Unit 10 Corpora and applied linguistics

10.1 Introduction

We have so far introduced most of the important concepts and practices in corpus linguistics related either to key issues like corpus design, markup, annotation, and the multilingual dimension, or to ancillary issues such as making statistical claims, using ready-made and DIY corpora and copyright clearance in corpus building. The final unit of Section A considers how corpora can be used in language studies. According to Leech (1997b: 9), corpus analysis can be illuminating 'in virtually all branches of linguistics or language learning' (see also Biber, Conrad and Reppen 1998: 11). One of the strengths of corpus data lies in its empirical nature, which pools together the intuitions of a great number of speakers and makes linguistic analysis more objective (cf. Biber et al 1998; McEnery and Wilson 2001: 103; though see unit 12 for a discussion of this claim). In this unit we will consider the use of corpus data in a number of areas of linguistics. Units 10.2 – 10.8 are concerned with the major areas of linguistics where corpora have been used while units 10.9 – 10.14 discuss other areas which have started to use corpus data. In unit 10.15, we will also discuss the limitations of using corpora in linguistic analysis.

10.2 Lexicographic and lexical studies

Corpora have proven to be invaluable resources for lexicographic and lexical studies. While lexicographers, even before the advent of modern corpora, used empirical data in the form of citation slips (e.g. Samuel Johnson's *Dictionary of the English Language*), it is corpora that have revolutionized dictionary making so that it is now nearly unheard of for new dictionaries and new editions of old dictionaries published from the 1990s onwards not to be based on corpus data. Corpora are useful in several ways for lexicographers. The greatest advantage of using corpora in lexicography lies in their machine-readable nature, which allows dictionary makers to extract all authentic, typical examples of the usage of a lexical item from a large body of text in a few seconds. The second advantage of the corpus-based approach, which is not available when using citation slips, is the frequency information and quantification of collocation which a corpus can readily provide. Some dictionaries, e.g. COBUILD 1995 and Longman 1995, include such frequency information. Information of this sort is particularly useful for materials writers and language learners alike (see case study 1 for a discussion of using corpora to improve learner dictionaries). A further benefit of using corpora is related to corpus markup and annotation. Many available corpora (e.g. the BNC) are encoded with textual (e.g. register, genre and domain) and sociolinguistic (e.g. user gender and age) metadata which allows lexicographers to give a more accurate description of the usage of a lexical item. Corpus annotations such as part-of-speech tagging and word sense disambiguation also enable a more sensible grouping of words which are polysemous and homographs. Furthermore, a monitor corpus allows lexicographers to track subtle change in the meaning and usage of a lexical item so as to keep their dictionaries up-to-date. Last but not least, corpus evidence can complement or refute the intuitions of individual lexicographers, which are not always reliable (cf. Sinclair 1991a: 112; Atkins and Levin 1995; Meijs 1996; Murison-Bowie 1996: 184) so that dictionary entries are more accurate. The observations above are line with Hunston (2002: 96), who summarizes the changes

brought about by corpora to dictionaries and other reference books in terms of five 'emphases':

- an emphasis on frequency;
- an emphasis on collocation and phraseology;
- an emphasis on variation;
- an emphasis on lexis in grammar;
- an emphasis on authenticity.

An important area of lexicographic study is loanwords. Lexicographers have traditionally used their intuitions as criteria to decide whether to include or exclude such lexical borrowings in a dictionary. Podhakecka and Piotrowski (2003) used corpus data to evaluate the treatment of 'Russianisms' in English. Their findings, which are based on a comparison of Russian loanwords in the BNC and *Oxford English Dictionary* (OED, electronic version) as well as *Longman Dictionary of Contemporary English* (2nd edition 1987 and 3rd edition 1995), are both expected and unexpected. On the one hand, they found that half of the 360 Russian loanwords they studied occurred only once in the BNC and very few items were really frequent. This finding is hardly surprising. What is unexpected is that the items selected by the OED on the basis of etymology exhibit the same type of distribution as items selected on the basis of frequency in the BNC. This finding suggests that intuition and corpora do not always lead to different conclusions (cf. unit 1.5). While Podhakecka and Piotrowski (2003) follow the traditional approach to loanword studies by analyzing loanwords as singly occurring items out of context, Kurtböke and Potter (2000) demonstrated, on the basis of their study of a number of English loans in a corpus of Turkish and a number of Italian loans in a corpus of English, that collocational patterns growing around loanwords are significant and should be included in the treatment of loanwords. They also found that '[a]ssimilation criteria based on frequency counts have proved to be less reliable than previously thought, and alternative criteria such as metaphor should also be taken into account' (Kurtböke and Potter 2000: 99).

In addition to lexicography, corpora have been used extensively in lexical studies (e.g. Nattinger and DeCarrico 1992; Schmitt 2004). The focus of corpus-based lexical studies is collocation and collocational meaning, i.e. semantic prosody and semantic preference.

Collocation has been studied for at least five decades. The term *collocation* was first used by Firth (1957) when he said 'I propose to bring forward as a technical term, meaning by *collocation*, and apply the test of *collocability*' (Firth 1957: 194). According to Firth (1968: 181), 'collocations of a given word are statements of the habitual or customary places of that word.' Firth's notion of collocation is essentially quantitative (cf. Krishnamurthy 2000: 32). The statistical approach to collocation is accepted by many corpus linguists including, for example, Halliday (1966: 159), Greenbaum (1974: 82), Sinclair (1991a), Hoey (1991), Stubbs (1995), Partington (1998), McEnery and Wilson (2001), and Hunston (2002). All of these linguists follow Firth in that they argue that collocation refers to the characteristic co-occurrence of patterns of words. One assumes that Greenbaum's (1974: 82) definition of collocation – 'a frequent co-occurrence of two lexical items in the language' – only refers to statistically significant collocation. He reserves the terms *collocability* and *collocable* for potential co-occurrence, using *collocation* and *collocate* solely for words which frequently co-occur (*ibid*: 80). While Greenbaum's definition does not

tell us how frequent the co-occurrence of two lexical items should be to be considered as a collocation, Hoey (1991: 6-7) uses the term *collocation* only if a lexical item appears with other items 'with greater than random probability in its (textual) context.' The random probability can be measured using statistical tests such as the MI (mutual information), *t* or *z* scores (see units 6.5 and 17).

Yet not all linguists would agree with Hoey's approach. Herbst (1996: 382), for example, argues against the statistical approach to collocation, asserting that if in Berry-Rogghe's (1972) 7,2000-word corpus, 'the most frequent collocates of a word such as *house* include the determiners *the* and *this* and the verb *sell*, this is neither particularly surprising nor particularly interesting.' It is true that if we search a nominal node word such as *house*, it is to all intents and purposes inevitable that determiners like *the* and *this* will be close to the top of the frequency list of co-occurring words. The presence of determiners such as *the* and *this* tells us *house* is a noun. The collocation of a node word with a particular grammatical class of words (e.g. determiners) is normally referred to as *colligation*. The fact that grammatical words sit on the top of a frequency list does not devalue the worth of collocations derived on the basis of statistics. Rather it means that because of the high overall frequencies of such grammatical words, brought about by their frequent co-occurrence with nouns, we should be selective in our approach to any given list of collocates, being prepared, on principled grounds, to exclude such words from the list of significant collocates even though they are very frequent. WordSmith, for example, allows users to exclude such frequent items by setting an upper limit, e.g. 1% of running words, from the list of collocates.

The task of determining frequency of co-occurrence manually is a daunting task, so it is no surprise that 'collocation is among the linguistic concepts which have benefited most from advances in corpus linguistics' (Krishnamurthy 2000: 33-34) in the age of the computer; the calculation of collocation statistics from electronic corpora is now a relatively trivial task given suitable software. Yet as well as being made easier to calculate, computerized corpora have freed linguists and lexicographers from an over reliance on intuition in the study of collocation. While some examples of collocation can be detected intuitively (cf. Deignan 1999: 23), 'particularly for obvious cases of collocation: *news* is *released*, *time* is *consumed*, and *computer programs run*' (Greenbaum 1974: 83), intuition is typically a poor guide to collocation. Greenbaum recognized this, and tried to address this problem by pooling the intuitions of large numbers of native speakers; he elicited data on collocation from a number of informants 'to provide access to the cumulative experience of large numbers of speakers' (*ibid*). He had to do this because no appropriate corpus resources were available when he undertook his work in the early 1970s. In those introspection-based elicitation experiments, he found it quite unsurprising that 'people disagree on collocations' (*ibid*). Intuition, as stated, is often a poor guide to collocation, 'because each of us has only a partial knowledge of the language, we have prejudices and preferences, our memory is weak, our imagination is powerful (so we can conceive of possible contexts for the most implausible utterances), and we tend to notice unusual words or structures but often overlook ordinary ones' (Krishnamurthy 2000: 32-33). Partington (1998: 18) also observes that 'there is no total agreement among native speakers as to which collocations are acceptable and which are not.' As Hunston (2002: 68) argues, whilst 'collocation can be observed informally' using intuition, 'it is more reliable to measure it statistically, and for this a corpus is essential.' This is because a corpus can reveal such probabilistic semantic patterns across many

speakers' intuitions and usage, to which individual speakers have no access (Stubbs 2001a: 153).

Shifting from form to meaning, Stubbs (2002: 225) observes that 'there are always semantic relations between node and collocates, and among the collocates themselves.' The collocational meaning arising from the interaction between a given node word and its collocates might be referred to as *semantic prosody*, 'a form of meaning which is established through the proximity of a consistent series of collocates' (Louw 2000: 57). The primary function of semantic prosody is to express speaker/writer attitude or evaluation (Louw 2000: 58). Semantic prosodies are typically negative, with relatively few of them bearing an affectively positive meaning. For example, Sinclair (1987, 1991a) observes that *HAPPEN* and *SET in* habitually collocate with nouns indicating unpleasant situations. However, it is also claimed that a speaker/writer can violate a semantic prosody condition to achieve some effect in the hearer – for example irony, insincerity or humour can be diagnosed by violations of semantic prosody according to Louw (1993: 173). Semantic prosody is strongly collocational in that it operates beyond the meanings of individual words. For example, both *personal* and *price* are quite neutral, but when they co-occur, a negative prosody may result: *personal price* most frequently refers to something undesirable, as demonstrated by all such examples from the BNC (2) and the Bank of English (18).

It would appear, from the literature published on semantic prosody (including semantic preference), that it is at least as inaccessible to a speaker's conscious introspection as collocation is (cf. Louw 1993: 173; Partington 1998: 68; Hunston 2002: 142). Yet as corpora have grown in size, and tools for extracting semantic prosodies have been developed, semantic prosodies have been addressed much more frequently by linguists (e.g. Louw 1993, 2000; Stubbs 1995; Partington 1998; Hunston 2002). The profiles of the semantic prosodies of many words and phrases have been revealed, e.g. in addition to those mentioned above, it has been suggested that *CAUSE* (Stubbs 1995), *COMMIT* (Partington 1998: 67), *PEDDLE/peddler* (*ibid*: 70-72), *dealings* (*ibid*: 72-74), *END up verbing* (Louw 2000: 54), *a recipe for* (Louw 2000: 63), *GET oneself* verb*ed* (*ibid*), *FAN the flame* (Stubbs 2001b: 445), *signs of* (*ibid*: 458), *ripe for* (*ibid*: 457), *underage* and *teenager(s)* (*ibid*: 454), *SIT through* (Hunston 2002: 60-62), and *bordering on* (Schmitt and Carter 2004: 8) typically carry an unfavourable affective meaning whereas *PROVIDE* (Stubbs 1995) and *career* (Stubbs 2001b: 459) have a positive prosody.

It might be argued that the negative (or less frequently positive) prosody that belongs to an item is the result of the interplay between the item and its typical collocates. On the one hand, the item does not appear to have an affective meaning until it is in the context of its typical collocates. On the other hand, if a word has typical collocates with an affective meaning, it may take on that affective meaning even when used with atypical collocates. As the Chinese saying goes, 'he who stays near vermilion gets stained red, and he who stays near ink gets stained black' – one takes on the colour of one's company – the consequence of a word frequently keeping 'bad company' is that the use of the word alone may become enough to indicate something unfavourable (cf. Partington 1998: 67).

In Stubbs' (2002: 225) comment cited above, the meaning arising from the common semantic features of the collocates of a given node word can be referred to *semantic preference*, which is defined 'by a lexical set of frequently occurring collocates [sharing] some semantic feature' (*ibid*: 449). For example, Stubbs (2001c: 65) observes that *large* typically collocates with items from the same semantic set

indicating ‘quantities and sizes’ (e.g. *number(s)*, *scale*, *part*, *quantities*, *amount(s)*) while Partington (2004: 148) notes that ‘absence/change of state’ is a common feature of the collocates of maximizers such as *utterly*, *totally*, *completely* and *entirely*.

