4 The use of corpora in language studies

4.1. CORPORA AS SOURCES OF EMPIRICAL DATA

In this and the following chapter we examine the roles which corpora may play both in the study of language itself and in the development of computational tools for processing natural language. This chapter is concerned with their roles in the study of language.

The importance of corpora in language study is closely allied to the importance more generally of empirical data. Empirical data enable the linguist to make statements which are objective and based on language as it really is rather than statements which are subjective and based upon the individual's own internalised cognitive perception of the language. The use of empirical data also means that it is possible to study language varieties such as dialects or earlier periods in a language for which it may not be possible to use a rationalist approach. But empirical linguistic research may be carried out without using a corpus. Although many researchers will refer to their data as a corpus, frequently these data do not fit the definition of a corpus in the sense that we have tended to use that term in this book – as many other corpus linguists have – for a body of text which is carefully sampled to be maximally representative of a language or language variety. These other data should more properly be thought of as collections of texts. Corpus linguistics proper, therefore, should be seen as a subset of the activity within an empirical approach to linguistics: corpus linguistics necessarily entails an empirical approach, but empirical linguistics need not entail the use of a corpus.

In the sections which follow, we consider the roles which corpora may play in a number of different fields of study in which language is the central object. In these brief discussions we focus on the conceptual issues of why corpus data are important in particular areas and how they can contribute to the advancement of knowledge in those areas, also providing real examples of the use of corpora in each area. In view of the huge amount of corpus-based linguistic research, the examples given are necessarily selective and the student should

consult the further reading for additional examples.

4.2. CORPORA IN SPEECH RESEARCH

The basic importance of corpora in speech research revolves around two main features. The first is that the corpus provides a *broad sample* of speech, that is, one which extends over a wide selection of variables (such as speaker age, sex and class) and across a variety of genres or activity types (such as conversation, news reading, poetry reading, liturgy, legal proceedings and so on). This breadth of coverage has two benefits. First, taking the corpus as a whole, it means that generalisations about spoken language can be made, which simply would not be possible with more restricted samples: in the case of the corpus, one is looking at features across as wide and representative a sample of the entire spoken language as it was feasible to collate. Second, by taking the constituent sections of the corpus either individually or together in the form of smaller subcorpora, it is possible to use the texts representing various variables, genres and activity types to study speech *variation* within a language.

The second important benefit of a spoken corpus is that – with a few exceptions – it provides a sample of *naturalistic* speech rather than speech which has been elicited under artificial conditions. The naturalistic nature of such data means that findings from the corpus are more likely to reflect language as it is actually used in 'real life', since these data are much less likely to be subject to additional production monitoring by the speaker, for example, trying to suppress a regional/social accent.[1]

On a more purely practical level, because the (transcribed) corpus has usually been enhanced with prosodic and other annotations, it is easier to carry out large scale quantitative analyses than with fresh raw data. Where the data have been enriched with other annotations such as parts of speech and syntactic structure, it is also possible to bring these into relationship with the phonetic annotations to study the interrelationships of linguistic levels. Furthermore, it is increasingly the case that the actual spoken data are also made available in high-quality form, so that instrumental analyses of the corpus may be carried out.

At least partly as a result of the fact that much phonetic corpus annotation has been at the level of prosody (see section 2.2.2.3(g)), most phonetic and phonological research using corpora has, as one might expect, tended to focus upon issues of prosody rather than upon other levels of speech. The work in these areas which has been carried out to date on spoken corpora can be broadly divided into three types.

The first type of work has used the corpus to look at the nature of prosody and how the prosodic elements of speech relate to other linguistic levels. In the past many theories have been adduced about how such constructs as the boundaries of intonation groups are motivated. The use of a spoken corpus enables the researcher either (1) quantitatively to test out such hypotheses

onreal data to see whether or not they appear to hold or (2) to generate hypotheses inductively from the corpus which may then be tested on further data.

An example of the first of these research paradigms is Wilson's (1989) study of prepositional phrases and intonation group boundaries. He hypothesised that postmodification of a noun by a prepositional phrase (e.g. *the man with the telescope*) would constitute a barrier to the placing of an intonation group boundary between the head noun and the preposition, since the prepositional phrase forms part of a larger noun phrase: an intonation group boundary would on the other hand be more likely to occur between a verb and a preposition where a prepositional phrase functions as an adverbial (e.g. *She ran with great speed*). These hypotheses were tested on a subsample of the Lancaster/IBM Spoken English Corpus and were found generally to hold, suggesting that there is indeed a relationship between the syntactic cohesiveness of a phrase and the likelihood of a prosodic boundary.

An example of the second paradigm is the work of Altenberg (1990), also on intonation group boundaries. Unlike Wilson, Altenberg did not start off with a hypothesis but instead generated a detailed account of the relationships between intonation group boundaries and syntactic structures from a monologue from the London-Lund corpus. From the results of this analysis he devised a set of rules for predicting the location of such boundaries, which were then applied by a computer program to a sample text from outside the corpus (a text from the Brown written corpus). When the sample text was read aloud, the predictions were found to identify correctly 90 per cent of the actual intonation group boundaries. In Altenberg's case, therefore, the research progressed from *analysing corpus data* to the generation of hypotheses to the testing of the hypotheses on more corpus data, rather than progressing from *theory* to the generation and testing of hypotheses.

A second type of work has looked at the basis of the prosodic transcriptions which are typically encoded within spoken corpora and used by researchers. Prosodic transcription raises the question of how far what is perceived and transcribed relates to the actual acoustic reality of the speech. Looking at the overlap passages of the Lancaster/IBM Spoken English Corpus, where the same passages were prosodically transcribed independently by two different phoneticians, Wilson (1989) and Knowles (1991) both found significant differences in the perception of intonation group boundaries, which suggested either that individual perception of the phonetic correlates of such boundaries differed or that other factors were affecting the placement of boundaries in the transcription. Wichmann (1993) looked more closely at the differences in the transcription of tones rather than boundaries. Looking at the transcription of falling tones in the corpus, she found that in the overlap passages there were major discrepancies in the perception of such tones. The transcribers seemed to have different notions of pitch height in relation to preceding syllables

which was also sometimes overridden according to the level of a given tone in the speakers' overall pitch range, and the results of a perception experiment by Wichmann suggested that there is in fact no real *perceptual* category of high and low. Such studies seem to suggest, therefore, that, in comparison to other forms of annotation such as part of speech, prosodic annotation is a much less reliable guide, at least to what it claims to depict.

The third type of work with speech corpora has looked at the typology of texts from a prosodic perspective. A good example of this is Wichmann's (1989) prosodic analysis of two activity types in the Lancaster/IBM Spoken English Corpus – poetry reading and liturgy. Considering Crystal and Davy's (1969) suggestion that a high frequency of level tones is especially characteristic of liturgy, she made a count of the distribution of level tones in all the text categories in the corpus. This count showed that, whilst liturgy did have a high proportion of level tones, this was not markedly the case and in fact the highest number of level tones was to be found in poetry reading. Looking in more detail at poetry reading and liturgy, Wichmann found that in the liturgical passages the highest concentration of level tones was in the prayer, whilst in the poetry reading the level tones tended to cluster in a final lyrical section of the poem which was included in the corpus. Wichmann suggests that, in the context of the prayer reading, the listener may be assumed to constitute an audience rather than the addressee (which is God), whilst, in the case of the lyric poetry, the reading is more of a performance than an act of informing. In contrast, the narrative section of the same poem could be considered to be an act of informing and this in fact showed a much more conversational typology of tones. Wichmann links these observations about the nature of the speaker/hearer roles to the prosodic patterns which were discovered and, on the basis of these results, argues that, contrary to the generalisation proposed by Crystal and Davy, the intonation patterns are not related to activity type (such as liturgy) but rather to the discourse roles of the hearer such as audience and addressee. In this study, therefore, we see clearly how corpus data can be of value in challenging and amending existing theories.

4.3. CORPORA IN LEXICAL STUDIES

Lexicographers made use of empirical data long before the discipline of corpus linguistics was invented. Samuel Johnson, for example, illustrated his dictionary with examples from literature, and in the nineteenth century the *Oxford English Dictionary* made use of citation slips to study and illustrate word usage. The practice of citation collecting still continues, but corpora have changed the way in which lexicographers – and other linguists interested in the lexicon – can look at language. (A detailed account of the use of corpora in lexicography is provided by Ooi (1998).)

