'the shooting of projectiles from weapons by the enemy in an armed conflict'. This collocation is (under normal circumstances) monosemous, and therefore a unit of meaning.

In the venerable field of phraseology, people have always been aware that language is full of units of meaning larger than the single word. When we hear 'She has not been letting the grass grow under her feet', we do not expect that to be literally true. Rather we have learned that the phrase 'not let the grass grow under one's feet' is an idiom, a unit of meaning which, according to the *New Oxford Dictionary of English* (*NODE*), means 'not delay in acting or taking an opportunity'. Indeed, the idiomaticity of language is a favourite topic of the discourse community. People like to talk about idioms; we feel that they are an important part of our cultural heritage. There is many a book explaining their origins, and there is hardly a dictionary that would dare to leave them out. Over the last century, we have come up with ever more refined typologies of idioms. Rosamund Moon's excellent study *Fixed Expressions and Idioms in English* (1998) provides a thorough corpus-based analysis of the phenomenon of idiomatic language. While some idioms are more or less inalterable ('it's raining cats and dogs'), others are somewhat ('a skeleton in the closet', 'a skeleton in the cupboard'). Most idioms oscillate between the two extremes of invariance and alterability. If we probe too deeply, our 'intuition' will often desert us. Are 'figments of imagination' an idiom, or can there be other figments? Does figment have a meaning of its own? We have to look in a corpus (here the British National Corpus) to find that there are indeed other figments, namely 'figments of linguistic bewitchment' and 'figments of fiction'. In the singular as well, there are some deviations from the prototypical collocate *imagination*: 'a figment of his own mind; a figment of my neurosis; a figment of its leaders' fantasies; a figment of his own name'. But these are four instances (i.e. less than 5 per cent) out of fifty-eight occurrences.

Idioms have found their way into bilingual dictionaries as well. The *Wildhagen Héraucourt German–English Dictionary* tells us that the English equivalent of *wie ein Blitz aus heiterem Himmel* [literally: like a bolt from a serene sky] is 'like a bolt from the blue'. Idioms feature rather prominently in foreign-language learning – with the result that speakers of English as a second language tend to overuse those they have learned, such as 'it's raining cats and dogs' (an idiom not greatly used by native speakers).

Modern linguistics has taught us that there is, indeed, a range of lexical constituents that can lay claim to being a unit of meaning. There are bound morphemes which have a meaning only by virtue of being part of a larger constituent (as the plural *-s* in English); there are free morphemes whose meanings seem to be rather invariable; there are words; and there are idioms including proverbs making up a full sentence. We have also learned that the borderlines between them are areas of contention. But while we would never doubt that morphemes are linguistic constructs, we have come to accept the ontological reality of the word (1.1).

Today, when we hear 'word', we normally think first of 'an element of speech', as the second sense given in the *OED* is circumscribed. If we believe Jack Goody (Goody 2000), this concept is foreign to oral societies. That is not so astonishing. In spoken language we normally do not insert a pause between words. Neither were the Greeks and Romans of antiquity in the habit of putting spaces between their written words. Where the space is inserted is largely a matter of convention, and not always well-established convention. Look in any large English dictionary for entries beginning with *half*. One dictionary has *half brother* as two words, another gives it a hyphen: *half-brother*. One has *halfback* as a single word, another has it with a hyphen. And so on. What is *linguistique de corpus* in French is *corpus linguistics* in English and *Korpuslinguistik* in German. There is no cogent reason other than tradition why there should be no space between the elements of German compounds, i.e why it is *Korpuslinguistik* rather than *Korpus Linguistik*.

Other modern languages missed the chance to define words by spaces. When it was recognised that in most cases it did not make sense to define a single Chinese character as a word and it became accepted that most Chinese words would consist of two or even three characters, it became a problem to identify words in a sentence. It is often the case that Chinese sentences can be cut up into words in different ways as long as we apply nothing but formal rules and leave out what they mean. Thus, in Chinese-language processing, there is still no segmentation software that is entirely reliable. How could it be different? We find cases of doubt in practically all Western languages. The problem of where there should be spaces and where not featured prominently in the German spelling reforms introduced in the mid-1990s.

Listening to foreign languages which we do not understand makes us even more aware of this problem. How do we know where a word begins and where it ends? Normally, people do not mark word boundaries phonetically. How do we know if two occurrences of the same concatenation of phonemes are occurrences of the same word

(e.g. *no* versus *know*)? How can someone who does not speak English find out that *a* and *an* are two variants of the same word, the indefinite article, or that *the* in *the enemy* and *the* in *the friend* are variants of the same definite article, even though they are usually pronounced differently?

Languages in written form seem, at first, to simplify matters for us, particularly if they are written in the Latin alphabet. There we find spaces between the words. But how reliable are they? We have already seen some variation with words beginning with *half*. Is *half time* the same as *halftime* and *half-time*? Some dictionaries distinguish the musical term *half tone* from the printing term *halftone*. If *corpus linguistics* is one word in German (*Korpuslinguistik*), why is it two words in English? Or is it not? Do compounds consist of two words, or are they, in spite of the space between the two elements they consist of, just one word? Are words in languages as Hungarian or Welsh, which often seem to consist of a rather large number of elements, words in the same sense as English words? One Finnish word *talossanikin* means 'also in my house', which is translated as four words in English. In Chinese, we find the same spaces between all characters and there is no special indicator telling us which characters belong together or where a word begins or where it ends. In order to identify words we have to rely on wordlists and dictionaries. But they are the more or less arbitrary results of lexicographers at work. What are we left with once we take away the spaces between words?

We have always known that there are units of meaning larger than the single word. From early childhood, we are made aware of them. A phrase like 'to turn a blind eye to something' has become part of our cultural heritage. It is an idiom, and idioms have always been listed in our dictionaries. Yet we are not so readily aware that large portions of our texts are also made up from larger, often rather complex units of meaning, like *weapons of mass destruction* or *friendly fire*. For the most part, these are absent from our dictionaries. With our ingrained focus on the single word, that is not surprising. The larger units escape the attention of even experienced and well-trained lexicographers. They do not catch the eye when we come across them. Before the advent of corpora and of corpus linguistics, we did not even have a methodology to detect them. Neither standard linguistics nor Chomskyan linguistics can identify these units of meaning.

What is it then that makes the single word continue to be such an attractive unit in linguistics? Words seem to be almost ideal units for grammars, particularly grammars that do not touch on meaning. Noam Chomsky's *Syntactic Structures* (1956) is a good example. Here we

find sentences (S), non-terminal symbols such as noun phrases (NP) and verb phrases (VP), and terminal symbols such as nouns (N), adjectives (Adj), determiners (Det), verbs (V), etc. Grammatical rules, starting with the S-symbol, generate strings of terminal symbols. In principle, we can insert the corresponding lexical elements in the slots provided by these symbols. Those lexical elements are single words. Up to a point, such a grammar seems to work, particularly for non-inflecting languages with a strict word order. We run into real trouble only when we demand that the sentences generated by this grammar make sense, that the sentences can be interpreted semantically. For a meaning-free grammar, the single word seems to be indeed the lexical element *par excellence*. In language learning, meaning-free grammars are good enough for constructing grammatical sentences in the target language, regardless of what they mean.

It is meaning, not grammar, that casts a shadow over the single word. A glance at any monolingual or bilingual dictionary confirms that the main problem of single words, from a semantic perspective, is their polysemy, their ambiguity and their fuzziness. For the verb *strike*, the *NODE* lists eleven senses. One of them is 'make (a coin or medal) by stamping metal'. As a sub-sense of this we find 'reach, achieve, or agree to (something involving agreement, balance, or compromise): the team has struck a deal with a sports marketing agency'. Though we might, upon consideration, come to accept this sense as a metaphorisation of striking coins, the actions seem to have hardly anything in common. The *strike* in *strike a deal* means something else than the *strike* in *strike coins*, and something different from the other ten senses ascribed to it in the dictionary entry. Indeed one could easily maintain that it has no meaning of its own; together with *deal* it does mean something, namely 'reach an agreement'. This is the gist of John Sinclair's article (1996) 'The empty lexicon'. Once we have identified semantically relevant collocates of words like *strike* (*a blow, a deal, oil*, etc.), their ambiguity and fuzziness disappears. The collocation *strike a deal* is as monosemous or unambiguous as anyone could wish. Even though neither the *NODE* nor the *Longman Dictionary of English Idioms* (1979) list *strike a deal* as an idiom, it seems to belong in this category. In the British National Corpus (BNC) there are twenty-five occurrences of *struck a deal*. The absence of *strike a deal* from larger dictionaries and specialised idiom dictionaries illustrates that the recognised lists of idioms, those we are aware of as part of our cultural heritage, represent no more than the tip of an iceberg. Time and again, corpus evidence suggests that there are many more semantically relevant collocations than dictionaries tell us.

What about the sense of *strike* described in the *NODE* as 'discover (gold, minerals, or oil) by drilling or mining'? In the Bank of English, there are 23,096 occurrences of *struck*. In a random sample of 500 occurrences, we find 7 instances for this sense of *strike*, 4 of 'struck gold', 2 of 'struck oil', and 1 of 'struck paydirt'. All of these citations represent metaphorical usage. Here are two examples:

> Dixon, who, together with the unfailing Papa San, struck gold with 'Run The Route'.
>
> telephone franchises. No one has struck paydirt yet, although the Bells have captured business

The example of *strike* 'discover by drilling or mining' shows that there is no obvious feature to tell us whether we should analyse a phrase as consisting of two separate lexical items (*strike* and *gold*) or whether we should analyse it as a collocation, i.e. as one lexical item (*strike gold*). It is not a question of ontological reality, of what there is, but a question of expediency. Carrying things to extremes and replacing most single words in our dictionaries by collocations would mean that these dictionaries would have to become much more voluminous. We would have to account for *strike a chord, strike a balance, strike a blow, strike a pose, strike a note, strike fear, strike terror, strike home, strike someone (as)* and possibly some others. If we leave things as they are, we find *strike a coin* and *strike a deal* belonging to the same sense category. Expediency alone, however, seems to be unsatisfactory. Aren't there any more plausible arguments?

3.3 Words, idioms and collocations

Let us look at another example in more detail. Some grammatical patterns are particularly prone to form collocations, such as nouns modified by adjectives. This fact has not escaped the attention of lexicographers. However, without the application of the methodology developed for corpus linguistics, it seems to be left to the whims of dictionary-makers what they decide to include. For the adjective *false*, the *American Heritage Dictionary* (4th edition, 2000) lists these collocations: *false alarm, f. arrest, f. consciousness, f. fruit, f. imprisonment, f. indigo, f. ipecac, f. memory syndrome, f. miterwort, f. pregnancy, f. pretense, f. rib, f. Solomon's seal, f. spikenard, f. start.* The *NODE* lists these collocations: *false acacia, f. alarm, f. bedding, f. card, f. colour, f. coral snake, f. cypress, f. dawn, f. economy, f. face, f. friend, f. fruit, f. gharial, f. helleborine, f. memory, f. move, f. oxlip, f. pretences, f. rib, f. scorpion, f. start, f. step, f. sunbird, f. teeth, f.*

topaz, f. vampire. Even if we acknowledge the differences between American and British English, there are surprisingly few overlaps: *f. alarm, f. fruit, f. memory, f. pretense/pretences, f. rib, f. start.* A random sample of 50 citations from the BNC attests *false alarm, f. dawn, f. pretences, f. start, f. teeth,* but in addition many other collocates of *false*: *assumptions, cheerings, claims, complaints, confidence, declarations, decisions, denial, distinctions, echo, enquiries, expectations, formastation* (!), *hopes, idea, information, market, money, position, proportion, readings, reasoning, report, take, testimony, theory, tradition, understanding, witness.* Which of these co-occurrences should be described as two separate lexical items, which as a single lexical item? How many senses should we ascribe to *false*? Does *false* in *false alarm* mean something different from *false* in *false echo* or *false witness*? Or would it make things easier to say that it does not really matter what *false* means in these instances and we should rather try to describe what *false alarm, false echo* and *false witness* mean? Which cases should we describe as collocations, which as a combination of (one meaning of) *false* with (one meaning of) the noun in question?

Within the confines of one language it is impossible to come up with clear criteria. But once we bring in a second language, we suddenly find the arguments we have been looking for. The *Wildhagen–Héraucourt* dictionary tells us that these are all possible German equivalents of *false*: 1. *falsch, unrichtig, irrig; ungesetzlich, widerrechtlich;* 2. *unwahr; trügerisch, täuschend; verräterisch, treulos; untreu;* 3. *falsch, gefälscht; unecht; nachgemacht; vorgetäuscht; blind; vorgeblich; Falsch-, Schein-; irrig, so genannt.* How helpful is such an entry? The senses are being distinguished by the different sets of equivalents. But some of the equivalents occur in more than one set. Does that mean that the equivalents themselves are polysemous, or just that the sense categories are fuzzy? (Note that *falsch* is the first and, implicitly, most significant equivalent for both sense 1 and sense 3!) For those who know some German it is also immediately obvious that the words we find within a given sense are far from synonymous; we cannot simply substitute them for each other in various contexts. Why then are we given three senses, and not one, or maybe ten or twenty? If we speak German well, the list of words will help us to choose the one that fits best into a given context. If we do not know German that well, how are we to choose the appropriate equivalent?

Naturally, the lexicographers are aware of their predicament. If they want to cater to native-English speakers with a cursory knowledge of German they have to deliver more. They have to give the translation equivalents not of *false* but of *false* in combination with the nouns it co-occurs with. They have to provide the translations for the collocations

of *false*. Some of these collocations are listed as additional information within a given sense category. For sense 1 we find: *false quantity, false arrest, false imprisonment.* For sense 2 we find: *false mirror, false oath* [the equivalent given is *Meineid*], *false pretences, false swearing.* For sense 3 we find *false coin, teeth, hair*; and the idiom *to sail under false colours.* There is also a subsequent section called *Verbindungen* ['collocations'] with more phrases: *false alarm, f. bottom, f. cap, f. door, f. key, f. ogive, f. shame, f. report, f. step, f. take-off.*

Looking at the *Oxford–Duden* (compiled 1990, i.e. *c.* 50 years after the first edition of the *Wildhagen*), *false* is again divided up into three senses. Again slightly abridged, sense 1 is *falsch*; *Fehl-* (*Fehldeutung, ...*); *Falsch-* (*Falschmeldung, ...*); *treulos*; *gefälscht*; sense 2: (*sham*) *falsch*; *künstlich*; *geheuchelt*; *gekünstelt*; sense 3: (deceptive) *falsch*; *unberechtigt*; *trügerisch.* There is no way to map these three senses on to the three senses of the *Wildhagen–Héraucourt.* The users are left in doubt whether the division into senses in either of the dictionaries reflects the way *false* is being used in English or the hypothesis that there are three different main translation equivalents of *false* in German. Neither claim seems to be particularly helpful or supported by evidence. It just happens that *Cobuild* (the *Collins Cobuild English Language Dictionary*) also divides *false* up into three senses, identified as (1) 'incorrect', (2) 'artificial' and (3) 'insincere'. The *Oxford–Duden treulos* (sense 1), however, does not sail under 'incorrect'; neither does *geheuchelt* (sense 2) travel under 'artificial', nor *unberechtigt* (sense 3) under 'insincere'. As to German ways of negotiating word meanings, it would be next to impossible to claim that *treulos* and *gefälscht* belong to the same category, or *künstlich* and *geheuchelt*, or *unberechtigt* and *trügerisch.*

However, this dictionary entry could give us some ideas on how to be more helpful to its users. For translating into our own native language we might welcome a list of all relevant equivalents (in order of frequency or alphabetic order) so that we might choose among them on the basis of our linguistic competence. For translating into a language other than our own, a language where we do not have a comparable competence, we would, first of all, need a default translation. In the case of *false*, that is easy. According to the *Oxford–Duden*, the first equivalent in each of the three sense categories is *falsch.* This is no doubt the most common equivalent, being closely related to it etymologically. This translation equivalent should be used whenever *false* is not followed by a noun that is given in a subsequent list of collocations. In bilingual lexicography, we can define a collocation as a phrase that cannot be translated using the default translations offered for its components. Thus, users do not need to be told that the equivalent of

false teeth is *falsche Zähne*, because that would be the default translation anyway. (Actually the German is more commonly *Gebiss*, a word used more often than *dentures* is in English.) But they do need to know that the equivalent of *false coin* is *Falschmünze* (as opposed to *falsche Münze*). How do we arrive at such a list of collocations? If we compare the lists we find in the *American Heritage Dictionary*, in the *NODE* and in the *Wildhagen–Héraucourt* there is only a relatively small overlap.

This is an indication that without suitable corpora, lexicographers are at a loss when it comes to collocations. Even though they are aware of the problem, their findings will be always accidental. Leaving aside, for the moment, the problem of identifying semantically relevant collocations in a monolingual context, we can sketch now what we have to do from a bilingual perspective. We have to look at a corpus. It should be big enough to mirror the kind of language we find in books, newspapers and 'educated speech', i.e. the kind of language we tend to teach in language teaching, and it would yield many more collocations than lexicographers can think of. We would then have to find translations for them. All of those for which the default translation of its elements would be wrong would be entered into the dictionary. We will certainly end up with different sets for each language. In German, a *false alarm* is a *blinder Alarm* (not a *falscher Alarm*); thus this phrase counts as a collocation and belongs in the dictionary. In French, however, it is *alarme fausse*, i.e. the default translation of *alarm* and *false*; and we do not have to treat it as a collocation. If, due to size, not all collocations can be entered into the dictionary, frequency would be an important parameter. We might do without the *false Solomon's seal* and without the *false coral snake*. They seem to be more part of terminology than of the general vocabulary, anyway. *False dawn*, on the other hand, is relatively frequent and would count as a collocation, from a German perspective. The *Oxford–Duden* tells us that its equivalent is: *Zodiakallicht*; (fig.) *Täuschung*. But is *false dawn*, from a monolingual perspective, really a unit of meaning, a single lexical item, or just the combination of two separate lexical items? Can we apply the default meaning test in a monolingual environment?

The *NODE* describes *false dawn* as 'a transient light which precedes the rising of the sun by about an hour, commonly seen in Eastern countries'. According to this definition, *false dawn* seems to be a single lexical item. For we cannot deduce from the meaning of *false* (or from any of the senses a monolingual dictionary may give) and from the meaning of *dawn* (or any of its dictionary senses) that it precedes sunrise by about an hour and that it is specific to Eastern countries (whichever might be meant). But are these really essential or just

ornamental features? If they are essential, then *false dawn* is a unit of meaning. For users of the *American Heritage Dictionary*, *false dawn* is described as 'resembling but not accurately or properly designated as the time each morning at which daylight first begins'. This is something that I would be able to deduce from my knowledge of *false* and *dawn*. Here, we are not told that it precedes the real dawn and that it is more commonly found in Eastern countries than elsewhere. If *false dawn* is nothing else, then it is not a unit of meaning. For in this definition, a false dawn resembles a dawn. It is an 'incorrect' dawn. To resolve the issue of the two definitions, let us have a look at the BNC. In the BNC, we find eighteen occurrences of *false dawn*. Just two of them refer to a meteorological situation:

> ... it was not until another hour had passed and the moon was paling in the night of the false dawn that they were at last among strange scattered rocks...

> It was a false dawn, replaced soon after by a now starless night that was blacker than the previous hours.

Neither of these citations mentions an Eastern country, and neither refers to a sunrise occurring an hour later. If these instances are representative, then the *American Heritage Dictionary* seems more reliable, and *false dawn* is not a lexical item. But what about the other occurrences? All of them refer to situations in social life that initially seem to be better than is recognised later. Most commonly these situations refer to economic enterprises. These are some typical citations:

> It is our belief that Christmas will prove to be yet another false dawn as far as reawakening consumer confidence is concerned.

> The organisation's chief executive was optimistic that the latest figures did not merely represent another false dawn.

> Unhappily, it was a false dawn.

Google confirms the BNC evidence. It lists 14,000 hits for the expression 'false dawn'. Among the first 40 hits there is not a single instance where *false dawn* means what the *NODE* says. The first four citations refer to an economic entity called False Dawn. The subsequent instances refer again to situations that appear to be better than later recognised, e.g. the headline: 'Another false dawn for Africa?'. Of course, this meaning of *false dawn* can easily be explained as a metaphor of the *American Heritage Dictionary*'s *false dawn*. Important as this

issue of metaphorisation is for lexicography, this is not the place to pursue it. Dissatisfied with either of the two definitions, we again checked Google, this time for '"false dawn" night morning'. Under www.space.com/spacewatch/zodiacal_light/ we found this definition, which ties in nicely with the German equivalent *Zodiakallicht*:

> At certain times of year in the right locations, a faint cone of light appears in the predawn sky for lucky viewers in dark locations. This eerie glow is the Zodiacal Light.
>
> It is best seen before daybreak, generally two to three hours before sunrise in the eastern sky. But it's also visible in the west at certain times of year. Over the centuries countless individuals have been fooled into thinking the Zodiacal Light was the first vestige of morning twilight. In fact, the Persian astronomer, mathematician and poet Omar Khayyam, who lived around the turn of the 12th Century, made reference to it as a 'false dawn' in his one long poem, The Rubaiyat.

If this is what false dawn means, then it is a unit of meaning that cannot be reduced to a combination of any of the dictionary senses of false and dawn. It is a unit of meaning in its own right, a collocation not just on the basis of the frequency of co-occurrence of its elements, but also on the basis of semantic relevance.

When it comes to word-meaning, we are in dire straits. Native speakers understand the meanings of (the more frequent) words of their language. But they are less competent in describing these meanings. This incompetence seems to be shared, to some extent, by the lexicographers. Whatever the reason may be, this may explain why linguistics as we know it has been preoccupied with grammar. Rules are more elegant than the intricacies of meaning. Rules have explanatory power, they create clarity and understanding, and they provide us with instructions on what to do. There is, however, no rule which could tell us how many senses a word has. The decisions taken are arbitrary. At first glance, it is hard to decide whether it is simply that linguistics has never developed a satisfactory method for dealing with the meanings of words, or whether the situation we are confronted with defies any methodology. Worst of all, the division into different senses seems not to reflect properly how people understand these words when they read them in a text. Experiments have shown that neither lay native speakers nor speakers for whom English is a second language nor trained linguists can easily agree which dictionary sense they should assign to the word in question (Fellbaum 1998; cf. also Edmonds 2002).

How does it come about that highly reputable dictionaries leave such a lot to be desired? There might be a better explanation than

incompetence. When we encounter an ambiguous word in a sentence, we normally do not ask ourselves which sense it is used in. Perhaps our understanding of fuzzy words such as *friendly* does not imply putting a given usage into a given pigeonhole. Perhaps our understanding of words is mostly based not on our capacity to categorise, but on our faculty to draw on analogies and to discover resemblances.