Semantic preference and semantic prosody are two distinct yet interdependent collocational meanings (see unit 13.3). According to Sinclair (1996, 1998) and Stubbs (2001c), semantic prosody is a further level of abstraction of the relationship between lexical units: collocation (the relationship between a node and individual words), colligation (the relationship between a node and grammatical categories), semantic preference (semantic sets of collocates) and semantic prosody (affective meanings of a given node with its typical collocates). Partington (2004: 151) notes that semantic preference and semantic prosody have different operating scopes: the former relates the node item to another item from a particular semantic set whereas the latter can affect wider stretches of text. Semantic preference can be viewed as a feature of the collocates while semantic prosody is a feature of the node word. On the other hand, the two also interact. While semantic prosody ‘dictates the general environment which constrains the preferential choices of the node item’, semantic preference ‘contributes powerfully’ to building semantic prosody (Partington 2004: 151).

There are different opinions regarding whether or not semantic prosody is a type of connotative meaning. Partington (1998: 68), Stubbs (2001a: 449) and Hunston (2002: 142) appear to take it for granted that semantic prosody is connotational. However, Louw (2000: 49-50) explicitly argues that ‘semantic prosodies are not merely connotational’ as ‘the force behind SPs [semantic prosodies] is more strongly collocational than the schematic aspects of connotation.’ In our view, connotation can be collocational or non-collocational whereas semantic prosody can only be collocational.

It is important to note that lexical studies also include morphological analysis, at the sub-lexical level, of the internal structure of a word in terms of its root, prefix and suffix, where appropriate. While there is presently no morphemically annotated corpus available, University College London (UCL) is currently planning to integrate the morphological annotation into the already POS tagged and syntactically parsed version of ICE-GB, the British component of the ICE corpus. Such morphological analysis not only greatly benefits morphologists, syntacticians and lexicographers, it is also useful for language learners. For example, Gries (2003) shows that there are some important semantic and distributional differences in adjective pairs ending with *-ic* and *-ical*, which language learners may find useful in distinguishing between the two.

10.3 Grammatical studies

Along with lexicographic and lexical studies, grammar is another area which has frequently exploited corpus data. This is because a balanced representative corpus not only provides a reliable basis for quantifying syntactic features of a language or language variety, it is also useful in testing hypotheses derived from grammatical theory. Corpora have had such a strong influence on recently published reference grammar books (at least for English) that ‘even people who have never heard of a corpus are using the product of corpus-based investigation’ (Hunston 2002: 96).

If *A Comprehensive Grammar of the English Language* (i.e. Quirk et al 1985) is viewed as a milestone in the study of English grammar, it is fair to say that the recently published *Longman Grammar of Spoken and Written English* (i.e. LGSWE, Biber et al 1999) is a new milestone. Based entirely on the 40-million-word Longman

Spoken and Written English Corpus, the new grammar gives 'a thorough description of English grammar, which is illustrated throughout with real corpus examples, and which gives equal attention to the ways speakers and writers actually use these linguistic resources' (Biber et al 1999: 45).

The new corpus-based grammar is unique in many different ways, for example, by exploring the differences between written and spoken grammars and taking register variations into account. The coverage given by the grammar to spoken English is particularly important. While grammatical studies have traditionally focused on written language, the availability of spoken English corpora (see unit 7.5) has provided unprecedented insights in spoken grammar. Recent studies have claimed that the traditional sentence-based grammar is inadequate in describing spoken language. In addition to Biber et al (1999) discussed above, further work has been undertaken by Carter, Hughes and McCarthy at the University of Nottingham (the so-called Nottingham school). In a series of studies based on the CANCODE corpus (Carter and McCarthy 1995; McCarthy and Carter 1997; Hughes and McCarthy 1998; McCarthy 1998), the Nottingham school approaches spoken grammar from the perspective of discourse, as in McCarthy's (1988: 78) words, 'discourse drives grammar.' This approach has allowed the authors to discover many features of spoken grammar (e.g. initial ellipsis and topics in pre-clause slots and tails in post-clause slots) that are absent or marginalized in written grammars. For example, Carter and McCarthy (1995: 152-153) find that while indirect speech has been thoroughly covered in traditional grammars focusing on backshift in the sequencing of tenses, a frequent reporting phenomenon in spoken discourse such as reporting verbs like *SAY* and *TELL* occurring in the past progressive appears to have been overlooked. While the simple past form of a reporting verb gives more authority to the contents of the utterance, the past progressive form of a reporting verb focuses more on the event of uttering *per se*. As such, the authors suggest that the grammar of spoken language is radically different from that of written language and needs to be modelled on the basis of spoken data with no prior assumption that spoken and written grammars share the same framework. The Nottingham school's focus on the 'differentness' of spoken and written grammars is in contrast with Biber et al's (1999) position, which focuses on 'sameness' and uses the same framework to describe spoken and written language. The 'sameness' approach is taken by yet another influential English grammar, Huddleston and Pullum (2000), which comments that while '[t]here are significant and interesting differences between spoken and written language' (*ibid*: 11), '[s]harp divergences between the syntax of speech and the syntax of writing […] are rare to the point of non-existence' (*ibid*: 13). Readers are advised to refer to Leech (2000) for a good review of corpus-based research in spoken English grammar and a comparison of the 'sameness' and 'differentness' approaches to spoken grammar.

We noted in unit 10.2 that the corpus-based approach to grammar has led to a focus on lexis in grammatical studies. While lexical information forms, to some extent, an integral part of the grammatical description in Biber et al (1999), it is the Birmingham school (e.g. Sinclair, Hunston, Francis and Manning) that focuses on lexis in their grammatical descriptions (the so-called 'pattern grammar', e.g. Hunston and Francis 2000). In fact, Sinclair et al (1990) flatly reject the distinction between lexis and grammar. These authors have given prominence to the close association between pattern and meaning, as embodied in the Collins COBUILD Grammar Patterns series (e.g. Francis et al 1996; 1997; 1998). Francis et al (1998), for example, present over 200 patterns on the basis of their study of 10,000 nouns and adjectives in the Bank of English and relate these patterns to meaning. While pattern grammars focusing on the

connection between pattern and meaning challenge the traditional distinction between lexis and grammar, they are undoubtedly useful in language learning as they provide 'a resource for vocabulary building in which the word is treated as part of a phrase rather than in isolation' (Hunston 2002: 106).

10.4 Register variation and genre analysis

We noted in unit 2 that corpus design typically relies on external criteria, or situational parameters in Biber's (1993) terms (see unit 11.2). These parameters define register in terms of the social or communicative context of their use. In Biber's works (Biber 1988: 170; Biber et al 1999: 25), the terms *register* and *genre* appear to be used interchangeably. While there are other possible definitions of *register* (cf. Paolillo 2000: 217-218) and *genre* (as used in critical discourse analysis, where a genre is defined as 'a socially ratified way of using language in connection with a particular type of social activity (e.g. interview, narrative, exposition)', see Fairclough 1995: 14), we adopt a rather loose definition of these terms in this book so that they are less exclusive.

The corpus-based approach is well suited for the study of register variation and genre analysis because corpora, especially balanced sample corpora, typically cover a wide range of registers or genres. Oh (2000), for example, used the 2.4-million-word Switchboard corpus of informal telephone conversation and the Brown corpus (see unit 7.4) to explore the similarities and differences between *actually* and *in fact* in written and spoken American English. He found that *actually* was 3.7 times more frequent than *in fact* in spoken discourse, and *actually* also showed a greater affinity with utterance-medial position in both written and spoken discourse.

The most powerful tool for approaching register and genre variation is perhaps the multi-feature/multi-dimensional analytical framework (i.e. MF/MD; see unit 14.2 for an overview and case study 5 for its application) established in Biber (1988), which presents a full analysis of 21 genres in spoken and written British English on the basis of 67 functionally related linguistic features in 481 texts from the LOB and London-Lund corpora.

The MF/MD approach is based on *factor analysis*. Factor analysis is commonly used in the social and behavioural sciences to summarize the interrelationships among a large group of variables in a concise fashion (cf. Biber 1988: 64). Biber (1988: 63, 79) used factor analysis in concert with frequency counts of linguistic features to identify sets of features that co-occur in texts with a high frequency. He referred to these sets of features as *dimensions* or *constructs*. Biber used factor analysis to reduce 67 linguistic features to 7 dimensions or factors (see unit 14.2 for further discussion). As these factors underlie linguistic features, they are conceptually clearer than the many features considered individually.

Biber (1988) used a whole chapter (chapter 5) to give a technical description of factor analysis. As we will only apply the dimensions established by Biber (see case study 5), the issue of how these factors were computed will not be our concern in this book. Nevertheless, a brief, non-technical account of how factor analysis works will prove helpful to understanding Biber's MF/MD approach.

Factor analysis starts with a simple correlation matrix of all linguistic features, on the basis of which the factorial structures are established and the factor *loading* or *weight* of each linguistic feature is computed. A factor loading or weight indicates the degree to which one can generalize from a given factor to an individual linguistic feature (Biber 1988: 81). A loading can be positive or negative, indicating the direction of

correlation. The greater the absolute value of a loading a linguistic feature has on a factor, the more representative the feature is of the dimension. Biber (1988: 87) decided that only the important or salient loadings should be interpreted as part of each factor. All features having loadings with an absolute value less than 0.30 were excluded as unimportant. Due to the large number of features loading on most of the factors, Biber used a conservative cut-off point of 0.35 to decide which features were to be included in the computation of factor scores (Biber 1988: 93). When a feature has a salient loading (above 0.35) on more than one factor, it was included in the factor score of the factor on which it had the highest loading so as to ensure that each feature was included in the computation of only one factor score (Biber 1988: 93). Using the procedure above, Biber identified seven dimensions or factors:

- Factor 1: informational versus involved production;
- Factor 2: narrative versus non-narrative concerns;
- Factor 3: explicit versus situation-dependent reference;
- Factor 4: overt expression of persuasion;
- Factor 5: abstract versus non-abstract information;
- Factor 6: online informational elaboration;
- Factor 7: academic hedging.

Of these, Factors 1, 3 and 5 are associated with 'oral' and 'literate' differences in English (Biber 1988:163; Biber and Finegan 1989: 489). As the factorial structure of Factor 7 was not strong enough for a firm interpretation, it was not discussed in detail in Biber (1988).

Using the MF/MD approach, Biber (1988) was able to describe the similarities and differences of various genres in spoken and written English with reference to the different dimensions. For example, Biber (1988: 165-166) finds that along Dimensions 1 and 5, personal letters and spontaneous speech demonstrate quite similar factor scores but differ considerably along Dimensions 3 and 6. Likewise, while face-to-face conversation differs markedly from official documents along Dimensions 1, 3 and 5, they are quite similar along Dimension 4. As such, Biber (1988: 169) concludes that:

> Each dimension is associated with a different set of underlying communicative functions, and each defines a different set of similarities and differences among genres. Consideration of all dimensions is required for an adequate description of the relations among spoken and written texts.

While the MF/MD analytical framework was originally developed to compare written and spoken registers in English, this approach has been used extensively in (1) synchronic analyses of specific registers and genres (Biber 1991; Biber and Finegan 1994a; Conrad 1994; Reppen 1994; Tribble 1999) and author styles (Biber and Finegan 1994b; Connor-Linton 1988; Watson 1994); (2) diachronic studies describing the evolution of registers (Biber and Finegan 1989, 1992; Atkinson 1992, 1993) and exploring the differences between literary and non-literary genres in Early Modern English (Taavitsainen 1997); (3) register studies of non-western languages (Besnier 1988; Biber and Hared 1992, 1994; Kim and Biber 1994) and contrastive analyses (Biber 1995b). In addition, the MF/MD approach has also been applied in addressing corpus design issues (e.g. Biber 1993; see unit 11.2) and the definitional issues of registers/genres and text types (e.g. Biber 1989). Unit 14.2 will further discuss Biber's MF/MD approach.