Corpora, and other (non-representative) collections of machine readable text, now mean that the lexicographer can sit at a computer terminal and call

up all the examples of the usage of a word or phrase from many millions of words of text in a few seconds. This means not only that dictionaries can be produced and revised much more quickly than before – thus providing more up-to-date information about the language – but also that the definitions can (hopefully) be more complete and precise, since a larger sample of natural examples is being examined. To illustrate the benefits of corpus data in lexicography we may cite briefly one of the findings from Atkins and Levin's (1995) study of verbs in the semantic class of 'shake'. In their paper, they quote the definitions of these verbs from three dictionaries – *the Longman Dictionary of Contemporary English* (1987, 2nd ed.), the *Oxford Advanced Learner's Dictionary* (1989, 4th ed.) and the *Collins COBUILD Dictionary* (1987, 1st ed.). Let us look at an aspect of just two of the entries they discuss – those for *quake* and *quiver.* Both the Longman and COBUILD dictionaries list these verbs as being solely intransitive, that is they never take a direct object; the Oxford dictionary similarly lists *quake* as intransitive only, but lists *quiver* as being also transitive, that is, it can sometimes take an object. However, looking at the occurrences of these verbs in a corpus of some 50,000,000 words, Atkins and Levin were able to discover examples of both *quiver* and *quake* in transitive constructions (for example, *It quaked her bowels*; *quivering its wings*). In other words, the dictionaries had got it wrong: both these verbs can be transitive as well as intransitive. This small example thus shows clearly how a sufficiently large and representative corpus can supplement or refute the lexicographer's intuitions and provide information which will in future result in more accurate dictionary entries.

The examples extracted from corpora may also be organised easily into more meaningful groups for analysis, for instance, by sorting the right-hand context of a word alphabetically so that it is possible to see all instances of a particular collocate together. Furthermore, the corpora being used by lexicographers increasingly contain a rich amount of textual information – the Longman-Lancaster corpus, for example, contains details of regional variety, author gender, date and genre – and also linguistic annotations, typically part-of-speech tagging. The ability to retrieve and sort information according to these variables means that it is easier (in the case of part-of-speech tagging) to specify which classes of a homograph the lexicographer wants to examine and (in the case of textual information) to tie down usages as being typical of particular regional varieties, genres and so on.

It is in dictionary building that the concept of an open-ended monitor corpus, which we encountered in Chapter 2, has its greatest role, since it enables the lexicographer to keep on top of new words entering the language or existing words changing their meanings or the balance of their use according to genre, formality and so on. But the finite sampled type of corpus also has an important role in lexical studies and this is in the area of quantification. Although frequency counts, such as those of Thorndike and Lorge (1944),

predate modern corpus linguistic methodologies, they were based on smaller, less representative corpora than are available today and it is now possible to produce frequencies more rapidly and more reliably than before and also to subdivide these across various dimensions, according to the varieties of a language in which a word is used. More frequency information is already beginning to appear in published dictionaries. The third edition of the Longman Dictionary of Contemporary English (1995), for example, indicates whether a word occurs amongst the 1,000, 2,000 or 3,000 most frequent words in spoken and written English respectively (see also Summers 1996). Moreover, frequency data need not apply solely to word forms: West (1953), working with a number of human analysts and an early, non-machine-readable corpus, produced a dictionary of word *sense* frequencies which has not yet been superseded. There is now an increasing interest in the frequency analysis of word senses and it should only be a matter of time before corpora are being used to provide word sense frequencies in dictionaries. Indeed, a project has been underway for over a decade at the Christian-Albrechts-Universität Kiel and Bowling Green State University in Ohio, which aims ultimately at the production of a broadly sense-ordered frequency dictionary for medieval German epic (Schmidt 1991): although based on the accumulation of analyses of individual literary texts rather than on a sampled corpus, this project is representative of the kinds of information which are increasingly becoming an issue in lexicography.

The ability to call up word combinations rather than individual words and the existence of tools, such as mutual information for establishing relationships between co-occurring words (see section 3.4.4), mean that it is also now feasible to treat phrases and collocations more systematically than was previously possible. For instance, again in the third edition of the Longman Dictionary of Contemporary English, bar charts are now provided showing the collocational preferences of certain key words. The verb *try*, for example, has a chart showing the frequencies of different grammatical patterns: try to do something, try something, try and do something, try doing something and so on. Word co-occurrences are important for a number of reasons, as we have already suggested in Chapter 3. For instance, a phraseological unit may constitute a piece of technical terminology or an idiom, and collocations are important clues to specific word senses. Techniques for identifying such combinations in text corpora mean that, like individual words, they can now be better treated in dictionaries and in machine-readable terminology banks for professional technical translators. The experiments of Church *et al.* (1991) and Biber (1993a), which we have already discussed in Chapter 3, are clear examples of how such technology may benefit lexicographers in crafting definitions from corpus data. Biber's study, for instance, shows how factor analysis of frequent collocates can help empirically to group together individual senses of words: this may enable the lexicographer to organise his or her concordance data more swiftly and usefully into sense groups, with the possible additional pay-off of being

able more easily to provide the sorts of sense frequency data that we discussed in the previous paragraph. Church *et al.*'s study shows a further way in which co-occurrence data can perhaps be used, that is to add greater delicacy to definitions: *strong* and *powerful*, for example, are often treated in dictionaries almost as synonyms, but the identification of differences in collocation can enable the lexicographer to draw these important distinctions in their usage as well as identifying their broad similarity in meaning.

As well as word meaning, we may also consider under the heading of lexical studies corpus-based work on **morphology** (word structure). The fact that morphology deals with language structure at the level of the word may suggest that corpora do not have any great advantage here over other sources of data such as existing dictionaries or introspection. However, corpus data do have an important role to play in studying the frequencies of different morphological variants and the productivity of different morphemes. Opdahl (1991), for example, has used the LOB and Brown corpora to study the use of adverbs which may or may not have a *-ly* suffix (e.g. *low/lowly*), finding that the forms with the *-ly* suffix are more common than the 'simple' forms and that, contrary to previous claims, the 'simple' forms are somewhat less common in American than in British English. Bauer (1993) has also begun to use data from his corpus of New Zealand English for morphological analysis. At the time that he wrote his paper his corpus was incomplete and so his results are suggestive rather than definitive, but they demonstrate the role which a corpus can play in morphology. One example which Bauer concentrates on is the use of strong and weak past tense forms of verbs (e.g. *spoilt* (strong) vs. *spoiled* (weak)). In a previous elicitation study amongst New Zealand students, Bauer had concluded that the strong form was preferred by respondents to the weak form, with the exceptions of *dreamed* and *leaned*. The written corpus data, on the other hand, suggested that the weak form, with the exception of *lit*, was preferred to a greater degree than the elicitation experiment had suggested. Bauer wonders how far this difference between the elicitation experiment and the texts of the written corpus may be due to editorial pressure on writers to follow the more regular spelling variant, a non-linguistic factor which was not present in his elicitation experiment, and he looks forward to testing this theory in relation to the New Zealand spoken corpus, which also lacks this editorial constraint. Here, then, we see how a corpus, being naturalistic data, can help to define more clearly which forms are most frequently used and begin to suggest reasons why this may be so.

4.4. CORPORA AND GRAMMAR

Grammatical (or syntactic) studies have, along with lexical studies, been the most frequent types of research which have used corpora. What makes corpora important for syntactic research is, first, their potential for the representative quantification of the grammar of a whole language variety and, second, their

role as empirical data, also quantifiable and representative, for the testing of hypotheses derived from grammatical theory.

Until the last quarter of the twentieth century, the empirical study of grammar had to rely primarily upon qualitative analysis. Such work was able to provide detailed descriptions of grammar but was largely unable to go beyond subjective judgements of frequency or rarity. This is even the case with more recent classic grammars such as the *Comprehensive Grammar of the English Language* (Quirk *et al.* 1985), whose four authors are all well-known corpus linguists. But advances in the development of parsed corpora (see Chapter 2) and tools for retrieval from them mean that quantitative analyses of grammar may now more easily be carried out. Such studies are important, because they can now at last provide us with a representative picture of which usages are most typical and to what degree variation occurs both within and across varieties. This in turn is important not only for our understanding of the grammar of the language itself but also in studies of different kinds of linguistic variation and in language teaching (see sections 4.7, 4.8, 4.9 and 4.11).

Most smaller-scale studies of grammar using corpora have included quantitative data analyses. Schmied's (1993) study of relative clauses, for example, provides quantitative information about many aspects of the relative clauses in the LOB and Kolhapur corpora. However, there is now also a greater interest in the more systematic treatment of grammatical frequency and at least one current project (Oostdijk and de Haan 1994a) is aiming to analyse the frequency of the various English clause types. Oostdijk and de Haan have already produced preliminary results based upon the syntactically parsed Nijmegen corpus and they plan to extend this work in the near future to larger corpora. The Nijmegen corpus is only a small corpus of some 130,000 words, but, with the completion of the British National Corpus and the International Corpus of English, the stage seems set for much more intensive treatments of grammatical frequency.

