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