Lexical bundles, also called lexical chains or multiword units, are closely associated with collocations and have been an important topic in lexical studies (e.g. Stubbs 2002). More recently, Biber found that lexical bundles are also a reliable indicator of register variation (e.g. Biber and Conrad 1999; Biber 2003). Biber and Conrad (1999), for example, showed that the structural types of lexical bundles in conversation are markedly different from those in academic prose. Biber's (2003) comparative study of the distribution of 15 major types of 4-word lexical bundles (technically known as 4-grams) in the registers of conversation, classroom teaching, textbooks and academic prose indicates that lexical bundles are significantly more frequent in the two spoken registers. The distribution of lexical bundles in different registers also varies across structural types. In conversation, nearly 90% of lexical bundles are declarative or interrogative clause segments. In contrast, the lexical bundles in academic prose are basically phrasal rather than clausal. Of the four registers in Biber's study, lexical bundles are considerably more frequent in classroom teaching because this register uses the types of lexical bundles associated with both conversation and academic prose.

10.5 Dialect distinction and language variety

A language variety can be broadly defined as a variant of a language that differs from another variant of the same language systematically and coherently. Varieties of a language may include, for example, the standard language (standardized for the purposes of education and public performance), dialects (geographically defined), sociolects (socially defined), idiolects (unique to individual speakers) and jargons (particular to specific domains). This book defines both *language variety* and *dialect* geographically. We refer to national variants (e.g. 'world Englishes' such as British English and American English; see unit 7.6) as language varieties and regional variants (e.g. variants in the south and north of Britain) as dialects while other variants such as sociolects, idiolects and jargons are considered as *language variations*. So-called standard varieties of a language such as Standard English in Britain and *putonghua* ('common language') in China are simply particular dialects that have been given legal or quasi-legal status and are typically used for education (e.g. teaching the language as a foreign language) and public purposes. While a standard language is usually based on the regional dialect of a capital city, it is also marked socially. For example, while accents or dialects usually tell us where a speaker is from, RP (the notional standard form of spoken British English) is a regionally neutral accent which tells us only about a speaker's social or educational but not regional background. Even though RP originated in the South East of England, it has developed to be regionally neutral but socially marked. We do not consider standard languages as dialects as defined in this book. However, our decision is one of convenience and should not be taken to imply that we do not conceive of standard varieties as dialects. Similarly, while we will use language variety as a term that encompasses language varieties such as pidgins and creoles, we appreciate once again that we are taking something of a terminological short cut in doing so.

Variations in dialects and language varieties are commonly found in pronunciation, spelling and word choice while grammatical differences are relatively few. Dialects typically vary quantitatively rather than qualitatively (cf. Bauer 2002: 108). It would appear that core grammatical structures are relatively uniform across dialects and language varieties of English (cf. Biber et al 1999: 19-21; Huddleston and Pullum 2000: 4). For example, Biber et al (1999: 398-399) observe that the *got*/*gotten*

alternation represents an important difference between American English and British English: while the pattern *have* + *gotten* rarely occurs in British English, it is very frequent in American English, especially in conversations and when the combination expresses a true perfect meaning. In contrast, the use of *do* following an auxiliary (e.g. *I'm not sure that I'll go, but I **may do***) is uncommon in American English (see Huddleston and Pullum 2000: 5). Biber (1987) also finds that nominalizations, passives and *it*-cleft structures are more frequent in American English whereas time/place adverbials, and subordinator deletion occur more frequently in British English. In case study 2 in Section C of this book, we will see that American English shows a strong preference for bare infinitives following HELP (e.g. *helped him get to his feet* and *helped finance the project*). Hundt (1998: 32), on the basis of a comparison of the frequencies and proportions of regular and irregular past tense forms of various verbs in WWC and FLOB, finds that New Zealand English (56.4% regular) and British English (68.7% regular) differ significantly in this respect. She also notes that 96.7% of the relevant verb forms are regular in the Brown corpus of American English, concluding that there is a difference between the three varieties of English, with New Zealand English being closer to British English. Readers must note that in Hundt's study, the difference between American English and the other two varieties might be attributed to language change (though further research is required to make it clear), as the Brown corpus sampled texts in 1961 whereas the sampling periods of WWC and FLOB are closer to each other yet much later than Brown. Tagnin and Teixeira (2003) show, on the basis of a comparable corpus of cooking recipes in four language varieties (British vs. American English, Brazilian vs. European Portuguese), that the differences between the two varieties of Portuguese at various levels (lexical, syntactic and textual) are much more marked than those between British and American English.

10.6 Contrastive and translation studies

This section takes linguistic comparisons one step further from dialects and varieties of the same language to different languages. This involves the use of multilingual corpora (see unit 5). There are two major types of linguistic investigations based on multilingual corpora: contrastive and translation studies.

As Laviosa (1998a) observes, 'the corpus-based approach is evolving, through theoretical elaboration and empirical realisation, into a coherent, composite and rich paradigm that addresses a variety of issues pertaining to theory, description, and the practice of translation.' Corpus-based translation studies come in two broad types: theoretical and practical (Hunston 2002: 123). With reference to theory, corpora are used mainly to study the translation process by exploring how an idea in one language is conveyed in another language and by comparing the linguistic features and their frequencies in translated L2 texts and comparable L1 texts. In the practical approach, corpora provide a workbench for training translators and a basis for developing applications like machine translation (MT) and computer-assisted translation (CAT) systems. In this section, we will discuss how corpora have been used in each of these areas.

Parallel corpora are a good basis for studying how an idea in one language is conveyed in another language (see case study 6). Xiao and McEnery (2002a), for example, used an English-Chinese parallel corpus containing 100,170 English words and 192,088 Chinese characters to explore how temporal and aspectual meanings in English were expressed in Chinese. In that study, the authors found that while both

English and Chinese have a progressive aspect, the progressive has different scopes of meaning in the two languages. In English, while the progressive canonically (93.5%) signals the ongoing nature of a situation (e.g. *John is singing*, Comrie 1976: 32), it has a number of other specific uses 'that do not seem to fit under the general definition of progressiveness' (Comrie 1976: 37). These 'specific uses' include its use to indicate contingent habitual or iterative situations (e.g. *I'm taking dancing lessons this winter*, Leech 1971: 27), to indicate anticipated happenings in the future (e.g. *We're visiting Aunt Rose tomorrow, ibid*: 29) and some idiomatic use to add special emotive effect (e.g. *I'm continually forgetting people's names, ibid*) (cf. Leech 1971: 27-29). In Chinese, however, the progressive marked by *zai* only corresponds to the first category above, namely, to mark the ongoing nature of dynamic situations. As such, only about 58% of situations referred to by the progressive in the English source data take the progressive or the durative aspect, either marked overtly or covertly, in Chinese translations. The authors also found that the interaction between situation aspect (i.e. the inherent aspectual features of a situation, e.g. whether the situation has a natural final endpoint; see unit 10.9) and viewpoint aspect (e.g. perfective vs. imperfective; see unit 15.3) also influences a translator's choice of viewpoint aspect. Situations with a natural final endpoint (around 65%) and situations incompatible with progressiveness (92.5% of individual-level states and 75.9% of achievements) are more likely to undergo viewpoint aspect shift and be presented perfectively in Chinese translations. In contrast, situations without a natural final endpoint are normally translated with the progressive marked by *zai* or the durative aspect marked by *-zhe*.

Note, however, that the direction of translation in a parallel corpus is important in studies of this kind. The corpus used in Xiao and McEnery (2002a), for example, is not suitable for studying how aspect markers in Chinese are translated into English. For that purpose, a Chinese-English parallel corpus (i.e. L1 Chinese plus L2 English) is required.

Another problem which arises with the use of a one-to-one parallel corpus (i.e. containing only one version of translation in the target language) is that the translation only represents one individual's introspection, albeit contextually and cotextually informed (cf. Malmkjær 1998). One possible way to overcome this problem, as suggested in Malmkjær, is to include as many versions of a translation of the same source text as possible in a parallel corpus. While this solution is certainly of benefit to translation studies, it makes the task of building parallel corpora much more difficult. It also reduces the range of data one may include in a parallel corpus, as many translated texts are only translated once. It is typically only literary works which have multiple translations of the same work available. These works tend to be non-contemporary and the different versions of the translation are usually spaced decades apart, thus making the comparison of these versions problematic.

The distinctive features of translated language can be identified by comparing translations with comparable L1 texts, thus throwing new light on the translation process and helping to identify translation norms. Laviosa (1998b), for example, in her study of L1 and L2 English narrative prose, finds that translated L2 language has four core patterns of lexical use: a relatively lower proportion of lexical words over function words, a relatively higher proportion of high-frequency words over low-frequency words, a relatively greater repetition of the most frequent words, and less variety in the words that are most frequently used. Other studies show that translated language is characterized, beyond the lexical level, by nominalization, simplification (Baker 1993, 1999), explication (i.e. increased cohesion, Øverås 1998) and

sanitization (i.e. reduced connotational meanings, Kenny 1998). As we will see in case study 6, the frequency of aspect markers in Chinese translations is significantly lower than that in the comparable L1 Chinese data. As these features are regular and typical of translated language, further research based upon these findings may not only uncover the translation norms or what Frawley (1984) calls the 'third code' of translation, it will also help translators and trainee translators to become aware of these problems.

The above studies demonstrate that translated language represents a version of language which we may call 'translationese'. The effect of the source language on the translations is strong enough to make the L2 data perceptibly different from the target L1 language. As such, a uni-directional parallel corpus is a poor basis for cross-linguistic contrast. This problem, however, can be alleviated by the use of a bi-directional parallel corpus (e.g. Maia 1998; Ebeling 1998), because the effect of translationese may be averaged out to some extent. In this sense, a well-matched bi-directional parallel corpus can become the bridge that brings translation and contrastive studies together. To achieve this aim, however, the same sampling frame must apply to the selection of source data in both languages. Any mismatch of proportion, genre, or domain, for example, may invalidate the findings derived from such a corpus.

While we know that translated language is distinct from the target L1 language, it has been claimed that parallel corpora represent a sound basis for contrastive studies. James (1980: 178), for example, argues that 'translation equivalence is the best available basis of comparison', while Santos (1996: i) claims that 'studies based on real translations are the only sound method for contrastive analysis.' Mauranen (2002: 166) also argues, though not as strongly as James and Santos, that translated language, in spite of its special features, 'is part of natural language in use, and should be treated accordingly', because languages 'influence each other in many ways other than through translation' (*ibid*: 165). While we agree with Mauranen that 'translations deserve to be investigated in their own right', as is done in Laviosa (1998b) and McEnery and Xiao (2002), we hold a different view of the value of parallel corpora for contrastive studies. It is true that languages in contact can influence each other, but this influence is different from the influence of a source language on translations in respect to immediacy and scope. Basically, the influence of language contact is generally gradual (or evolutionary) and less systematic than the influence of a source language on the translated language. As such, translated language is at best an unrepresentative special variant of the target language. If this special variant is confused with the target L1 language and serves alone as the basis for contrastive studies, the results are clearly misleading. This may have long-term adverse effects because contrastive studies are 'typically geared towards second language teaching and learning' (Teich 2002: 188). We would not want to misrepresent an L1 by teaching the translationese approximation of it. But parallel corpora still have a role to play in contrastive analysis. Parallel corpora can serve as a useful starting point for cross-linguistic contrasts because findings based on parallel corpora invite 'further research with monolingual corpora in both languages' (Mauranen 2002: 182). In this sense, parallel corpora are 'indispensable' to contrastive studies (*ibid*).

With reference to practical translation studies, as corpora can be used to raise linguistic and cultural awareness in general (cf. Hunston 2002: 123; Bernardini 1997), they provide a useful and effective reference tool and a workbench for translators and trainees. In this respect even a monolingual corpus is helpful. Bowker (1998), for example, found that corpus-aided translations were of a higher quality with respect to

subject field understanding, correct term choice and idiomatic expressions than those undertaken using conventional resources. Bernardini (1997) also suggests that traditional translation teaching should be complemented with what she calls 'LCC' (large corpora concordancing) so that trainees develop 'awareness', 'reflectiveness' and 'resourcefulness', the skills that 'distinguish a translator from those unskilled amateurs.'

In comparison to monolingual corpora, comparable corpora are more useful for translation studies. Zanettin (1998) demonstrates that small comparable corpora can be used to devise a 'translator training workshop' designed to improve students' understanding of the source texts and their ability to produce translations in the target language more fluently. In this respect, specialized comparable corpora are particularly helpful for highly domain-specific translation tasks, because when translating texts of this type, as Friedbichler and Friedbichler (1997) observe, the translator is dealing with a language which is often just as disparate from their native language as any foreign tongue. Studies show that translators with access to a comparable corpus with which to check translation problems are able to enhance their productivity and tend to make fewer mistakes when translating into their native language. When translation is from a mother tongue into a foreign language, the need for corpus tools grows exponentially and goes far beyond checking grey spots in L1 language competence against the evidence of a large corpus. For example, Gavioli and Zanettin (1997) demonstrate how a very specialized corpus of text on the subject of hepatitis helps to confirm translation hypotheses and suggest possible solutions to problems related to domain-specific translation.