As explained in Chapter 1, there has since the 1950s been a division in linguistics between those who have taken a largely rationalist view of linguistic *theory* and those who have carried on *descriptive* empirical research with a view to accounting fully for all the data in a corpus. Often these approaches have been presented as competitors but they are in fact not always as mutually exclusive as some would wish to claim: there is a further, though not at present very large, group of researchers who have harnessed the use of corpora to the *testing* of essentially rationalist grammatical theory rather than to pure linguistic description or the inductive generation of theory.

One example of this kind of rationalist-to-empiricist approach to grammar is provided by the exchange of papers between the team of Taylor, Grover and Briscoe, and Geoffrey Sampson. Taylor, Grover and Briscoe (1989) had produced an automatic **parser** for English in the form of a generative grammar, not directly based on empirical data, which they wanted to test on corpus

data. Independently, Sampson (1987) had manually analysed a set of noun phrases in the LOB corpus, concluding that a generative grammar could not be used successfully to parse natural English text, since the number of different constructions occuring in natural texts is so large as to make it impossible to account for them all in a set of grammatical rules. Nevertheless, Taylor, Grover and Briscoe used their grammar to analyse a superset of Sampson's data and obtained a success rate of 87.97 per cent of noun phrase types correctly parsed. Extrapolating this to the number of tokens of each type in the data, their success rate would have amounted to 96.88 per cent. If Sampson were right about the deficiencies of generative grammars, they argued, the remaining examples would be expected primarily to be single, syntactically odd types. However, the failures of the grammar could in fact be classified quite easily and tended to represent oversights in devising the set of grammar rules rather than syntactic oddities. Taylor, Grover and Briscoe were thus able to argue that, at least as far as noun phrases go, Sampson had considerably overstated the degree to which a generative grammar cannot account for the constructions to be found in natural texts.[2] The matter remains controversial, however: Sampson (1992) continues to argue that the rule set for accurate parsing is too open-ended and irregular to be practicable.

One team whose work, in part, addresses this problem is that led by Jan Aarts at Nijmegen University. Here, rationalist and empiricist approaches to grammar are combined within a paradigm based upon the development of primarily rationalist formal grammars and the subsequent testing of these grammars on real-life language contained in computer corpora (Aarts 1991). A formal grammar is first devised by reference both to the linguists' introspection and to existing accounts of the grammar of the language, in a similar manner to the research practices employed by rationalist grammarians such as Chomsky. However, in contrast to such purely rationalist grammatical research, this grammar is then loaded into a computer parser and is run over a corpus to test how far it accounts for the data within the corpus. (See Chapter 5 for a discussion of corpus based parsing technologies.) On the basis of the results of this corpus parsing experiment, the grammar is then modified to take account of those analyses which it missed or got wrong. This does not mean amending it to deal with every single instance in the corpus, just the most important or frequent constructions. By proceeding in this way, it is possible to investigate the degree to which basically rationalist grammars can account for corpus data and how far they need to be amended to handle the data.

One final note may be added to our brief consideration of corpora and grammar and that is a case in which corpus-based approaches to grammar have actually dovetailed with grammatical theory. Michael Halliday's theory of systemic grammar is based upon the notion of language as a paradigmatic system, that is, as a set of choices for each instance from which a speaker must select one. Such a set of choices is inherently probabilistic, that is to say in each

situation various choices are more or less likely to be selected by the speaker. Halliday (1991) uses this idea of a probabilistically ordered choice to interpret many aspects of linguistic variation and change in terms of the differing probabilities of linguistic systems. For example, it may be that written English prefers *which* to *that* as a relativiser. We might quantify this statement using a corpus and say that *which* is 39% probable as opposed to *that* which is 12% probable.[3] But it may be that, in contrast to writing, conversational speech shows a greater tendency towards the use of *that*, so that *which* is only 29% probable whereas *that* is 18% probable.[4] It is one of Halliday's suggestions that the notion of a **register**, such as that of conversational speech, is really equivalent to a set of these kinds of variations in the probabilities of the grammar. Halliday is enthusiastic about the role which corpora may play in testing and developing this theory further, in that they can provide hard data from which the frequency profiles of different register systems may be reconstructed. Here, then, in contrast to the frequent hostility between grammatical theory and corpus analysis, we see a theoretician actively advocating the use of corpus data to develop a theory.

4.5. CORPORA AND SEMANTICS

We have already seen how a corpus may be used to look at the occurrences of individual words in order to determine their meanings (*lexical semantics*), primarily in the context of lexicography. But corpora are also important more generally in semantics. Here, their main contribution has been that they have been influential in establishing an approach to semantics which is objective and which takes due account of indeterminacy and gradience.

The first important role of the corpus, as demonstrated by Mindt (1991), is that it can be used to *provide objective criteria for assigning meanings to linguistic items*. Mindt points out that most frequently in semantics the meanings of lexical items and linguistic constructions are described by reference to the linguist's own intuitions, that is, by what we have identified as a rationalist approach. However, he goes on to argue that in fact semantic distinctions are associated in texts with characteristic observable contexts – syntactic, morphological and prosodic – and thus that, by considering the environments of the linguistic entities, an empirical objective indicator for a particular semantic distinction can be arrived at. Mindt presents three short studies in semantics to back up this claim. By way of illustration, let us take just one of these – that of 'specification' and futurity. Mindt is interested here in whether what we know about the inherent futurity of verb constructions denoting future time can be shown to be supported by empirical evidence; in particular, how far does the sense of futurity appear to be dependent on co-occurring adverbial items which provide separate indications of time (what he terms 'specification') and how far does it appear to be independently present in the verb construction itself? Mindt looked at four temporal constructions – namely, *will*, *be going to*,

the present progressive and the simple present – in two corpora – the Corpus of English Conversation and a corpus of twelve contemporary plays – and examined the frequency of specification with the four different constructions. He found in both corpora that the simple present had the highest frequency of specification, followed in order by the present progressive, *will* and *be going to*. The frequency analysis thus established a hierarchy with the two present tense constructions at one end of the scale, often modified adverbially to intensify the sense of future time, and the two inherently future-oriented constructions at the other end, with a much lesser incidence of additional co-occurring words indicating futurity. Here, therefore, Mindt was able to demonstrate that the empirical analysis of linguistic contexts *is* able to provide objective indicators for intuitive semantic distinctions: in this example, inherent futurity was shown to be inversely correlated with the frequency of specification.

The second major role of corpora in semantics has been in establishing more firmly the notions of *fuzzy categories* and *gradience*. In theoretical linguistics, categories have typically been envisaged as hard and fast ones, that is, an item either belongs in a category or it does not. However, psychological work on categorisation has suggested that cognitive categories typically are not hard and fast ones but instead have fuzzy boundaries so that it is not so much a question of whether or not a given item belongs in a particular category as of how often it falls into that category as opposed to another one. This has important implications for our understanding of how language operates: for instance, it suggests that probabilistically motivated choices of ways of putting things play a far greater role than a model of language based upon hard and fast categories would suggest. In looking empirically at natural language in corpora it becomes clear that this 'fuzzy' model accounts better for the data: there are often no clear-cut category boundaries but rather gradients of membership which are connected with *frequency* of inclusion rather than simple inclusion or exclusion. Corpora are invaluable in determining the existence and scale of such gradients. To demonstrate this, let us take a second case study from Mindt. In this instance, Mindt was interested in the subjects of verb constructions with future time reference, specifically the distinction between subjects that do or do not involve conscious human agency, which theory had previously identified as an important distinction. As a rough correlate of this distinction, Mindt counted the frequency of personal and non-personal subjects of the four future time constructions in the same two corpora referred to above. He found that personal subjects occurred most frequently with the present progressive, whereas the lowest number of personal subjects occurred with the simple future, refuting previous theoretical claims. *Will* and *be going to* had only a small preference (just 2–3%) for personal subjects, and the rank order was the same for both corpora. So this case study seemed to suggest that there is a semantic relationship correlating the type of agency with the verb form used for future time reference. But note that none of the constructions occurred solely with

either conscious or non-conscious agency: rather, Mindt's analysis showed that the present progressive was simply *more likely* to have a personal subject than the other constructions and especially the simple future. In other words, the data formed a gradient of membership for the two fuzzy sets of verb constructions related to conscious agency and to non-conscious agency, on which some constructions are seen to belong more often than others in one of the two sets. What is important to recognise here is that the definition of this gradience of membership was only possible using corpus-based frequency data: a purely theoretical linguistic approach might conceivably not have recognised the fuzziness and indeterminacy present in such circumstances and instead might well have attempted, wrongly, to shoehorn a given construction into just one clear-cut category.