Standard linguistics and Chomskyan (or post-Chomskyan) linguistics have not been strong in lexicography. With the demise of philology, the study of the meanings of words has more or less ceased to be a serious academic topic. There is still academic lexicology, and there is semantics; but lexicology has never questioned the categorical approach to word meaning. Rather than describing the meaning of a lexical item as a whole, it has sought to decompose it into more basic semantic features or categories. Many lexicologists still insist that once we get our categories right, better dictionaries will emerge. Semantics, these days, is predominantly cognitive semantics. Cognitive semantics wants to extend Chomsky's claim of the sameness of all languages to meaning as well. These semanticists say that, in principle, we all share the same language, the so-called language of thought, these days often called 'mentalese' (Fodor 1975, Pinker 1994). When we speak, they say, we translate an expression in mentalese into a natural language, and as hearers, we re-translate the natural language expression we hear back into mentalese. This is how Steven Pinker describes this universal mental language:

> People do not think in English or Chinese or Apache; they think in a language of thought. This language of thought probably looks a bit like all these languages; presumably it has symbols for concepts, and arrangements of symbols ... [C]ompared with any given language, mentalese must be richer in some ways and simpler in others. It must be richer, for example, in that several concepts must correspond to a given English word like stool or stud. ... On the other hand, mentalese must be simpler than spoken languages; conversation-specific words and constructions (like a and the) are absent, and information about pronouncing words, or even ordering them, is unnecessary.
>
> (Pinker 1994, pp. 81–2)

Pinker does not tell us, however, how many different concepts correspond to *friendly*, and so we are not told what the universal solution to the categorisation of word meanings would look like. It seems that the universality of mentalese is achieved by getting rid of everything which is language specific. There are many languages that do not feature articles, so mentalese does not have them; and languages come

up with different word orders, so mentalese does not have information about word order. Pinker is by no means alone in his putting his faith in mental representations. He is supported by, among others, Dan Sperber and Deirdre Wilson who discuss the following options: '[T]here are fewer concepts than words', 'there is roughly a one-to-one mapping between words and concepts', and '[m]ost mental concepts do not map into words' (Sperber and Wilson 1998, pp. 186–7). Concepts are more angelic than the earthly words of our natural languages; they seem to avoid the unpleasantness of dealing with the many unpredictable idiosyncrasies of words we find in all the human languages. For cognitive linguists, a word has as many senses as there are concepts into which it translates. Unfortunately there is no dictionary of concepts that lexicographers can consult. Rather, it is the other way around. The so-called conceptual ontologies, which are still popular in artificial intelligence, should be, as their proponents claim, in theory language independent. How would that be possible? How could we describe the content of a concept without using language? As it is, conceptual ontologies borrow heavily from dictionaries, and there is little hope that it could ever be the other way around. Semantics and lexicology, as they are practised today in the academic world, contribute very little towards an improvement of our dictionaries.

Standard linguistics has brought about better grammars. While it has also brought about a noticeable improvement of dictionaries, particularly of bilingual dictionaries, modern lexicography still falls short of answering our enquiry into the meanings of words in a satisfactory way. The vast majority of people, however, who listen to other people or read their texts, or who try to tell something to other people, do this because they want to understand or be understood. They do not analyse a sentence for the beauty of its syntactic construction, or because they are hunting for a rare species of a verb form. They may not even know that the sentence they have just uttered was in the passive voice. All they want to be sure about is that they, or their listeners, got the meaning right. And here the linguists seem to be unable to help them. They can tell you that 'Paul loves Mary' is (roughly) equivalent to 'Mary is loved by Paul'. But when asked what love means, linguists will refer Mary and Paul to their poor cousins, the lexicographers, who write in the dictionary (in this case the *Cobuild English Dictionary for Advanced Learners*): 'If you love someone, you feel romantically or sexually attracted to someone' or: 'You say that you love someone when their happiness is very important to you, so that you behave in a kind and caring way to them.' If Mary is being told by Paul that he loves her, she finds it important to know what he means by love. She does not have to

be aware that the dictionary could inform her about the many senses of this word, which for her is just fuzzy. For her the question is: does Paul only want to go to bed with her, or is he also willing to do the dishes? If Mary grew up in a Western country where English is the native language, she perhaps would not have a problem understanding Paul. But if she came from an Islamic or Hinduistic culture, she might not be acquainted with our kind of love talk. Standard linguistics will not be able to help her. Something new is needed. When we want to find out how language is being used, what words, sentences, texts mean, we have to analyse texts. Looking at the scripts of soap operas, Hollywood movies, novels and magazines read by young people, we can find out what normally happens after a lad says 'I love you'. It is from these soaps, movies, stories, alongside the examples set by his peers, that Paul has learned when to use the phrase himself.

3.4 Corpus linguistics: a different look at language

What is language? Is it the miraculous language faculty we all are born with, which, once it is awakened by verbal contact with native speakers, empowers us to become native speakers as well, and which requires but minimal input to tune the innate mechanism to the specifics of that language? Is it our competence to come up with grammatical sentences that have never been said or heard before? Is there an innate language organ, just as there is an innate capability to see and distinguish colours? If this is what language is, then we have to study it as a feature of the human mind and we do not have to be aware of the rules. They are wired into our brain, and we follow them unconsciously. We also do not have to learn what words mean. Once we are exposed to a word, we relate it to the mental concept into which it translates.

Or is language an acquired skill enabling us to take an active part in verbal communication? Can we learn a language in the same way as we learn to tie our shoelaces, to play chess or to solve equations? This is how we learn to speak a foreign language. We are taught the grammatical and inflectional rules, we are taught the equivalents of the words of our own language in that new language, and vice versa, and in the end we can produce utterances in the new language that comply with what we have learned. It does not really matter if the language we learn really exists, in the sense that there are native speakers. Learning French is hardly different from learning Esperanto, and, in principle, it should not be too different from learning a programming language. If this is what language is, then we take it to be the accumulation of all the instructions needed to speak it competently. If this is what language is,

language is not a feature of the mind. Once we have accumulated all the instructions, then there is nothing new to learn about the language.

Or is language something tangible, namely the accumulation of all the acts of communication that took place in a language community, in the same way that British architecture can be seen as the sum of all the buildings that were built in Britain and that we know about? Is the language of the Etruscans or of the Mayans what remains of their texts, or is it the sum of all the acts of communication that ever took place in Etruscan or Mayan? If we accept the latter position, then we can never hope to understand Etruscan or Maya fully. If English is the totality of all acts of communication of the English-language community, of all the texts that exist or have existed at a given time, then language is not a feature of the mind. It is something that exists, in some physical way, something that remains of the recent and the more remote past, something that keeps on growing and developing. If this is the English language, then most of it is lost – most spoken texts, except the very few that were recorded, and many written texts, except those that survive in libraries or in some kind of accessible archive. If we have to restrict our study of English to what is still accessible because it was recorded and preserved, then our picture of English will certainly be much larger than we can ever hope to come to terms with; but it will never be the full picture.

Language is a human faculty which children acquire naturally without being given instructions; it is a set of rules we have learned, from forming plural nouns, to using words in the appropriate order, to following the conventions of letters or essays or reports, and it is a long list of words we have learned (from the simplest of everyday vocabulary to learning that 'an apophthegm is a concise maxim, like an aphorism'). It is also the sum of all texts in that language. In *Macbeth*, IV, iii, 220, Shakespeare uses the verb *dispute* in the sense of 'revenge'. Nobody uses the word like that any more. But this usage has not exactly disappeared. Shakespeare's texts are still a part of our discourse. We read them, we watch his plays, we discuss his language. Thus there are different ways to look at language. It is up to us to decide how we want to study it. It depends on which aspect of language we are interested in. If we want to find out what is common to all languages, we should embrace Chomskyan linguistics. If we want to find out if a French sentence is structured grammatically, we should rely on standard linguistics. If we want to find out what words, sentences and texts mean, we should opt for corpus linguistics.

Corpus linguistics sees language as a social phenomenon. Meaning is, like language, a social phenomenon. It is something that can be

discussed by the members of a discourse community. There is no secret formula, neither in natural language nor in a formal calculus, that contains the meaning of a word or phrase. There is no right or wrong. What I call *a weapon of mass destruction* differs probably a lot from what President George W. Bush calls *a weapon of mass destruction.* What I call a *baguette* is not the same as what many supermarkets sell as a *baguette.* What I call *love* may not be what my partner calls *love.* Different people paraphrase words or phrases in different ways. They do not have to agree. In a democracy, everyone's opinion is as good as anyone else's.

Meaning is what can be communicated verbally. If you do not know what *apophthegm* means, you can ask your fellow members of the English discourse community. Many may not be quite sure themselves, and they may refer you to the dictionaries. Someone may quote Samuel Johnson's famous apophthegm 'Patriotism is the last refuge of a scoundrel', and perhaps from then on you will not forget what the word means. The meaning of *apophthegm* for you, then, is the sum of all you have heard from the people you have asked plus all of what you have found in the dictionaries. There is certainly more to the meaning of *apophthegm.* There are more dictionaries that you could consult, there are more people you could ask, there are more texts you could find in libraries and archives containing the word embedded in various contexts. The full meaning of the word is only available once all occurrences of the word in the texts of the English discourse community have been taken into account. All citations together (plus what people tell you when you ask them) are everything one can know about the meaning of *apophthegm.* There is nothing else that could tell us what this word means. And all of it is verbal communication.

The perspective of Chomskyan and cognitive linguistics represents a very different view of language. In that perspective, language is a psychological, a mental phenomenon. Both views are, of course, legitimate, and they are complementary. Corpus linguistics deals with meaning. Cognitive linguistics is concerned with understanding. Meaning and understanding can easily be confused, but it pays to keep them apart. Understanding is something personal, an act that we carry out, both as speakers and as hearers. For cognitive linguists, understanding means translating a word, a sentence, a text into the language of thought, into mentalese. But there remain many unsolved questions. Are all mental concepts universal, including 'bureaucracy' and 'carburettor', which seem to be rather culture specific? Chomsky thinks there are good arguments to believe that all concepts, including those we are not yet aware of (like future neologisms) are innate (Chomsky 2000, p. 65). Others, like Anna Wierzbicka, think that only a limited

number of basic or primitive concepts are universal and that culture-specific concepts are compositional, in the sense that they are composed of basic concepts. These complex concepts are not universal (Wierzbicka 1996). Jerry Fodor, however, rejects the idea of compositionality (Fodor 1998; Fodor and Lepore 2002) (see also 2.9).

The unresolved question of the nature of mental concepts is only one of the problems cognitive linguists are confronted with. The other main problem is that of the Aristotelian qualia. Daniel Dennett defines qualia as 'the way things seem to us'. Qualia are 'ineffable' (i.e. they cannot be described), they are 'intrinsic' (internal to the mind) and 'private' (known only to oneself) (Dennett 1993, pp. 65, 338ff.). The image the word *primrose* evokes in my mind is different from the image the same word evokes in your mind. The affective qualities that go with it, i.e. what you feel when you hear the word *primrose*, is something you cannot fully convey to other people. It is difficult to see how the assumption of a universal conceptual basis can be reconciled with the view that understanding is a first-person experience that defies communication. But even if there were a consensus among cognitive linguists about how understanding works, it would still be necessary to set it apart from meaning. Meaning is what we trade in when we communicate; by exchanging content we share it. Thus, cognitive linguistics and corpus linguistics have a different focus of interest. The cognitive sciences are concerned with what happens in the mind in the process of encoding and decoding a message. Corpus linguistics is concerned with the message itself.

Corpus linguistics can tell us more about meaning than either Chomskyan linguistics or standard linguistics. Even so, corpus linguistics can never give us the full picture. If meaning is not a formula, an unambiguous expression in some symbolic calculus (which was what many of the adherents of analytic philosophy were hoping for), if meaning is neither a mental image informed by ineffable qualia, nor a universal concept in a language of thought we know nothing about, if meaning is what can (and must be) conveyed verbally, then meaning is something we can talk about only in natural language. In all probability, we know what the word *school* means not because at some point in our past we looked it up in the dictionary. We know what it means because someone, or, more probably, a number of people, must have told us, in the course of our childhood, what it meant. The people who told us must have learned it the same way. This process, or rather activity, of conveying the meaning has been repeated generation after generation ever since there were schools. If we assemble everything that has been said, in this discourse, about schools, then we have the

meaning of *schools*. Not everyone will paraphrase the word *school* for us in the same words. It could well emerge that the common denominator is very small. A good collection of quotations will show this diversity. The following citations are a selection taken from the Bank of English, a 450-million word corpus of English language:

and offers an after- school club. There are infant and
them in detention after school. Yet pupils in adjoining
having a tough time at school and came home in tears again
as they can, because school fees are so unpredictable.
he was sent to boarding school in England, where he was a
small private day school in California. There were
children' s camps during school holidays, which include
at eleven to a grammar school. The rest stayed on at
And, I' m still in high school!'' While rewarding the first
university medical school but it could be rented or
Oxford, said that more school sport is the answer to the
career after leaving music school to start the family, saw it
we are a caring sort of school that looks after everybody' s
written by Head of School, Heather Dixon. 'The two-day
like some kind of prep school, with its Standing Committee
currently still at primary school, later gained a place at
I' ll have to go to public school. Iz and Jude say the teachers
The boy, now 15, skipped school for a year as he took orders
is practical: 'In Sunday School they told us what you do.
last night demanded that the school council and head nun Mother
teenagers. The four go to school, do homework and finish
said: 'I used to walk to school with Lisa and her children.

Corpus linguistics studies languages on the basis of discourse. English discourse is the totality of texts produced, over centuries, by the members of the English discourse community. Even if we confine ourselves to the texts that have been preserved, this discourse is much too large to make it, *in toto*, the object of our research. It will never be possible to study all extant texts. All corpus linguistics can do is to work with a (suitable) sample of the discourse. Such a sample is called the corpus. Because we can never access the whole discourse and not even all extant texts, we can never be sure that what we have assembled as the meaning of a word like *school* will be the full picture. Even more important is the fact that the picture we can deduce from the corpus is full of contradictions. Some like school; others hate it. Some find it useful; for others it is a waste of time. For all lexical items that are worth thinking and talking about, there is hardly a common denominator, there is little agreement. The discourse is not nearly as streamlined as dictionaries want to make us believe. Some lexicographers seem to think that because what we find in our corpus is nothing but an arbitrary and accidental collection of occurrences, this evidence has to be

checked by what *school* is in reality, that it is dangerous to rely only on discourse evidence. But if there is a reality outside of the discourse, it has to be turned into a text, it has to become a part of the discourse, so that it can be communicated.

We should not, therefore, believe that, if we import information which is not found in our corpus, we are importing discourse-external, factual knowledge. We must not mistake for reality what is outside of our corpus. It is still the discourse. We find, for example, in many dictionaries the custom of adding the Latin name of plant species. Thus the *NODE* tells us the species name of the elm tree is *ulmus*. This has nothing to do with reality. It is information copied from other texts, from Linnaeus's classification of plants and animals (2.8 above). This taxonomy is actually a part of discourse and can be discussed in discourse. But isn't this classification, as many people believe, including philosophers of language, a mirror of reality? Isn't a species the same as the natural kind these philosophers (and many cognitive linguists with them) take for granted? Isn't it a fact that there is a species called *elm* or *ulmus* which would still exist even if there were no humans to give it a name? Isn't it true that a tree either is an elm or it is not, regardless of what you or I happen to believe? Is the category species a concoction of the members of the discourse community, or are there, out there in whatever reality may be, entities that can be classified as belonging to this species or that?

Ernst Mayr, a leading biologist and evolutionist, is deeply sceptical about the reality of natural kinds. He recalls, in his recent book *What Evolution Is*, the history of the species concept:

> Traditionally, any class of objects in nature, living or inanimate, was called a species if it was considered to be sufficiently different from any other similar class ... Philosophers referred to such species as 'natural kinds' ... This typological concept is in conflict with the populational nature of species and with their evolutionary potential.
>
> (Mayr 2002, pp. 165–8)

It seems that the concept of species is, after all, being discussed in uncountable contributions to the discourse. A query in Google for 'definition + species' yields 735,000 hits. The concept of species or category allows us to put items into a pigeonhole because they share features we think are important. It is a useful device. But we must not forget that we decide which features are so important that the items sharing them belong in the same pigeonhole. George Lakoff, a cognitive linguist widely known for his work on metaphors, gave one of his books the title *Women, Fire, and Dangerous Things*, because one of the

four noun classes in the Australian language Dyirbal includes females, fire and dangerous animals (among other things; see Lakoff 1987, pp. 92–104).

The discussion about whether there are elms because we have agreed on calling something an *elm*, or whether we call something *elms* because elms exist in reality goes back to a disagreement between Plato and Aristotle. Platonic realism tells us that there are natural kinds, and we cannot do better but acknowledge them and give them names. According to this view, we would not be able, in the long run, to cope with reality, unless we find out and accept what nature really is. This nature exists independently of our giving names to the entities that it comprises. Aristotelian nominalism disagrees. It holds that people are free to put some things into one pigeonhole and other things into another pigeonhole. It is humans who invent categories to make sense of reality; it is not that they discover categories when they investigate reality. We find it important to distinguish oranges from lemons. Yet for some of us, mandarins, satsumas, tangelos and tangerines are all the same. Do they belong to different categories? Is a morello just a kind of cherry or is it a different fruit?

Wherever in the world analytic philosophy prevails, it seems to go hand in hand with some version or other of realism. Actually, this is not surprising. For analytic philosophers, the important question is this: what has to be the case to make a sentence such as 'this is an elm' or 'this is a morello' true? What makes such a sentence coincide with reality? But to ask this presupposes that there are things out there that are elms. We would have to redefine our concept of truth if elms could be anything that we agree on calling *elms*. Cognitive linguistics holds that if not words then certainly concepts are locked onto things out there in what is called reality (Fodor 1994). Thus cognitive linguistics shows itself to be an offspring of analytic philosophy.

For realists it is therefore very important that the things words stand for really exist and are not just chimeras like the Nazi concept of race. John Searle, a highly distinguished scholar within the philosophy of mind community, tells us in his recent book *Mind, Language and Society*: 'Among the mind-independent phenomena in the world are such things as hydrogen atoms, tectonic plates, viruses, trees and galaxies. The reality of such phenomena is independent of us' (Searle 1998, pp. 13–14). Can we be sure of this? Two hundred years ago, people had never heard about hydrogen atoms, tectonic plates or viruses. But they thought they knew, as a fact, that there was phlogiston, a combustible matter that escapes into the air whenever something is burning. Will we, in another two hundred years, still be happy to describe certain

macromolecular structures with an ability to replicate as viruses? Or, for that matter, can we be so sure about the reality of trees? Are there irrefutable criteria to distinguish trees from shrubs or bushes? The *NODE* calls the hazel 'a temperate shrub or a small tree', for the *Cobuild* it is only 'a small tree'. For Germans, it is either a bush (*Haselnussbusch*) or a shrub (*Haselstrauch*), but never a tree. What we call a tree depends, it seems, more on decisions taken by the language community than on facts.

In the Middle Ages a meeting of bishops declared rabbits to be fish. This gave them permission to have rabbit on their Friday menu. Today we are wiser. We know that rabbits belong to the category of rodents. But is this category more real than a category grouping together things that a good Catholic could eat on a Friday? That rabbits belong to the category of rodents seems to be scientifically true, whereas the category of things permitted as food for Fridays is entirely arbitrary and no longer widely accepted. But the Linnean system of classifying plants and animals in terms of relationship and ancestry is not perennial; it became accepted in the Western world in the course of the nineteenth century, and perhaps it will be superseded one day by a new classification based on DNA. Which categorial systems refer more directly to reality, if it is possible to ask such a question?

So if we do not find in our corpus something that tells us what a word means, where are the facts that determine that word's meaning? Facts, as we have seen, only become facts once they are introduced into the discourse. They may be, for all we know, external to the discourse. But it is up to the members of the discourse community to introduce into the discourse what they deem to be facts. The vast majority of things we think are facts, or what we think we know to be true, are things that we have never encountered or investigated personally but have been told about in discourse. Some people say they know, as a fact, that there are weapons of mass destruction in Iraq. They have never been there; they have never investigated the existence or non-existence personally; and they are relying on texts that are part of the discourse. For any one of us (perhaps other than a leader like the president of the United States of America) it is quite impossible to establish a fact without having it negotiated by the discourse. It is the discourse that decides whether a phenomenon is real or not. There may be plenty of facts outside the discourse, but the only facts we can talk about are the ones that have been introduced into the discourse.

It therefore seems obvious that the only source we can ever hope to access about the meaning of a word is the discourse. We cannot hope to make the discourse as a whole accessible to our lexicographic

enquiries, but we can compile larger and larger corpora, and we can also use the ever-growing Internet as a virtual corpus. Nevertheless, as new words and phrases are coined day by day, it is conceptually impossible to come up with a corpus that comprises the whole vocabulary of a discourse community. There will always be words which are not contained in our corpus. And there is always the chance to add to our corpus the texts in which these words occur. When it comes to the meaning of words, corpus linguists have to consult their corpus, amend it, consult it again, and so forth, in a Sisyphean effort. What corpus linguists make out as the meaning of words, can, thus, never be more than an approximation. A different, a larger corpus can always come up with new paraphrases that were missing from the original corpus.

All communication acts together constitute the discourse of a given discourse community. There is, you could say, a discourse community of all people speaking English. It has existed for centuries, ever since English was around. In it we have the texts written by Geoffrey Chaucer, William Shakespeare, Elizabeth Gaskell and Sylvia Plath, and all the other texts we find in our libraries and archives. We have lost, of course, all the oral communication acts (with the exception of some recent ones) because they could not be recorded, and we have lost most of the unprinted written material, because it was thrown away. All those texts are part of the discourse. We can never study all of it, not even what is extant.

Noam Chomsky and many of his followers have dismissed the corpus as the source of our linguistic knowledge. Language, they say, is productive. With limited means, a finite vocabulary and a manageable set of rules, our language faculty empowers us to generate an infinite number of utterances. All the time things are being said that have not been said before. Corpus research, they claim, will only tell us what people have said so far. It will not tell us what people are going to say tomorrow. That is certainly true. Corpus linguistics cannot predict language change any better than meteorologists can predict the weather of tomorrow or of next week. When Ted Levitt used *globalization* in the title of an article 'The globalization of markets' he published in the *Harvard Business Review* in 1983, he could not have known, and linguists were not able to predict, that globalisation would become a keyword of the 1990s.

Generative linguists, however, are not, as we have seen, very much concerned with semantic change. They are interested in grammar. Of course, grammar also changes over time. If we regard quotatives as part of grammar and not of the lexicon, then it is an example of grammatical change that it is now possible to say: 'He comes into the room

and he is like "It's much too hot for me in here", and he turns on the air'. Our old grammars do not list the construction *be like* + direct speech. But is this what the generative grammarians have in mind? What they mean by the generative force of grammar is that using the very same grammar (the grammar of the ideal native speaker) we can produce an infinite set of sentences. This is certainly a true claim, even though Chomsky also admits that 'expressions of natural languages are often unparseable (not only because of length, or complexity in some sense independent of the nature of the language faculty)' (Chomsky 2000). Whatever conforms to rules (some expressions apparently do not) will not be better confirmed by looking at data. More empirical evidence will not make us wiser. Once we have found out that sound travels in standard air at a speed of 330 metres per second, there is no point in examining ever more sound events. If you have learned to inflect Lithuanian nouns with their seven cases correctly, there is absolutely no need to study the inflections of Lithuanian nouns in a corpus. If you know for sure that split infinitives are 'illegal', no amount of split infinitives in your corpus will make them legal. Corpus linguistics should keep its hands off grammar, to the extent that the rules we find in our grammar books are indisputable. (They are not always, though.)

Therefore, in this sense, corpus linguistics is no help when it comes to studying the grammar of a language of which the rules have already been 'discovered'. (However, are these 'discovered' rules always adequate?) But it can tell us more about the meaning of words than standard or Chomskyan linguistics. It extracts from the discourse all that we can find out about meaning. Natural human language is unique in this respect. It is the discourse community that negotiates how words should be used and what they mean. The result of these negotiations is not always agreement. Some people may say that *weapons of mass destruction* is a neutral and unbiased expression; others may say it is derogatory because you only use it for the weapons of your enemy. There seems to be no common understanding what these weapons of mass destruction exactly are, and, consequently, what the phrase *weapons of mass destruction* means. Do cluster bombs belong in that set? What about depleted uranium? We only have to look at the recent discourse to find numerous citations in which people are keen to tell us what they think weapons of mass destruction are. A search in the Bank of English on weapons of mass destruction shows us that they stand against the conventional weapons and most commonly mean biological, chemical and nuclear weapons, as in the following citations:

> Terrorists were seeking weapons of mass destruction: chemical, biological and nuclear.
>
> ... Bush's policy goal of regional security and stability meant eradicating Iraq's capability to build weapons of mass destruction – chemical, biological, and nuclear –...
>
> The Security Council is still not satisfied that all weapons of mass destruction, notably biological and chemical arms, have been purged from Iraq...
>
> The evidence that it is assembling biological, chemical and other weapons of mass destruction is overwhelming.