While monolingual and comparable corpora are of use to translation, it is difficult to generate 'possible hypotheses as to translations' with such data (Aston 1999). Furthermore, verifying concordances is both time-consuming and error-prone, which entails a loss of productivity. Parallel corpora, in contrast, provide '[g]reater certainty as to the equivalence of particular expressions', and in combination of suitable tools (e.g. ParaConc, see case study 6), they enable users to 'locate all the occurrences of any expression along with the corresponding sentences in the other language' (*ibid*). As such, parallel corpora can help translators and trainees to achieve improved precision with respect to terminology and phraseology and have been strongly recommended for these reasons (e.g. Williams 1996). A special use of a parallel corpus with one source text and many translations is that it can offer a systematic translation strategy for linguistic structures which have no direct equivalents in the target language. Buyse (1997), for example, presents a case study of the Spanish translation of the French clitics *en* and *y*, where the author illustrates how a solution is offered by a quantitative analysis of the phonetic, prosodic, morphological, semantic and discursive features of these structures in a representative parallel corpus, combined with the quantitative analysis of these structures in a comparable corpus of L1 target language. Another issue related to translator training is translation evaluation. Bowker (2001) shows that an evaluation corpus, which is composed of a parallel corpus and comparable corpora of source and target languages, can help translator trainers to evaluate student translations and provide more objective feedback.

Finally, in addition to providing assistance to human translators, parallel corpora constitute a unique resource for the development of machine translation (MT) systems. Starting in the 1990s, the established methodologies, notably, the linguistic rule-based approach to machine translation, were challenged and enriched by an approach based on parallel corpora (cf. Hutchins 2003: 511; Somers 2003: 513). The new approaches,

such as example-based MT (EBMT) and statistical MT, were based on parallel corpora. To take an example, EBMT works by matching any sentence to be translated against a database of aligned texts previously translated to extract suitable examples which are then combined to generate the correct translation of the input sentence (see Somers: *ibid*). As well as automatic MT systems, parallel corpora have also been used to develop computer-assisted translation (CAT) tools for human translators, such as translation memories (TM), bilingual concordancers and translator-oriented word processor (cf. Somer 2003; Wu 2002).

The main concern of this section is the potential value of parallel and comparable corpora to translation and contrastive studies. Parallel corpora are undoubtedly a useful starting point for contrastive research, which may lead to further research in contrastive studies based upon comparable corpora. In contrast, comparable corpora used alone are less useful for translation studies. Nonetheless, they certainly serve as a reliable basis for contrastive studies. It appears then that a carefully matched bi-directional parallel corpus provides a sound basis for both translation and contrastive studies. Yet the ideal bi-directional parallel corpus will often not be easy, or even possible, to build because of the heterogeneous pattern of translation between languages and genres. So we must accept that, for practical reasons alone, we will often be working with corpora that, while they are useful, are not ideal for either translation or contrastive studies. We will return to the exploitation of the use of parallel and comparable corpora in units 15.2-15.3 in Section B and case study 6 in Section C of this book.

10.7 Diachronic study and language change

This section shifts our focus from the synchronic studies discussed in the previous sections to diachronic studies and language change. The nature of diachronic study determines its reliance on empirical historical data. Diachronic study is perhaps one of the few areas which can only be investigated using corpus data. This is because the intuitions of modern speakers have little to offer regarding the language used hundreds or even tens of years before.

We noted in unit 7.7 that while a number of corpora (e.g. LOB vs. FLOB, and Brown vs. Frown) are suitable for the diachronic study of English, the most famous corpus of this kind is the Helsinki corpus, produced by the English Department of the University of Helsinki. Following the creation of the corpus, the analysis of the corpus was carried out on their subsequent project 'English in transition: Change through variation', which produced three volumes of studies: *Early English on the Computer Age: Exploration through the Helsinki Corpus* (Rissanen, Kytö and Palander-Collin 1993), *English in Transition: Corpus-based Studies in Linguistic Variation and Genre Styles* (Rissanen, Kytö and Heikkonen 1997a), and *Grammaticalization at Work: Studies of Long-term Developments in English* (Rissanen, Kytö and Heikkonen 1997b). The Helsinki corpus not only sampled different periods covering one millennium, it also encoded genre and sociolinguistic information (e.g. author rank, sex and age, cf. Rissanen et al 1997a: 3). This allowed the authors of these volumes to go beyond simply dating and reporting change by combining diachronic, sociolinguistic and genre studies.

Peitsara (1993), for example, in her study of prepositional phrases introducing agency in passive constructions in the Early Modern and Modern English (ca. 1350-1640) components in the Helsinki corpus, finds that while at the beginning of the period *by* and *of* were equally frequent, by the end of the period, *by* had gained prominence to

the extent that it was three times more frequent than *of* by the 15th century. This trend accelerated over time, so that by the 16th century it was eight times more frequent than *of*. Furthermore, she notes that such a contrast was particularly marked in the genre of official documents and correspondences. Likewise, based on the Corpus of Early English Correspondence (developed at the University of Helsinki), Nevalainen (2000) observes that in Early Modern English, female authors led the move in replacing the verbal suffix *-th* with *-s* and using *you* in subject position whereas male authors took the lead in replacing double negation with single negation.

Such findings can only be made via the use of properly composed diachronic corpora. This research, and much more beside (see units 15.4 and 15.5), has been enabled by the production of diachronic corpora.

10.8 Language learning and teaching

The early 1990s saw an increasing interest in applying the findings of corpus-based research to language pedagogy. This is apparent when one looks at the published literature. In addition to a large number of journal articles, at least nine single-authored or edited volumes have recently been produced on the topic of teaching and language corpora: Wichmann et al (1997), Kettemann and Marko (2002), Burnard and McEnery (2000), Aston (2001), Hunston (2002), Granger et al (2002), Tan (2002), Aston et al (2004) and Sinclair (2004b). These works cover a wide range of issues related to using corpora in language pedagogy, e.g. corpus-based language description, corpus analysis in classroom, and learner corpora (cf. Keck 2004).

In the opening chapter of *Teaching and Language Corpora* (Wichmann et al 1997), Leech noted that a convergence between teaching and language corpora was apparent. That convergence has three focuses, as noted by Leech (1997b): the direct use of corpora in teaching (teaching about, teaching to exploit, and exploiting to teach), the indirect use of corpora in teaching (reference publishing, materials development, and language testing), and further teaching-oriented corpus development (LSP corpora, L1 developmental corpora and L2 learner corpora). These three focuses of convergence are worthy of note.

Of these focuses, perhaps the most relevant for this book are ‘teaching about’ and ‘teaching to exploit’. ‘Teaching about’ means teaching corpus linguistics as an academic subject like other sub-disciplines of linguistics such as sociolinguistics or discourse analysis. Corpus linguistics has now found its way into the curricula for linguistics and language related degree programs at both postgraduate and undergraduate levels. ‘Teaching to exploit’ means providing students with ‘hands-on’ know-how, as emphasized in this book, so that they can exploit corpora for their own purposes. Once the student has acquired the necessary knowledge and techniques of corpus-based language study, learning activity may become student centred. If ‘teaching about’ is viewed as being associated typically with students of linguistics and languages, ‘teaching to exploit’ relates to students of all subjects which involve language study/learning. ‘Exploiting to teach’ means using a corpus-based approach to teaching language and linguistics courses, which would otherwise be taught using non-corpus-based methods. As for the indirect use of corpora in language teaching, we have already noted in units 10.2 and 10.3 that corpora have revolutionized reference publishing in a manner such that people who have never heard of a corpus are using the products of corpus research. As we will see later in this section, corpora also have a lot to offer in terms of syllabus design, materials development and language testing. Finally, teaching-oriented corpora are particularly useful in teaching

languages for specific purposes (LSP corpora) and in research on L1 (developmental corpora) and L2 (learner corpora) language acquisition. In the remainder of this section, we will introduce the uses of corpora in a number of areas in language pedagogy, including syllabus design and materials development, using corpora in classroom, teaching domain-specific language and professional communication, teacher training, language testing, as well as learner corpus research.

While corpora have been used extensively to provide more accurate descriptions of language use (see units 10.2-10.5), a number of scholars have also used corpus data directly to look critically at existing TEFL (Teaching English as a Foreign Language) syllabuses and teaching materials. Mindt (1996), for example, finds that the use of grammatical structures in textbooks for teaching English differs considerably from the use of these structures in L1 English. He observes that one common failure of English textbooks is that they teach 'a kind of school English which does not seem to exist outside the foreign language classroom' (1996: 232). As such, learners often find it difficult to communicate successfully with native speakers. A simple yet important role of corpora in language education is to provide more realistic examples of language usage. In addition, however, corpora may provide data, especially frequency data, which may further alter what is taught. For example, on the basis of a comparison of the frequencies of modal verbs, future time expressions and conditional clauses in corpora and their grading in textbooks used widely in Germany, Mindt (*ibid*) concludes that one problem with non-corpus-based syllabuses is that the order in which those items are taught in syllabuses 'very often does not correspond to what one might reasonably expect from corpus data of spoken and written English', arguing that teaching syllabuses should be based on empirical evidence rather than tradition and intuition with frequency of usage as a guide to priority for teaching (1996: 245-246; see discussion below).

Hunston (2002: 189) echoes Mindt suggesting that 'the experience of using corpora should lead to rather different views of syllabus design'. The type of syllabus she discusses extensively is a 'lexical syllabus', originally proposed by Sinclair and Renouf (1988) and outlined fully by Willis (1990). According to Sinclair and Renouf (1988: 148), a lexical syllabus would focus on '(a) the commonest word forms in a language; (b) the central patterns of usage; (c) the combinations which they usually form.' While the term may occasionally be misinterpreted to indicate a syllabus consisting solely of vocabulary items, a lexical syllabus actually covers 'all aspects of language, differing from a conventional syllabus only in that the central concept of organization is lexis' (Hunston 2002: 189). Sinclair (2000: 191) would say that the grammar covered in a lexical syllabus is 'lexical grammar', not 'lexico-grammar', which attempts to 'build a grammar and lexis on an equal basis.' Indeed, as Murison-Bowie (1996: 185) observes, 'in using corpora in a teaching context, it is frequently difficult to distinguish what is a lexical investigation and what is a syntactic one. One leads to the other, and this can be used to advantage in a teaching/learning context.' Sinclair and his colleagues' proposal for a lexical syllabus is echoed by Lewis (1993, 1997a, 1997b, 2000), who provides strong support for the lexical approach to language teaching.

While syllabus design and materials development are closely associated with what to teach, corpora have also provided valuable insights into how to teach. The issue of how to use corpora in the language classroom has been discussed extensively in the literature. With the corpus-based approach to language pedagogy, the traditional 'three P's' (Presentation – Practice – Production) approach to teaching may not be entirely suitable. Instead, the more exploratory approach of 'three I's' (Illustration –

Interaction – Induction) may be more appropriate, where 'illustration' means looking at real data, 'interaction' means discussing and sharing opinions and observations, and 'induction' means making one's own rule for a particular feature, which 'will be refined and honed as more and more data is encountered' (see Carter and McCarthy 1995: 155). This progressive induction approach is what Murison-Bowie (1996: 191) and Aston (1997) would call the interlanguage approach: partial and incomplete generalizations are drawn from limited data as a stage on the way towards a fully satisfactory rule. While the 'three I's' approach was originally proposed by Carter and McCarthy (1995) to teach spoken grammar, it may also apply to language education as a whole, in our view.

It is certainly clear that the teaching approach focusing on 'three I's' is in line with Johns' (1991) concept of 'data-driven learning (DDL)'. Johns was perhaps among the first to realize the potential of corpora for language learners (e.g. Higgins and Johns 1984). In his opinion, 'research is too serious to be left to the researchers' (1991: 2). As such, he argues that the language learner should be encouraged to become 'a research worker whose learning needs to be driven by access to linguistic data' (*ibid*). Indeed, as Kennedy (2003) observes, language learning is a process of learning 'explicit knowledge' with awareness, which requires a great deal of exposure to language data. Data-driven learning can be either teacher-directed or learner-led (i.e. discovery learning) to suit the needs of learners at different levels, but it is basically learner-centred. This autonomous learning process 'gives the student the realistic expectation of breaking new ground as a "researcher", doing something which is a unique and individual contribution' (Leech 1997b: 10).