4.6. CORPORA IN PRAGMATICS AND DISCOURSE ANALYSIS

Closely related to pragmatics, spoken language analysis and sociolinguistics is discourse analysis.

Discourse analysis is another area where the 'standard' corpora (such as the BNC) have been relatively little used. This is largely because discourse analysts are normally interested in looking at the discursive practices associated with particular social practices (cf. Fairclough 1993: 226); hence, although corpora such as the BNC may contain some relevant material (e.g. news reporting on a given topic or event), the likelihood of this is rather haphazard and the corpus is unlikely to contain sufficient relevant data for a given project. Nevertheless, there *are* important points of contact between corpus linguistics and discourse analysis.

First, the application of computer-aided corpus analysis methods has a considerable history within certain traditions of discourse analysis. For example, as early as the 1970s Michel Pécheux was making use of a rather sophisticated form of automatic parsing as an adjunct to a Marxist theory of discourse: an important part of his methodology involved transforming the sentences in a corpus into sets of simpler structures and then using distributional procedures to look for repeated patterns of equivalence and substitution. (See Thompson 1984: 238–47, for more details on this work.) Also in France, there is a long history, associated with a group of scholars at the ENS St Cloud, of using word-frequency statistics and multivariate statistical analyses (so-called *lexicometry*) to examine the discourse of political texts (Bonnafous and Tournier 1995). And, to give one final example, in English linguistics there has been an increase of interest in using computer tools for collocation analysis and concordancing as an aid to discourse analysis (cf. Hardt-Mautner 1995). So, methodologically, there is a strong tradition of symbiosis between the two disciplines.

Second, although little used in this way at present, the standard corpora have important potential in discourse analysis as control data. If a discourse analyst

finds a set of features to be important in a given sample of texts, it may be reasonable to ask whether these features are actually tied to the specific social practices concerned or whether they arise through more general social practices – that is, those that lead to the formation of genres. The standard corpora, since they contain a wide variety of genres and text types, may be used alongside specialist discourse analysis corpora in order to discover how far certain features are distinctive of the discourse under examination and how far they occur elsewhere in the language as a whole. For instance, Myers (1991b) looked at cohesion devices in a sample of professional and popular articles on molecular genetics. He found that the two text types tended to use very different kinds of devices: professional articles made much more use of lexical cohesion than did the popularisations. Using a standard corpus, this work could be extended to see how far the findings are unique to biological science (or laboratory science in general) and how far they could be extended to other areas such as history or music; in other words, is this a more general feature of professional versus popular discourse or does it tell us something more specific about particular kinds of science?

4.7. CORPORA AND SOCIOLINGUISTICS

Sociolinguistics shares with historical linguistics, dialectology and stylistics the fact that it is an empirical field of research which has hitherto relied primarily upon the collection of research-specific data rather than upon more general corpora. But such data are often not intended for quantitative study and hence also are often not rigorously sampled. Sometimes the data are also elicited rather than naturalistic data. What is important about a corpus is that it can provide what these kinds of data do not provide – a representative sample of naturalistic data which can be quantified. Although corpora have not as yet been used to a great extent in sociolinguistics, there is evidence of an increasing interest in their exploitation in this field.

Most sociolinguistic projects with corpora to date have been relatively simple lexical studies in the area of language and gender. An example of this is the study by Kjellmer (1986), who used the Brown and LOB corpora to examine masculine bias in American and British English. He looked specifically at the occurrence of masculine and feminine pronouns and at the occurrence of the lexical items *man/men* and *woman/women*. Kjellmer found that the frequency of the female items was much lower than the male in both corpora, but that female forms were more frequent in British English than in American English. He also found differences in male/female ratios by genre. In general, women featured more strongly in imaginative rather than informative prose, with romantic fiction, unsurprisingly, having the highest proportion of women. However, although these frequency distributions were unsurprising, Kjellmer found that his other hypothesis – that women would be less 'active', that is, would be more frequently the objects rather than the subjects of verbs – to be

falsified: in fact women and men had similar subject/object ratios.

Holmes (1994) has looked more critically at the methodology of these kinds of studies. Taking three gender-related lexical issues – the frequency of *Ms* as compared with *Miss/Mrs*; the use of 'sexist' suffixes; and the use of generic *man* – Holmes makes two important methodological points. First, she shows that it is important when counting and classifying occurrences to pay attention to the context and whether real alternatives are in fact available. For instance, whilst there is a non-gender marked alternative for *policeman* or *policewoman* (namely *police officer*), there is no such alternative for the *-ess* form in *Duchess of York*. The latter should therefore be excluded from counts of 'sexist' suffixes when looking at this kind of gender bias in writing. Second, Holmes points out the difficulty of classifying a form when it is actively undergoing semantic change. As an example she takes the word *man*. She argues that, whilst it has clearly a single male referent in a phrase such as *A 35 year old man was killed* and is clearly generic (referring to mankind) in phrases such as *Man has engaged in warfare for centuries*, it is difficult to decide in phrases such as *We need the right man for the job* whether *man* is intended to refer solely to a male person or whether it is not gender specific and could reasonably be substituted by *person*. Because of the drift towards 'non-sexist' writing, it becomes harder to decide whether or not the usage is generic, because an alternative is available which was not previously available. These simple but important criticisms by Holmes should incite a more critical approach to data classification in further sociolinguistic work using corpora, both within and without the area of gender studies.

Although, as stated, relatively little work on sociolinguistics has hitherto been carried out using the standard corpora, it seems likely that it will increase in quantity in the near future. Yates (1993), for example, has shown how quantifiable measures can be generated from the Hallidayan theory of language as social semiotic and has carried out studies using a body of computer-mediated communication (e.g. e-mail) to examine the nature of this new genre in a sociolinguistic perspective, focusing on issues such as the presentation of self, literacy practices and the presentation of knowledge. The greater expansion of sociolinguistic corpus work seems to be hampered really by only three practical problems: the operationalisation of sociolinguistic theory into measurable categories which can be applied to corpora; the absence of sociolinguistic information encoded in current corpora; and the lack of sociolinguistically motivated sampling. Yates's work is an example of how the first problem may be addressed. The situation as regards the nature of the corpora is also changing. In historical corpora, sociolinguistic variables such as social class, sex of writer and educational background are now being encoded. The Helsinki diachronic corpus and the Lampeter Corpus of Early Modern English Tracts both encode data such as these. With modern language corpora, the Longman-Lancaster corpus already contains header fields for gender of writer and,

furthermore, the spoken part of the BNC has been collected using demographic market research techniques for age and social class as well as geographical location. These sociolinguistically annotated corpora should enable corpus-based work on social variation in language to begin in earnest.

4.8. CORPORA IN STYLISTICS AND TEXT LINGUISTICS

Researchers in stylistics are typically more interested in individual texts or the *oeuvres* of individual authors than in more general varieties of a language and, hence, although they may be interested in computer assisted textual analysis, they tend not to be large scale users of corpora as we have defined them. Nevertheless, some stylisticians are interested in investigating broader issues such as genre, and still others have found corpora to be important sources of data in their research.

The concept of style is based on the assumption that authors have a choice between different ways of putting things, for instance, between using technical or non-technical vocabulary, between using long and short sentences or between using co-ordination and subordination. The definition of what constitutes an author's individual style thus depends at least in part in examining the degree by which the author leans towards particular ways of putting things. To arrive at a picture of how far this constitutes a stylistic trend requires comparisons to be made not only internally within the author's own work but also with other authors or the norms of the language or variety as a whole. Furthermore, as Leech and Short (1981) point out in their monograph on style in fiction, stylistics often demands the use of quantification to back up such judgements which may otherwise appear to be subjective rather than objective.

Corpora, as standard representative samples of varieties or languages, form a particularly valuable basis for comparisons between an author, text or collection of texts and a particular variety of a language. One example of the use of a corpus for this purpose is the Augustan Prose Sample collected by Milic. Milic's primary aim in collecting this corpus was to have a basis for a quantitative comparison of Jonathan Swift's prose style. The corpus contains 80,000 words, which were sampled from various genres of published texts. Texts were sampled so as to give as regular as possible a distribution of samples from throughout the period 1675–1725. Also, rather than concentrating on the best-known writers of the period, Milic chose texts which could be taken to represent a broad range of the sorts of things which educated people were reading at this time. The intention was therefore to arrive at a representative sample of the published English of this period so as to have a normative sample with which to compare Swift's style.