But the corpus tells us much more than that, it shows us how black and white our world picture is. It tells us that indeed when we talk or write about the weapons of mass destruction, we often mean Iraqi (or other enemy) weapons, that it is very often Iraq or Baghdad that is developing, producing, building, acquiring these weapons, and that it is the United Nations who is banning or trying to eliminate them from the Middle East.

The discourse is full of paraphrases of words and of comments concerning their meaning and the connotations that come with them. Aren't these explanations the kind of information we would like to find when we look up a word or a phrase in the dictionary? Once we take the view that the meaning of words is what members of the discourse community proffer as their meaning, the distinction lexicographers have become attached to, namely the distinction between lexical knowledge and encyclopaedic knowledge, dissolves. Encyclopaedic knowledge is part of our discourse just as much as whatever dictionaries offer as word meanings. The meaning of the phrase *weapons of mass destruction* is what people tell us *weapons of mass destruction* are. Similarly, the true meaning of *water* is not, as the famous American philosopher Hilary Putnam wants us to believe, what water is 'in reality', but what people tell us water is (Putnam 1975, pp. 215–71).

Corpus linguistics questions the position of the word as the core unit of language. The word is not inherent to language. The Greek word *logos* which we usually believe to be the equivalent of *word* means primarily 'speech' or the 'act of speaking', then 'oral communication', and also an 'expression'. Where it does mean 'word', it means first of all the 'spoken word' (as opposed to *rhema* or *onoma*). Latin *verbum* also means first of all 'expression', 'speech' and 'spoken word'. When we think today of *word*, it seems to be much less a transitory sound event than the written word, something that can easily be identified because it is preceded and followed by a space, a space we normally do not speak or hear. Spaces between written words are a

relatively recent invention. It was the monks in the medieval *scriptoria* who introduced them because it made it easier to copy texts. Words are what constitute dictionary entries, and because *weapons of mass destruction* is not a single word, it is hidden away in the dictionary, if it occurs at all. In the *NODE*, the phrase is found under the entry for destruction: 'the action or process of killing or being killed: weapons of mass destruction'.

3.5 A brief history of corpus linguistics

Corpus linguistics is a fairly new approach to language. It emerged in the 1960s, at the same time as Noam Chomsky made his impact on modern language studies. His *Syntactic Structures* appeared in 1957, and while it quickly became a widely discussed text, it was only the publication in 1965 of his *Aspects of the Theory of Syntax* and the subsequent reception of this work that provoked the revision of the standard paradigm in theoretical linguistics. Yet while language theory became increasingly interested in language as a universal phenomenon, other linguists had become more and more dissatisfied with the descriptions they found for the various languages they dealt with. Some of the grammar rules in these descriptions were so obviously violated in all (written) texts that they could not be adequate. Certain features of the language were insufficiently described. For example, there had always been a distinction between transitive verbs and intransitive verbs. This is not enough, however, to describe the number and quality of objects or complements that can depend on a verb. These objects include the direct object, various kinds of indirect objects, prepositional objects and clausal objects, among others. They have to be properly kept apart if we want to describe grammatical structure accurately. For instance, if a verb is turned from active into passive voice, some objects can disappear while others will become subjects. In the 1950s, details such as these raised empirical questions which could not be answered by introspection alone. Real language data were needed.

In the English-speaking world, the first large-scale project to collect language data for empirical grammatical research was Randolph Quirk's Survey of English Usage which later led to what became the standard English grammar for many decades: *A Comprehensive Grammar of the English Language* (Quirk *et al.* 1985). The project kicked off in the late 1950s. It formed a reference point for anyone interested in empirical language studies, including the Brown Corpus to be mentioned below. But at the time, the Survey did not consider computerising the data. This happened much later, in the mid-1980s,

in Quirk and Greenbaum's subsequent project now known as the International Corpus of English (ICE) (http://www.ucl.ac.uk/english-usage/ice/).

Quirk's Survey was a mixture of spoken and written data; there were about 500,000 words of spoken English within a total of one million words. The spoken component was actually the first to be put on a computer, by Jan Svartvik, and became, in the late 1970s, the London Lund Corpus. It was transcribed in an elaborate way, with much phonological and even phonetic information. It became the first spoken corpus widely available for use, published as a book, though unfortunately still not available as a soundtrack (Svartvik 1990).

The Survey was mostly interested in grammar, not in meaning. Nevertheless, it was one of the very few projects working on empirical data. Due to the pervasiveness of the Chomskyan paradigm, it became increasingly difficult in the 1960s to find acceptance of this kind of data-oriented language research. The Survey was the exception in Britain at that time. Later, in the 1970s, this strand of research was to be taken up by a number of Scandinavian linguists, most of them based in Bergen, Lund and Oslo.

The second data-oriented project in the 1960s was the Brown Corpus, named after Brown University in Providence, Rhode Island, where it was compiled by Nelson Francis and Henry Kučera. The corpus consists of one million words, taken in samples of 2,000 words from 500 American texts belonging to 15 text categories as defined by the Library of Congress. The Brown Corpus was a carefully organised corpus, very easy to use, and proofread until it was almost free of mistakes. So is the similarly composed corpus of British English, the LOB (Lancaster–Oslo–Bergen)–Corpus from the 1970s (Johansson *et al.* 1978). Later, both corpora were manually tagged with part-of-speech information. While it was at first hoped that these corpora would answer questions concerning both the grammar and the lexicon, it was soon realised that a corpus of one million words cannot contain more than a tiny fraction of the whole vocabulary. After the Brown Corpus was compiled and the proofreading was completed, it seemed that linguists, at least in America, lost interest in it. It hardly played a role in transatlantic linguistics, even though it became a popular resource in European linguistics. The LOB–Corpus was exploited in subsequent corpus studies, for research into grammar and, more importantly, into word frequency, but not into meaning, mostly in co-operation between British and Scandinavian scholars, including Geoffrey Leech, Knut Hofland and Stig Johansson.

It seems it was Nelson Francis who was the first to apply the term

corpus to his electronic collection of texts. John Sinclair believes this is how the new usage may have originated:

> There is a story that Jan Svartvik tells about him [Nelson Francis] coming to London with a tape containing the Brown Corpus or part of it and meeting Randolph Quirk there in the mid sixties. Nelson threw this rather large and heavy container, as tapes were then, on Quirk's desk and said: 'Habeas corpus'. Francis also uses *corpus* in the title of his collection of texts, i.e. the Brown University Corpus, and as such it is referred to in the OSTI Report.
> (Interview with John Sinclair in Krishnamurthy 2003)

A third, and certainly most important, early corpus project was English Lexical Studies, begun in Edinburgh in 1963 and completed in Birmingham. The principal investigator was John Sinclair. It was he who first used a corpus specifically for lexical investigation, and it was he who took up the novel concept of the collocation, introduced in the 1930s by Harold Palmer and A. S. Hornby in their *Second Interim Report on English Collocations* (1933), and then taken up by J. R. Firth in his paper 'Modes of meaning' (Firth 1957). This project investigated, on the basis of a very small electronic text sample of spoken and written language, amounting to not even one million words, the meaning of 'lexical items', a term that included collocations. John Sinclair's final report, *English Lexical Studies* (often referred to as the OSTI-Report), was distributed in no more than a handful of typewritten copies in 1970. It was often referred to in later studies, but has only recently been published properly for the very first time, by the Birmingham University Press (Krishnamurthy 2003). At the time, Sinclair had not yet completely abandoned the notion of the word as the unit of meaning, but he was keen to modify the traditional view of the word as the core unit. Still, while the project participants explored the relationship between the word and the unit of meaning, there was no clear appreciation of semantic units as multi-word units with their variations stretching across the phrases. A beginning had nevertheless been made.

Unfortunately, in the 1970s, 1980s and even 1990s, the quest for meaning all but disappeared from the agenda of the newly established corpus research. This is not as astonishing as it sounds. After all, compiling corpora, particularly larger ones, posed a host of problems, mostly technical ones, but also the still popular question of representativeness. Was there a corpus that could be said to represent the discourse? Was it possible to define text types, domains or genres in general terms? Was there a recipe for the composition of what came to

be called a reference corpus? How important was size? What was the role of special corpora?

Standardisation also became an issue of overriding importance for the 1980s and 1990s. How should corpora be encoded? Was it permissible to add corpus-external information in the form of annotation or tagging? Could there be a common tag-set for all languages? Wouldn't using annotated corpora mean that you only extract from them what you first added to them, thus perpetuating possible misconceptions?

Then there is the question of frequency. With corpora, it was, for the first time, possible to come up with lists of the most frequent words accounting for the basic vocabulary. Everything could be counted and compared: verb–complement constructions, the distribution of the various relative pronouns, or the position of adjectival modifiers in late Middle English noun phrases. Register variation of different Englishes is still a common topic of many corpus studies. Frequency information could also shed new light on grammatical rules. It became possible to investigate the relationship between rare events and a decrease of linguistic competence, of what one could say and what one would say. In this sense, frequency data could be used to revise our view of syntax.

If we look at the papers from the 13th and 14th International Conferences on English Language Research on Computerised Corpora (Aarts *et al.* 1992; Fries *et al.* 1993), organised by the venerable ICAME association, these were very much the topics presented there. The papers deal with creating corpora, with corpus design questions, with annotation, with language varieties and with parsing techniques. Among the thirty-eight papers presented at the two conferences, perhaps four or five focus on collocational aspects of language and only one explicitly deals with semantic issues: Willem Meijs on 'Analysing nominal compounds with the help of a computerised lexical knowledge system'. Here, too, then, we learn very little about extracting meaning from the corpus, and more about assigning predefined semantic features from a conceptual ontology to collocations found in the corpus.

It is not astonishing that the final report *Towards a Network of European Reference Corpora* (finally published in 1995) of the 1991/92 European Commission project talks about user needs, corpus design criteria, encoding, annotation and even knowledge extraction, but does not touch on meaning as a possible focus of corpus research (Calzolari *et al.* 1995). Even the introductions to corpus linguistics which appeared in the 1990s refrain from devoting much space to the corpus-oriented study of meaning. Tony McEnery and Andrew Wilson

(McEnery and Wilson 1996) may serve as one example. Forty pages of their book are devoted to encoding, twenty pages deal with quantitative analysis, twenty-five pages describe the usefulness of corpus data for computational linguistics and thirty pages cover the use of corpora in speech, lexicology, grammar, semantics, pragmatics, discourse analysis, sociolinguistics, stylistics, language teaching, diachrony, dialectology, language variation studies, psycholinguistics, cultural anthropology and social psychology. The final twenty pages present a case study on sub-languages and closure. In Graeme Kennedy's introduction to corpus linguistics (Kennedy 1998) thirty pages out of three hundred are devoted to 'lexical description', including twelve pages on collocation. Unsurprisingly, for Kennedy lexical description seems to be more or less synonymous with frequency information. In their book of similar size *Corpus Linguistics: Investigating Language Structure and Use* (also 1998) Douglas Biber, Susan Conrad and Randi Reppen again have about thirty pages on 'lexicography'. The two basic questions they address are: 'How common are different words? How common are the different senses for a given word?' (Biber *et al.* 1998, p. 21). This looks like frequency analysis together with the belief that word senses are somehow discourse external and can be assigned to lexical items. But at least they mention, on two pages, the relevance of the context for determining senses. The rest of the section is devoted to an investigation into the distribution of the word *deal*, with its various senses, over the registers of different text genres. In the absence of an introduction dealing explicitly with matters of meaning, John Sinclair's *Corpus, Collocation, Concordance* (1991) filled the gap, until Michael Stubbs' *Words and Phrases: Corpus Studies of Lexical Semantics* was published in 2001.

There was, however, a large corpus-based dictionary project, the *Collins Cobuild English Language Dictionary*, conceived and designed in the mid-1970s and published in 1987, under the guidance of John Sinclair. The story of this venture is told in *Looking Up: An Account of the Cobuild Project in Lexical Computing*, also published in 1987. This was the first ever general language dictionary based exclusively on a corpus. Therefore, the corpus had to be big enough to include all the lemmas and all the word senses the dictionary assigned to these lemmas. A consequence is that rare words, like *apo(ph)thegm*, are missing. They were not in the corpus. However, except in cases of doubt the lexicographers did not use corpus information to carve up the meaning of a word into its senses; rather, the corpus was used in the first place to validate the lexicographers' decision and to provide examples. More could not be done with this corpus of 18.3 million words (Birmingham

Collection of English Text), then the largest general language corpus in the world. From today's point of view, collocations are not given the prominence they ought to have. Dictionary publishers have not been keen on collocation dictionaries. In many ways, the *Cobuild* dictionary is still unique. While it encouraged other dictionary makers to include more corpus evidence, there is still no other dictionary exclusively based on a corpus.

Elena Tognini-Bonelli distinguishes between the corpus-based and the corpus-driven approaches (Tognini-Bonelli 2001). Linguistic findings (including the contents of dictionaries) are corpus based if everything that is being said is validated by corpus evidence. Findings are corpus driven if they are extracted from corpora, using the methodology of corpus linguistics, then intellectually processed and turned into results. This is a crucial distinction. The corpus-based approach will deliver only results within the framework of standard linguistics. It can show that one of the five senses normally listed for *friendly* does not occur at all in the corpus, and that in addition to the five senses, there is another usage that has been overlooked by other dictionaries. It will not show that you can get rid of most of the ambiguity by identifying the collocates of *friendly* and making these collocations your lemmas. If corpus linguistics is really going to complement standard linguistics rather than just extend it, it must follow the corpus-driven, not the corpus-based approach. This is what we aim to demonstrate in the following chapter.

4 Directions in corpus linguistics

Wolfgang Teubert and Anna Čermáková

4.1 Language and representativeness

Ever since linguists started using corpora they have been thinking hard about how corpora should be composed. The corpus should represent the discourse, or some predefined section of it. What the Brown Corpus represented was the English language of the year 1961, in print, as catalogued by the Library of Congress. In this corpus, each publication is assigned to one of fifteen content categories. The catalogue for the publications of 1961 represents this discourse. It tells us how many texts were published within each of the categories, and these figures were used as guidelines to select the texts. From each of the 500 texts chosen, a 2,000-word sample was then entered into the corpus. This selection process can be operationalised, turned into unambiguous, clear instructions, and is therefore objective. But is the corpus representative?

It represents, in a rather loose way, the Library of Congress catalogue. That is not the same, though, as the discourse constituted by all the printed publications of the USA in 1961. The fifteen categories into which the catalogue entries are divided are arguable. You could have more or fewer, and the subject fields could be defined quite differently. A few centuries ago, there would have been a category for alchemy and one for astrology, but none for economics. The whims of people change. Depending on the number and content of these basic categories, one might come up with an entirely different selection of texts for our corpus, a selection which was in every respect as objective as that of the Brown Corpus.

Then there is the question of readership. In a catalogue, a newspaper with a circulation of several million copies has an entry comparable to a book printed in 120 copies. But is the number of readers important? What really determines the importance of a text: who wrote it? How many copies circulated? How many people read it? Is it right to include only printed and published texts and thus to exclude perhaps more

than 90 per cent of what makes up the discourse of any given year: informal conversations within the family, in schools, in bars, cafes and clubs, with friends on an outing, at the workplace; the letters we receive, the advertisements we read, the reports, minutes and memos we find on our desks, to name but a few?

There are wider questions. Are English texts published outside of the USA, but found on the shelves of the Library of Congress, part of the American discourse? What about books published by Americans who live outside the USA? Does American English include, for example, the English spoken by immigrants in the USA or the English of Puerto Ricans? What exactly is the discourse?

A language, a discourse, consists of the totality of verbal interactions that have taken place and are taking place in the community where this language is spoken. This community we call the discourse community. Language communities can be small. Some are so small, in fact, that their languages have become endangered, or even extinct as with many of the Uralic (Finno-Ugrian) languages. There are (or were) languages spoken by only a dozen people or even less. Manx, for example, died out a few decades ago; at the end, there was only one (native) speaker left, conversing in Manx only with the handful of linguists specialising in this Celtic language. Other language communities are so large and diverse, like the community of English-language speakers, that it does not seem proper from a sociological perspective to call them communities at all.

The totality of the verbal interactions of a specific language community includes idiolects, sociolects, dialects, regional variants, languages for special purposes, eighteenth-century language and contemporary language, female language and male language, slang and jargon, and innumerable other kinds of language we can sometimes distinguish.

Languages and discourse communities do not exist as such. They are social constructs. We construe them to suit our purposes. Until the dissolution of the old Yugoslavia, most of us believed there was a language called Serbo-Croatian. Now there are books telling us that such a language never existed, and that Serbian and Croatian were always distinct languages. Nowadays words considered to be originally Serbian (or even Turkish) are purged from Croatian and replaced by newly coined words built from 'purely' Croatian morphemes. Half a century ago the Northern Indian lingua franca Hindustani (a pidgin that become a Creole) was replaced by Hindi and Urdu. Both of these were originally at least as artificial as Hindustani, yet today, thanks to massive political intervention, they are irrefutably natural languages in their

own right and to a large extent mutually incomprehensible. Germans normally do not understand spoken Swiss German, but tradition has it that it is the same language. Slovaks and Czechs do not need interpreters to understand each other, but historical and political circumstances have enforced the notion that they are two separate languages. There is no formula telling us what a language is and what a language community is. It is up to us to design our formula in agreement with our intentions. We define languages and language communities according to experience, according to what seems useful at a given time.

Discourse communities may be social constructs, but we do experience them as real. The members of a discourse community negotiate who belongs to it and who does not. There are thousands of texts telling us, as a 'fact', how many speakers there are for English, or Chinese or Manx. The discourse itself is unfathomable, inexhaustible, and as a whole, inexplorable. Perhaps we can approach the conundrum of representativeness more easily if we approach it from the other end, from the corpus.

In the words of John Sinclair (1991) a corpus is 'a collection of naturally occurring language text, chosen to characterize a state or variety of a language'. The texts are all samples, cross-sections of the discourse. But sampling the discourse can mean different things. If we look at the discourse of written English texts, we could, if we chose to, say that a representative corpus is one that reflects the frequencies and proportions of all the twenty-six letters plus the special characters like punctuation marks and the space in between words. Is that what we should call a representative corpus?

Perhaps, though, we are interested not so much in the frequency of letters as in the frequency of words. Let us assume, again, that there are half a million words in English (the number does not matter, really, because we will not agree on the definition of word). Some of these half-million words are very frequent, such as the function words (*a, the, to,* etc.; see 1.1); some of them are quite frequent (say, the 20,000 or so headwords you would find in a typical pocket dictionary); and the rest of them are rather less frequent. The most frequent word in English is the definite article *the* and nearly all of the most frequent hundred words are function words such as pronouns and prepositions. Among the most frequent words there are only a very few nouns and verbs which can be said to have a meaning of their own, and all of these words are highly ambiguous or fuzzy (words like *thing* or *set*). All words are part of the discourse; they all have been used at least once. But no matter how large our sample of the discourse is, we will miss most of them. There is no occurrence of my favourite word *apophthegm* in the

450-million-word Bank of English. This is purely accidental. There are, on the other hand, many thousands of other words nobody has ever heard of, words occurring only once in the corpus, for example *abelch*, *airpad*, *eurocrisis* and *keyphone*. Such a word, for which we have no more than one citation, is called a *hapax legomenon* (Greek: 'read only once'). Some of these words may be misspellings but many may be real words.

Therefore, talking about the frequency of words, we just may be able to say that a corpus represents a discourse, inasmuch as the 10,000 most frequent words of the discourse are also the 10,000 most frequent words of the corpus. The presence or absence of words less frequent is as unpredictable as the winning numbers in a lottery. But even if we only consider the most frequent part of the vocabulary we find ourselves at a loss.

In whatever way we look at the question of representativeness, we will always have to define what it is that our corpus is to represent. As long as we have not defined what the discourse is which we want to represent, we just do not know what the 10,000 most frequent words are. Nor do we know how different domains (such as politics, gardening, property law or rugby) are distributed over the discourse. The same is true for genres, based on text-external classification (fiction, newspaper language, academic writing, appliance instructions, poetry), or text types, based on text-internal features (containing first person singular, past tense, passive, quotations, etc.), and for registers (e.g. formal, informal, technical, derogatory, vulgar language).

There have been many attempts to define the discourse, and the catalogue of the Library of Congress is just one example. You might want to compile a corpus representing the discourse of Australian English of the year 2000. Since we cannot hope to have easy access to spoken texts, let us restrict our discourse to written texts. Let us also exclude, for the moment, unpublished written texts. We thus narrow our definition of the discourse for which we want to compile a corpus down to the totality of written English texts, published in Australia in the chosen year. Is that what we want? Let us further assume we have agreed on what to do with texts by Australians published outside Australia and texts by non-Australian English writers published within Australia, and that we have agreed upon the relationship of writing texts (sampling authors) to reading texts (sampling readers). There are still other parameters such as gender, educational background and age of writer and/or of reader. Probably for a country like Australia, some linguists are also interested in the ethnic backgrounds of the writers/readers of texts. These are parameters defining the discourse community, and not the discourse. Are these parameters we should be interested in?

In any case, we are only justified in claiming that a given corpus is representative of a discourse, however we have defined it, if we have, at least in principle, access to all the texts the discourse consists of. Only then will we have all the relevant information concerning the parameters mentioned above, and only then can we be sure that the corpus we compile as a sample of this discourse is representative, at least in respect to the parameters mentioned above. But if this utopia came true, we could well do without the corpus. We would already have the discourse as a whole and would not need to sample it. We would have no need to work with a sample. We could work with the whole discourse. Perhaps in a decade or two it will be possible to access all the written texts published in Australia in a given year. But presumably it will never be possible to enumerate all the spoken and written texts of the Australian discourse of a given year as a whole. This is why it does not make much sense to talk about representativeness.

4.2 Corpus typology

However, this might not be at all what we mean when we discuss the representativeness of a corpus of British, Australian or American English. Fortunately, these discourse communities have discussed at length what they mean by standard English. There is a good measure of agreement about what kind of English we should teach to foreigners. Of course, these attitudes have changed in the course of history. A century ago, languages were mostly taught on the basis of 'good' literature, including novels and non-fiction books on certain cultural and historical topics. Today we think the register to be taught should also include the kind of spoken language used by the educated middle classes. This is not the most numerous segment of society, but, in the eyes of the discourse community, it is the most prominent or significant. We could also define as standard English the private annual reading load of educated middle-class citizens. This might consist of a larger share of broadsheets and a smaller one of tabloids, amounting to probably 50 per cent of the total, perhaps another 10 per cent consisting of the weekly and monthly periodicals we subscribe to, perhaps 25 per cent consisting of fiction and non-fiction books we read through the year, and the remaining 15 per cent an odd mixture of brochures, instruction leaflets and sundry printed material we come across, from tax forms and telephone bills to tourist brochures and theatre programmes. We could include, in proportion, children's literature. There is no cogent reason to exclude our professional reading load, but many people, including many linguists, would think that

these texts are too diverse and too far away from what we usually read that they would not belong to any common ground. The language of a medical journal, of aircraft maintenance manuals or the customs regulations, for example, is not part of my linguistic repertoire. Our corpus thus comes to consist of what the members of the discourse community have agreed to be representative of standard (British, Australian or American) English. For the reasons discussed above, we should not call a corpus which represents such a socially accepted standard a representative corpus. These days, corpus linguists prefer to call such a corpus a reference corpus.