Johns (1991) identifies three stages of inductive reasoning with corpora in the DDL approach: observation (of concordanced evidence), classification (of salient features) and generalization (of rules). The three stages roughly correspond to Carter and McCarthy's 'three I's'. The DDL approach is basically different from the 'three P's' approach in that the former is bottom-up induction whereas the latter is top-down deduction. The direct use of corpora and concordancing in the language classroom has been discussed extensively in the literature (e.g. Tribble 1990, 1997, 2000, 2003; Tribble and Jones 1990, 1997; Flowerdew 1993; Karpati 1995; Kettemann 1995, 1996; Wichmann 1995; Woolls 1998; Aston 2001; Osborne 2001), covering a wide range of issues including, for example, underlying theories, methods and techniques, and problems and solutions.

In addition to teaching English as a second or foreign language in general, a great deal of attention has been paid to domain-specific language use and professional communication (e.g. English for specific purposes and English for academic purpose). For example, Thurstun and Candlin (1997, 1998) explore the use of concordancing in teaching writing and vocabulary in academic English; Hyland (1999) studies the metadiscourse in introductory course books (see unit 14.3); Thompson and Tribble (2001) examine citation practices in academic text; Koester (2002) argues, on the basis of an analysis of the performance of speech acts in workshop conversations, for a discourse approach to teaching communicative functions in spoken English; Yang and Allison (2003) study the organizational structure in research articles in applied linguistics; Carter and McCarthy (2004) explore, on the basis of the CANCODE corpus, a range of social contexts in which creative uses of language are manifested; Hinkel (2004) compares the use of tense, aspect and the passive in L1 and L2 academic texts.

There are two other areas of language education in which corpora have recently been used: language teacher training and language testing. For learners to benefit from the

use of corpora, language teachers must first of all be equipped with a sound knowledge of the corpus-based approach. It is unsurprising to discover then that corpora have been used in training language teachers (e.g. Conrad 1999; Allan 1999, 2002; Seidlhofer 2000, 2002; O'Keeffe and Farr 2003). Another emerging area of language pedagogy which has started to use the corpus-based approach is language testing. Alderson (1996) envisaged the possible uses of corpora in this area: test construction, compilation and selection; test presentation; response capture; test scoring, and calculation and delivery of results. He concludes that '[t]he potential advantages of basing our tests on real language data, of making data-based judgments about candidates' abilities, knowledge and performance are clear enough. A crucial question is whether the possible advantages are born out in practice' (Alderson 1996: 258-259). The concern raised in Alderson's conclusion appears to have been addressed satisfactorily. Choi, Kim and Boo (2003) find that computer-based tests are comparable to paper-based tests. A number of corpus-based studies of language testing have been reported. For example, Coniam (1997) demonstrated how to use word frequency data extracted from corpora to generate cloze tests automatically. Kaszubski and Wojnowska (2003) presented a corpus-based program for building sentence-based ELT exercises – TestBuilder. The program can process raw and POS tagged corpora, tagged on the fly by a built-in part-of-speech tagger, and uses this as input for test material selection. Indeed, corpora have recently been used by major providers of test services for a number of purposes: 1) as an archive of examination scripts; 2) to develop test materials; 3) to optimize test procedures; 4) to improve the quality of test marking; 4) to validate tests; and 5) to standardize tests (cf. Ball 2001; Hunston 2002: 205). For example, the University of Cambridge Local Examinations Syndicate (UCLES) is active in both corpus development (e.g. Cambridge Learner Corpus, Cambridge Corpus of Spoken English, Business English Text Corpus and Corpus YLE Speaking Tests) and the analysis of native English corpora and learner corpora. At UCLES, native English corpora such as the BNC are used 'to investigate collocations, authentic stems and appropriate distractors which enable item writers to base their examination tasks on real texts' (Ball 2001: 7); the corpus-based approach is used to explore 'the distinguishing features in the writing performance of EFL/ESL learners or users taking the Cambridge English examinations' and how to incorporate these into 'a single scale of bands, that is, a common scale, describing different levels of L2 writing proficiency' (Hawkey 2001: 9); corpora are also used for the purpose of speaking assessment (Ball and Wilson 2002; Taylor 2003) and to develop domain-specific (e.g. business English) wordlists for use in test materials (Ball 2002; Horner and Strutt 2004).

One of the most exciting recent developments in corpus-based language studies has been the creation and use of learner corpora in language pedagogy and interlanguage studies. At the pre-conference workshop on learner corpora affiliated to the International Symposium of Corpus Linguistics 2003 held at the University of Lancaster, the workshop organizers, Tono and Meunier, observed that learner corpora are no longer in their infancy but are going through their nominal teenage years – they are full of promise but not yet fully developed. In language pedagogy, the implications of learner corpora have been explored for curriculum design, materials development and teaching methodology (cf. Keck 2004: 99). The interface between L1 and L2 materials has been explored. Meunier (2002), for example, argues that frequency information obtained from native speaker corpora alone is not sufficient to inform curriculum and materials design. Rather, 'it is important to strike a balance between frequency, difficulty and pedagogical relevance. That is exactly where

learner corpus research comes into play to help weigh the importance of each of these' (Meunier 2002: 123). Meunier also advocates the use of learner data in the classroom, suggesting that exercises such as comparing learner and native speaker data and analyzing errors in learner language will help students to notice gaps between their interlanguage and the language they are learning. Interlanguage studies based on learner corpora which have been undertaken so far focus on what Granger (2002) calls 'Contrastive Interlanguage Analysis (CIA)', which compares learner data and native speaker data, or language produced by learners from different L1 backgrounds. The first type of comparison typically aims to identify under or over-use of particular linguistic features in learner language while the second type aims to uncover L1 interference or transfer. In addition to CIA, learner corpora have also been used to investigate the order of acquisition of particular morphemes (see case study 3). Readers can refer to Granger et al (2002) for recent work in the use of learner corpora, and read Granger (2003b) for a more general discussion of the applications of learner corpora.

Before we close the discussion in this section, it is appropriate to address some objections to the use of corpora in language learning and teaching. While frequency and authenticity are often considered two of the most important advantages of using corpora, they are also the motivation for criticism of the corpus-based approach from language pedagogy researchers. For example, Cook (1998: 61) argues that corpus data impoverishes language learning by giving undue prominence to what is simply frequent at the expense of rarer but more effective or salient expressions. Widdowson (1990, 2000) argues that corpus data is authentic only in a very limited sense in that it is de-contextualized (i.e. traces of texts rather than discourse) and must be re-contextualized in language teaching. Nevertheless, it can also be argued reasonably that:

> on the contrary, using corpus data not only increases the chances of learners being confronted with relatively infrequent instances of language use, but also of their being able to see in what way such uses are atypical, in what contexts they do appear, and how they fit in with the pattern of more prototypical uses. (Osborne 2001: 486)

This view is echoed by Goethals (2003: 424), who argues that 'frequency ranking will be a parameter for sequencing and grading learning materials' because '[f]requency is a measure of *probability* of usefulness' and '[h]igh-frequency words constitute a core vocabulary that is useful above the incidental choice of text of one teacher or textbook author.' Hunston (2002:194-195) observes that '[i]tems which are important though infrequent seem to be those that echo texts which have a high cultural value', though in many cases 'cultural salience is not clearly at odds with frequency.' While frequency information is readily available from corpora, no corpus linguist has ever argued that the most frequent is most important. On the contrary, Kennedy (1998: 290) argues that frequency 'should be only one of the criteria used to influence instruction' and that '[t]he facts about language and language use which emerge from corpus analyses should never be allowed to become a burden for pedagogy.' As such, raw frequency data is often adjusted for use in a syllabus, as reported in Renouf (1987: 168). It would be inappropriate, therefore, for language teachers, syllabus designers, and materials writers to ignore 'compelling frequency evidence already available', as pointed out by Leech (1997b: 16), who argues that:

> Whatever the imperfections of the simple equation 'most frequent' = 'most important to learn', it is difficult to deny that frequency information becoming

available from corpora has an important empirical input to language learning materials.

If we leave objections to the use of frequency data to one side, Widdowson (1990, 2000) also questions the use of authentic texts in language teaching. In his opinion, authenticity of language in the classroom is 'an illusion' (1990: 44) because even though corpus data may be authentic in one sense, its authenticity of purpose is destroyed by its use with an unintended audience of language learners (cf. Murison-Bowie 1996: 189; see units 12.2 and 16.2 for further discussion). The implication of Widdowson's argument is that only language produced for imaginary situations in the classroom is 'authentic'. However, as argued by Fox (1987), invented examples often do not reflect nuances of usage. That is perhaps why, as Mindt (1996: 232) observes, students who have been taught 'school English' cannot readily cope with English used by native speakers in real conversation. As such, Wichmann (1997: xvi) argues that in language teaching, 'the preference for "authentic" texts requires both learners and teachers to cope with language which the textbooks do not predict.'

In conclusion, it is our view that corpora will not only revolutionize the teaching of subjects such as grammar in the 21st century (see Conrad 2000), they will also fundamentally change the ways we approach language pedagogy, including both what is taught and how it is taught.

We have so far reviewed the applications of corpora in major areas of language studies. The sections that follow will discuss other areas which have started to use corpus data.

10.9 Semantics

We have already touched upon semantics at the lexical level when we discussed semantic prosody/preference and pattern meanings in unit 10.2. But corpora are also more generally important in semantics in that they provide objective criteria for assigning meanings to linguistic items and establish more firmly the notions of fuzzy categories and gradience (see McEnery and Wilson 2001: 112-113), as demonstrated by Mindt (1991). This section considers semantics in more general terms, with reference to the two functions of corpus data as identified above by McEnery and Wilson (*ibid*).

Corpora have been used to detect subtle semantic distinctions in near synonyms. Tognini-Bonelli (2001: 35-39), for example, finds that *largely* can be used to introduce cause and reason and co-occur with morphological and semantic negatives, but *broadly* cannot; yet while *broadly* can be used as a discourse disjunct for argumentation and to express similarity or agreement, *largely* cannot. Gilquin (2003) seeks to combine the corpus-based approach with the cognitive theory of frame semantics in her study of the causative verbs *GET* and *HAVE*. The study shows that the two verbs have a number of features in common but also exhibit important differences. For example, both verbs are used predominantly with an animate causer. Yet while with *GET* the causee is most often animate, the frequencies of animate and inanimate causees are very similar with *HAVE*. Nevertheless, when causees are expressed as an object (i.e. not demoted), the proportion of animates and inanimates is reversed, with a majority of animates with *GET* and a predominance of inanimates with *HAVE*. While Tognini-Bonelli (2001) and Gilquin (2003) can be considered as examples of assigning meanings to linguistic items, Kaltenböck (2003) further exemplifies the role of corpus data in providing evidence for fuzzy categories and gradience in his study of the syntactic and semantic status of anticipatory *it*. Kaltenböck finds that both the

approach which takes anticipatory *it* to have an inherent cataphoric function (i.e. referring *it*) and the view that it is a meaningless, semantically empty dummy element (i.e. prop *it*) as have been proposed previously are problematic as they fail to take into account the actual use of anticipatory *it* in context. The analysis of instances actually occurring in ICE-GB showed very clearly that delimiting the class of *it*-extraposition (and hence anticipatory *it*) is by no means a matter of 'either-or' but has to allow for fuzzy boundaries (Kaltenböck 2003: 236): 'anticipatory *it* takes an intermediate position between prop *it* and referring *it*, all of which are linked by a scale of gradience specifying their scope of reference (wide vs. narrow)' (Kaltenböck 2003: 235). The functionalist approach taken by Kaltenböck (2003) is in sharp contrast to the purely formalist approach which, relying exclusively on conceptual evidence, identifies anticipatory *it* as meaningless. Kaltenböck argues that:

> the two approaches operate with different concepts of meaning: a formalist will be interested in the meaning of a particular form as represented in the speaker's competence, while for the view expressed here 'meaning' not only resides in isolated items but is also the result of their interaction with contextual factors. (Kaltenböck 2003: 253)

Let us now turn to a core area of semantics – aspect. According to Smith (1997: 1), 'aspect is the semantic domain of the temporal structure of situations and their presentation.' Aspect has traditionally been approached without recourse to corpus data. More recently, however, corpus data has been exploited to inform aspect theory. Xiao and McEnery (2004a), for example, have developed a corpus-based two-level model of situation aspect, in which situation aspect is modelled as verb classes at the lexical level and as situation types at the sentential level. Situation types are the composite result of the rule-based interaction between verb classes and complements, arguments, peripheral adjuncts and viewpoint aspect at the nucleus, core and clause levels. With a framework consisting of a lexicon, a layered clause structure and a set of rules mapping verb classes onto situation types, the model was developed and tested using an English corpus and a Chinese corpus. The model has not only provided a more refined aspectual classification and given a more systematic account of the compositional nature of situation aspect than previous models, but it has also shown that intuitions are not always reliable (e.g. the incorrect postulation of the effect of external arguments). We will return to discuss aspect in unit 15.3 of Section B and case study 6 of Section C of this book.