Another level of stylistic variation, which may often not explicitly be called 'stylistics', is the more general variation between genres and channels. Corpora have found a particular role in examining the stylistics of these more general constructs. One of the most common uses of corpora has been in looking at

the differences between spoken language and written language. For instance, Altenberg (1984) has examined the differences in the ordering of cause–result constructions and Tottie (1991) has examined the differences in negation strategies. Other work has looked at variation between particular genres, using subsamples of corpora as the database. Wilson (1992b), for example, looking at the usage of *since*, used the learned, Belles-Lettres and fiction genre sections from the LOB and Kolhapur corpora, in conjunction with a sample of modern English conversation and the Augustan Prose Sample, and found that causal *since* had evolved from being the main causal connective in late seventeenth century writing to being particularly characteristic of formal learned writing in the twentieth century. A project has also begun at Lancaster University to build a speech presentation corpus (Leech, McEnery and Wynne 1997: 94–100). Speech presentation (i.e., how spoken language is represented in written texts) is an area of stylistics to which much attention has been given. The speech presentation corpus will contain a broad sample of direct and indirect speech from a variety of genres and allow researchers to look for systematic differences between, for example, fictional and non-fictional prose.

Allied to their use for comparing genres, corpora have been used to challenge, empirically, existing approaches to text typology. Variation studies, and also the sampling of texts for corpora, have typically been based on external criteria such as channel (e.g., speech and writing) or genre (e.g., romantic fiction, scientific writing). However, there is now a large body of work that addresses textual variation from a language internal perspective. Biber (1988), for example, looking initially at the variation between speech and writing, carried out factor analyses of 67 linguistic features across text samples from 23 major genre categories taken mostly from the LOB and London-Lund corpora. From these analyses, Biber extracted 5 factors which, by reference to the highest loaded items on each, he interpreted as representing particular dimensions of linguistic variation. For instance, Biber's Factor 2 is taken to represent a distinction between narrative and non-narrative: past tense verbs, third person pronouns and verbs with perfective aspect receive high positive loadings on this factor. In the factor analysis, each genre sample also received a factor score on each dimension so that, taken together, it is possible to see how genres differ from one another and on what dimensions. Having once arrived at this 5-factor framework, it is then possible to use it to score other texts. For example, Biber and his collaborators have already applied the framework to the historical development of English genres, to the examination of primary school reading materials and to texts in other languages (cf. Biber 1995). What is important about Biber's work from a methodological point of view is that it enables a broad, empirically motivated comparison of language variation to be made, rather than the narrow single-feature analyses which have often been used to add a little bit at a time to our understanding of variation. It is also a very clear example of how the quantitative empirical analysis of a finite

sampled corpus may contribute new approaches to old issues.

Kay Wikberg has also in recent years looked critically at the genre classifications in corpora. One of the issues with which Wikberg has been concerned has been with the non-homogeneous nature of some corpus genre categories. Looking at category E of the LOB corpus ('skills and hobbies'), for example, Wikberg (1992) found that the distributions of certain high frequency words, together with other textual features, supported a division of this category into two subcategories of procedural discourse (that is, discourse which describes to the reader how to do something – for example, a recipe or instruction manual) and non-procedural discourse. These findings have prompted Wikberg to suggest that corpus compilers should pay greater attention to text typology in constructing corpora and that users should pay more attention to the possibilities of intra-category variation in analysing the results of corpus-based investigations.

Work such as that of Biber and Wikberg has two important roles to play. First, it provides an incentive to stylistic analysis which is not only empirical and quantitative but which also takes greater account of the general stylistic similarities and differences of genres and channels rather than how they differ on individual features. Second, it should be influential in developing more representative corpora. Culturally-motivated and, hence, possibly artificial notions of how language is divided into genres can be replaced or supplemented by more objective language-internal perspectives on how and where linguistic variation occurs. Indeed, in this latter context it should be noted that Biber and Finegan have applied Biber's multidimensional model to the building of their ARCHER historical corpus.

Text typology is in fact a cross-over area between stylistics and text linguistics – that discipline which is concerned with the structure of texts above the level of the sentence. It is thus worth noting briefly some of the other text-linguistic work that has been done with corpora.

The fact that most corpora do not contain whole texts does not facilitate their use for the study of internal text structuring. However, two important areas of corpus-based text-linguistic research have been text segmentation and the study of anaphoric reference.

4.9. CORPORA IN THE TEACHING OF LANGUAGES AND LINGUISTICS[5]

Resources and practices in the teaching of languages and linguistics typically reflect the division in linguistics more generally between empirical and rationalist approaches. Many textbooks contain only invented examples and their descriptions are based apparently upon intuition or second-hand accounts; other books – such as the books produced by the Collins COBUILD project – are explicitly empirical and rely for their examples and descriptions upon corpora or other sources of real life language data.

Corpus examples are important in language learning as they expose students at an early stage in the learning process to the kinds of sentences and vocabulary which they will encounter in reading genuine texts in the language or in using the language in real communicative situations. The importance of such empirical data also applies as much in the teaching of linguistics as it does in the teaching of foreign languages. In our own teaching, we have found that students who have been taught using traditional syntax textbooks, which contain only simple example sentences such as *Steve puts his money in the bank* (from Horrocks 1987), often find themselves unable to analyse longer, more complex corpus sentences such as *The government has welcomed a report by an Australian royal commission on the effects of Britain's atomic bomb testing programme in the Australian desert in the fifties and early sixties* (from the Spoken English Corpus). It is this latter kind of sentence, however, which a prospective linguist would need to be able to analyse since such sentences reflect some of the sorts of language which he or she would encounter in analysing other texts.

However, corpora, much more so than other sources of empirical data, have another important role in language pedagogy which goes beyond simply providing more realistic examples of language usage. A number of scholars have used corpus data to look critically at existing language teaching materials. For example, Kennedy (1987a, 1987b) has looked at the ways of expressing quantification and frequency in ESL (English as a second language) textbooks; Holmes (1988) has looked at ways of expressing doubt and certainty in ESL textbooks; Mindt (1992) has looked at future time expressions in German textbooks of English; and Ljung (1990) has looked at the vocabulary in Swedish textbooks of English. The methodologies adopted by these scholars are broadly similar: they analyse the relevant constructions or vocabularies both in the sample textbooks and in standard corpora of English such as the LOB corpus and the London-Lund corpus, then they compare their findings between the two data sets. Most of these studies have found that there exist considerable differences between what textbooks are teaching and how native speakers actually use language as evidenced in the corpora. Some textbooks have been found to gloss over important aspects of usage or variations in usage, and sometimes textbooks may even foreground less frequent stylistic choices at the expense of more common ones. The more general conclusion which scholars such as Mindt and Kennedy have drawn from these exercises is that non-empirically based teaching materials can be positively misleading and that corpus studies should be used to inform the production of materials, so that the more common choices of usage are given more attention than those which are less common. This is the kind of approach which the COBUILD team at Birmingham have adopted in producing their materials (e.g., Sinclair 1987).

One particular type of foreign language teaching is that which comes under the heading of 'languages for special purposes'. This refers to the

domain-specific teaching of a language for a particular area of application, for example, the teaching of medical English to medical students. Following similar logics to those espoused by Mindt and others, some researchers have built corpora of particular varieties of English with the aim of exploiting them in language teaching for specific purposes. One example of such a corpus is the Guangzhou Petroleum English Corpus, a corpus of approximately 411,000 words of English sampled from the petrochemical domain. A similar kind of corpus has been constructed at the Hong Kong University of Science and Technology, where a 1,000,000-word corpus of English has been built, sampled from the kinds of computer science textbooks which students of that subject are likely to use. Such corpora can be used to provide many kinds of domain-specific material for language learning, including quantitative accounts of vocabulary and usage which address the specific needs of students in a particular domain more directly than those taken from more general language corpora.

Corpora have been used not only in language teaching but also in the teaching of linguistics. Excellent examples of such use are the courses on varieties of English and corpus linguistics which John Kirk has run for several years at the Queen's University of Belfast (Kirk 1994a). For the entire student assessment on these courses, Kirk uses projects rather than traditional essays and exams. Kirk requires his students to base their projects on corpus data which they must analyse in the light of a model. The students are then required to draw their own conclusions about the data. Among the models which Kirk's students have chosen to apply to various samples of corpus data are Brown and Levinson's politeness theory, Grice's co-operative principle and Biber's multidimensional approach to linguistic variation. The students must then submit their final project as a properly word-processed research report. In taking this approach to assessment, therefore, Kirk is using corpora not only as a way of teaching students about variation in English but also to introduce them to the main features of a corpus-based approach to linguistic analysis.