Today, corpus linguists would expect a national language reference corpus to comprise between 50 and 500 million words, if not more. There are perhaps one or two dozen languages for which reference corpora of this (or larger) size already exist or are under construction. For German, there is the IDS (Institut für Deutsche Sprache) corpus with more than a billion words; there is the Språkbanken Swedish corpus of 75 million words and the Czech National Corpus of 100 million words; and there are two large reference corpora for English, the 100-million word British National Corpus and the 450-million word Bank of English. Reference corpora of different languages are comparable if they are similar in size and if their composition is similar in respect to genres and/or other parameters. The PAROLE corpora of all official EU languages, for example, are comparable in this sense, but, given today's standards, they are rather small – not more than 20 million words per language.

Reference corpora are being used for a multitude of purposes. Reference corpora contain the standard vocabulary of a language. They are the corpus linguist's main resource to learn about meaning. If they are large enough, they reveal the contexts into which words are usually embedded, and with which other words they form collocations. Only in corpora of this size can we detect these units of meaning that are so much more telling than single words, with their ambiguities and fuzziness. We need reference corpora, the larger the better, for investigating lexical semantics. A typical reference corpus will represent what the discourse community agrees to be what a fairly educated member of the middle class would read outside of work, mostly in printed form, but also handwritten or typed; and, in principle at least, it should also contain a sample of what they would hear, in conversation, at more formal social events, or on the radio. It is carefully construed, with a deliberate composition. The British National Corpus of 100 million words, compiled in the early 1990s, is a good example (http://www.hcu.ox.ac.uk/BNC/).

Reference corpora, however, also serve another purpose. They can be used as benchmarks for special corpora. Whenever we do not want to look at standard language as a whole but at some special phenomenon we happen to be interested in, we usually have to compile a corpus that fits our research focus. Such a corpus is called a special corpus. Special corpora are sometimes quite small, under a million words, though they can be much bigger of course. Let us assume, for the moment, we are interested in the collocation *friendly fire*. Our research questions are: how quickly did this neologism spread after it was first coined in 1976? How was it paraphrased? Are people aware of the inherent irony? Are there different usages? Does it occur only in talking about the military, or are there other domains in which we now use *friendly fire*? When did the expression pick up in British English? What was *friendly fire* called before the expression was coined? What happened to that word? If this is the set of questions to which we want to find answers, we have to compile an appropriate corpus. It must include texts from 1976 onwards. In order to find out the frequency over longer periods of time, we must set up subcorpora for different phases, say for 1976–8, 1980, 1985, 1990 and 2000. These subcorpora have to be identical in size and in composition so that it really makes sense to compare frequencies. The easiest way to come up with such a set of comparable subcorpora is to take newspapers. So perhaps we should take *USA Today*, the *Washington Post*, the *Los Angeles Times*, the *Burlington Intelligencer* and the *Springfield Examiner*, for the years we want to look at. Where do we find these newspapers and their text files? Some papers publish all the year's texts in annual CDs, which can be bought. In other cases, we may have to contact the publisher. If we cannot get a paper, we must find a suitable substitute. We should look around for databanks on the Internet containing this kind of material. Do we need the whole newspaper? In fact, we could reduce the size of the corpus by selecting only those articles in which the phrase occurs. Or should we just take the sentences containing *friendly fire*? Better not, for it may well be that relevant semantic information may be found in the wider context. Can we leave out certain sections, such as sport? Not if they contain the phrase. We then have to compile a very similar corpus of British newspapers, to be able to find similarities and differences. If our British corpus includes the (London) *Times* of 2 November, 2001, we will find the following citation:

> Blair's war effort is put under friendly fire. Labour rebels, disgruntled backbenchers, forced the first Commons vote on the conflict in Afghanistan.

This citation tells us a number of things: first of all, *friendly fire* is also used in British English. Second, the phrase is also used in the political domain. Third, if used outside of the military domain, we also find a metaphorical usage. These backbenchers do not use guns or missiles; they use their right to vote. And finally, there is a semantic difference that comes with metaphorisation: this friendly fire is no longer accidental; it is intentional.

There is no standard recipe for the composition of a special corpus. All we have to do is to draw up a set of hypotheses that will guide us in defining the special corpus we need. It may well turn out that in the course of our inquiry, our hypotheses have to be modified. This may mean it becomes necessary to extend our corpus. It is always possible to add to it. Corpora are by no means sacrosanct. They are the corpus linguists' creation, and they can do with them whatever they deem reasonable.

An alternative to the reference corpus is the opportunistic (or cannibalistic) corpus. The opportunistic corpus does not claim to represent a language or to mirror a discourse; an opportunistic corpus is based on the assumption that each and every corpus is imbalanced. Once we take for granted that corpora are inherently imbalanced, we are free to tackle the problem of representativeness or balance from a different angle. This new perspective is the strict separation of corpus compilation from corpus application. The opportunistic corpus is the result of collecting all the corpora one can lay hands upon. Almost all of these corpora will be special corpora; but there may also be a few that call themselves reference corpora. The larger the opportunistic corpus is, the better it is. But the best opportunistic corpus is also the one that is documented in the most comprehensive way.

Therefore we first have to define the genres, domains and text types, and in doing so we have to take into consideration two aspects. One aspect is what possible users of our opportunistic corpus might want to look for. What are the genres, domains and text types that have been discussed and analysed by linguists? This should be the starting point of our own classification. The other aspect is what kind of information we can hope to find in any corpus we want to add to our opportunistic corpus. Will there be any genre/domain/text type information, either in the texts or attached to them, that we can retrieve automatically and include in the documentation files? This is an important question because the integration of a new corpus into an existing opportunistic corpus should be done as automatically as possible. What we also need is, for each text, the date of its publication, the name and details of the author, its title and all the other information one would like to put into

a bibliography. An ideal opportunistic corpus is a corpus in which this kind of information is available for each text of each of the corpora it is composed of.

Once there is a sufficiently large opportunistic corpus available, people who want to use corpora for their research can query the documentation in order to identify the texts they would like to use. Someone working on the vocabulary of Victorian novels would select all the relevant novels they would find in the opportunistic corpus, and leave the rest aside. Another project might be an exploration of the special language of sport. Opportunistic corpora will contain a lot of newspaper material, and again the thorough documentation of all texts will make it very simple to select the ones we are interested in, i.e. the sport sections of the newspapers, plus whatever other material is classified as belonging to the domain of sports. There will always be research topics for which an opportunistic corpus does not provide the basis. The larger it grows, however, the wider will be the variety of research purposes it fits. Indeed, whenever people maintaining an opportunistic corpus come across some special corpus which could be useful for a future research agenda, it should be added. Opportunistic corpora are principally open-ended. The corpus holdings of the Mannheim Institut für Deutsche Sprache, now running at more than two billion words and still growing rapidly, are currently the largest opportunistic corpus.

The monitor corpus is a corpus that monitors language change. It is, in principle, regularly updated and open-ended. Corpus linguistics is particularly interested in lexical change, such as:

- the change of frequency of words or other units of meaning (compounds, multi-word units, collocations, set phrases), which is often indicative of a change in meaning or a change in the domains in which words are used;
- the occurrence of new words;
- the occurrence of new larger units of meaning;
- changing context profiles, i.e. changes in the frequencies of words occurring in the contexts of words or other units of meaning.

The introduction of new words into the discourse is the most obvious, but not the most frequent lexical change. Studies undertaken for a German newspaper (the *Süddeutsche Zeitung*, Munich) have shown that the majority of new character strings in between blanks, i.e. strings that have not been previously registered, are first typing errors and then names of persons, organisations and geographical units, then

abbreviations. The small remainder (about fifteen items per day) consists of previously unrecorded forms of words already registered, ad hoc compounds (which are written in one word in German) and every now and then a true neologism.

Monitor corpora should, as much as possible, adhere to the same initial composition. As far as they consist of newspapers and periodicals, this should not be difficult. Newspapers tend to develop their own unique styles, and this style manifests itself in a specific vocabulary. Comparing this week's *Daily Telegraph* with last week's *Guardian* yields unreliable evidence. What is new for one paper may have been another paper's common usage for a long time. A corpus of nothing but newspapers and periodicals seems to be somewhat unsatisfactory. We would like to include other genres, as well. But what do we achieve, in terms of documenting lexical change, by randomly selecting, say, ten fiction and ten non-fiction books per annum? Novels as well as popular science or history books tend to have a specific topic. A single book on tennis would change the frequency of tennis terminology for this monitor corpus year to such an extent that it would bias the results. Such a slant could be set off only if we added not ten but hundreds of books per year. A compromise would be to include a book review journal like the *Times Literary Supplement.* A book review will normally contain the new vocabulary that comes with the book, but not to such an extent that it will bias frequency counts. Unfortunately, large-scale monitor corpora reflecting what is seen as standard written language are still not available. However, this situation will change over the next few years.

A parallel corpus, sometimes also called a translation corpus, is a corpus of original texts in one language and their translations into another (or several other languages). Reciprocal parallel corpora are corpora containing original texts and translated texts in all languages involved. Sometimes parallel corpora contain only translations of the same texts in different languages, but not the text in the original language. Such a corpus can tell us how the English we find in translations differs from authentic English. Sometimes it is not known – or it is thought to be irrelevant – which text is the original and which text is the translation. For example, we are not told in which language legal documents issued by the European Commission are drawn up. It used to be mostly French, but more recently the final version has often been in English, and it can well happen that previously working versions were drafted in other languages. The same is true for texts issued by the Vatican. These days a long time has passed since Latin was the original version. The languages are mostly Italian and French but also

English, Spanish and German, and the Latin version is added at a later stage. These parallel corpora cannot tell us how French texts are translated into English, but they can show in which cases the word *travail* (or its plural *travaux*) is equivalent to *work*, and in which cases to *labour* or other expressions.

Parallel corpora are repositories of the practice of translators. The community of translators from language A to language B and vice versa know a lot more about translation equivalence than can be found in any (or all) of the bilingual dictionaries for these languages. Even the largest bilingual dictionary will present only a tiny segment of the translation equivalents we find in a not too small parallel corpus. Because the ordering principle of printed dictionaries is alphabetical, based on mostly single-word entries, bilingual dictionaries do not record larger and more complex units of meaning in a methodical way. Neither do they tell us which of the equivalents they offer belong in which contexts. This is one of the reasons why bilingual dictionaries do not help us to translate into a language we are not very familiar with. The user is left with many options and hardly any instructions for selecting the proper equivalent. From parallel corpora we can extract a larger variety of translation equivalents embedded in their contexts, which make them unambiguous. This is what makes parallel corpora so attractive. Working with parallel corpora lets us do away with ambiguity, with being given alternatives between which we have to choose. We will identify monosemous units of meaning in one language and find the equivalents in the other language(s). Now, when we have to translate a given word, we will compare this word with the words we find in its company, with the words in the units of meaning we have extracted from the parallel corpus. The closest match usually renders the correct translation equivalent.

For most applications, parallel corpora will have to be aligned so that a unit in one language corresponds to the equivalent unit in another language. The standard unit of alignment is still the sentence. In the beginning, parallel corpora were sentence aligned by hand. For alignment is not a trivial task. First, it is not always easy to identify sentence endings automatically. Full stops can also designate abbreviations, some of which can occur either within a sentence or at the end, such as *etc.* In some languages, full stops can also indicate ordinal numbers (2. = 2nd). Second, sentences are by no means stable units. One sentence in the source language can correspond to two or more sentences in the target language; or two source language sentences can be subsumed in one target language sentence. Anyone who has closely compared texts and their translations will have noticed that sometimes

sentences are plainly omitted in the translation, or that new sentences are introduced, it seems almost at the translator's whim. This is why, even though there are various tools available to align corpora on the sentence level, alignment is a time-consuming process involving substantial human intervention. This is one of the reasons why there are still only a few parallel corpora of considerable size (say, more than 5 million words per language).

It is even trickier to align corpora on the lexical level. Ideally, one would like to see each unit of meaning in the source corpus linked to the equivalent unit in the target corpus. (Source and target, in this context, do not refer to the language of the original and the translated text; rather, the source language is the language which we choose as our point of departure, while the target language is the one in which we want to find equivalents.) Our results will, to a certain extent, depend on this directionality. It is well known that bilingual dictionaries are not reversible. Whether the results extracted from a parallel corpus are reversible, and to what extent, is still unknown. Lexical alignment uses statistical procedures and/or lexicon look-up. Neither is very reliable. It is because bilingual dictionaries (and the lexicons derived from them) are not very instructive that we turned to parallel corpora in the first place. Hence the lexical alignment we start with is only tentative; and all it tells us is what could be an equivalent of the source unit. We will still be given both *work* and *labour* and also some other words like *employment* or *job* as equivalent of French *travail*. If we want to find out more, we have to look at the contexts in which *travail* is embedded when it is translated as *work*, as opposed to the contexts in which *travail* is embedded when it is translated as *labour*. Thus, if we have to translate *travaux* followed by the adjective *préparatoires*, our parallel corpus tells us that this phrase is never translated as *preparatory labours* but always as *preparatory work* (with a singular phrase in English corresponding to the French plural phrase).

Recently it has become quite common among corpus linguists to consult the Internet as a virtual corpus. This is particularly useful when we want to find out if a word or a phrase we have heard really exists and in which kinds of texts it occurs. Whenever we cannot find evidence of words or units of meaning in our classic corpora, we can turn to the Internet. There are many commercial browsers we can use, like Altavista or Google, and they all have their advantages and disadvantages. The Internet is larger than any existing library, and if a word is in current use, we are bound to find it there. What we do not know, however, is how the Internet is composed in terms of the parameters mentioned earlier. Frequencies of occurrence have to be carefully

interpreted. The Internet can be seen neither as a sample of a middle-class person's private reading load nor as a sample of text production *in toto*. So far, there are hardly any transcripts of spoken language on the Internet, and the written language we find is a reflection of what kind of texts different people put on the Internet. Some texts exist only there; others are copied from other written material. And here too we must be careful; not all copies are perfect clones of the original texts.

For practical purposes, the Internet, even if we restrict ourselves to the freely accessible websites, is, at any given time, if not infinite then certainly inexhaustible. No browser can claim to cover more than a selection. Such a selection is usually so big that by the time we have extracted all citations for a given keyword (or larger unit of meaning), some texts queried will already have been taken from the servers they were on, while others, new ones, will have been added. The Internet is a virtual corpus, and, like the discourse of any language community, we cannot expect to access it as a whole. Normally, if someone wants to use the Internet as a source, they should, therefore, download all the texts they are working with, and compile them in a special corpus; and they should document them with their web addresses and other bibliographic information and the date of the download.

4.3 Meaning in discourse

> 'When I use a word,' Humpty Dumpty said in a rather scornful tone, 'it means just what I choose it to mean – neither more nor less.'
>
> 'The question is,' said Alice, 'whether you can make words mean so many different things'.
>
> 'The question is,' said Humpty Dumpty, 'which is to be the master – that's all'.
>
> (Lewis Carroll, *Alice Through The Looking Glass*)

It is not Humpty Dumpty as an individual but the discourse community as a whole (or at least sufficiently significant fractions of it) that decide what a word means. Individual members of this community who want to say something but are dissatisfied with the words they find and how they are being used, have two options. They can either introduce a new word (e.g. *Eurotrash*, the name of a popular series on British television) into the discourse – which happens relatively rarely – or they can try to change the meaning of an existing word, by using it in a new context. Shifts in the meaning often start off in slang: today's slang use of *wicked* means 'good', while the 'proper' meaning is quite the opposite. The meaning of the word *tart* as 'woman of loose morals' has become so

dominant that bakeries and cafes sometimes see themselves pressed to use new (and no doubt more elegant) names like *gateau* and *torte*. Another example is the *cold caller* (most neologisms appear to be collocations of some sort or other), who is not a caller in the cold, but someone who is paid for calling people they do not know to try to sell things to them. This new usage is now beginning to be registered in the dictionaries.

To show how meaning is constituted in discourse we will present one word in detail. We want to show that there is no magic formula, inside or outside of the discourse, no concept or no feature of 'reality' that we can identify as the thing the word stands for. Words are symbols. But they do not stand for something unequivocally assigned to them by some infallible deity for a shorter or longer eternity. A word in a text refers to (or is a trace of) previous occurrences of the same word, in the same text, or in all previous texts to which the present text, sometimes explicitly, but mostly implicitly, refers. It refers to all that has been said about that word previously. As we know, people do not always agree with each other. For corpus linguists, this is good news. For then we find a discussion going on in the discourse community, with various factions making different claims about what they consider to be reality. If this controversial feature can be subsumed under one concept, then each faction will try to define this concept as it suits their views. They will volunteer to present their views in the form of paraphrases of the linguistic expression in question to the linguists on silver trays.

The word we have chosen for the investigation here is *globalisation* (or in its alternative spelling, *globalization*). This word is a derivation from the adjective *global*, which has been part of the English language for many centuries without changing its meaning in a noticeable way. From this adjective, it was always possible to form the word *globalisation*, and, though not very frequently, it happened now and then. Not all dictionary-makers have registered that somewhere in the 1990s this word suddenly embarked on an unprecedented career. Thus, the relatively recent *NODE* has only a very short entry for *globalize*:

> **globalize** (also **-ise**) verb
> develop or be developed so as to make possible international influence or operation.

In the same entry, *globalization* is mentioned only as a derived noun, meaning nothing more than 'the process or activity of making something global'. Globalisation was not always as popular as it is today. In 1983, the economic scientist Ted Levitt entitled one of his articles for the *Harvard Business Review* 'The globalization of markets'. This journal

is obligatory reading for all leading experts, and Ted Levitt is a well-known figure. There are only nine sentences with the actual word *globalization* in his ten-page article. The article does not define *globalization* as a term, nor does it introduce it as a new word, rather it is used assuming everyone understands it. It seems to have been this very article that determined the current usage of *globalisation* that has made the term a household word in disciplines such as economics and social and political studies. As terms need to be defined, specialised terminological dictionaries do so, and give us a host of definitions. However, in general language lexicography, *globalisation* is not recognised as a new word by a number of reference dictionaries. We have already mentioned *NODE* but we do not find *globalisation* in *The Oxford Dictionary of New Words* (1997) either. The *Macmillan English Dictionary for Advanced Learners* (2002), however, tells us that *globalization* is 'the idea that the world is developing a single economy and culture as a result of improved technology and communications and the influence of very large MULTINATIONAL companies'.

What does the corpus tell us about the meaning of *globalisation/globalization*? First of all, it tells us that most citations in British sources of the Bank of English use the form with *s*, while almost all citations in the US part of the sources use the form with *z* (although *z* is also the preferred form for some British publishers). It is useful to look at them separately, because it just may happen that British *globalisation* has a meaning that differs from American *globalization*.

But before we look at this example in more detail, we will explore the notion of meaning as usage and paraphrase.

4.4 Meaning as usage and paraphrase

How does corpus linguistics deal with meaning? Meaning, as has been said before, is in the discourse. But how do we look for it there? How do we find it? There are two main aspects to meaning. Meaning is usage and paraphrase. Usage and paraphrase reflect the two ways we deal with language. We can participate in the discourse as speakers and as hearers.

Knowing the usage of a word or other lexical item lets us participate successfully in discourse. To make ourselves understood as speakers we must use linguistic items according to the expectations of hearers. These expectations are based on what has been said before. Since everybody has heard time and again the phrase 'the increasing globalisation of the financial markets', it will offend no one if we use *globalisation* with *increasing* as an adjectival modifier, and with *of the*

financial markets as a genitive or prepositional modifier, depending on what you prefer to call it. There are other adjectives that often modify *globalisation*, and there are other nouns we frequently find in the genitive phrase modifiers. We also find that *globalisation* can be part of a genitive phrase modifying another noun, such as *the effects of globalisation*. This kind of contextual data determines the usage of a word. For nouns such as *globalisation*, we should also know of which verb phrases it can be the subject, and which verbs it can complement. Not all the information we find is relevant for establishing the usage of a lexical item. But the words, together with their phrasal positions, which occur with a satisfactory frequency and with a defined statistical frequency in the context of the lexical item in question (i.e. *globalisation*), make up its usage. As long as we, as discourse participants, stick to the established usage, we cannot go wrong. But once we say 'the green globalisation of our forgotten noses' we will have lost our audience. Nobody will listen to us, and we will be told we are talking nonsense. There is nothing remotely similar to this phrase in established usage that could serve as an analogy for interpreting it. Usage is something that can be established by a computer. That means in order to deal with the usage of a lexical item we do not necessarily have to understand it, in the sense that we would be able to paraphrase it. Computers can create texts fully complying with the established usage of all the lexical items that form the text. They might be indistinguishable from texts of certain politicians, but this does not imply that the computers knew what they were saying. In this sense, usage is meaning only in a very twisted way.

Usage, however, is what we have to learn as discourse participants. It is what comes naturally to native speakers, it seems. Those who are striving to acquire English as a second language have to learn consciously that bereavement in the context of guilt is expressed by *grief*, while bereavement in the context of sadness is always felt as *sorrow*, and not the other way around. Usage is therefore something we have to cope with as members of the discourse community, as little as it may help us with the understanding of a lexical item.

Determining the usage of lexical items and coping with it are essential to the methodology of corpus linguistics. It is as close as computational methods can hope to approximate the mystery of meaning. Whenever we want the computer to identify a word as part of a unit of meaning in a corpus, it will be done through established usage. The profile of usage generated by the computer (which itself is incapable of understanding, incapable of what is sometimes called 'interpreting symbols') is like the fingerprint of a word as part of a unit

of meaning: it identifies the person without telling you who the person is. The usage profile is the device in computation that can resolve ambiguity.

Meaning, we have said, is usage and paraphrase. The computer can give us the usage profile of a unit of meaning without knowing what it means. But the computer does not know and cannot know what it means. Indeed it seems as if humans are the only species of the kingdom of machines or animals who have a tendency to think about what something is about. When they see, in springtime, bees flying from blossom to blossom, they believe they know that this is about making honey, and has the additional fortuitous effect that the blossoms get fertilised. They see these two aspects as the meaning of why bees fly. The bee does not know why it is flying (and probably does not care about it). And while a computer, programmed for this task, may have no problem in translating the sentence 'Bees fly from flower to flower to produce honey' into the French sentence '*Les abeilles volent de la fleur à la fleur pour produire le miel*' (translation produced by the Altavista browser) we would not believe for a minute that this machine or the program it runs on has any idea about bees, flowers or honey. Only humans can appreciate aboutness. Only they can deal with signs. For something like flying bees is 'about' something else only when we take this something (flying bees) as a symbol for something else (producing honey). But is not the flower then a symbol for the bee signifying that it will find sugar there? I don't think it is. I don't think the bee will reason along the lines of 'oh, over there I can see (or smell) a flower – I take this to mean that I'll find sugar there'. An appreciation of aboutness presupposes consciousness, awareness. Only humans can be conscious in this sense. This unique human mental ability to find out consciously (rather than randomly) what something is about is called intentionality. In the philosophy of mind there has been a long debate about whether intentionality really is a human trait or perhaps just an illusion. If it were nothing but an illusion, then we could say the human mind is, in principle, the same as a computer, only more complex. But if intentionality exists, computers will never become like humans, and we will always be able to pull the plug when we feel like it. The issue of intentionality is very skilfully discussed in John Searle's book *Intentionality* (Searle 1983).