The examples cited above demonstrate that corpora do have a role to play in the study of meaning, not only at the lexical level but in other core areas of semantics as well. Corpus-based semantic studies are often labour-intensive and time-consuming because many semantic features cannot be annotated automatically (consider e.g. causer vs. causee and animate vs. inanimate in Gilquin's (2003) study of causative *GET/HAVE*). Yet the interesting findings from such studies certainly make the time and effort worthwhile. In the next section we will review the use of corpora in pragmatics.

10.10 Pragmatics

As noted in unit 4.4.6, pragmatics is strongly – though not exclusively – associated with spoken discourse. This is hardly surprising considering that written registers tend to be referentially explicit whereas spoken registers typically 'permit extensive reference to the physical and temporal situation of discourse' (Biber 1988: 144). In Kennedy's (1998: 174) words, 'What we say and how we say it is influenced by who we are talking to and where the interaction is taking place.' Until the mid-1990s

corpus-based pragmatic studies were severely constrained because there was only one reasonably large, publicly available corpus which was sufficiently marked up for prosodic and discourse features, the London-Lund Corpus (i.e. LLC, see unit 7.5) (cf. Kennedy 1998: 174). It is, therefore, unsurprising that earlier corpus-based work on pragmatics was based fairly exclusively on the LLC. For example, Svartvik (1980), on the basis of a sample of 45,000 words from the LLC, found that the discourse marker *well* is an important device which allows the speaker time to think online while keeping a turn in conversation. The pragmatic functions of *well* include polite disagreement, qualified refusal, reinforcement, modification, indirect and partial answers, and delaying tactics. Aijmer's (1987) study of *oh* and *ah* in a 170,000-word sample from the LLC provides a full account of the major pragmatic functions of the two 'disjunct markers' (Jefferson 1978: 221). Tottie (1991), on the basis of a comparison of the LLC and the Santa Barbara Corpus of Spoken American English (i.e. SBCSAE, see unit 7.5), finds that American speakers use backchannel agreement markers (e.g. *yeah*, *sure* and *right*) three times as frequently as British speakers.

Aijmer (1987: 63) notes that one of the pragmatic functions of *oh* and *ah* is to signal 'a shift or development to something not foreseen by the speaker', thus construing what comes afterwards as 'topically not coherent' (Jefferson 1978: 221). Discourse markers such as *anyway*, *however* and *still* help to establish coherence in spoken discourse (see Lenk 1995, 1998a, 1998b). Lenk (1998b), for example, uses the LLC and SBCSAE corpora to investigate how *however* and *still* are involved in the process of achieving conversational coherence. It was found that 'the function of both of these discourse markers is to connect parts of the discourse that are not immediately adjacent, or that are not topically related' (Lenk 1998b: 256). Nevertheless, while '*however* closes digressions that are relevant to the development of the main topic, or that bear interactional significance', '*still* closes off subjective comments within a quasi-objective narration or presentation of facts' (Lenk 1998b: 256). It is also interesting to note that *however* is used as a discourse marker only in British English (Lenk 1998b: 251).

Spoken language, and face-to-face conversation in particular, takes place on the basis of a shared context, avoids elaboration or specification of reference, and reflects the needs for real-time processing (Leech 2000). It is, therefore, hardly surprising that conversation is more vague than most written genres. Vagueness is pervasive in conversation where it plays an important role. The most obvious reason for using vague expressions is uncertainty at the time of speaking. In this case, vagueness allows speakers to maintain fluency even though they lack information about a given quantity, quality or identity, or, when such information is potentially available, they cannot access or process it in time. However, speakers may still choose to be vague even when they could in principle be more precise. This is because vague language can serve a number of pragmatic functions. Jucker, Smith and Lüdge (2003), for example, analyze the vague additives (i.e. approximators, downtoners, vague category identifiers and shields) and instances of lexical vagueness (i.e. vague quantifying expressions, vague adverbs of frequency, vague adverbs of likelihood, and placeholder words) in a corpus of semi-controlled spoken interactions between students in California. They find that vagueness is an interactional strategy which plays an important role in managing conversational implicature. First, vague expressions may serve as focusing devices, directing the hearer's attention to the most relevant information. Second, vague expressions of quantities provide information about the significance of the quantity and may provide a reference point in terms of a scale. Third, vague expressions may also convey several aspects of propositional

attitude (e.g. conveying different levels of certainty regarding the propositional content, conveying the newsworthiness or expectedness of a statement, and conveying evaluative meaning). Finally, vague expressions may serve various social functions (serving as politeness strategies, softening implicit complaints and criticisms, and providing a way of establishing a social bond). As such, vague language helps to 'guide the hearer towards the best interpretation of the speaker's intention' (Jucker, Smith and Lüdge 2003: 1766).

Similarly, Drave (2002) studies vague language (VL) in intercultural conversations. The corpus he used was the Hong Kong Corpus of Conversational English (HKCCE), a corpus consisting of 98,310 words of native speaker English (NSE) and 84,208 words of English produced by native speakers of Cantonese (NSC). Drave (2002: 27) finds that vague language can be used in naturally occurring conversations strategically for promoting politeness and intersubjectivity and for managing asymmetries of knowledge, particularly in intercultural interaction. It was found that while quantitatively NSE seems to be 'vaguer' than NSC, the two groups do not differ qualitatively, 'with very few VL items being used exclusively by one group or the other and the rank orders of VL items of the most frequent items virtually identical' (Drave 2002: 29).

McEnery, Baker and Cheepen (2002) explored the relationship between directness and lexical markers of politeness with reference to operator requests to 'hold the line', using a corpus of telephone-based transactional dialogues. They found that of the various types of request strategies (bare imperative, deletion, conditional *if*, prediction and question), only the bare imperatives were unambiguously direct while all of the other types were to some extent indirect imperatives. It is also interesting to note that while bare imperatives are the most common request strategy, they are typically softened by mitigators such as *please* and *just* (McEnery, Baker and Cheepen 2002: 64-65).

While politeness strategies are particularly important in transactional dialogues as explored by McEnery, Baker and Cheepen (2002), conversation is not always polite. Complaining is unavoidable. Laforest (2002) presents an interesting study which characterizes 'the complaint/complaint-response sequence in everyday conversations between people who are on intimate terms' (Laforest 2002: 1596), in this case, peer family members (i.e. husbands/wives and brothers/sisters). The complaints exchanged between people who are not peers (i.e. parents vs. children) were excluded in order to neutralize the variation introduced by a difference in hierarchical position between interactants. The data used in this study was taken from a corpus of about 50 hours of family conversations recorded in Montréal. The complaints analyzed in this study illustrated the numerous ways in which speakers expressed dissatisfaction with the behaviour of people close to them. They had preferential realization patterns that could be linked in part to the intimacy of the relationship between the interactants: in many ways, they were uttered without the special precautions generally associated with face-threatening acts (FTAs) outside the private sphere (Laforest 2002: 1617-1618). Laforest found that while the complainees most often reject the blame levelled at them, well characterized arguments are virtually absent from the corpus; the entry into the argument is negotiated in the speech turns that follow the complaint/response sequence, and the argument only breaks out if the complainer questions the value of the complainee's response. The study also shows that both complainer and complainee use various strategies for avoiding an argument and, more often than not, succeed in doing so (Laforest 2002: 1596).

Nowadays, pragmatic studies are more varied than before. One area of increasing interest is historical pragmatics which, like general diachronic studies, depends heavily upon corpus data. For example, Arnovick (2000) examines the speech event of parting, focusing on the development of *Goodbye*, which was originally an explicit blessing *God be with you*. She finds that the formal development from *God be with you* to *Goodbye* is linked to functional shifts. In the English Drama section of the Chadwyck-Healey corpus, the original form, which appeared in closing sections of dialogue in Early Modern English, was used as a blessing as well as a greeting at parting while the contracted form became stronger in the force of the polite closing greeting. Arnovick's study shows that the end of the 17th century and the beginning of the 18th century marked a crucial period during which the blessing declined and the closing form *Goodbye* increased in frequency. Jucker and Taavitsainen (2000) undertake a diachronic analysis of one particular speech act, i.e. insults, through the history of English on the basis of a corpus composed of both literary and non-literary data. Their analysis of written materials of the past periods indicates an evident bias towards the conventionalized insults. Most early examples are found in literary texts, which reflect generic conventions of the time and the culture that gave rise to these literary forms. Jacobsson (2002) used a pilot version of the Corpus of English Dialogues (CED, see unit 7.7) to study gratitude expressions such as *Thank you* and *Thanks* in Early Modern English. The author found that while these expressions themselves were probably the same in the Early Modern period as they are today, they 'had not developed the discourse-marking features of today's British English; nor is it possible to see the complex patterns of thanking in different turn-positions in the CED material' (Jacobsson 2002: 78). Biber (2004) explores, on the basis of the ARCHER corpus (see unit 7.7), the patterns of historical change in the preferred devices used to mark stance across the past three centuries. He finds that of the grammatical categories marking stance, modal verbs have undergone a decrease in use whereas other devices such as semi-modals, stance adverbials, and stance complement clause constructions have all increased in use across the historical periods in his study (Biber 2004: 129).

Pragmatics is an area in which more and more corpus data is being used. However, meanings dependent upon pragmatics cannot easily be detected automatically. As in semantics, the automatic extraction is not likely unless the corpora used for such studies have been annotated manually with the required analyses.

10.11 Sociolinguistics

While sociolinguistics has traditionally been based upon empirical data, the use of standard corpora in this field has been limited. The expansion of corpus work in sociolinguistics appears to have been hampered by three problems: the operationalization of sociolinguistic theory into measurable categories suitable for corpus research, the lack of sociolinguistic metadata encoded in currently available corpora, and the lack of sociolinguistically rigorous sampling in corpus construction (cf. McEnery and Wilson 2001: 116).

Corpus-based sociolinguistic studies have so far largely been restricted to the area of gender studies at the lexical level. For example, Kjellmer (1986) compared the frequencies of masculine and feminine pronouns and lexical items *man/men* and *woman/women* in the Brown and LOB corpora. It was found that female items are considerably less frequent than male items in both corpora, though female items were more frequent in British English. It is also interesting to note that female items were

more frequent in imaginative (especially romantic fiction) than informative genres. Sigley (1997) found some significant differences in the distribution patterns of relative clauses used by male and female speakers/writers at different educational levels in New Zealand English. Caldas-Coulthard and Moon (1999) found on the basis of a newspaper corpus that women were frequently modified by adjectives indicating physical appearance (e.g. *beautiful*, *pretty* and *lovely*) whereas men were frequently modified by adjectives indicating importance (e.g. *key*, *big*, *great* and *main*). Similarly, Hunston (1999b) noted that while *right* is used to modify both men and women, the typical meaning of *right* co-occurring with men is work-related ('the right man for the job') whereas the typical meaning of *right* co-occurring with women is man-related ('the right woman for this man'). Hunston (2002: 121) provided two alternative explanations for this: that women are perceived to be 'less significant in the world of paid work', or that 'men are construed as less emotionally competent because they more frequently need "the right woman" to make their lives complete.' In either case, women are not treated identically (at least in linguistic terms) in society. Holmes (1993a, 1993b, 1993c, 1997) has published widely on sexism in English, e.g. the epicene terms such as *-man* and *he*, gender-neutral terms like *chairperson*, and sexist suffixes like *-ess* and *-ette*. Holmes and Sigley (2002), for example, used Brown/LOB and Frown/FLOB/WWC to track social change in patterns of gender marking between 1961 and 1991. They found that

> while women continue to be the linguistically marked gender, there is some evidence to support a positive interpretation of many of the patterns identified in the most recent corpora, since the relevant marked contexts reflect inroads made by women into occupational domains previously considered as exclusively male. (Holmes and Sigley 2002: 261)

While Holmes and Sigley (2002) approached gender marking from a diachronic perspective, Baranowski (2002) approached the issue in a contrastive context. Baranowski investigated the epicene pronominal usage of *he*, *he or she* and singular *they* in two corpora of written English (one for British English and the other for American English), and found that the traditional form *he* is no longer predominant while singular *they* is most likely to be used. The form *he or she* is shown to be used rather rarely. The study also reveals that American writers are more conservative in their choice of a singular epicene pronoun. In gender studies like these, however, it is important to evaluate and classify usages in context (cf. Holmes 1994), which can be time-consuming and hard to decide sometimes.