One further application of corpora in language and linguistics teaching is their role in computer-assisted language learning. A computer system based upon a parsed corpus database is already being used in the teaching of English grammar at the University of Nijmegen (van Halteren 1994). Recent work at Lancaster University has also been looking at the role of corpus-based computer software for teaching undergraduates the rudiments of grammatical analysis (McEnery and Wilson 1993). This software – Cytor – reads in an annotated corpus (currently either part-of-speech tagged or parsed) a sentence at a time, hides the annotation and asks the student to annotate the sentence him- or herself. The student can call up help in the form of a list of tag mnemonics with examples or in the form of a frequency lexicon entry for a word giving the possible parts of speech with their frequencies. Students can

also call up a concordance of similar examples. Students are given four chances to get an annotation right. The program keeps a record of the number of guesses made on each item and how many were correctly annotated by the student. In the Lent Term of 1994, a preliminary experiment was carried out to determine how effective the corpus-based CALL system was at teaching the English parts of speech. A group of volunteer students taking the first-year English Language course were split randomly into two groups. One group was taught parts of speech in a traditional seminar environment, whilst others were taught using the CALL package. Students' perfomance was monitored throughout the term and a final test administered. In general the computer-taught students performed better than the human-taught students throughout the term, and the difference between the two groups was particularly marked towards the end of the term. Indeed, the performance of CALL students in a final pen and paper annotation test was significantly higher than the group taught by traditional methods (McEnery, Baker and Wilson 1995).

The increasing availability of multilingual parallel corpora makes possible a further pedagogic application of corpora, namely as the basis for translation teaching. Whilst the assessment of translation is frequently a matter of style rather than of right and wrong, and therefore perhaps does not lend itself to purely computer-based tutoring, a multilingual corpus has the advantage of being able to provide side-by-side examples of style and idiom in more than one language and of being able to generate exercises in which students can compare their own translations with an existing professional translation or original. Such an approach is already being pioneered at the University of Bologna using corpora which, although they do not contain the same text in more than one language, do contain texts of a similar genre which can be searched for relevant examples (Zanettin 1994).

Parallel corpora are also beginning to be harnessed for a form of language teaching which focuses especially on the problems that speakers of a given language face when learning another. For example, at Chemnitz University of Technology work is under way on an internet grammar of English aimed particularly at German-speaking learners. An example of the sort of issue that this focused grammar will highlight is **aspect**, an important feature of English grammar but one which is completely missing from the grammar of German. The topic will be introduced on the basis of relatively universal principles (reference time, speech time and event time) and the students will be helped to see how various combinations of these are encoded differently in the two languages. The grammar will make use of a German–English parallel corpus to present the material within an explicitly contrastive framework. The students will also be able to explore grammatical phenomena for themselves in the corpus as well as working with interactive online exercises based upon it (Hahn and Schmied 1998).

4.10. CORPORA IN HISTORICAL LINGUISTICS

Empirically based textual research is a *sine qua non* of historical linguistics. Historical linguistics can also be seen more specifically as a species of corpus linguistics, since the extant texts of a historical period or a 'dead' language form a closed corpus of data which may only be extended by the (re-)discovery of previously unknown manuscripts, books or inscriptions. Indeed, sometimes it is possible to use most or all of the entire closed corpus of a language for research: this can be done, for example, for ancient Greek using the *Thesaurus Linguae Graecae* corpus, which contains most of extant ancient Greek literature. But in practice historical linguistics has not tended to conform strictly to a corpus linguistic paradigm. Given that the entire population of texts can be very large, historical linguistics has tended to take a more selective approach to empirical data, simply looking for evidence of particular phenomena and making, at most, rather rough estimates of frequency. Mostly there has been no real attempt to produce *representative* samples – which is what corpora, as we have defined them, are – and to provide hard frequencies based on those: such frequency analyses have been largely confined to studies of individual literary texts or authors.

However, in recent years there has been a change in the way that some historical linguists have approached their data, which has resulted in an upsurge in strictly corpus-based historical linguistics and the building of corpora for this purpose. The Augustan Prose Sample, referred to above in section 8 on stylistics, is one example of a corpus which aims to represent a particular historical period and variety of the English language, but perhaps the most widely known English historical corpus is the diachronic Helsinki corpus.

The Helsinki corpus is a corpus of approximately 1.6 million words of English dating from the earliest Old English period (before AD 850) to the end of the Early Modern English period (defined by the compilers as being 1710). The corpus provides a continuous diachronic picture of the language between those dates: it is divided into three main periods – Old English, Middle English and Early Modern English – and each of these periods is divided into a number of 100-year subperiods (or 70-year subperiods in the case of Early Modern English and the second half of Middle English). The Helsinki corpus also aims to be representative not only in terms of chronological coverage, but also in terms of the range of genres, regional varieties and sociolinguistic variables such as gender, age, education and social class which are represented. For each text sample, the information on all these criteria is contained in COCOA format at the beginning of the sample to enable the researcher to select information from the corpus according to his or her specific needs. The Helsinki team have also produced 'satellite' corpora of early Scots and early American English to stand alongside the historical 'English' English corpus (cf. Kytö 1991, Meurman-Solin 1993).

But the Helsinki team are not the only team involved in historical corpus

building: indeed, this has almost become a growth industry. A few examples of other English historical corpora that have recently been developed are the Zürich Corpus of English Newspapers (ZEN) (a corpus covering the period from 1660 up to the establishment of *The Times* newspaper), the Lampeter Corpus of Early Modern English Tracts (a sample of English pamphlets from between 1640 and 1740, all taken from the collection at the library of St David's University College, Lampeter) and the ARCHER corpus (a corpus of British and American English between the years 1650 and 1990).

The actual work which is carried out on historical corpora is qualitatively very similar to that which is carried out on modern language corpora, although, in the case of corpora, such as the Helsinki corpus, which provide diachronic coverage rather than a 'snapshot' of the language at a particular point in history, it is also possible to carry out work on the evolution of the language through time. As an example of this latter kind of work, one may take Peitsara's (1993) study of prepositional phrases denoting agency with passive constructions. She made use of four subperiods from the Helsinki corpus covering late Middle and Early Modern English (*c.* 1350–1640) and calculated the frequencies of the different prepositions introducing such agent phrases. The calculation showed that throughout the period the most common prepositions introducing agent phrases were *of* and *by*, but that, from being almost equal in frequency at the very beginning of the period (a ratio of 10.6:9), *by* rapidly gained precedence so that by the fifteenth century it is three times more common than *of* and by 1640 around eight times as common. Peitsara also made use of the text type information, showing that, whilst by the end of the period up to half of the individual texts contained agent phrases introduced by more than one preposition type, some texts showed an unusual tendency to use just one type. This was particularly marked in documents, statutes and official correspondence and it is suggested that this may be a result of bilingual influence from French. Individual authors of texts within categories are also shown to differ in their personal and stylistic preferences.

This kind of quantitative empirical study, by providing concrete data which it is now at last possible to obtain through the use of a computer corpus, can only help our understanding of the evolution of a language and its varieties: indeed, it has a particular importance in the context of Halliday's (1991) conception of language evolution as a motivated change in the probabilities of the grammar. But it is important, as Rissanen (1989) has pointed out, also to be aware of the limitations of historical corpus linguistics. Rissanen identifies three main problems. First, there is what he calls the 'philologist's dilemma', that is the danger that the use of a corpus and computer to extract specific data may supplant the in-depth knowledge of language history which is to be gained from the study of the original texts in their context. This is, however, not a danger inherent in corpus-based research *per se* but in an overreliance on corpora in training researchers. Second, there is the 'God's truth fallacy', which is the

danger that a corpus may be used to provide representative conclusions about the entire language period without understanding the limitations of the corpus with regard to what it does and does not contain in terms of genres and so on: all corpus researchers need to keep in perspective what their corpora can reasonably be taken to represent. Finally, there is the 'mystery of vanishing reliability', by which Rissanen means that the more variables which are used in sampling and coding the corpus – periods, genres, sociological variables and so on – the harder it is to represent each fully and to achieve statistical reliability. As he suggests, the most effective way of solving this problem is to build larger corpora which are proportionate to the distinctions which are encoded in them, but with current corpora it is simply necessary in designing research projects to be aware of the statistical limitations of the corpus. Rissanen's reservations are valid and important but should not be taken to diminish the value of corpus-based historical linguistics (he is, after all, one of the most well-known historical corpus linguists): rather, they should be considered as warnings of possible pitfalls which need to be taken on board by scholars, since they are surmountable with the use of appropriate care.