Only something that is a sign can mean something, because only a sign can signify something other than itself. We can say that our life has a meaning only if we take it to be a sign, a symbol for something else. Things, events, processes which we do not interpret as signs do not mean anything. However, there are different kinds of signs: symptoms,

icons and symbols in the narrow sense. What we have mentioned above, the flying bees as signs that they are on their way to produce honey, is a symptom. It is something we can figure out, not because flying bees bear some resemblance to a honey jar, but because the discourse is full of stories about bees flying around to produce honey. What we did, when we saw those bees, was to remember those stories and to use our common sense to infer that they were up to honey-making. The second kind of sign are icons, signs that somehow give you a visual (or oral or tactile) clue of what they, as signs, stand for. A big picture at the roadside representing a honey pot will probably mean that there is someone there who wants to sell their honey. Icons are signs that we interpret in terms of their resemblance to whatever is indicative of some thing, act or event. Interpreting icons is, again, largely an application of common sense, together with memory of other instances to which this instance may be analogous. Finally, symbols in the narrow sense are signs to which a meaning has been assigned arbitrarily. The vocabulary of a language is commonly seen as a set of such symbols. Some people may speak as if it is the dictionary or lexicographers who assign meanings to words. But the lexicographers only document the meanings that are already assigned. It is the discourse community that assigns meanings to words (or, rather, to lexical items). Members of this community may not always be happy with the meanings they find assigned, and they are free to change the assignments, for the sign, the lexical item, does not resemble what it stands for. If, in England, the word *robin* stands for one kind of bird, that does not prevent Americans from deciding that in their variety of English the word stands for quite a different bird. It is not particularly surprising that words such as *beech*, *bream*, *grasshopper* and *magpie* have different meanings in different parts of the English-speaking world. With icons, of course, it cannot be so simple. A placard representing a jar of honey can hardly indicate that you can buy green asparagus. Flying bees will never signify that coal is being mined. A good guide to signs is Rudi Keller's book *A Theory of Linguistic Signs* (1998).

Meaning is an aspect of signs, of symptoms, icons and symbols. Meaning is one of the aspects, form being the other. Meaning and form are inseparable. Once you take away the form, the meaning vanishes. This is why it is wrong to look at language as a system into which you can encode a message and from which you can decode a message. There is no message without form. Thus it is wrong to say the text contains a meaning; the text is the meaning.

There are many theories about meaning, and almost all claim that the meaning of a linguistic unit is something outside of the discourse.

Some say the meaning is what the linguistic unit refers to, out there in some discourse-external reality; others say that the meaning corresponds to some representation we have in our minds; some say meaning is represented by semantically interpreted logical calculus or some other formal system. Reality will hardly do, as we have pointed out above. It is, before it is ordered and structured by language, amorphous and chaotic. Mental representations only multiply our problems. For either these representations are also signs, in which case their meaning is inseparable from their form, and to get at their meaning, we would have to come up with yet another representation, and still another one on top of that, and so on. (John Searle, in his book *The Rediscovery of the Mind*, 1992, refers to this phenomenon as the 'homunculus' problem, while for Daniel Dennett, in his book *Consciousness Explained*, 1993, it is the problem of the 'central meaner'. Both scholars agree that translating the meaning of a linguistic unit into a mental representation is nothing but a fallacy.) Or, if mental representations are translations of meaning into an expression of some formal system, we are still no better off. For how would we know what the expressions of such a formal system mean? We could only explain them in natural language, and then we are back at square one. Of course we can translate any sentence or text into an artificial language such as Esperanto. But to understand that Esperanto sentence or text we would have to re-translate it into a language we are familiar with. We just cannot escape the prison of natural language. All these attempts to approach meaning are like burning wood in a stove: if we succeed, we are left with nothing. Once the form is burnt, the meaning has vanished.

Our point of departure was that a sign is something that stands for something else. But if, as we have been arguing, it does not stand for something in some reality not affected by the discourse, if it does not stand for a representation in the mind, if it does not stand for some expression in a formal linguistic system, then what does it stand for? When you are asked what a cantaloupe is, or a unicorn, what comes to your mind? You may remember market stalls where you have seen heaps of cantaloupes, you may remember eating cantaloupe, and the taste and texture of the fruit perhaps come to your mind. You may remember stories you were told about unicorns, or you may have read about them. You may also remember what you have told other people about them. You may remember a picture in a children's book, or a Burgundian wall hanging, or a little illustration in a medieval manuscript, depicting a unicorn. Isn't that what meaning is, what the linguistic signs *cantaloupe* and *unicorn* stand for?

If we were to take these memories to be the meaning of lexical units, then we would adopt the position of cognitive linguistics. Our mental representations would not be as orderly as Anna Wierzbicka (1996) would like them to be, and certainly these representations would not be anything like universal, because everyone's memories are unique, and they would contain ineffable qualia like your or my taste experience of cantaloupes. Your memories form your understanding of cantaloupes and unicorns, and they would be reflected in your response to the question what these items are. But these memories are private and individual, and therefore they cannot be the meaning of the signs *cantaloupe* and *unicorn*.

Language is a social, not a psychological phenomenon, and so is meaning. The meaning of *cantaloupe* and *unicorn* is what is said about them in the discourse. Your response to the question what these items are will be a new contribution to the discourse, and it may well contain statements that have not been said before. If your audience is happy with them, they may remember them, and they may even repeat them in suitable situations. In ten or twenty years, you might even find traces of it in new editions of dictionaries. On the other hand, your audience will also compare what you say with what they have heard before. If what you say disagrees with their memories, they will ask other members of the discourse community. Unless they quote from dictionaries, it is highly improbable that those who volunteer their view of cantaloupes and unicorns will repeat anyone else's statements *verbatim*, word for word. If your audience is still unsatisfied with what they are being told, they might resort to querying libraries, archives and even the Internet. In the end, they will come up with a host of reports on cantaloupes and on unicorns. Some of these reports look more or less like definitions; some like more technical explanations; many are just stories. A lot will overlap each other, while some reports will not be supported by any others. There could be a lot of disagreement. Should we trust Caesar when he claims there were real unicorns in the Teutonic forests, for example? Many of the statements will in fact not be directly about cantaloupes and unicorns but about what other people have said about them, about how reliable or credible their claims are.

We call all these statements, definitions, explanations, and stories that focus on cantaloupes or on unicorns, paraphrases. What all the paraphrases of the word *unicorn* will definitely have in common is that they paraphrase the word *unicorn*. That alone is what keeps them together. Otherwise they may be as different as they come. This set of paraphrases then is the meaning of the lexical item *unicorn*. It cannot be reduced to a simple formula. It is fuzzy, vague, full of contradictions;

some of it may be true and some of it may be wrong. It is not the linguist's task to filter out what they think is right. This is what the linguistic sign *unicorn* stands for: the set of paraphrases dealing with unicorns. This is what the word *unicorn* is about.

Meaning as paraphrase thus shows us another way of identifying units of meaning. In this perspective, a unit of meaning is whatever we find paraphrases for in the discourse. Usage profiles can be handled efficiently by computers. Paraphrases, on the other hand, have to be interpreted. They have to be understood. This is something computers cannot do. Therefore they will never know what cantaloupes and unicorns are.

After these rather lengthy remarks about usage and paraphrase, we can finally return to the case study of the meaning of the word *globalisation*. In the following section, we show how a computer can create a usage profile for us.

4.5 Globalisation

To start, we have analysed a sample of 200 citations (from the Bank of English) of *globalisation*, and asked the computer to give us the most frequent collocates of the word. Among the most frequent collocates are the following words: *anti, world, against, means, economic, international* and *business*, as illustrated by the following concordance lines:

```
   Forum, the main anti- globalisation umbrella group at the
 defended against anti- globalisation protesters by one of
than we used to. Despite globalisation, the world has at the
  debate, related to the globalisation of the world economy,
  of the protest against globalisation is, however, mistakenly
    has held our against globalisation of culture for a long time
    is the argument that globalisation means economic problems
  artistic strategies. Globalisation here means the same old
            and economic globalisation was kidding themselves
economic policy is the globalisation of markets. If economic
  change in the wake of globalisation, international competi-
    related has been the globalisation of international manu-
       in business law globalisation will cause even worse
illogical attacks on the globalisation of business
```

As we have said before, it is useful in this case to look separately at the spelling variant with *z* in order to establish whether *globalization* is used

differently. Already the total number of actual occurrences in the Bank of English is quite different: *globalization* occurs 468 times, while the total number of the citations for *globalisation* is 1,447. This can be easily explained by the fact that the data in the Bank of English are 'biased' towards British English in which the spelling with *s* is more common. But that is not the whole truth. If we look at the sources of the citations, we can see that citations for *globalisation* come mainly from newspaper texts and citations for *globalization* are mainly from books, many of them American (but also some British. The most frequent collocates for *globalization* (based on a sample of 200) are only partly the same as above. We do find such collocates as: *economic, markets, world, investment, financial, international* but we do not find *anti* or *against.*

To find some of the paraphrases of *globalisation* we looked up a sample of what the corpus tells us about 'what globalisation is':

degradation. But globalisation is a fact and, by
rapidly changing world. Globalisation is a much overused word
on the world stage.'' Globalisation is a trend that many
technological change. Globalisation is a catch-all to
and access to capital. Globalisation is a redistribution
conventional wisdom on globalisation is a relic of the
war and the economics of globalisation is a story which gets
the poor even poorer. Globalisation is a fancy euphemism
problem in particular. Globalisation is a market-led process,
in inverted commas. Globalisation is a term that Giddens
at Conduit, explains: 'Globalisation is a trend that everyone
major issues. Still, globalisation is a process that has
you realise that globalisation is an accepted phenomenon
and benefiting from, globalisation is an open society, in
may be in danger. Globalisation is an opportunity and a
apreciation is that globalisation is an unstoppable force,
lot of the criticism of globalisation is based on ignorance
Mr Rubin argues globalisation is both good and inevi-
Precisely because globalisation'' is demonised as an
Etzioni argue that globalisation is destroying communities
euro' s arrival is that globalisation is here to stay as
the allegation that globalisation is inherently harmful.
those who say that globalisation is just a bigger market,
- with mixed results. Globalisation is like a giant wave,
The reality is that globalisation is not inevitable, it is
tourists. All the same, globalisation is not to be resisted,
best way to deal with globalisation is not to fight it but
however, that globalisation is not a painless exercise
trend towards increasing globalisation is not easily
past few months show, globalisation is not a one-way street.
must demonstrate that globalisation is not just a code word
a sounder basis for globalisation is required. If neither

voiced an anxiety that globalisation is robbing nations
left in the age of globalisation is tearing apart even
the global economy. Globalisation is the big issue of our
humbly born. To be anti- globalisation is to march, under the

As we can see globalisation 'is' and 'is not' lots of things. The discourse takes notice of the fact that, if there is a relationship between the word *globalisation* and some discourse-external reality, it is not an easy one. We are told that *globalisation* is a much overused word, a story, a fancy euphemism, a term, and not just a code word or a one-way street. This shows how we can use natural language to talk about language. It is something we cannot do in formal languages like mathematics, logical calculi, or programming languages. There you have to move outside the system to be able to talk about it. But where should we move from our language? Indeed every other language, every formal language, can be defined only by natural language. There is no formal algorithm, no calculus that would not need this kind of definition or explanation in a natural language. And therefore we have to use our own language to discuss it. If we look at the context left to *globalisation*, we see that people discuss the way in which other people use *globalisation*. Someone explains, you realise, there is an appreciation and also a criticism, people argue, demonstrate or say that *globalisation* is this or that.

But what is the thing behind the word? Globalisation is: a fact, a trend, a process, a phenomenon, an opportunity, an unstoppable force; it is both good and inevitable; is here to stay; is inherently harmful; is like a giant wave; is not inevitable; is not to be resisted; is not a painless exercise; and is the big issue. This looks confusing. Is it not inevitable, or is it here to stay? Is it harmful, or is it good? For whom is it an opportunity, for whom is it a painful exercise? Can we understand the word on the basis of these citations? Can we define it?

Our corpus is the Bank of English, with about 450 million words. The texts it contains were spoken or written mainly in the last twenty years and it is strongly biased towards British English. When we go through the concordance lines of *globalisation* we notice there is a lot of debating, worrying and talking about globalisation; there are many protests, demonstrations and campaigns against globalisation, and we find many instances of anti-globalisation and anti-globalisation protesters; globalisation is economic and increasing, there is an ongoing globalisation process; people talk about an age or era of globalisation, about the benefits, the challenge, the impact, the pressures, the forces and effects of globalisation. This all sounds familiar. Globalisation is often also connected with emotions, and these are predominantly

negative ones. About one-third of the citations are of neutral tone and only about one-tenth of them can be considered to have a positive tone. This occurs mainly in the context of business, politics and new technology.

We get a slightly different picture if we look at *globalization*. Not all of these instances are American English, for the *OED* prefers the spelling with *z*. Here is a sample of the citations:

the impact of economic globalization on the world of work
overwhelmed by the rapid globalization of economic
trends, including the globalization of markets.
us a picture of how this globalization of financial markets
elsewhere in the world, globalization holds the promise of
This is the world of new globalization of borders easily
itself a result of the globalization of investment. While
illustrates cost of the globalization of investment. It also
technology and increasing globalization challenging the way
Third, there has been a globalization of the international
In this new form of globalization, the international

The most apparent differences from the previous citations are the matter-of-fact tone and the different genre from which these citations come.

Let's compare our corpus evidence with Ted Levitt's use of the word *globalization* in 1983. The following are the citations from his article:

> The **globalization** of markets is at hand.
>
> Nor is the sweeping gale of **globalization** confined to these raw material or high-tech products, where the universal language of customers and users facilitates standardization.
>
> The theory holds, at this stage in the evolution of **globalization**, no matter what conventional market research and even common sense may suggest about different national and regional tastes, preferences, needs, and institutions.
>
> Barriers to **globalization** are not confined to the Middle East.
>
> It orchestrates the twin vectors of technology and **globalization** for the world's benefit.
>
> The differences that persist throughout the world despite its **globalization** affirm an ancient dictum of economics – that things are driven by what happens at the margin, not at the core.
>
> To refer to the persistence of economic nationalism (protective and subsidized trade practices, special tax aids, or restrictions for home producers) as a barrier to the **globalization** of markets is to make a valid point.

> Two vectors shape the world – technology and **globalization**.
>
> Given what is everywhere the purpose of commerce, the global company will shape the vectors of technology and **globalization** into its great strategic fecundity.

Levitt's first use of the word *globalization* in his article (apart from the title) is in the following sentence: *The globalization of markets is at hand.* In the Bank of English there are eleven occurrences of *globalisation of markets* and five of *globalization of markets* as illustrated by the following lines:

```
of stocks. Also, as globalisation of markets continues to
a mystery where the globalisation of markets is taking us. Its
  adjustment to the globalisation of markets and the influence
    into account the globalisation of markets and of the features
      because of the globalisation of markets, which will have
        rapidly with globalisation of markets. Kevin Bales reports
```

Levitt talks about 'the globalization of markets' being 'at hand'. All the evidence from the Bank of English shows that 'globalisation of markets' is now an established notion. It tells us what is happening with the 'globalisation of markets' presently: it 'grows' and we have to take it 'into account'.

Levitt mentions *globalization* for the second time when he talks about how 'globalization' influences world business; the expression he uses is the 'gale of globalization'. In the Bank of English we find no occurrences of 'gale of globalisation'. If we look up what comes as a 'gale of', we find it is mostly 'giggles' and 'laughter' and 'wind', but we do also find *gale of economic change, energy* or *modernity,* notions closely related to globalisation.

The third occurrence of *globalization* in Levitt's text is in the phrase 'the evolution of globalization', which has no occurrences in our corpus. This may correspond to our earlier finding that 'globalisation' in our contemporary language use is an established notion. Levitt mentions 'barriers to globalization' twice in his text. There are no citations for exactly the same wording in our corpus, but if we look up 'barriers to' we can see that most frequently we talk about 'barriers to entry' and 'barriers to trade', again facts closely related with globalisation. Levitt strongly associates 'technology and globalization' (three occurrences in his text); and indeed, judging from the evidence from Bank of English, globalisation is still very much associated with technology both directly and indirectly.

pressure from technology and globalisation 'the half-life
benefit from new technology and globalisation than others,
arthritic. When technology and globalisation demanded changes
INFORMATION technology and globalisation are the driving
arguing that technology and globalisation are tending to
of clothes. Technology and globalisation mean that every-
the impact of technology and globalisation, where business
needs. Technology and globalisation have revolu-
Why? 'Because of technology and globalization, everyone has

The last occurrence of 'globalization' in Levitt's text is the following:

> The differences that persist throughout the world despite its globalization affirm an ancient dictum of economics – that things are driven by what happens at the margin, not at the core.

A search for 'globalisation' and 'despite' yielded the following lines:

insular **despite** the globalisation that affects
But, **despite** all of this, globalisation has been the source
despite the drift towards globalisation, national policies
despite the increasing globalisation of the capital
despite pressures of globalisation, it will take a

What we present here is based on insufficient evidence. We do not know if there are differences between American and British usage, as we do not have a comparable corpus of American English. Our citations are not classified according to domain, genre, text type or publication date. We cannot see if there was a change of meaning, and we do not know whether *globalisation* is used differently in texts written for the public at large from texts addressed at a professional audience. The structure of the Bank of English makes it hard to extract the information needed for this kind of classification. But the evidence we have for example clearly suggests there is a difference in tone, depending on the genre: texts from newspapers often have a sceptical tone, while the tone of professional and academic writing is more 'matter of fact'. This is not surprising; in fact it could have been expected and it is most probably also true of many other words.

This, then, is how and what corpus linguistics can contribute to the meaning of the word *globalisation*. It is the evidence of the corpus citations of *globalisation* within their contexts, condensed and brought into some kind of contingent order. It is much more than we would find in any dictionary, and, at the same time, it does not have the coherence of an encyclopaedic article. It is obviously in many points contradictory; it is nothing like a definition. Is this the meaning of *globalisation*? Globalisation has become such an important fact of our society that social scientists have felt the need to define it. 'Globalisation' has thus established itself as a term. Let's have a look at its complex definition in the *Oxford Dictionary of Sociology* (1998). Is it so very different from some of our citations?

> **globalization, globalization theory**
> Globalization theory examines the emergence of a global cultural system. It suggests that global culture is brought about by a variety of social and cultural developments: the existence of a world-satellite information system; the emergence of global patterns of consumption and consumerism; the cultivation of cosmopolitan life-styles; the emergence of global sport such as the Olympic Games ... the spread of world tourism; the decline of the sovereignty of the nation-state ... More importantly, globalism involves a new consciousness of the world as a single place. ... Perhaps the most concise definition suggests that globalization is 'a social process in which the constraints of geography on social and cultural arrangements recede and in which people are becoming increasingly aware that they are receding' (Malcolm Waters, Globalization, 1995) ... Contemporary globalization theory argues that globalization comprises two entirely contradictory processes of homogenization and differentiation; and that there are powerful movements of resistance against globalization processes ... It is undoubtedly true that, on a planet in which the same fashion accessories (such as designer training-shoes) are manufactured and sold across every continent, one can send and receive electronic mail from the middle of a forest in Brazil, eat McDonald's hamburgers in Moscow as well as Manchester, and pay for all this using a Mastercard linked to a bank account in Madras, then the world does indeed appear to be increasingly 'globalized'. However, the excessive use of this term as a sociological buzzword has largely emptied it of analytical and explanatory value ...

As globalisation is a worldwide phenomenon, we need not restrict ourselves to the English-language evidence. There are also some interesting data available about *Globalisierung*, its German equivalent, based on the analysis of one single newspaper, the *Tageszeitung*. It was only in the year 1996 that *Globalisierung* gained ground. From 1988 until the end of 1995 we find altogether about 160 citations. Then, suddenly, for the year 1996 the figure jumps up to 320, and from then

on remains more or less on the same level. Before 1996, *Globalisierung* was used as an *ad hoc* formation derived from *global*, without any specific meaning other than 'the action or process of something turning global'. In each instance, it was necessary to specify what was being globalised, and therefore *Globalisierung* never came alone, but was always in the company of modifiers (such as in *die Globalisierung der modernen Lebensweise*, the globalisation of the modern way of life). For the year 1996, when the word suddenly made a jump in frequency (indicating, among other things, that a change of meaning, from something more general to something more specific, had occurred), we find a large number of citations in which *Globalisierung* comes without modifiers, but with explanations or paraphrases (*in der Tat bedeutet Globalisierung Amerikanisierung*, indeed, globalisation means Americanisation). By the time that everyone is supposed to understand what *Globalisierung* means, this percentage goes down again, and we find again many modifiers. Before 1996, the modifiers appeared to be a random lot – everything could be connected to *Globalisierung*. After 1996, the same modifiers recurred time and again; the new meaning of *Globalisierung* associated the word mostly with finance, trade, technology, the economy at large, and the workforce, indeed remarkably similar to the modifiers we find with *globalisation* in the Bank of English.

We deliberately chose the word *globalisation* as a neologism. Must we not assume that all words once were neologisms? Before Lucullus brought the cherry (*cerasus*) from Persia to Rome, there was no need for a name for it. It makes sense to assume that the introduction of neologisms into the discourse always occurs along the same lines. As long as the word (or larger unit of meaning) is still new, it needs to be explained. Not everyone understands the word in the same way. Explanations also serve the purpose of negotiating the meaning of a word within the discourse community. It can take a while before the larger part of the audience has come to some kind of tacit agreement. From then on, only those discourse participants who object to the agreement will come up with new paraphrases. In principle, we can say that once the meaning of a new word has become uncontroversial it can be used to paraphrase other units of meaning.

The word *globalisation* shows nicely that the discourse community is not at all homogenous. The discourse community is, it should be remembered, identical with the society whose language we are dealing with, and it is as multifaceted as this society is. So while for one section of the discourse community the meaning of *globalisation* has become established and accepted, there are other sections still in disagree-

ment. This is why the paraphrases and explanations we find in the corpus do not have a common denominator and why, at times, they even contradict each other. Contradictory evidence, however, is not what we have come to regard, in traditional lexicography, as the meaning of a word. Now, once we have presented the citations, we must sift the evidence and write up a coherent, concise definition that we can put into the dictionary. Now we must find the magic formula.

There is no secret formula; and there is no overt formula, unless we leave it to a committee of experts (on what?) to define *globalisation* once and for all, so that anyone who thereafter uses *globalisation* differently will be reprimanded. But if this were to happen, *globalisation* would have ceased to be a natural language word; it would have become a term in a formal system of terminology. Corpus linguistics has nothing to contribute to standardised terminology. No contribution to the discourse will ever change the meaning of a standardised term, because it has no other meaning than the definition assigned to it. *Ascorbic acid, DNA* or *electrolytic rectifier* are terms with relatively fixed meaning, hardly variable in any imaginable context.

Why do new words occur? Why do other words disappear? Will *globaloney* be as widely used in twenty years' time as *globalization* is today? In December 2002 *Newsweek* published an article 'The new buzzword: globaloney'. The article begins: 'So far as we can tell, congresswoman Clare Boothe Luce coined the term "globaloney" in 1943 to trash what Vice President Henry Wallace liked to call his "global thinking" ...' (Miller 2002).

There are only five citations for *globaloney* in the Bank of English and they all are from British sources.

in impressive-sounding globaloney are provided by
and more earlier. For all the globaloney to be found in modern
LSE lecture last week – is globaloney. Much of the talk about
nosed divi, talkin a load o' globaloney (Fings ain' t what they
Business, 1993. Hirst, P. Globaloney in Prospect,

It is only recently that *globaloney*, as in the example below from the *Mail on Sunday* (22 July 2001), has started appearing more frequently in the press as a reaction to the phenomenon of globalisation. We have to wait to see whether it catches on.