In addition to sexism, femininity and sexual identity are two other important areas of gender studies which have started to use corpus data. For example, Coates (1999) used a corpus of women's (and girls') 'backstage talk' to explore their self-presentation in contexts where they seem most relaxed and most off-record, focusing on 'those aspects of women's backstage performance of self which do not fit prevailing norms of femininity' (Coates 1999: 65). Coates argued that the backstage talk 'provides women with an arena where norms can be subverted and challenged and alternative selves explored' while acknowledging 'such talk helps to maintain the heteropatriarchal order, by providing an outlet for the frustrations of frontstage performance.' Thorne and Coupland (1998: 234) studied, on the basis of a corpus of 200 lesbian and gay male dating advertisement texts, a range of discursive devices and conventions used in formulating sexual/self-gendered identities. They also discussed these discourse practices in relation to a social critique of contemporary gay attitudes, belief and lifestyles in the UK. Baker (2004) undertook a corpus-based keyword analysis of the debates over a Bill to equalize the age of sexual consent for

gay men with the age of consent for heterosexual sex at sixteen years in the House of Lords in the UK between 1998 and 2000 (see unit 21.5 for further discussion of keywords). Baker's analysis uncovered the main lexical differences between oppositional stances and helped to shed new light on the ways in which discourses of homosexuality were constructed by the Lords. For example, it was found that *homosexual* was associated with acts whereas *gay* was associated with identities. While those who argued for the reform focused on equality and tolerance, those who argued against it linked homosexuality to danger, ill health, crime and unnatural behaviour.

While corpus-based sociolinguistic research has focused on language and gender, corpora have also started to play a role in a wide range of more general issues in sociolinguistics. For example, Banjo (1996) discussed the role that ICE-Nigeria is expected to play in language planning in Nigeria; Florey (1998: 207) drew upon a corpus of incantations 'in order to address the issue of the extent to which specialised sociocultural and associated linguistic knowledge persists in a context of language shift'; Puchta and Potter (1999) analyzed question formats in a corpus of German market research focus groups (i.e. 'a carefully planned discussion designed to obtain perceptions on a defined area of interest in a permissive, non-threatening environment', see Krueger 1994: 6) in an attempt 'to show how elaborate questions [i.e. 'questions which include a range of reformulations and rewordings' (Puchta and Potter 1999: 314)] in focus groups are organized in ways which provide the kinds of answers that the focus group moderators require' (Puchta and Potter 1999: 332); de Beaugrande (1998: 134) drew data from the Bank of English to show that terms like *stability* and *instability* are not 'self-consciously neutral', but rather they are socially charged to serve social interests. Dailey-O'Cain (2000) explored the sociolinguistic distribution (sex and age) of focuser *like* (as in *And there were **like** people blocking, you know?*) and quotative *like* (as in *Maya's **like**, 'Kim come over here and be with me and Brett'*) as well as attitudes towards these markers.

In a more general context of addressing the debate over ideal vs. real language, de Beaugrande (1998: 131) argues that sociolinguistics may have been affected, during its formative stages, as a result of the long term tradition of idealizing language and disconnecting it from speech and society. Unsurprisingly, sociolinguistics has traditionally focused on phonological and grammatical variations in terms of 'features and rules' (de Beaugrande 1998: 133). He observes that the use of corpus data can bring sociolinguistics 'some interesting prospects' (de Beaugrande 1998: 137) in that '[r]eal data also indicate that much of the socially relevant variation within a language does not concern the phonological and syntactic variations' (de Beaugrande 1998: 133). In this sense, the

> corpus can help sociolinguistics engage with issues and variations in usage that are less tidy and abstract than phonetics, phonology, and grammar, and more proximate to the socially vital issues of the day […] Corpus data can help us monitor the ongoing collocational approximation and contestation of terms that refer to the social conditions themselves and discursively position these in respect to the interests of various social groups. (de Beaugrande 1998: 135)

With the increasing availability of corpora which encode rich sociolinguistic metadata (e.g. the BNC), the corpus-based approach is expected to play a more important role in sociolinguistics. To give an example of this new role, readers will have an opportunity to explore, in case study 4 in Section C, the patterns of swearing in modern British English along such dimensions as sex, age, social class of speakers and writers encoded in the BNC.

10.12 Discourse analysis

Closely allied with sociolinguistics is discourse analysis (DA), especially critical discourse analysis (CDA), which is mainly concerned with the studies of ideology, power and culture (cf. Fairclough 1995). While both corpus linguistics and DA rely heavily on real language, Leech (2000: 678-680) observes that there is 'a cultural divide' between the two: while DA emphasizes the integrity of text, corpus linguistics tends to use representative samples; while DA is primarily qualitative, corpus linguistics is essentially quantitative; while DA focuses on the contents expressed by language, corpus linguistics is interested in language *per se*; while the collector, transcriber and analyst are often the same person in DA, this is rarely the case in corpus linguistics; while the data used in DA is rarely widely available, corpora are typically made widely available. It is also important to note that some terms used in DA are defined differently from corpus linguistics. Apart from *genre* as noted previously, for example, *keywords* in DA refers to words that have a particular significance in a given discourse. The cultural divide, however, is now diminishing. McEnery and Wilson (2001: 114) note that there are some important 'points of contact' between DA and corpus linguistics: the common computer-aided analytic techniques, and the great potential of standard corpora in DA as control data. Because the corpus-based approach tends to obscure 'the character of each text as a text' and 'the role of the text producer and the society of which they are a part' (Hunston 2002: 110), some DA authors have avoided using corpus data. For example, Martin (1999: 52) argues that analyzing a lot of text from a corpus simultaneously would force the analyst to lose 'contact with text.' Yet Stubbs (1997) and de Beaugrande (1999, 2001), among many others, have insisted that corpora are indeed useful for studies of this kind.

Specialized corpora are particularly useful in discourse analysis and most of the recently published studies of ideology and culture are based on specialized corpora. Political discourse is perhaps the most important and most widely used data in discourse analysis. This is perhaps because politics is '[o]ne area of social life in which the increasing salience of discourse has been especially apparent' (Johnson, Culpeper and Suhr 2003: 41). For example, Sotillo and Starace-Nastasi (1999) undertook a critical discourse analysis on the basis of a corpus of 123 Letters to the Editors (LEs) of two weekly newspapers written by candidates for political office, their supporters, and opponents in an American working class town. They found that gender and class markers were salient in the discourse of LEs. With regard to class, there is an antagonistic dialogue between residents of the third ward (working class) and those of the second and first wards (middle class): middle-class residents of the first and second wards remain unsympathetic to the concerns of third-ward residents, especially to their claims of a deteriorating quality of life. With respect to the saliency of gender in LEs, qualitative differences were found between males and females in writing style, lexical and syntactic choices, and tone of communication. For example, men used more qualifiers and intensifiers than women, and women writers of LEs were often less confrontational and more conciliatory than their male counterparts in their criticism of those in power.

Teubert (2000) studied the language of Euroscepticism in Britain on the basis of a corpus of texts downloaded from websites which take an antagonistic attitude towards the European Union. Corpus analysis techniques like collocation and phraseology enabled Teubert to make explicit what was implied but not stated by Eurosceptics: only Britain in the whole of Europe is a true democracy with a truly accountable

government (Teubert 2000: 76-77). Similarly, Fairclough's (2000) comparative analysis of keywords (in the sense as used in corpus linguistics) in a corpus of the British Prime Minister Blair's speeches and other documents from New Labour and a corpus of documents from Old Labour Party showed that the party has changed its ideology, as reflected by its language.

Johnson, Culpeper and Suhr (2003) explored discourses of political correctness ('PC') in a corpus of articles gathered from three broadsheet newspapers in the UK between 1994 and 1999. Their frequency and (statistically defined) keyword analyses showed that while the overall frequency of so-called 'PC'-related terms ('political correctness', 'politically correct', etc.) generally declined in the five-year period, there was an interesting link between the frequency of such terms and the ways in which they have been drawn upon as a means of framing debates over Blair and the Labour Party throughout the period in question.

Saraceni (2003) analyzed two corpora of interviews and speeches related to the war in Iraq in an attempt 'to understand the extent to which, at least in linguistic terms, the ideas of Blair and Bush may not be as alike as one might be tempted to believe.' His analysis revealed some important differences in the ways in which Blair and Bush use language and in what they actually say. While Bush's rhetoric is typically right wing, Blair's discourse is more enigmatic, lacking many of the characteristics of right-wing rhetoric but not typical of left-wing rhetoric either (Saraceni 2003: 12). The marked contrast between words and actions in this case is a good example of a complex issue which the corpus-based approach alone cannot resolve.

Partington (2003) provides a full, corpus-based account of the discourse of White House press briefing, in an attempt 'to show how it is possible to use concordance technology and the detailed linguistic evidence available in corpora to enhance the study of the discourse features of a particular genre of the language' (Partington 2003: 3). The major corpus resource used by him is a corpus consisting of 48 briefings, amounting to approximately 250,000 words. The work presented in Partington (*ibid*) represents an unusual contribution to corpus-based discourse analysis because a large part of the book is devoted to devising 'a suitable methodology to study features of interaction in large bodies of texts, in corpora' (Partington 2003: 5). Such methodologies are particularly important in the context of most studies in discourse analysis undertaken so far having been based on corpora of a number of single texts (e.g. Pardo 2001).

In addition to political discourse, corpora have been used in analyzing a number of other types of discourse, for example, academic discourse (e.g. Piper 2000), business discourse (e.g. Koller 2004), everyday demotic discourse (Carter and McCarthy 2004), legal discourse (e.g. Graham 2001), media discourse (e.g. Downs 2002; Moore 2002; Pan 2002; Page 2003), medical discourse (e.g. Salager-Meyer, Ariza and Zambrano 2003), and workshop discourse (e.g. Holmes and Marra 2002).

The works reviewed so far are all based mainly on specialized corpora, though some of them (e.g. Piper 2000; Johnson, Culpeper and Suhr 2003; Partington 2003) have used general corpora such as the BNC for comparative purposes. In contrast, there has been far less work in discourse analysis that is based directly on general corpora. There are a number of reasons for this. First, most discourse analysts prefer to study whole texts – general corpora are typically composed of samples. Second, with a few exceptions (e.g. the BNC), most general corpora have not encoded variables required for discourse analysis (e.g. metadata relating to the language producer). Third, most general corpora have not included spoken data for spoken discourse analysis yet, as Partington (2003: 262) observes, some linguists only consider spoken language as

discourse. Finally, the field of discourse analysis has historically been accustomed to analyzing a small number of single texts whereas general corpora provide a much larger number of texts. There are, however, a number of studies which are based on general corpora. For example, Stubbs (1996) gives numerous examples of what he calls 'cultural keywords' in the Bank of English; de Beaugrande (1999) compared the ideologies as reflected by 'liberal' and its derivatives (e.g. 'liberalism', 'liberalization') in the UK and the US-based corpus resources as well as in the Corpus of South African English (i.e. CSAE, which was originally developed as part of the ICE corpus).

In conclusion, while the corpus-based approach to discourse analysis is still in its infancy, corpora (either specialized or general) do present a real opportunity to discourse analysis, because the automatic analysis of a large number of texts at one time 'can throw into relief the non-obvious in a single text' (Partington 2003: 7). As de Beaugrande (1999) comments:

> Obviously, the methods for doing a 'critical discourse analysis' of corpus data are far from established yet. Even when we have examined a fairly large set of attestations, we cannot be certain whether our own interpretations of key items and collocations are genuinely representative of the large populations who produced the data. But we can be fairly confident of accessing a range of interpretative issues that is both wider and more precise than we could access by relying on our own personal usages and intuitions. Moreover, when we observe our own ideological position in contest with others, we are less likely to overlook it or take it for granted. (de Beaugrande 1999: 287)

10.13 Stylistics and literary studies

Style is closely allied to registers/genres and dialects/language varieties (see unit 14), because stylistic shifts in usage may be observed with reference to features associated with either particular situations of use or particular groups of speakers (cf. Schilling-Estes 2002: 375). In this section, we will consider only what Carter (1999: 195) calls 'literary language'. Literariness is typically present in, but not restricted to literary texts. However, given that most work in stylistics focuses upon literary texts, the accent of this section will fall upon literary studies.

Stylisticians are typically interested in individual works by individual authors rather than language or language variety as such. Hence while they may be interested in computer-aided text analysis, the use of corpora in stylistics and literary studies appears to be limited (cf. McEnery and Wilson 2001: 117). Nevertheless, as we will see shortly, corpora and corpus analysis techniques are useful in a number of ways: the study of prose style, the study of individual authorial styles and authorship attribution, literary appreciation and criticism, teaching stylistics, and the study of literariness in discourses other than literary texts have all been the focus of corpus-based study.