4.11. CORPORA IN DIALECTOLOGY AND VARIATION STUDIES

We have already seen the use of corpora to compare varieties in the sense of genres and channels. In this section we shall be concerned with geographical variation. Corpora have long been recognised as a valuable source for comparison between language varieties as well as for the description of those varieties themselves, since they are defined as constituting maximally representative samples of the respective language varieties. Indeed, certain corpora have even tried to follow as far as possible the same sampling procedures as other corpora in order to maximise the degree of comparability. Thus the LOB corpus contains broadly the same genres and sample sizes as the Brown corpus and is sampled from the same year, 1961. The Kolhapur Indian corpus is also broadly parallel to Brown and LOB, although in this case the sampling year was 1978.

One of the earliest pieces of work with the LOB and Brown corpora was the production of a word frequency comparison of American and British written English, showing the several differences which exist between those two national varieties. In the years following this work, the corpora have been used as the basis for the comparison of more complex aspects of language such as the use of the subjunctive (Johansson and Norheim 1988). The increase in the number of corpora of national varieties of English has since led to further comparative studies between, for instance, Australian and New Zealand English (Bauer 1993). The completion of the International Corpus of English (ICE), which in fact is a collection of many 1,000,000-word corpora of English from countries where English is a first or major language, will provide an even further range and motivation for comparative studies. For example, Josef

Schmied and his colleagues are using the East African component of ICE to look at the English spoken and written in Kenya and Tanzania (e.g., Schmied 1994).

One of the roles for corpora in national variation studies has been as a test-bed for two theories of language variation: Quirk *et al.*'s (1985: 16) 'common core' hypothesis – namely, that all varieties of English have central fundamental properties in common which differ quantitatively rather than qualitatively – and Braj Kachru's conception of national varieties as forming many unique 'Englishes' which differ in important ways from one another. To date, most work on lexis and grammar in the Kolhapur Indian corpus, studied in direct comparison with Brown and LOB, has appeared to provide support for the common core hypothesis (Leitner 1991). However, there is still considerable scope for the extension of such work and the availability of the ICE subcorpora should provide a wider range of data to test these hypotheses.

Compared to 'national variety', 'dialect' is a notoriously tricky term in linguistics, since dialects cannot readily be distinguished from languages on solely empirical grounds. However, the term 'dialect' is most commonly used of sub-national linguistic variation which is geographically motivated. Hence Australian English might not be considered expressly to be a dialect of English, whereas Scottish English, given that Scotland is a part of the United Kingdom, might well be so regarded; a smaller subset of Scottish English – for example, the English spoken in the Lowlands – would almost certainly be termed a 'dialect'. Taking 'dialect' to be defined in this way, it is the case that rather few dialect corpora exist at the present time. However, two examples are the Helsinki corpus of English dialects and John Kirk's Northern Ireland Transcribed Corpus of Speech (NITCS). These corpora both consist of spontaneous conversations with a fieldworker: in Kirk's corpus, as the name suggests, from Northern Ireland, and in the Helsinki corpus from several English regions.

Dialectology is a firmly empirical field of linguistics but has tended to concentrate on elicitation experiments and less controlled sampling rather than using corpora. The disadvantage of this approach is that elicitation experiments tend to concentrate on vocabulary and pronunciation, whereas other aspects of dialects, such as syntax, have been relatively neglected. The collection of stretches of natural spontaneous conversation in a corpus means, however, that these aspects of the language are now more amenable to study. Moreover, because the corpus is sampled so as to be representative, quantitative as well as qualitative conclusions can be drawn from it about the target population as a whole and the corpus can also be compared with corpora of other varieties.

This ability to make comparisons using dialect data has opened up a new avenue of research for dialectologists, namely, the opportunity to examine the degree of similarity and difference of dialects as compared with 'standard' varieties of a language. A particularly good example of the latter type of research is the work carried out by John Kirk on the identity of Scots. Kirk has used corpora of Scots – both dramatic texts and natural conversations – alongside

corpora of 'standard' British English such as LOB, London-Lund and *English* dramatic texts – to examine to what extent Scots deviates from 'standard' English on various elements. He has also begun to make a three-way comparison with Northern Irish English using his NITCS. In a recent study, Kirk (1993) examined the modal verb *will* in the NITCS, a corpus of Edinburgh speech, a corpus of Scots plays, the Lancaster/IBM Spoken English Corpus and a corpus of English plays. He classified the occurrences of root and epistemic *will* according to whether the following verb was stative or dynamic. The statistics were subjected to a VARBRUL analysis (see section 3.4.6) to discover which variables were significant in accounting for variation in the data. Here Kirk found, in accordance with previous studies that he has carried out on other aspects of usage, that regionality is not a significant factor but that text type is a significant factor. Again, then, this appears to support the common core hypothesis for dialects as well as for national Englishes, but it has to be said that the systematic study of dialects at linguistic levels above the word has only recently begun in earnest: we may expect to see more of this type of work as more corpora along the lines of the NITCS are built.

4.12. CORPORA IN PSYCHOLINGUISTICS

Psycholinguistics is almost by definition a laboratory subject. In order to test hypotheses about how language is processed in the mind, it is necessary to measure correlates of mental processes such as the length of time it takes to position a syntactic boundary in reading or how eye movements change during reading or even to look at which parts of the brain itself function at different points in the language understanding process through the use of imaging equipment. However, corpora do have parts to play in psycholinguistics.

One important use for corpora is as a source of data from which materials for laboratory experiments may be developed. For instance, as Schreuder and Kerkman (1987) have pointed out, frequency – that is, the familiarity of words to speakers of a language – is an important consideration in a number of cognitive processes such as word recognition. The psycholinguist should not therefore go blindly into experiments in areas such as this with only intuitive notions of frequency to guide the selection and analysis of materials. Properly sampled corpora are able to provide psycholinguists with more concrete and reliable information about frequency, including, as more annotated corpora become available, the frequencies of different senses and parts of speech of ambiguous words. Basic word frequency data from the Brown and London-Lund corpora, accompanied by specific psycholinguistic data from other sources, are already being marketed directly at psycholinguists by Oxford University Press in the form of a lexical database, *The Oxford Psycholinguistic Database* (see Wilson 1992a).

An example of a more direct role for a corpus in psycholinguistics is Garnham *et al.*'s (1981) study, which used the London-Lund spoken corpus to

look at the occurrence of speech errors in natural conversational English. Before this study was carried out, nobody quite knew how frequent speech errors were in everyday language for the simple reason that such an analysis required adequate stretches of natural conversation and previous work on speech errors had been based upon the gradual *ad hoc* accumulation of data from many different sources. The spoken corpus, however, was able to provide exactly the kind of data which was required. Although, given that the London-Lund corpus cannot be obtained in the form of original sound recordings, it was not possible to state with certainty that every single speech error in the corpus had been accounted for, Garnham's study was able for the first time to classify and count the frequencies of different error types and hence provide some estimate of the general frequency of these in relation to speakers' overall output. This was a contribution to the study of language processing which could not have been achieved without a corpus of natural spoken language.

Another role for corpora lies in the analysis of language pathologies, where an accurate picture of the abnormal data must be constructed before it is possible adequately to hypothesise and test what may be wrong with the human language processing system, and in the analysis of language development. Although little such work has been done with rigorously sampled corpora to date, it is important to stress their *potential* for these analyses (see Perkins 1995). Studies of the language of linguistically impaired people and also analyses of the language of children who are developing their (normal) linguistic skills, lack the quantified representative descriptions which are available, or at least are becoming available, for normal adult language. There has certainly during the last decade been a move towards the empirical analysis of machine-readable data in these areas: at Reading University, for example, a corpus of impaired and normal language development has been collected (Fletcher and Garman 1988); the Polytechnic of Wales (POW) corpus is a corpus of children's language; and the CHILDES database contains a large amount of normal and impaired child language data in several different languages. However, only a few of these data collections are properly sampled corpora in the sense of LOB or the BNC. Nevertheless, the interest in computer-aided analysis in these fields must mean that it will only be a matter of time before corpus collection in these areas becomes more statistically rigorous.