> These 'anti-globalist' demonstrators are globalists themselves. Whether they are hippy-dippy groups concerned with peace, the environment and Third World debt, or old-fashioned leftie groups with a tradition of small-scale violence, they all want to replace the present globalisation with their own form of globaloney.

Words are different from terms. A word we find in the text resonates with an infinite number of previous citations, many of them shared between the speaker and the hearer, because they have grown up in the same discourse community and have been exposed to many of the same texts. It is in the light of these previous citations that the speaker uses the word and that the hearer will understand it. It is linguists who define the discourse community. They decide how deeply they want to dig into the past, and they define where the lines are drawn at the fringes. But even given these limitations, this discourse is just not available *in toto*, neither to the linguists nor to any members of the discourse community. None of them will ever be able to capture the full meaning of a word (or a larger unit of meaning). When we read or listen to texts we have not previously encountered, we may well be confronted with citations showing new semantic aspects. What we find out about the meaning of a word will never be more than an approximation.

4.6 What corpus linguistics can tell us about the meaning of words

In 1976, a relatively unknown author by the name of Courtlandt D. B. Bryan writes a novel about an American soldier in Vietnam killed accidentally by fire from American forces, and he calls this novel *Friendly Fire*. This phrase immediately catches on, and now, also due to two wars against Iraq, there is hardly anyone left in the English-speaking world who does not know what *friendly fire* means. During the last war against Iraq, journalists started using more widely another expression for 'friendly fire': *blue on blue*. During the war, the *Guardian* featured a series of articles on 'The language of war'. *Blue on blue* had an entry of its own explaining its meaning and origin:

> Blue on blue, which made its debut yesterday after the downing of an RAF Tornado by an American Patriot missile, comes from wargaming exercises where the goodies are blue and – in a hangover from cold war days – the baddies are red. Replaces the older term 'friendly fire' which, as Murphy's Laws of Combat eloquently note, isn't.
>
> (Stuart Millar, *Guardian*, 24 March 2003)

This is nothing that could have been predicted by linguists. As we have seen for *globalisation*, there is no rule that can predict the emergence of

new expressions. And there is no rule which tells us whether an expression will catch on or not. *Friendly fire* is a fairly recent addition to our vocabulary. The title of this 1976 novel quickly entered the general discourse. It replaced the military term *fratricide*, which we also find in French. But *fratricide* is a word meaning 'the killing of one's brother (or sister)'. As such, it is rare and smacks of erudition. *Friendly fire*, on the other hand, has a familiar ring, in spite of being a neologism. With each subsequent war, it became more popular. In the 450 million words of the Bank of English, there are 267 occurrences of this phrase. Here are a few citations:

those men died from friendly fire, a phenomenon which he said
Author C.D.F. Bryan < Friendly Fire> came to the five-day
may have been killed in friendly fire. General Johnston also said
Relatives of the friendly fire victims are angrily accusing
mistakes, accidents 'friendly fire', including Private Errol

Do lexicographers regard *friendly fire* as a unit of meaning? The largest online English dictionary is WordNet, an electronic database that has been compiled for some years now – and is still being compiled – at Princeton University under the guidance of Christiane Fellbaum. WordNet is more than a traditional dictionary. It systematically lists relations between each entry and other entries, such as synonymy, hyponymy, meronymy and antonymy. It organises the senses which it assigns to its entries as 'synsets' (sets of synonyms), where each synset is defined as a list of all entries sharing this particular meaning. All synsets or senses come with glosses and often also with an example. For several years now, WordNet has been listing collocations as well. But we did not find an entry for *friendly fire*. There could have been several possible reasons. The phrase was too new, or it was not frequent enough, or it was thought not to be a unit of meaning. The third of these reasons turned out to be the case.

The adjective *friendly* has four senses in WordNet:

1. friendly (vs. unfriendly) – (characteristic of or befitting a friend; 'friendly advice'; 'a friendly neighborhood'; 'the only friendly person here'; 'a friendly host and hostess')
2. friendly – (favorably disposed; not antagonistic or hostile; 'a government friendly to our interests'; 'an amicable agreement')
3. friendly (vs. unfriendly) – ((in combination) easy to understand or use; 'user-friendly computers'; 'a consumer-friendly policy'; 'a reader-friendly novel')

4. friendly (vs. hostile) – (of or belonging to your own country's forces or those of an ally; 'in friendly territory'; 'he was accidentally killed by friendly fire')

This entry shows that a deliberate decision was made not to enter *friendly fire* as a collocation. For the compilers of WordNet, the phrase is a combination of two units of meaning. Are they right? Is there a separate sense of *friendly* accounting for cases such as *friendly fire* and *friendly territory*? Are there other phrases where we find this sense of *friendly*, such as *friendly houses, friendly planes, friendly newspapers*? Friendly houses seems to belong to synset 1 (cf. *friendly neighbourhood*), *friendly newspapers* seems to belong to synset 2 ('favourably disposed'). So perhaps there are really only two instances for the fourth synset. The antonym of *friendly territory* (Google: 5,130 hits) is sometimes *hostile territory* (Google: 27,800 hits), but more often *enemy territory* (Google: 239,000 hits). The antonym of *friendly fire* (Google: 150,000 hits) is sometimes *hostile fire* (Google: 30,300 hits), but again more often *enemy fire* (83,300 hits). Both antonyms should be mentioned in the entry. The question is whether it makes sense to construe a sense that is limited to two instances.

Let us now have a look at *fire* in WordNet. The noun *fire* has eight senses in WordNet.

1. fire – (the event of something burning (often destructive); 'they lost everything in the fire')
2. fire, flame, flaming – (the process of combustion of inflammable materials producing heat and light and (often) smoke; 'fire was one of our ancestors' first discoveries')
3. fire, firing – (the act of firing weapons or artillery at an enemy; 'hold your fire until you can see the whites of their eyes'; 'they retreated in the face of withering enemy fire')
4. fire – (a fireplace in which a fire is burning; 'they sat by the fire and talked')
5. fire, attack, flak, flack, blast – (intense adverse criticism; 'Clinton directed his fire at the Republican Party'; 'the government has come under attack'; 'don't give me any flak')
6. ardor, ardour, fervor, fervour, fervency, fire, fervidness – (feelings of great warmth and intensity; 'he spoke with great ardor')
7. fire – (archaic) once thought to be one of four elements composing the universe (Empedocles))
8. fire – (a severe trial; 'he went through fire and damnation')

The sense we are interested in is, of course, sense 3. Here we find the phrase *enemy fire* in an example. Adding up the glosses for sense 4 of *friendly* and sense 3 of *fire*, we obtain, *mutatis mutandis*, 'the act of firing weapons ... at our own or our allies' forces'. This is an appropriate definition. Is WordNet right to deny *friendly fire* the status of a unit of meaning? While other dictionaries have nothing equivalent to WordNet sense 4 of *friendly*, some of them list *friendly fire* as a separate entry, recognising the phrase as a unit of meaning, e.g. *NODE*: [Military] 'weapon fire coming from one's own side that causes accidental injury or death to one's own people'. Both options seem legitimate. The disadvantage of the first alternative is that it introduces a polysemy which does not exist if we accept the unit of meaning as a solution. In the context of *fire*, *friendly* can only mean sense 4, and in the context of *friendly*, *fire* can only mean sense 3. But multiplying the four senses of *friendly* with the eight senses of *fire*, we end up with thirty-two combinations, out of which we have to select the only one possible. So, if we accept Ockham's razor as the underlying principle for constructing a semantic model ('Entities are not to be multiplied without necessity'), the interpretation of *friendly fire* as a unit of meaning is obviously preferable.

From a methodological point of view, it makes sense to put *friendly fire* down as a unit of meaning because it simplifies the linguist's task to account for what a text, a sentence, a phrase mean. It is more convenient to treat the phrase as a collocation than to describe it as the contingent co-occurrence of two single words. This aspect is particularly important for the computational processing of natural language, for example for machine translation. Computers do not ask whether the meaning of *friendly fire* (or of *false dawn*) is something that cannot be inferred from the meaning of the parts they are constituted of. We use computers that do not understand what people talk about. We want them to facilitate the translation of sentences in which we encounter these and comparable phrases. The above meaning has been discussed in terms of usage and of paraphrase. Usage is something computers can cope with. If *friendly fire* is used in a unique way and not in any of the other thirty-one ways suggested by WordNet, then it is simpler to deal with it as a unit in its own right, as a lexical item that just happens to be composed of two words. But usage does not tell us how we understand the phrase. When we want to communicate with other members of the discourse community about how we understand *friendly fire*, we have to paraphrase it. Whether a given paraphrase, i.e. the interpretation of a phrase, is acceptable to the discourse community has to be left to the members of that community.

The question is, therefore, whether *friendly fire* is a unit of meaning also from the perspective of meaning as paraphrase. The answer to this question is simple. It is a unit of meaning if we find paraphrases telling us how others understand it, and thus, how we would do better to understand it as well. In the *NODE*, we have already found one paraphrase. That this is more than the concoction of an assiduous lexicographer is confirmed by a glance at the Bank of English. Among the citations of *friendly fire* there are about a dozen that comment on the phrase, try to explain it, circumscribe it or downright paraphrase it, for example:

> The United States Defence Department says an investigation has shown that about one out of every four Americans killed in battle during the Gulf War died as a result of '**friendly fire**' – in other words, they were killed by their own side.
>
> Whether called fratricide, amicicide, blue on blue, **friendly fire**, or – as in official U.S. casualty reports from Vietnam – 'misadventure', the phenomenon had become all too commonplace on twentieth-century battlefields.
>
> In Vietnam, the Americans coined the phrase '**friendly fire**', a monstrous use of the language, as if any such fire could be regarded as friendly.
>
> And the other problem, low visibility increases the risk of **friendly fire** – a term that means mistakenly shooting at your own side.

We learn that friendly fire is a 'phrase', a 'term', that it constitutes a 'monstrous use of language', that the Americans introduced it into the discourse in their Vietnam war, and that it means your troops are 'killed by their own side'. Paraphrases of this kind abound when a new unit of meaning, be it a single word or a collocation, enters the discourse. Then people must be told about it. As we have seen, the first evidence of *friendly fire* is probably the title of the 1976 novel. Unfortunately there are no corpora that could verify an assumption that, during that and the subsequent year, there was an abundance of paraphrases. Here again, a bilingual perspective might prove useful. What happens when translators are confronted with a lexical item for which they cannot find a translation equivalent because it has not been translated before?

Corpus linguistics tells us that translation equivalence is not something that latently always exists and just has to be discovered. Translation equivalence has to be construed. As with meaning, this construal is a communal activity, only it does not involve a discourse community of a specific language such as English, but the community of bilingual speakers of the two languages involved. One translator will come up

with a proposal, which is then negotiated with the other members of that community, until agreement is reached and every translator starts using the same equivalent, or until several equivalents are considered acceptable and translators choose among them. It seems as if in the case of *friendly fire* translators had to start from scratch. Apparently there was never a fixed expression in German as an equivalent of *fratricide, blue on blue* or *friendly fire.*

Friendly fire is a phrase which is worth looking at from a bilingual perspective. What does the bilingual perspective add to the issue? This relatively new expression became more frequent only in the course of the first Gulf war, when more British soldiers were killed by friendly (mostly American) fire than by enemy fire. It was only then that the phrase began to be translated into other languages, German among them. How was it translated?

The second edition of the *Oxford–Duden,* published in 1999, acknowledges *friendly fire* as a single lexical item and gives it a separate entry. The translation equivalent it proposes is *eigenes Feuer* ('own fire'). Other translation equivalents which we find in Google and in various corpora are *freundliches Feuer, befreundetes Feuer* and the English collocation *friendly fire,* as a borrowing into German. Most of the texts we find there are texts originally written in German, not translations from the English. Still we have to assume that the concept 'friendly fire' did not exist before it was introduced into the German discourse via translations. For neither of the German equivalents mentioned above occur in the older texts of our corpora. Thus all four German options have to be seen as the results of translations.

It is noteworthy that there is, in Google, only one occurrence of *durch befreundetes Feuer* ('through/by fire of our friends'). We might have expected more, given that *befreundet* is the standard translation for the fourth meaning of *friendly* in WordNet, where we find *friendly fire* together with *friendly territory.* Indeed, *friendly territory* is *befreundetes Territorium* in German. This is a first indication that translators understand *friendly fire* as a collocation and not as a contingent combination of two single words. We can be sure that *befreundetes Feuer* will never become the default equivalent of *friendly fire.* For the phrase *durch freundliches Feuer* we find forty-eight occurrences in Google. This is a second indication that translators see *friendly fire* as a true collocation. For *freundliches Feuer* (*freundlich* being the default translation of *friendly*) would normally – without English influence – never mean 'soldiers killed by their own side' but something quite different, as in this singular Google citation:

Ihre nachtschwarzen Augen leuchteten jedoch in freundlichem Feuer, als sie in die Runde ihrer Amazonenkriegerinnen sah. ('Yet her nightblack eyes glowed in a friendly fire, as she was glancing at the round of her Amazon warriors.')
(www.silverbow.de/kilageschichte.htm)

As a single lexical item, as a unit of meaning, however, *freundliches Feuer* can mean anything the discourse community accepts. Before this can happen, however, people have to do a lot of explaining. This becomes evident from the two examples taken from Google:

Es gab 120 Verletzte durch 'freundliches Feuer' – also Treffer durch die eigenen Leute. ('There were 120 wounded from "friendly fire" – i.e. hits by one's own people.')
(www.stud.uni-goettingen.de/~s136138/ pages/read/depleted.html)

Natürlich haben die amerikanischen Militärs auch einige elektronische Mittel erfunden, um den 'Fratrizid', wie der Tod durch 'freundliches Feuer' im offiziellen Jargon auch genannt wird, möglichst auszuschließen. ('Of course, the American military have invented some electronic gadgets to rule out "fratricide", as death by "friendly fire" is often called in official jargon.')
(www.ish.com/_1048075934919.html)

In the first example the audience is told explicitly, in the form of a paraphrase, what *friendly fire* means. In both instances we find *freundliches Feuer* in quotation marks, making the audience aware that it is a new expression, and that this expression has to be understood as a unit of meaning. The next few years will show whether *freundliches Feuer* will become the default translation of friendly fire.

More frequent is *eigenes Feuer*, with 107 hits in Google for the phrase *durch eigenes Feuer* ('through/by own fire'). Two examples are presented which show that this phrase is the result of English interference:

Das Verteidigungsministerium in London hat Berichte bestätigt, nach denen durch 'eigenes Feuer' in der Nähe von Basra ein britischer Soldat getötet und fünf weitere verletzt worden sind. ('The Ministry of Defence in London has confirmed reports that near Basra one British soldier was killed and five more were wounded by "friendly fire".')
(www.tagesschau.de/aktuell/meldungen/0,1185,OID1725410_TYP1_THE1687956_NAVSPM3~1664644_REF,00.html)

Man kann es sich leicht vorstellen, dass es für die Moral eines militärischen Verbandes die schlimmste Erfahrung ist, wenn ein Kamerad durch eigenes Feuer, durch friendly fire, ums Leben kommt. ('It is easy to imagine that it is the worst experience for the morale of a military unit when a comrade dies from one's own fire, from *friendly fire*.')
(www.dradio.de/cgi-bin/es/neu-kommentar/609.html)

It seems strange indeed that the expression *eigenes Feuer*, which is very easy to understand, is put in quotation marks, but it shows that the speaker uses it as a translation of *friendly fire*. This becomes even more evident in the second example where the perfectly transparent *eigenes Feuer* is paraphrased by the much less familiar *friendly fire*. There seems to be a certain uneasiness to represent the concept expressed in English by a single unit of meaning, by a decomposable adjective+noun phrase, i.e. by two separate words. Therefore it is still doubtful whether *eigenes Feuer* will become the German default equivalent. Even though it seems to be more common, its other disadvantage is that it sounds less like *friendly fire* than the option *freundliches Feuer*.

However, the most frequent equivalent we find is the borrowing *friendly fire*. There are, in Google, 459 hits for '*durch friendly fire*'. Again we notice that in most citations, the collocation is put into quotation marks, indicating the novelty and strangeness of the expression. Here are two examples from the Österreichisches Zeitungskorpus (ÖZK; 'Austrian Newspaper Corpus'), a 500-million word corpus covering the 1990s:

> *Und fast schon ans Zynische grenzt jene Bezeichnung, welche die Militärsprache für den irrtümlichen Beschuß der eigenen Leute kennt. Man nennt das friendly fire – freundliches Feuer.* ('And that name borders almost on cynicism which military jargon uses for mistaken fire on one's own people. They call it friendly fire – *freundliches Feuer*.')

> *An dieser Frontlinie beobachten wir auch immer wieder das, was die Militaristen 'friendly fire' nennen, nämlich Verluste in den eigenen Reihen durch fehlgeleitete Geschosse aus den eigenen, nachfolgenden Linien. Was die Haider-Diskussion anlangt, hat sich dieses Phänomen sogar zu einer Art intellektueller Selbstschußanlage verfestigt.* ('At this frontline, we keep seeing what the military call "friendly fire", i.e. losses in one's own lines from badly aimed shots from one's own rear lines. As for the discussion about Mr Haider, this phenomenon has become firmly established as a kind of intellectual automatic firing device.')

Paraphrases reveal whether a phrase has become a fixed expression, a collocation, a unit of meaning. The paraphrases in these two examples do not tell us what *friendly* means, they explain what *friendly fire* is. While we have learned above to establish, whenever expedient, collocations or fixed expressions on the basis of usage, paraphrases will tell us whether indeed they are understood as units of meaning. There is one more indicator for a true collocation: its availability for metaphorisation processes. The second example demonstrates that friendly fire in German can now be used to refer to internecine warfare. As a

metaphor, *friendly fire* loses the feature of 'accidental fire'; instead it refers to consciously hostile actions within a group. Here is another example, taken from Google:

> *Nicht alle 'Liberalen' sind eingeschwenkt. Aber das friendly fire schmerzt besonders. Merkels Kandidatur ist streitbesetzt.* ('Not all "liberals" [within the Christian Democratic Party] could be won over. But the friendly fire is particularly painful. [Party chair] Merkel's candidature is controversial.')
> (www.zeit.de/2001/51/Politik/print_200151_k-frage.html)

The same metaphorical usage is also found in English texts. Here is an example taken from Google:

> Defence Secretary Geoff Hoon faced questions about the deployment, why it happened so quickly, what his exit strategy was and how long it would last – all of which he had answered in previous exchanges.
>
> But his opposite number, Bernard Jenkin, offered his overall support for the operation.
>
> There was not even much friendly fire from Mr Hoon's own benches.
> (news.bbc.co.uk/hi/english/uk_politics/ newsid_1884000/1884226.stm)

In this section we have explored *friendly fire* in a monolingual and a bilingual context with the aim of finding criteria that set apart statistically significant, but contingent co-occurrences of two or more words from semantically relevant collocations, also called fixed expressions. There are two approaches. If we look at meaning from the perspective of usage, we find that there are good reasons of simplicity to assign collocation status to those expressions which, taken as a whole, are monosemous. The phrase *friendly fire* belongs here; a collocation analysis will reveal that it (almost) always occurs in comparable contexts. This perspective is decisive for the computational processing of natural language; as we will see, it facilitates computer-aided translation.

From the perspective of language understanding, the prime criterion for assigning collocation status to lexical co-occurrence patterns is paraphrase. If we find that a phrase is repeatedly paraphrased as a unit of meaning, we have a reason to assume that it is a single lexical item. A supporting criterion is that the phrase, as a whole, can be used in a metaphorical way. This is, as we have seen, the case for both *false dawn* and *friendly fire*. A third criterion is specific to a bilingual perspective. It seems that the translation equivalent of a true collocation is not what would be the most appropriate translation if each of the elements were translated separately. If it were, we would expect, as the equivalent of *friendly fire*, the German phrase *befreundetes Feuer*, for which we found only one occurrence. Rather, collocations are translated as a whole,

and it does not seem to matter whether the favoured equivalent makes any sense if interpreted literally as a combination of the elements involved. The phrase *freundliches Feuer* is, if taken literally, seriously misleading. For a new unit of meaning, this does not matter; the unit will mean whatever is acceptable to the discourse community. Finally, the high frequency of the English phrase *friendly fire* in German texts suggests that there is no acceptable autochthonous German equivalent and that the English phrase therefore has to be imported.

Is *friendly fire* a true collocation? 'True' collocations can be shown to be not only statistically significant but also semantically relevant. Semantic relevance can be demonstrated both for the methodological approach and for the theoretical approach to the definition of units of meaning. The analysis presented here has demonstrated that the concept of the unit of meaning as the criterion for fixed expressions is not arbitrary. Corpus linguistics can make an enormous impact on lexicography. It can change our understanding of the vocabulary of a natural language. We can overcome the unfortunate situation that most of the (more common) lexical items in the dictionaries are polysemous. The ambiguity we had to deal with in traditional linguistics will disappear once we replace the medieval concept of the single word by the new concept of a collocation or a unit of meaning. Instead of choosing among four senses for *friendly* and eight senses for *fire*, we end up with one single meaning for the fixed expression *friendly fire*.

4.7 Collocations, translation and parallel corpora

In this section, we will address the methodological aspect of working with collocations. Our aim is to demonstrate the impact which the appreciation of the collocation phenomenon can have on translation. As empirical bases, we will produce evidence from several parallel corpora. To work with these corpora, we have to align each text and its translation first on a sentence-to-sentence level and then on the level of the lexical item, be it a single word, or an idiom, or a 'true' collocation, in short, on the level of the unit of meaning.

All those who have ever translated a text into their own or a foreign language know that we do not translate word by word. Nevertheless, our traditional translation aid is the bilingual dictionary. Most entries, by far, are single words, and for most of the words we find many alternatives for how to translate them. In most cases, the dictionary cannot tell us which of the alternatives we have to choose in a particular case. This is why bilingual dictionaries are not very helpful when the target language is not our native language. We do not translate

single words in isolation but units that are large enough to be monosemous, so that for them there is only one translation equivalent in the target language, or, if there are more, then these equivalents will be reckoned as synonymous.

We call these units translation units. Are they the same as units of meaning? Not quite. Natural languages cannot be simply mapped onto each other. The ongoing negotiations among the members of a discourse community lead to results which cannot be predicted. Languages go different ways. They construe different realities. According to most monolingual English dictionaries, the word *bone* seems to be a unit of meaning, described in the *NODE* as 'any of the pieces of hard, whitish tissue making up the skeleton in humans and other vertebrates'. This accurately describes the way *bone* is used in English. From a German perspective, however, *bone* has, traditionally speaking, three different meanings; there are three non-synonymous translation equivalents for it. In the context of fish (or any of its hyponyms), Germans use the word *Gräte*. In the context of non-fishy animals, dead or alive, and of live humans, they call a bone *Knochen*. In the context of the bones of the deceased, the German word is *Gebeine*. For translating into German, the relevant unit of meaning therefore is *bone* plus all the context words that help to make the proper choice among the three German equivalents. What we come up with in our source text is (probably) not a fixed expression, a collocation of the type *false dawn* or *friendly fire*, but rather a set of words (collocates) we find in the close vicinity of *bone*. Thus in Google we find:

> The poor were initially buried in areas in the churchyard or near the church. From time to time, the bones (*Gebeine*) were dug up and then laid out in a tasteful and decorative manner in the charnel house.
> (death.monstrous.com/graveyards.htm)

> Then place trout on a plate and run a knife along each side of ... Sever head, fins and remove skin with a fork. All you have left is great eating with no bones (*Gräten*).
> (www.mccurtain.com/kiamichi/troutbonanza.htm)

> We expect a person to say she feels terrible after breaking a bone (*Knochen*).
> (www.myenglishteacher.net/unexpectedresults.html)

The word in italics indicates the appropriate German translation in each case. A suitable parallel corpus would give us a sufficient number of occurrences for each of the three translation equivalents. Once we have found all the instances of *Gräte(n)* we can then search for *bone(s)* in the aligned English sentence and set up the collocation profile of

bone when translated as *Gräte*. Such a collocation profile is a list of all words found in the immediate context of the keyword (*bone* in our case), listed according to their statistical significance as collocates of the keyword. The collocation profile of bone as the equivalent of *Gräte* will contain words like *trout, salmon, eat, fin, remove,* etc. A dictionary of translation units would give, for each keyword which is ambiguous relative to the target language, the collocation profile going with each of the equivalents. The users then have to check which of the words contained in the collocation profiles occur in the context of the word they are about to translate, and the choice can then be made almost mechanically. These combinations of a keyword together with their (statistically significant) collocates are also called collocations. Thus we find two kinds of collocations: those which can be described as fixed expressions and to which a grammatical pattern can be assigned (*false dawn*: adjective + noun) and those of which we can say only that the collocates are found in the immediate context of the keyword (e.g. *trout* in the context of *bone*). Both kinds of collocations have in common that they are monosemous, either in a monolingual or in a bilingual perspective, and that they therefore represent units of meaning or translation units.

The parallel corpora we are working with have been compiled from selections of the legal documents issued by the European Commission and excerpts from the proceedings of the European Parliament, together with some reports issued by them. They do not talk much about bones. This is why we chose another keyword, French *travail/travaux*. We have included the plural *travaux* in our analysis, because the plural is often rendered as a singular when translated into English. The default translation is *Arbeit* in German, while for English there are two main translation equivalents: *work* and *labour*. When do we translate *travail/travaux* as *work*, when as *labour*? The parallel corpus allows us to set up the relevant collocation profiles, on the basis of an analysis of a context span of five words to the left and five words to the right of the keyword:

Travail/travaux translated as *work*	*Travail/travaux* translated as *labour*
Programme (410)	Marché (747)
Commission (255)	Ministre (170)
Conseil (212)	Marchés (151)
Cours (123)	Sociales (125)
Organisation (122)	Affaires (117)
Préparatoires (113)	Emploi (88)
Vue (109)	Forces (65)
Groupe (108)	Normes (60)

Temps (99)	Femmes (60)
Sécurité (97)	Sociale (50)

For each of the collocation profiles, we have selected the ten most frequent words (other than grammatical words like articles and prepositions) found in the context. The frequency of each item is given in brackets. The most amazing finding is that there is no overlap at all between the two profiles. This is striking evidence that *travail/travaux* occurs in different contexts when it is translated as *work* from those when it is translated as *labour*. Do the collocation profiles help with translation? Here are two French sentences, one in which *travaux* corresponds to *work*, one in which *travail* corresponds to *labour*:

WORK: La réforme du fonctionnement du Conseil soit opérée indépendamment des travaux préparatoires en vue de la future conférence intergouvernementale.

LABOUR: Le Comité permanent de l'emploi s'est réuni aujourdhui sous la présidence de M. Walter Riester, ministre fédéral du travail et des affaires sociales d'Allemagne.

Indeed, the collocation profile approach to translation seems to work. This has little to do with our human understanding of meaning. In the first example, we find *vue*, part of the fixed expression *en vue de*, a prepositional expression meaning 'in the face of'. This is in no way semantically connected with *travaux* meaning 'work'. That it is part of the profile is contingent to our corpus. Also, there seems no sound reason why *travaux* in the context of *Conseil* should be translated as *work* and not as *labour*. It just happens to be that way.

Again, in the second example there is no obvious reason why *emploi* would necessitate the equivalent *labour*. It just so happens that in eighty-eight cases where we find *emploi* close to *travail/travaux*, we find *labour* and not *work* in the translation. The real reason is a different one: *le ministre du travail* is a named entity in the form of a fixed expression for which the equivalent in English is 'Minister of Labour' or 'Secretary of Labour'. What we learn here is that the methodological approach to collocation analysis, the approach based on usage rather than on paraphrase, is a technical operation whose results do not map well onto human understanding.

Investigations of translation equivalence based on parallel corpora are still very much in their infancy. The collocation profiles have to become more refined. The goal is to increase their significance by allocating positions in grammatical patterns to the lexical elements they contain. For the time being our parallel corpora are too small for that. Once they can compare in size with our monolingual corpora we

may well find out that the kind of collocations which are not fixed expressions (like *travail/travaux* and its collocates as they appear in a collocation profile) can be better described as 'true collocations' conforming to a specific grammatical pattern. Thus, in the first sentence, we find *travaux préparatoires.* This phrase can be seen as a monosemous fixed expression, a unit of meaning, conforming to the adjective + noun pattern, and indeed it is (almost) always rendered as *preparatory work* in our parallel corpus.

Parallel corpora monitor the practice of translation. Because they often cannot rely on bilingual dictionaries, translators have to acquire a competence that is the result of experience and interaction with other members of the bilingual discourse community of which they are a part. In their work, they aim to reflect the conventions upon which this community has agreed. The methodology of corpus linguistics enables us to tap this expertise. Our goal is, as we have said above, to replace the single-word entries of current bilingual dictionaries with entries of translation units. The results can be impressive. In a final example, we will use a small French–German parallel corpus. The word we have chosen is *exclusion,* meaning roughly the same as its English counterpart. For the single word we will find an astonishing variety of equivalents. But this diversity disappears once we replace the single word by a collocation of which it is a part. In our example, the fixed expression is *exclusion sociale.* For it, we find only one German equivalent: *soziale Ausgrenzung.* From our bilingual perspective, this proves that *exclusion sociale* is, indeed, a 'true' collocation. It is monosemous; it is a unit of meaning.

To begin, here are some corpus extracts, in the form of a KWIC (key word in context)–concordance, demonstrating the diversity:

extraites pour la vente, à l'exclusion des activités de transformation
den Verkauf mit Ausnahme der Tätigkeiten zur Weiterverarbeitung
['with the exception of activities']

qui résulte de leur travail, à l'exclusion de l'irradiation résultant
wobei Bestrahlung durch Grundstrahlung unberücksichtigt bleiben
['remain ignored']

roïde, la peau ou le tissu osseux, à l'exclusion des extrémités désignées
so Bestrahlung anderer Organe oder Gewebe als Extremitäten
['other organs or tissues than']

des concertations qui débouchent sur l'exclusion de ceux qui sont
deren Ergebnis die Arbeitslosen ausgeschlossen werden
['are being excluded']

il nous manque le combat contre l'exclusion des travailleurs plus âgés
uns fehlt die Bekämpfung der Ausgrenzung von älteren Beschäftigten
['exclusion']

de viandes de gibier sauvage à l'exclusion des viandes de porcin sauvage
von Wildfleisch, ausgenommen Wildschweinfleisch, aus Drittländern
['except boar meat']

This is only a small selection of the variety encountered; all citations are taken from the first ten instances. All translations are perfectly viable. Within their contexts, they are certainly appropriate. Only one of them, we should add, features in the largest French–German dictionary, the *Sachs–Villatte* (1st edition 1979): *mit Ausnahme von/der* as the equivalent of the phrase *à l'exclusion de.* In our few lines, we have four occurrences of this French phrase; and each time it is translated differently. We also find *Ausschluss, Ausschließung, Verweisung,* but no *Ausgrenzung.* Traditional bilingual dictionaries also tend to overlook the fact that it often makes sense to translate a noun phrase (*sur l'exclusion de ceux*) by a verb phrase (*ausgeschlossen werden* ['are being excluded']).

Once we move on to the collocation *exclusion sociale,* the result is straightforward. In 29 of the total of 31 occurrences in our small corpus, we find *soziale Ausgrenzung* as the German equivalent. In the remaining two instances, the adjective has been turned into an adverb modifying the verb. This is a representative selection of our findings:

```
    diese Opfer sozialer Ausgrenzung für immer ausgeschlos
 und der Gefahr sozialer Ausgrenzung entgegengewirkt wird.
 Kampf gegen die soziale Ausgrenzung in ihren verschiedenen
das Problem der sozialen Ausgrenzung junger Leute
 Vermeidung der sozialen Ausgrenzung, sind in einer
   von Armut und sozialer Ausgrenzung ist.
 Bekämpfung der sozialen Ausgrenzung.
   Armut und der sozialen Ausgrenzung. In der EU leben
```

4.8 Conclusion: from meaning to understanding

From a corpus linguistics perspective, the meaning of a unit of meaning is what we can glean from the discourse. It is what we can find out about how a unit of meaning is being used. More important than the plain usage data are the paraphrases of a unit of meaning. They explain to us what this unit means; they attempt to define it; they tell us how this unit is semantically related to other units of meaning. A whole book can be a paraphrase. All those books about globalisation try to explain to their audiences what *globalisation* means. Indeed, the con-

flation of linguistic knowledge with encyclopaedic knowledge is one of the major axioms of corpus linguistics.

It is impossible to compile the complete meaning of a unit of meaning. We cannot have access to more than a tiny fraction of the discourse. Therefore we will never capture all the paraphrases that the discourse contains for a given unit of meaning. Corpora, be they as large as we might imagine, will only ever provide a glimpse of what has been said. This shouldn't deter us. The relevance principle of corpus linguistics assures us that whatever is thought to be important will be repeated in other texts. Once our corpus is large enough to display a certain saturation of paraphrases we can rest assured that what is missing is at least not the mainstream understanding of our unit of meaning.

It is unlikely that any two persons have been exposed to exactly the same discourse events. Once they discuss the meaning of a unit of meaning, their views are bound to differ. They may have heard some identical paraphrases or some that are similar, but each of them will also have heard paraphrases that the other person hasn't. Each of them will subscribe to some paraphrases and will object to others. This is why it is highly unlikely that two people will ever entirely agree on what a unit of meaning means. There is no one description that will completely cover what the unit means. The discourse community is a community of autonomous members. So if two persons want to achieve an agreement on what a unit of meaning such as *globalisation* means, they have to negotiate. The result of their negotiation won't necessarily be that there is only one way to paraphrase *globalisation*; they could also agree that there are two or three competing paraphrases, partially overlapping, partially contradicting each other.

Wouldn't that mean that such a unit of meaning has not one, but two, or three, or many meanings? Wouldn't that contradict our claim that units of meaning have only one meaning, and that, therefore, linguists shouldn't be concerned about lexical ambiguity? Whether a chunk, a conglomerate of words (or, for that matter a single word) is a unit of meaning is not a matter of identical paraphrases, it is a matter of usage. There might be a dozen different paraphrases for *globalisation*; as long as all occurrences of *globalisation* display the same usage pattern, it will continue to be counted as one unit of meaning. Only if two (or more) usage patterns emerge is there ambiguity. Then we are forced to add more lexical elements to the chunk or conglomerate, until again for this larger unit we find only one usage pattern.

Even if there are no two people for whom a unit of meaning means exactly the same, meaning is still a social and not a mental phenomenon. All the paraphrases of a unit of meaning are part of the same

discourse. But no member of the discourse community will have been exposed to all of them. If we ask any member of the discourse community what *globalisation* means, they might provide us with yet another paraphrase, and this paraphrase would, of course, also become part of the discourse and thus be available to other members of the discourse community. They would probably attempt to describe as closely as they could how they understood *globalisation.* But a paraphrase can never be more than one voice among many.

Paraphrases are exclusively verbal. They are part of the discourse. My understanding of a unit of meaning, however, is private. It normally involves a lot of what is not verbal and what cannot be easily verbalised. Your understanding of *globalisation* will originate from the paraphrases you have heard, but it will not stop there. As all these paraphrases tell you something different, you're forced to make up your own mind. While trying to make sense of these paraphrases, you'll use your own judgement. When some people paraphrase *globalisation,* you may have more or less strong reservations. When they tell you that globalisation leads to prosperity, you may associate that with an image of the poor in some underdeveloped country. Or you might think of Enron managers and of how they ruined the indigenous economy of the countries they did business with. However, you'll never be able to verbalise all the associations, all these flashes of memory that come to you whenever someone uses the word *globalisation* in your company. How one understands a unit of meaning will always remain a first-person experience, accessible only to that person, in the same way as emotions are. Only I can really know how I feel grief, no matter how hard I try to explain what I feel to others. Only I can know how I experience globalisation, when I am confronted with the word. People aren't machines. Even if they are fed with the same input they can come up with different conclusions.

The interesting question, then, is how do people develop their understanding of units of meaning? There was a time when we hadn't heard of the word *globalisation.* Today, when we hear it, we think we understand it, and our understanding of it encompasses a lot more than what any dictionary definition would contain. Where do these associations come from? How did we arrive at this complex, fuzzy network of associations and images?

We would like to investigate this quandary by probing into the word *truth.* What does it mean, and how does its meaning relate to our understanding of this word? We will start with the definition we find in the *NODE* (here, and in subsequent quotes, leaving out technical details, examples and further senses):

the quality or state of being true; that which is true or in accordance with fact or with reality; a fact or belief that is accepted as true

What does *true* mean?

in accordance with fact or reality; real or actual

How are *fact* and *reality* defined?

fact: a thing that is indisputably the case; the truth about events as opposed to the interpretation
reality: the world or state of things as they actually exist, as opposed to an idealistic or notional state of them

What does *be the case* mean?

be so

Finally, what is the meaning of *actual(ly)*?

existing in fact, typically as contrasted with what was intended, expected or believed

The definitions are, as we can see, to a large extent circular. This, in itself, is not surprising. All dictionary definitions have to be circular; they are using the words which also have to be defined in the dictionary. What is surprising is the close circuit. *Truth* is defined by *true* and by *fact* and *reality*; *true* is defined by *fact* and *reality*, and by *actual*; *fact* is defined by *be the case* (i.e. 'be so'), and by *truth*; *reality* is defined by *actual(ly)*; and *actual(ly)* is defined by *fact*. So *truth* is defined by *fact*, and *fact*, in turn, is defined by *truth*. Lexicographers normally try to avoid definitions with such close circles because they do not really help the user to understand the lexical item in question. However, in the case of words like *truth*, *fact* and *reality*, there seems to be no other way to proceed.

This set of definitions is not (and is not intended to be) equivalent to my (or anyone else's) understanding of the concept 'truth'. Truth, we no doubt all feel, is something immensely important and goes far beyond just being the case. Truth is a moral value, it is something people owe to each other, it is something very deep which needs to be explored responsibly, and it is not something we come across or appeal to when we deal with the mundane facts of everyday life like asking for a pint of ale.

Fortunately, the *NODE* gives us some hints that truth is not quite as simple as we have made it look in our summary of the dictionary

definitions. This gives credit to the exceptional quality of this dictionary. *Truth*, we are told, can also mean a 'belief that is accepted as true', *truth* stands in opposition to *interpretation*, and it refers to reality as opposed to 'idealistic or notional' things, while *actual(ly)* refers to facts as opposed to 'what was intended, expected or believed'. So truth is opposed to what is just 'notional' or 'believed', or a subjective 'interpretation', and it can also be a 'belief' that is accepted (by whom?) as the truth. So *truth* is more than 'what is the case'. People can have conflicting ideas about what is true. There is a tension that seems to go along with this word; and the dictionary makes us aware that truth is a contentious issue. This is shared by my understanding of *truth*.

A look at the American *Random House College Dictionary* of 1975 (two-thirds of the size of *NODE*) shows definitions for the words in question that are, on the surface, very similar to the *NODE*. This is what we find for *truth* (here again, and in subsequent quotes, we leave out technical details, examples and further senses):

> 1. true or actual state of the matter. 2. conformity with fact or reality; verity. 3. a verified or indisputable fact, proposition, principle. 4. state or character of being true.

These are the definitions for *fact* and *reality*:

> *fact*: 1. the quality of existing or being real. 2. something known to exist or have happened. 3. a truth known by actual experience or by observation. 4. something said to be true or to have happened.
> *reality*: 1. the state or quality of being real. 2. resemblance to what is real. 3. a real thing or a fact.

It seems we also have to take into consideration the adjective *real*:

> 1. true, not merely ostensible or nominal. 2. actual rather than imaginary, ideal or fictitious. 3. having actual, rather than imaginary, existence. 5. genuine, authentic.

The other two words asking for definitions are *verity* and *verify*:

> *verity*: 1. the state or quality of being true. 2. something that is true, as a principle, a belief, or statement.
> *verify*: 1. to prove the truth of, confirm. 2. to ascertain the truth, or correctness of. 3. to act as ultimate proof or evidence of; serve to confirm.

On the whole, the Random House definitions seem to profess a stronger realism than the *NODE* ones. The 'true or actual state of a

matter' is much more straightforward than 'the quality or state of being true; that which is true or in accordance with fact or with reality; a fact or belief that is accepted as true'. Something is true, or it is not. We are not made aware of the tension connected with *truth*. Where it comes in is in the first definition of *real*: 'true, not merely ... nominal'. But this allusion to the medieval battleground of realism versus nominalism presupposes an acquaintance with philosophy few people can claim; on others it is mostly lost. The discourse is brought in by the phrase 'indisputable fact', reminding us of the *NODE* phrase of something being 'indisputably the case'. It comes in much stronger in the definition 'something said to be true or to have happened'. But should we subscribe to this definition? Would we really say that a UFO incident was true because it is said by some people to have happened? The Random House definitions do not let us feel the tension that the *NODE* conveys, for instance with its definition for *actual*: 'existing in fact, typically as contrasted with what was intended, expected or believed'.

When we ask ourselves how we understand the word *truth*, or what *truth* means to us personally, the mundane dictionary definitions with their close circular definitions will be about the last thing that comes to our mind. Truth, we feel, is something very important, something that is frequently at stake. It is a moral value. The way we will have first learned about truth may easily have been in the context of lying. Our parents, rightly interested in our whereabouts, wanted to make sure we would tell them the truth, and this is why they taught us lying is wrong. It is strange that neither of the dictionaries mentions *lies* in their definitions of truth. It certainly plays a very prominent role in my understanding of *truth*. In the Catechism of the Catholic Church we read:

> (2483) Lying is the most direct offence against truth. To lie is to speak or act against the truth in order to lead into error someone who has the right to know the truth.

This is a somewhat jesuitical way of putting it, in spite of being the received wisdom. Parents, we are told, do have the right to know the truth; children don't. Again tension comes in. Truth is never simple.

Understanding is a first-person experience. We will never be able to convey fully, verbally or in any other way, to other people how we understand a unit of meaning, just as we are not able to let anyone else know exactly what kind and intensity of pain we suffer. Our understanding of any unit of meaning is not something static that could be put into words. When we hear a unit of meaning, or a text sequence, or when we want to use a unit of meaning within a textual sequence, there are memories that come up, memories of events to which we were a

witness or in which we played a part. Often what we think are genuine memories of an event itself are recollections of subsequent verbalisations of the event. All these memories involve images or other sensations, and while some of them refer to actual sensual data, others are largely imaginary. Another part of these memories will be memories of other people's contributions to the discourse, things we heard people say themselves, or things that were reported. These texts again, as we remember them, will evoke memories. It is our memory that forms our understanding. But we have little control over what we remember. Remembering is a combination of intention and randomness. It is not the result of an algorithmic procedure. Our understanding of a unit of meaning is nothing fixed. It depends on the situation, on how we feel, on what we want to do, and on innumerable other factors. Any new input will change our understanding. We never can understand a text when we read it a second time in the same way as we understood it as we read it the first time.

We would not know about truth without other people telling us what it is. It is the paraphrases, the explanations, the instructions we received from them that we remember and that evoke the memories of events we associate with them. These paraphrases are what constitutes for us the meaning of a unit of meaning. As we have said often before, we have all been exposed to different sets of such paraphrases, and therefore a word such as *truth* may mean different things to different people. But what is worth remembering is also worth repeating. Therefore many of the paraphrases will strike a familiar chord even if we have never read the texts in which they occur.

To find paraphrases of the unit of meaning *truth*, we searched the Bank of English (BoE), with its 420 million words, for sentences beginning with '*Truth is*'. This is a very common pattern for opening up a paraphrase. Altogether, we found 159 occurrences of this phrase. Compared to the total number of occurrences of the word *truth* in the BoE, 34,645, this is only a tiny fraction. As it turned out, about half of the citations were not paraphrases at all. They were sentences like: *Truth is most of us have mediocre souls.* However, the remaining paraphrases still represent something of a common denominator of what truth means to all of us. At first glance it seems amazing that so few of them refer to what the dictionaries tell us. Perhaps it's not so strange, though. The definitions we find in the dictionaries are normally not controversial. So there is no point discussing them in the discourse community. Here, now, is a selection of paraphrases for truth, ordered loosely into seven pigeonholes. We've left out from these corpus citations what we deemed to be accidental and irrelevant.

(1) Truth is an emotional phenomenon
Truth is a force which pierces your heart, Vysotsky said.
Truth is mostly subjective and that's good when you are talking about music.
Truth is an attribute of love. Love is not complete without truth.
The truth never hurts another person.

(2) Truth is a spiritual phenomenon
Truth is a totem to Murphy: artistic and spiritual truth, rather than mere accuracy.
Truth is always before us: the truth of God is bigger and smaller than all our formulations, however precious they may be.
Truth is one of the first casualties of secularism.
Truth is our king, the rest is nothing.
'Truth is our king.' Truth was holy, and cloud-cuckoo-land was silly, and blasphemy too.

(3) Truth is ugly
Truth is full of warts, and worse. It is a heap of dirt, sucked dry by Ariadne's kiss.
Truth is horrible. We live in an empty and meaningless cosmos where we can only expect to suffer.
Truth is not Beauty. It is something to be hidden in the deepest depths of one's inmost being.

(4) Truth is elusive
Truth is a black cat in a darkened room and justice is a blind bat, said Bertolt Brecht.
Great Britain spent centuries making modifications to the ancient system of trial by combat. Truth is immaterial and, often, so is justice.
Truth is the most fragile of ideas.

(5) Truth is relative
Truth is always relative.
Truth is an immensely personal matter – what is true for me is not necessarily true for you.
Truth is, in fact, a product of dispute.
Truth is sought in a joint quest and effort.
Truth is a victim of time.
Truth is something complicated, something to be sought out.
Truth is provisional, Mr Rushdie seems to be saying.

(6) Truth is absolute
Truth is absolute.
Truth is blindingly obvious once you've recognised it.
Truth is established rationally, by proof.

Truth is normatively consonant with warranted assertability.
Truth is truth, in Malaysia or in Manchester.

(7) Truth is a many splendoured thing
Truth is a difficult concept.
Truth is a problem.
Truth is at stake.
Truth is the main thing. Lenin said: More light! Let the party know everything!
Truth is the foundation of trust.
Truth is manly.
Truth is often stranger than fiction.
Truth is what the masses like.
Truth is not a priority.

All these statements are part of the meaning of truth. We could not have heard them all. But all are part of the discourse. Many of them will sound familiar. Google has 33,000 hits for *truth* + '*stranger than fiction*'. Similar figures would be found for many other paraphrases. Even the phrase 'Truth is normatively consonant with warranted assertability' is not as singular as it looks; Google has 292 hits for '*truth consonant warranted assertability*'. What has caught the attention of people will be endlessly repeated in the discourse. It will leave traces in many texts.