4.13. CORPORA AND CULTURAL STUDIES

It is perhaps now a commonplace in linguistics that texts contain the traces of the social conditions of their production. But it is only relatively recently that the role of a corpus in telling us about culture has really begun to be explored. After the completion of the LOB corpus of British English, one of the earliest pieces of work to be carried out was a comparison of its vocabulary with that of the earlier 'parallel' American Brown corpus (Hofland and Johansson 1982). This revealed interesting differences which went beyond the purely linguistic

ones such as spelling (e.g. *colour/color*) or morphology (e.g. *got/gotten*). Roger Fallon, in association with Geoffrey Leech, has picked up on the potential of corpora in the study of culture. Leech and Fallon (1992) used as their initial data the results of the earlier British and American frequency comparisons, along with the KWIC concordances to the two corpora to check up on the senses in which words were being used. They then grouped those differences which were found to be statistically significant into fifteen broad domain categories. The frequencies of concepts within these categories in the British and American corpora revealed findings which were suggestive not primarily of linguistic differences between the two countries but of cultural differences. For example, travel words proved to be more frequent in American English than in British English, perhaps suggestive of the larger size of the United States. Words in the domains of crime and the military were also more common in the American data and, in the crime category, 'violent' crime was more strongly represented in American English than in British English, perhaps suggestive of the American 'gun culture'. In general, the findings from the two corpora seemed to suggest a picture of American culture at the time of the two corpora (1961) that was more macho and dynamic than British culture. Although such work is still in its infancy and requires methodological refinement, it seems an interesting and promising line which, pedagogically, could also more closely integrate work in language learning with that in national cultural studies.

4.14. CORPORA AND SOCIAL PSYCHOLOGY

Although linguists have been the main users of corpora, we should observe that they have not been the sole users. Researchers in other fields which make use of language data have also in recent years taken an interest in the exploitation of corpus data. Perhaps the most important of these have been social psychologists.

Social psychologists are in a curious position. Unlike their colleagues in many other branches of psychology which rely on careful measurements carried out in laboratory conditions, they require access to *naturalistic* data which cannot be reproduced adequately in the laboratory owing to their spontaneous and context-governed nature. At the same time, however, they are also under pressure to quantify and test their theories rather than to rely solely upon qualitative data (although the latter do play a major part).

One area of research within social psychology is that of how and why people attempt to explain things. Explanations, or *attributions* as they are often called in social psychology, are important to the psychologist because they are revealing about the ways in which people regard their environment. To obtain data for studying explanations researchers have relied upon various sources of naturally produced texts such as newspapers, diaries, company reports and so on. However, these have tended to be written texts: spoken data are reported to have been used less frequently, although most everyday human interaction

takes place through the medium of speech.

Antaki and Naji (1987) hit upon the idea of using a spoken corpus – specifically the London-Lund corpus – as a source of data for explanations in everyday conversation. They took 200,000 words of conversation from the corpus and retrieved all instances of the commonest causal conjunction *because* (and its variant *cos*). An analysis of a pilot sample was used to arrive at a classification scheme for the data. The scheme was then used to classify all the explanations which had been retrieved from the corpus according to what was being explained – for example, 'actions of speaker or speaker's group', 'general states of affairs' and so on. A frequency analysis of the explanation types in the corpus showed that explanations of general states of affairs were the most common type of explanation (33.8%), followed by actions of speaker and speaker's group (28.8%) and actions of others (17.7%). Previous theory in this field had suggested that the prototypical type of explanation is the explanation of a person's single action. However, Antaki and Naji's findings appear to refute this notion. Although the data do not include all the causal statements in the corpus (for example, those introduced by *since*) and the London-Lund corpus is made up of conversations taken from a largely restricted social group (primarily middle-class academics), work such as Antaki and Naji's shows clearly the potential of corpora to test and modify theory in subjects which require naturalistic quantifiable language data, and one may expect more social psychologists to make use of corpora in the future.

4.15. CHAPTER SUMMARY

Corpora have a number of features which make them important as sources of data for empirical linguistic research. We have seen a number of these exemplified in this chapter in several areas of language study in which corpora have been, and may be, used. In summary, however, the main important advantages of corpora are:

1. *Sampling and quantification* Because the corpus is sampled to be maximally representative of the population, findings on that sample may be generalised to the larger population. Hence quantification in corpus linguistics is more meaningful than quantification in other forms of empirical linguistics because it can be assumed to tell us about a variety or a language, rather than just the sample which is being analysed.

2. *Ease of access* Using a corpus means that it is not necessary to go through a process of data collection: all the issues of sampling, collection and encoding have been dealt with by someone else. The majority of corpora are also readily available, often either free or at a low cost price to cover media and person time in handling distribution. But not only are the corpora themselves easy to obtain: once the corpus has been obtained, it is also easy to access the data within it. Because the corpus is in machine readable form, a concordance program can quickly extract frequency lists

and indices of various words or other items within it.

3. *Enriched data* Many corpora are now available already enriched with additional interpretive linguistic information such as part-of-speech annotation, grammatical parsing and prosodic transcription. Hence data retrieval from the corpus can be easier and more specific than with unannotated data.

4. *Naturalistic data* Not all corpus data are wholly unmonitored in the sense that the people producing the spoken or written texts included in the corpus are unaware until after the fact that they are being asked to participate in the building of a linguistic corpus. But, even with spoken data where surreptitious recording is no longer legally permitted, at least in the UK, the data are largely naturalistic, unmonitored and the product of real social contexts. Thus the corpus provides the most reliable source of data on language as it is actually used.

4.16. STUDY QUESTIONS

1. Carry out a literature search – perhaps using Altenberg's bibliography – on an area or topic in English linguistics which interests you. Try to look at as many of the studies you have found as you are able. Make a critical assessment of the differences that corpus analysis has made (if any) to our understanding of the area or topic you have chosen. How important has the use of corpus data been in that area?

2. As we have seen, open-ended monitor corpora have an important role in lexicography in that they enable researchers to keep abreast of new words and changes in the usage of existing words. Think about the kinds of issue which are important in some of the other areas of linguistics which we have looked at in this chapter. Do monitor corpora have any potential advantages over finite sampled corpora in other areas of language study?

3. In the discussions of sociolinguistics (section 4.7) and historical linguistics (section 4.10) we saw notes of caution by Holmes and Rissanen about the use of corpus data in specific circumstances. Thinking in a broader perspective, what dangers do you think are involved in corpus-based linguistics? How can these dangers be reduced?

4. What has the empirical analysis of corpora contributed to linguistic *theory* and what does it have the potential to contribute?

4.17. FURTHER READING

There is a huge amount of research literature which has made use of corpus data. For English corpus research, there is an extremely valuable bibliography collated by Bengt Altenberg. This was published in Johansson and Stenström

(1991) and is also available electronically from the ICAME file server (gopher: nora.hd.uib.no). More recent unpublished updates are also to be found on the ICAME server. Although not totally exhaustive, this bibliography contains the vast majority of published work using English language corpora.

For a more detailed overview of corpus-based projects than it has been possible to provide in this chapter, the student is recommended to look in the various specialist collections of papers. The festschrifts for Jan Svartvik (Aijmer and Altenberg 1991), Jan Aarts (Oostdijk and de Haan 1994b), Geoffrey Leech (Thomas and Short 1996) and Gunnel Tottie (Fries, Müller and Schneider 1997) contain papers across a broad range of fields, whilst the books of proceedings from annual ICAME conferences (Aarts and Meijs 1984, 1986, 1990; Johansson and Stenström 1991; Leitner 1992; Johansson 1982; Meijs 1987; Kytö, Ihalainen and Rissanen 1988; Souter and Atwell 1993; Aarts, de Haan and Oostdijk 1993; Fries, Tottie and Schneider 1994; Percy *et al.* 1996; Ljung 1997) provide a diachronic as well as a broad perspective of corpus-based research. For corpus-based historical linguistics see three recent specialist collections of papers: Kytö, Rissanen and Wright (1994); Hickey, Kytö and Lancashire (1997); and Rissanen, Kytö and Palander-Collin (1993). Papers, reviews and information are also to be found in the annual *ICAME Journal* (formerly *ICAME News*).

Foreign language corpus research is harder to track down, since it lacks the central organisation which English language research has. Students are advised to search keywords (e.g. *corpus/Korpus*) or citations of basic corpus linguistic texts in the standard linguistic bibliographies.

NOTES

1. It should be noted, however, that spoken corpus data are not always purely naturalistic. Naturalism is easier to achieve with pre-recorded material such as television and radio broadcasts. With spontaneous conversation, however, there are ethical and legal considerations which prevent the use of surreptitious recording without prior consent.
2. Taylor, Grover and Briscoe also argue that Sampson's argument about generative grammars is tied to the way he has chosen to analyse his data. However, these details are too complex to consider here; Sampson (1992) continues to dispute their claims.
3. Frequencies, slightly simplified, from Schmied (1993).
4. Frequencies, slightly simplified, from Pannek (1988).
5. See also McEnery and Wilson 1997.