As we see it, understanding a unit of meaning is a feature of our memories. Part of it is verbal input, what we have gleaned from the discourse. This is the part that constitutes what the unit means for each of us individually. It is what we can convey verbally, by repeating it verbatim or by rephrasing it. The other part of understanding is constituted by the memories that are evoked by hearing or saying a unit of meaning in a given situation. These memories are fuzzy and instable, they are full of holes and constantly shifting. They are true first-person experiences. Try as we can, we will never be able to relate them faithfully to others. This doesn't mean they cannot be verbalised. We will refer to them whenever we discuss truth with other members of the discourse community. These textual sequences will enrich the discourse on truth, and they may well change what *truth* means, for those who hear them and for ourselves. The third part of our understanding of a unit of meaning is our rationalisation of the verbal input and of the memories it evokes. We don't have to accept everything we're being told. We can form our own opinion, and that can differ more or less from the mainstream meaning of that unit. We can contribute our own paraphrase of *truth*. If it differs a lot from what others believe, they will probably reject it. Then it won't leave traces in subsequent texts. But

our understanding of paraphrase may just differ modestly from what *truth* means to other people. If it catches their attention, if it expresses an idea that lies in the air, if it reverberates the Zeitgeist, then it may be picked up by others, and it may even change the mainstream meaning.

For corpus linguistics, meaning is a social phenomenon. It is the members of the language community who negotiate what units of meaning mean. What a unit of meaning means is the result of a democratic process. Everyone has, or should have, a voice in it. Meaning is not a matter for experts, self-appointed or otherwise. We do not have to accept that the meaning of *murder* includes abortion. There is no truth in the matter of meaning, and there is no legitimate coercion to agree on a definition. We do not have to accept that *property* is an inviolate right. We can also say that all *property* is theft. Both views are equally legitimate. What we have to learn is what it takes to make our paraphrases palatable to the other members of the discourse community. Education is about learning to exercise one's rights as a free citizen in a responsible way. Corpus linguistics puts us into a position where we can inform ourselves what use others have made of language. This knowledge empowers us to contribute successfully to the discourse of which we are members.

This page intentionally left blank

Glossary

affix
a meaningful element which is typically found attached to a stem or base; for example, in English the word *unwanted* contains two affixes, the prefix *un-* and the suffix *-ed.*

alignment
the process of aligning equivalent units in bilingual or multilingual **parallel corpora**, so that a unit in one language corresponds to the equivalent unit in another language and both of them can be accessed or displayed at the same time.

annotation
corpus-external information added to a **corpus**, such as **tagging** or information identifying the origin and nature of the text.

antonymy
the relationship of oppositeness in meaning, as in English between the words *good* and *bad* or *buy* and *sell.*

cognate, cognate word
(1) a word related to one or more other words in the same language by derivation, as in English *thought* is a cognate of *think.*
(2) a word which shares a common ancestor with one or more other words, as with English *sleep,* Dutch *slaap* and German *Schlaf,* which are all considered to be descended from an ancestral Germanic form.

cognitive linguistics
a branch of linguistics or cognitive science which seeks to explain language in terms of mental processes or with reference to a mental reality underlying language.

collocate
a word repeatedly found in the close vicinity of a node word in texts; for example, in English the words *partial, lunar, solar* are collocates of the word *eclipse.*

collocation
the habitual meaningful co-occurrence of two or more words (a node word and its **collocate** or **collocates**) in close proximity to each other; as a lexical relationship, **collocation** can be defined quantitatively as the degree to which the probability of a word *y* occurring in text is increased by the presence of another word *x*.

collocation profile
a computer-generated list of all the **collocates** of a node word in a **corpus**, usually listed in the order of their statistical significance of occurrence.

concordance
a list of lines of text containing a node word, nowadays generated by computer as the principal output of a search of a **corpus** showing the word in its contexts and thus representing a sum of its usage; see also **KWIC**.

connotation
the emotional or personal associations of a word, often contrasted with **denotation**.

content word
a word with a relatively clear meaning of its own, in contrast to a **function word**.

corpus
a collection of naturally occurring language texts in electronic form, often compiled according to specific design criteria and typically containing many millions of words.

denotation
the central or core meaning of a word, sometimes claimed to be the relationship between a word and the reality it refers to, and often contrasted with **connotation**.

discourse
the totality of verbal interactions and activities (spoken and written) that have taken place and are taking place in a language community.

etymology
an account of the historical origin and development of a word.

fixed expression
a co-occurrence of two or more words which forms a unit of meaning.

function word
a word with a relatively general meaning serving to express functions such as grammatical relationships, as in English the words *for, to, the,* in contrast to a **content word**.

generative
(of a grammar or a finite set of formal rules) capable of generating an infinite set of grammatical sentences in a language.

hapax legomenon
a word or form found only once in a body of texts, for example in a **corpus** or in the works of a single author.

hyponymy
the relationship of meaning between specific and general words; for example, in English *rose* is a hyponym of *flower*

idiom
a type of **fixed expression** in which the meaning cannot be deduced from the meanings or functions of the different parts of the expression, as with the English idiom *kick someone upstairs* meaning 'move someone to what seems to be a more important post but with the motive of removing them from their current post'.

KWIC (short for **key word in context**)
a computer-generated set of **concordance** lines in which the node word is in the centre of each line.

lemma
a form which represents different forms of a lexical entry in a dictionary, as with the English lemma *bring* representing *bring, brings, bringing* and *brought.*

lexical item
a word understood as a unit of meaning rather than as a written or spoken form.

lexicogrammar
the **lexicon** and grammar of a language, taken together as an integrated system.

lexicon
the vocabulary or word stock of a language, usually understood as a lexical system or as part of **lexicogrammar.**

lexicology
the study of the **lexicon.**

lexicography
the art and science of dictionary-making.

mentalism
the belief in the reality of the human mind and in the possibility and importance of systematically investigating its nature.

meronymy
the relationship of meaning between part and whole, as in English between the words *arm* and *body* or *sole* and *shoe.*

monitor corpus
a **corpus** which contains specimens of language taken from different times (and is ideally regularly updated) and which thus assists the study of language change.

morpheme
the smallest element of language which carries a meaning or function, including **affixes** such as *pre-* or *-ed* as well as irreducible words such as *want* or *white.*

neologism
a new word, form, construction or sense introduced into **discourse** and ultimately into the language.

opportunistic corpus
a **corpus** which makes use of existing and readily available resources, does not claim to be representative, and reflects the assumption that every corpus is inevitably imbalanced.

paradigm
a set of forms, usually grammatically conditioned, based on a single **lexical item**, as in English the set *chase, chasing, chased* or *want, wanting, wanted.*

parallel corpus
a **corpus** which contains equivalent and usually **aligned** texts in two or more languages; it is sometimes called a **translation corpus** but does not always include the original text as well as translations of it.

parsing
grammatical analysis of a text, usually with the principal aim of identifying elements as subjects, nouns, verbs, and so on.

part of speech = word class

qualia
the felt qualities associated with experiences, such as the feeling of a pain, or the hearing of a sound, which are expressed by specific words.

reference corpus
a **corpus** which aims to be balanced and to reflect the contemporary language.

semantics
the systematic study of meaning in language.

special corpus
a **corpus** built for a special research purpose.

synonymy
the relationship of identity (or more realistically of near identity) in meaning, as in English between *dentures* and *false teeth* or *often* and *frequently.*

tagging
attaching grammatical labels, usually indicating **word classes**, to words in a **corpus**, usually by automatic methods.

term
a word with a meaning that is relatively precise and independent of the context, often subject to some special convention or regulation, as for example with technical terms defined by standards associations.

thesaurus
a reference work in which words are grouped by meaning rather than listed alphabetically.

translation corpus
a **corpus** which contains an original text and at least one translation of it into another language; see also **parallel corpus**.

word class
a small set of grammatical categories to which words can be allocated, varying from language to language but usually including such classes as noun, verb and adjective; also known as **part of speech**.

This page intentionally left blank

References

Aarts, J., P. de Haan and N. Oostdijk (eds), 1992, *English Language Corpora: Design, Analysis and Exploitation*, Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Rodopi, Amsterdam.

Adams, Douglas, 1983, *The Meaning of Liff*, Pan Books, London.

The American Heritage Dictionary, 2000, (4th edn), Houghton Mifflin, Boston.

Béjoint, Henri, 2000, *Modern Lexicography: an Introduction*, Oxford University Press, Oxford. (First published as *Tradition and Innovation in Modern English Dictionaries*, 1994.)

Biber, D., S. Conrad and R. Reppen, 1998, *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press, Cambridge.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan, 1999, *Longman Grammar of Spoken and Written English*, Pearson Education, Harlow, England.

Brewer's Dictionary of Phrase and Fable, 1999, Cassell, London (millennium edition, revised by Adrian Room, originally compiled by Ebenezer Cobham Brewer and published 1870).

Calzolari, N., M. Baker and T. Kruyt (eds), 1995, *Towards a Network of European Reference Corpora*, Linguistica Computazionale Vol. XI–XII, Giardini Editori e Stampatori, Pisa.

Carroll, Lewis, 1994, *Alice Through the Looking Glass*, Penguin Popular Classics, London.

Catechism of the Catholic Church, 1995, available online (www.christusrex.org).

Chambers's 20th Century Dictionary, 1983, edited by E. M. Kirkpatrick, Chambers, Edinburgh.

Chapman, R. W., 1948, *Lexicography*, Oxford University Press, London.

Chomsky, Noam, 1957, *Syntactic Structures*, HarperCollins Publishers, New York and Glasgow.

Chomsky, Noam, 1965, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Massachusetts.

Chomsky, Noam, 1966, *Cartesian Linguistics: a Chapter in the History of Rationalist Thought*, Harper & Row, New York and London.

Chomsky, Noam, 1993, *Rethinking Camelot: JFK, the Vietnam War, and US Political Culture*, South End Press, Boston.

Chomsky, Noam, 2000, *New Horizons in the Study of Language and Mind*, Cambridge University Press, Cambridge, Massachusetts.

Chomsky, Noam and Edward S. Herman, 1988, *Manufacturing Consent: the Political Economy of the Mass Media*, Pantheon Books, New York.

Collins COBUILD English Language Dictionary, 1987, editor-in-chief John Sinclair, HarperCollins, London.

Collins Dictionary of the English Language, 1979, edited by Patrick Hanks, William Collins, Glasgow.

Collins–Robert French Dictionary, 1998 (5th edn), HarperCollins, London.

Cowie, A. P., 1990, Language as words: lexicography, in N. E. Collinge (ed.), *An Encyclopedia of Language*, Routledge, London and New York.

Culler, Jonathan, 1976, *Saussure*, Fontana Modern Masters, William Collins, Glasgow.

Dennett, D. C., 1993, *Consciousness Explained*, Penguin, London.

Dictionary of Caribbean English Usage, 1996, edited by Richard Allsopp, Oxford University Press, Oxford.

Dictionary of Jamaican English, 1980 (rev. edn), compiled by Frederic G. Cassidy and Robert Le Page, Cambridge University Press, Cambridge.

Dictionary of Lexicography, 1998, compiled by R. R. K. Hartmann and Gregory James, Routledge, London.

Edmonds, P., 2002, Introduction to SENSEVAL, *ELRA Newsletter*, October 2002.

Eggins, Suzanne, 1994, *An Introduction to Systemic Functional Linguistics*, Pinter, London.

Fellbaum, Ch. (ed.), 1998, *WordNet: an Electronic Lexical Database*, MIT Press, Cambridge, Massachusetts.

Firth, J. R., 1957, *Papers in Linguistics 1934–1951*, Longman, London.

Fodor, J. A., 1975, *The Language of Thought*, MIT Press, Cambridge, Massachusetts.

Fodor, J. A., 1994, *The Elm and the Expert: Mentalese and its Semantics*, MIT Press, Cambridge, Massachusetts.

Fodor, J. A., 1998, *Concepts; Where Cognitive Science Went Wrong*, 1996 John Locke Lectures, Oxford University Press, Oxford.

Fodor, J. A. and E. Lepore, 2002, *The Compositionality Papers*, Oxford University Press, Oxford.

Fries, Charles C., 1940, *American English Grammar*, Appleton Century Crofts, New York.

Fries, U., G. Tottie and P. Schneider (eds), 1993, *Creating and Using English Language Corpora*, Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich, Rodopi, Amsterdam.

Goody, J., 2000, *The Power of Written Tradition*, Smithsonian Institute Press, Washington and London.

Green, Jonathon, 1996, *Chasing the Sun: Dictionary Makers and the Dictionaries They Made*, Henry Holt and Company, New York.

Haas, W., 1962, The theory of translation, *Philosophy* 37: 208–28.

Halliday, M. A. K., 1994a (2nd edn), *An Introduction to Functional Grammar*, Edward Arnold, London.

Halliday, M. A. K., 1994b, On language in relation to the evolution of human consciousness, in S. Allen (ed.), *Of Thoughts and Words – Proceedings of Nobel Symposium 92: The Relation Between Language and Mind*, World Scientific Publishing, Singapore and London.

Harris, Roy, 1987, *Reading Saussure: a Critical Commentary on the Cours de Linguistique Générale*, Duckworth, London.

Hartmann, R. R. K., 1983, *Lexicography: Principles and Practice*, Academic Press, London and New York.

Hartmann, R. R. K., 1986, *The History of Lexicography*, John Benjamins, Amsterdam and Philadelphia.

Hartmann, R. R. K., 2001, *Teaching and Researching Lexicography*, Longman Pearson Education, Harlow.

Hasan, Ruqaiya, 1987, Directions from structuralism, in N. Fabb, D. Attridge, A. Durant and C. MacCabe (eds), *The Linguistics of Writing: Arguments Between Language and Literature*, Manchester University Press, Manchester.

Householder, Fred W. and Sol Saporta (eds), 1962, *Problems in Lexicography*, Indiana University Press, Bloomington.

Hunston, Susan and Gill Francis, 2000, *Pattern Grammar: a Corpus-Driven Approach to the Lexical Grammar of English*, John Benjamins, Amsterdam and Philadelphia.

Jackson, H. and E. Ze Amvela, 1999, *Word Meaning and Vocabulary: an Introduction to Modern English Lexicology*, Cassell, London.

Johansson, S., G. Leech and H. Goodluck, 1978, *Manual of Information to Accompany the Lancaster–Oslo/Bergen Corpus of British English, for Use*

With Digital Computers, University of Oslo, Department of English, Oslo, available online (http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM).

Johnson's Dictionary: A Modern Selection by E. L. McAdam and George Milne, 1995, Cassell, London.

Keller, R., 1998, *A Theory of Linguistic Signs*, Oxford University Press, Oxford.

Kennedy, G., 1998, *An Introduction to Corpus Linguistics*, Longman, London and New York.

Krishnamurthy, R. (ed.), 2003, *English Collocation Studies: the OSTI Report*, University of Birmingham Press, Birmingham (new edition of Sinclair, J., S. Jones and R. Daley, 1970, English Lexical Studies: Report to OSTI on Project C/LP/08).

Lakoff, G., 1987, *Women, Fire, and Dangerous Things*, University of Chicago Press, Chicago.

Landau, Sidney I., 1989 (2nd edn), *Dictionaries: the Art and Craft of Lexicography*, Cambridge University Press, Cambridge.

Langenscheidts Großwörterbuch Französisch (Sachs–Vilatte), 1979, Teil 1: Französisch–Deutsch, Völlige Neubearbeitung, Teil 2: Deutsch–Französisch, Völlige Neubearbeitung 1968 mit Nachtrag 1979, Langenscheidt, Berlin and Munich.

Larousse, Pierre, 1865–76, *Grand Dictionnaire Universel du XIXe Siècle*, 15 vols, Librairie Larousse, Paris. (Supplements published 1878, 1890 and various editions published later.)

Levitt, T., 1983, The Globalization of Markets, *Harvard Business Review* 6 (3), May–June 1983.

Lewis, Charlton T. and Charles Short, 1879, *A Latin Dictionary: Founded on Andrews' Edition of Freund's Latin Dictionary, Revised, Enlarged and in great part Rewritten by Charlton T. Lewis and Charles Short*, Oxford University Press, Oxford. (Various editions published later.)

Liddell, Henry George and Robert Scott, 1843, *Greek–English Lexicon*, Oxford University Press, Oxford. (Various editions published later.)

Littré, Emile, 1863–73, *Dictionnaire de la Langue Française* (Supplement published 1878 and various editions published later.)

Longman Dictionary of Contemporary English, 1978, editor in chief Paul Procter, Longman, London.

Longman Dictionary of Contemporary English, 1987 (new edn), editorial director Della Summers, Longman, Harlow.

Longman Dictionary of English Idioms, 1979, Longman, Harlow and London.

Lyons, J., 1970, *Chomsky*, Fontana Modern Masters, William Collins, London.

Lyons, J., 1977, *Semantics*, 2 vols, Cambridge University Press, Cambridge.

McArthur, Tom (ed.), 1992, *The Oxford Companion to the English Language*, Oxford University Press, Oxford.

McDavid Jr, Raven I. and Audrey R. Duckert (eds), 1973, *Lexicography in English*, New York Academy of Sciences, Annals 211, New York.

McEnerry, T. and A. Wilson, 1996, *Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Macmillan English Dictionary for Advanced Learners, 2002, Macmillan Publishers, Oxford.

Macquarie Dictionary, 1997 (3rd edn), editor in chief Arthur Delbridge, Macquarie Library, Sydney.

Macquarie Concise Dictionary, 1998 (3rd edn), general editors A. Delbridge and J. R. L. Bernard, Macquarie Library, Sydney.

Malinowksi, B., 1935, *Coral Gardens and their Magic*, 2 vols, Allen & Unwin, London.

Martin, J. R., 1992, *English Text: System and Structure*, John Benjamins, Philadelphia and Amsterdam.

Mayr, E., 2002, *What Evolution Is*, Weidenfeld & Nicholson, London.

Millar, S., 2003, The Language of War, *Guardian*, 24 March 2003.

Miller, K. L., 2002, The New Buzzword: Globaloney, *Newsweek*, Special Edition, December 2002–February 2003.

Moon, R., 1998, *Fixed Expressions and Idioms in English. A Corpus-based Approach*, Clarendon Press, Oxford.

New English Dictionary on Historical Principles, 1884–1928, edited by James A. H. Murray, H. Bradley, W. A. Craigie and C. T. Onions, Clarendon Press, Oxford.

New Oxford Dictionary of English, 2001, Oxford University Press, Oxford.

New Shorter Oxford English Dictionary on Historical Principles, 1993 (rev. edn), 2 vols, edited by Lesley Brown, Clarendon Press, Oxford.

Oxford Dictionary of New Words, 1997, edited by E. Knowles and J. Elliott, Oxford University Press, Oxford.

Oxford Dictionary of Sociology, 1998, edited by G. Marshall, Oxford University Press, Oxford.

Oxford–Duden German Dictionary German–English/English–German, 1999 (2nd edn), Oxford University Press, Oxford.

Oxford English Dictionary, 1989 (revised edition of the *New English Dictionary on Historical Principles*), 20 vols, prepared by J. A. Simpson and E. S. C. Weiner, Clarendon Press, Oxford.

Palmer, H. and A. S. Hornby, 1933, *The Second Interim Report on English Collocations*, Kaitakusha, Tokyo.

Pavel, Silvia, and Diane Nolet, 2002, *Handbook of Terminology*, Ter-

minology and Standardization Translation Bureau, Ministry of Public Works and Government Services, Canada.

Pinker, S., 1994, *The Language Instinct: How the Mind Creates Language*, William Morrow, New York.

Putnam, H., 1975, The Meaning of 'Meaning', in *Mind, Language and Reality*, Philosophical Papers vol. 2.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik, 1985, *A Comprehensive Grammar of the English Language*, Longman, London.

Random House College Dictionary, 1975 (2nd edn), Random House, New York.

Robins, R. H., 1979 (2nd edn), *A Short History of Linguistics*, Longman, London.

Roget, Peter Mark, 1852, *Thesaurus of English Words and Phrases*, Longman, Brown, Green and Longman, London. (Various editions published later.)

Sager, Juan C., 1990, *A Practical Course in Terminology Processing*, John Benjamins, Amsterdam and Philadelphia.

Said, E. W., 1995, *Orientalism: Western Conceptions of the Orient*, Penguin, London.

Sampson, Geoffrey, 1980, *Schools of Linguistics: Competition and Evolution*, Hutchinson, London.

Sampson, G., 1997, *Educating Eve: The 'Language Instinct' Debate*, Cassell, London.

de Saussure, Ferdinand, 1960, *Course in General Linguistics*, Peter Owen, London (translated by Wade Baskin).

de Saussure, Ferdinand, 1972, *Cours de Linguistique Générale*, Editions Payot, Paris (édition critique préparée par Tullio de Mauro).

de Saussure, Ferdinand, 1983, *Course in General Linguistics*, Duckworth, London (translated by Roy Harris).

Searle, J. R., 1983, *Intentionality: An Essay in the Philosophy of Mind*, Cambridge University Press, Cambridge.

Searle, J. R., 1992, *The Rediscovery of the Mind*, MIT Press, Cambridge, Massachusetts.

Searle, J. R., 1998, *Mind, Language and Reality*, Basic Books, New York.

Sinclair, J. (ed.), 1987, *Looking Up: An Account of the Cobuild Project in Lexical Computing*, HarperCollins, London.

Sinclair, J., 1991, *Corpus, Collocation, Concordance*, Oxford University Press, Oxford.

Sinclair, J., 1996, The Empty Lexicon, *International Journal of Corpus Linguistics* 1: 99–119.

Sperber, D. and D. Wilson, 1998, The Mapping between the Mental

and the Public Lexicon, in P. Carruthers and J. Boucher (eds), *Language and Thought*, Cambridge University Press, Cambridge.

Strang, Barbara M. H., 1970, *A History of English*, Methuen, London.

Stubbs, M., 2001, *Words and Phrases: Corpus Studies of Lexical Semantics*, Blackwell Publishers, Oxford.

Svartvik, J. (ed.), 1990, *The London Corpus of Spoken English: Description and Research*, Lund Studies in English 82, Lund University Press, Lund.

Tognini-Bonelli, E., 2001, *Corpus Linguistics at Work*, John Benjamins, Amsterdam.

Warburg, Jeremy, 1968, Notions of correctness, supplement to Quirk, Randolph (2nd edn), *The Use of English*, Longman, London and Harlow.

Webster, Noah, 1828, *American Dictionary of the English Language.*

Wierzbicka, Anna, 1980, *Lingua Mentalis: the Semantics of Natural Language*, Academic Press, Sydney.

Wierzbicka, A., 1996, *Semantics. Primes and Universals*, Oxford University Press, Oxford.

Wildhagen, K. and W. Héraucourt, 1963–72, *English–German/German–English Dictionary*, 2 vols, Brandstetter, Wiesbaden.

Wright, Joseph, 1898–1905, *The English Dialect Dictionary, Being the Complete Vocabulary of all Dialect Words still in Use, or Known to have been in Use during the Last Two Hundred Years*, 6 vols, Henry Frowde, London.

Zgusta, Ladislaw, 1971, *Manual of Lexicography*, Mouton, The Hague.

Corpora

The Bank of English, http://titania.cobuild.collins.co.uk/

British National Corpus, http://www.natcorp.ox.ac.uk/

Brown Corpus, manual available at http://www.hit.uib.no/icame/brown/bcm.html

Czech National Corpus, http://ucnk.ff.cuni.cz/english/

IDS (Institut für Deutsche Sprache) corpus *COSMAS*, http://corpora.ids-mannheim.de/~cosmas/

International Corpus of English (ICE), http://www.ucl.ac.uk/english-usage/ice/

London Lund Corpus, http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

Språkbanken, http://spraakbanken.gu.se/

WordNet, http://www.cogsci.Princeton.edu/~wn/

This page intentionally left blank

Index

(words in bold can be found in the Glossary)