As noted in unit 4.4.7, one of the focuses in the study of prose stylistics is the representation of people's speech and thoughts. Leech and Short (1981) developed an influential model of speech and thought presentation, which has been used by many scholars for literary and non-literary analysis (e.g. McKenzie 1987; Roeh and Nir 1990; Simpson 1993). The model was tested and further refined in Short, Semino and Culpeper (1996), and Semino, Short and Culpeper (1997). Readers can refer to unit 4.4.7 for a description of the speech and thought categories in the model. Using this model, Semino, Short and Wynne (1999) studied hypothetical words and thoughts in contemporary British narratives; Short, Semino and Wynne (2002) explored the

notion of faithfulness in discourse presentation; Semino and Short (2004) provide a comprehensive account of speech, thought and writing presentation in fictional and non-fictional narratives.

The corpus-based approach has also been used to study the authorial styles of individual authors. Corpora used in such studies are basically specialized. For example, if the focus is on the stylistic shift of a single author, the corpus consists of their early and later works, or works of theirs that belong to different genres (e.g. plays and essays); if the focus is on the comparison of different authorial styles, the corpus then consists of works by the authors under consideration. However, as Hunston (2002: 128) argues, using a large, general corpus can provide 'a means of establishing a norm for comparison when discussing features of literary style.' The methodology used in studying authorial styles often goes beyond simple counting; rather it typically relies heavily upon sophisticated statistical approaches such as MF/MD (see unit 10.4; e.g. Watson 1994), Principal Component Analysis (e.g. Binongo and Smith 1999a), and multivariate analysis (or more specifically, cluster analysis, e.g. Watson 1999; Hoover 2003b). The combination of stylistics and computation and statistics has given birth to a new interdisciplinary area referred to as 'stylometry' (Holmes 1998; Binongo and Smith 1999b), 'stylometrics' (Hunston 2002: 128), 'computational stylistics' (Merriam 2003), or 'statistical stylistics' (Hoover 2001, 2002).

Watson (1994) applied Biber's (1988) MF/MD stylistic model in his critical analysis of the complete prose works of the Australian Aboriginal author Mudrooroo Nyoongah to explore a perceived diachronic stylistic shift. He found that Nyoongah has shifted in style towards a more oral and abstract form of expression throughout his career and suggested that this shift 'may be indicative of Nyoongah's steadily progressive identification with his Aboriginality' (Watson 1994: 280). In another study of Nyoongah's early prose fiction (five novels), Watson (1999) used cluster analysis to explore the notion of involvement, more specifically *eventuality* (certainty vs. doubt) and *affect* (positive vs. negative). The analysis grouped *Wildcat*, *Sand* and *Doin* into one cluster and grouped *Doctor* and *Ghost* into another cluster, which represent two very distinct styles. The first cluster is more affective and representative of informal, unplanned language, using more certainty adverbs, certainty verbs and expression of affect; in contrast, the second cluster is more typical of more structured, integrated discourse, highlighted by a greater use of adjectives, in particular doubt adjectives and negative affect adjectives, and a very low expression of affect.

Binongo and Smith (1999a, 1999b) applied Principal Component Analysis in their studies of authorial styles. In Binongo and Smith (1999b), for example, the authors studied the distribution of 25 prepositions in Oscar Wilde's plays and essays. They found that when the plays and essays are brought into a single analysis, the difference in genre predominates over other factors, though the distinction is not clear-cut, with a gradual change from plays to essays (Binongo and Smith 1999b: 785-786).

In addition to stylistic variation, authorship attribution is another focus of literary stylistics. In a series of papers published in *Literary and Linguistic Computing*, Hoover (2001, 2002, 2003a, 2003b) tested and modified cluster analysis techniques which have traditionally been used in studies of stylistic variation and authorship attribution. Hoover (2001) noted that cluster analyses of frequent words typically achieved an accuracy rate of less than 90% for contemporary novels. Hoover (2002) found that when frequent word sequences were used instead of frequent words, or in addition to them, in cluster analyses, the accuracy often improved, sometimes drastically. In Hoover (2003a), the author compared the accuracies when using

frequent words, frequent sequences and frequent collocations, and found that cluster analysis based on frequent collocations provided a more accurate and robust method for authorship attribution. Hoover (2003b) proposed yet another modification to traditional authorship attribution techniques to measure stylistic variation. The new approach takes into consideration locally frequent words, a modification which is justified when one considers that authorship attribution focuses on similarities persisting across differences whereas the study of style variation focuses on variations of authorial style. Lexical choice is certainly part of authorial style. The modified approach has achieved improved results on some frequently studied texts, including Orwell's *1984* and Golding's *The Inheritors*. Readers can refer to Haenlein (1999) for a full account of the corpus-based approach to authorship attribution.

Authorship is only one of the factors which affect stylistic variation. Merriam (2003) demonstrated, on the basis of 14 texts by three authors, that three other factors, proposed originally by Labbé and Labbé (2001), namely the vocabulary of the period, treatment of theme and genre, also contributed to intertextual stylistic variation.

Louw (1997: 240) observed that '[t]he opportunity for corpora to play a role in literary criticism has increased greatly over the last decade.' He reported on a number of examples from his students' projects which showed that 'corpus data can provide powerful support for a reader's intuition' on the one hand while at the same time providing 'insights into aspects of "literariness", in this case the importance of collocational meaning, which has hitherto not been thought of by critics' (Louw 1997: 247). Likewise, Jackson (1997) provided a detailed account of how corpora and corpus analysis techniques can be used in teaching students about style.

While we have so far been concerned with literary texts, literariness is not restricted to literature, as noted at the beginning of this section. Carter (1999) explored, using the CANCODE corpus, the extent to which typically non-literary discourses like everyday conversation can display literary properties. He concluded that:

> The opposition of literary to non-literary language is an unhelpful one and the notion of literary language as a yes/no category should be replaced by one which sees literary language as a continuum, a cline of literariness in language use with some uses of language being marked as more literary than others. (Carter 1999: 207)

10.14 Forensic linguistics

The final example of the use of corpora and corpus analysis techniques which we will consider in this section is forensic linguistics, the study of language related to court trials and linguistic evidence. This is perhaps the most applied and exciting area where corpus linguistics has started to play a role because court verdicts can very clearly affect people's lives. Corpora have been used in forensic linguistics in a number of ways, e.g. in general studies of legal language (e.g. Langford 1999; Philip 1999) and courtroom discourses (e.g. Stubbs 1996; Heffer 1999; Szakos and Wang 1999; Cotterill 2001), and in the attribution of authorship of linguistic evidence. For example, such texts as confession/witness statements (e.g. Coulthard 1993) and blackmail/ransom/suicide notes related to specific cases (e.g. Baldauf 1999) have been studied. Corpora have also been used in detecting plagiarism (e.g. Johnson 1997; Woolls and Coulthard 1998).

Legal language has a number of words which either say things about doing and happening (e.g. *intention* and *negligence*) or refer to doing things with words (e.g. *AGREE* and *PROMISE*). Such key words are central to an understanding of the law but

are often defined obscurely in statutes and judgments. Langford (1999) used corpus evidence to demonstrate how the meanings of words such as *intention, recklessness* and *negligence* can be stated simply and clearly in words that anyone can understand. When L2 data is involved, defining legal terms becomes a more challenging task. Dictionaries are sometimes unhelpful in this regard. Philip (1999) showed how parallel corpora, in this case, a corpus of European Community directives and judgements, could be used to identify actual translation equivalents in Italian and English.

Courtroom discourses are connected to the 'fact-finding' procedure, which attempts to reconstruct reality through language, e.g. prosecutor's presentation, the eyewitness's narratives, the defendant's defence, and the judge's summing up. As people may choose to interpret language in different ways according to their own conventions, experiences or purposes, the same word may not mean the same thing to different people. Unsurprisingly, the prosecutor and the defendant produce conflicting accounts of the same event. While the judge's summing up and the eyewitness's testimonies are supposed to be impartial, studies show that they can also be evaluative.

Stubbs (1996) gave an example based on his own experience in analyzing a judge's summing up in a real court case, which involved a man being accused of hitting another man. The judge's summing up used a number of words that had a semantic preference for anger, e.g. *aggravated, annoyed, irritation, mad* and *temper*. The judge also quoted the witness who claimed to have been hit, using the word *reeled* four times. The word *reeled* was used with reference to the person being hit falling backwards after he had allegedly been assaulted. If we look at how the word *REEL* is used in the BNC, we can see that it is often used to connote violence or confusion due to some sort of outside force. The word carries an implication that the man was struck or pushed quite violently and was therefore likely to be remembered by the jury because of the number of times it was repeated by the judge (who, being the most important person in the court room, holds a lot of power and may be assumed to be able to influence people), and because it paints quite a dramatic picture. Another unusual aspect of the judge's speech was his use of modal verbs, which are used typically to indicate possibility or give permission. The judge used two modal verbs in particular, *may* and *might*, a total of 31 times in his speech and the majority of these occurred in phrases such as *you may think that, you may feel, you may find* and *you may say to yourselves*… Stubbs found that only three of these could truly be said to indicate possibility. In the other cases it was used to signal what the judge actually thought about something. Given the importance of the judge in the courtroom, the implication of phrases such as *you may think* can become 'it would be reasonable or natural for you to think that…' or even 'I am instructing you to think that…'. Supported by corpus evidence, Stubbs claimed that in a number of ways, the judge was using linguistic strategies in order to influence the jury.

While the court imposes severe constraints on the witness's right to evaluate in their narratives, the overall evaluative point of the narration is perhaps most clear in this context. Heffer (1999) explored, on the basis of a small corpus of eyewitness accounts in the trial of Timothy McVeigh, the 'Oklahoma Bomber', some of the linguistic means by which lawyer and witness cooperate in direct examination to circumvent the law of evidence and convey evaluation. He found that while witnesses seldom evaluate explicitly, a combination of a careful examination strategy and emotional involvement can result in highly effective narratives replete with evaluative elements.

Cotterill (2001) explored the semantic prosodies in the prosecution and the defence presented by both parties in the O. J. Simpson criminal trial, drawing upon data from

the Bank of English. The prosecution repeatedly exploited the negative semantic prosodies of such terms as *ENCOUNTER*, *CONTROL* and *cycle of* in order to deconstruct the professional image of Simpson as a football icon and movie star wishing to 'expose' the other side of Simpson. Cotterill found that in the Bank of English, *ENCOUNTER* typically refers to an inanimate entity and collocates with such words as *prejudice*, *obstacles*, *problems*, *a glass ceiling* (used metaphorically to refer to a barrier in one's career), *hazards*, *resistance*, *opposition*, and *risks*, all of which are negative. The modifiers of *resistance* (*stiff*) and *opposition* (*fierce*) also indicate violence. An analysis of the agents and objects of *CONTROL* in the corpus was also revealing. Corpus evidence shows that the typical agents of *CONTROL* are authority figures or representatives from government or official bodies (e.g. police), while the objects of *CONTROL* often refer to something bad or dangerous (e.g. chemical weapons, terrorist activities). It appears then, in this context, that *CONTROL* is legitimate only when the controller has some degree of authority and when what is controlled is bad or dangerous. Cotterill (2001: 299) suggested that the prosecutor was constructing Simpson as a man who was entirely unjustified and unreasonable, and excessively obsessed with discipline and authority. Another group of collocates of *CONTROL* in the corpus refers to various emotional states or conditions. But in this context, it appears that women tend to control their emotions while men tend to control their temper. In this way, Simpson was portrayed as a violent and abusive husband who finally lost his temper and murdered his emotionally vulnerable wife. The corpus shows that *cycle of* collocates strongly with negative events and situations (e.g. *violence* and *revenge killings*), and cycles tend to increase in severity over a long period of time. These two characteristics were just what the prosecutor believed the Simpson case displayed (Cotterill 2001: 301). The defence attorney, on the other hand, attempted to minimize and neutralize the negative prosodies evoked by the prosecution through a series of carefully selected lexical choices and the manipulation of semantic prosodies in his response. For example, he repeatedly conceptualized Simpson's assaults as 'incidents' (a relatively more neutral term), and used a series of verbal process nominalizations (i.e. *dispute*, *discussion* and *conversation*) in his defence statement. *Incidents* only occur at random rather than systematically. The Bank of English shows that at the top of the collocate list of *incident* (MI, 4:4 window) is *unrelated*. The defence attorney used the term *incident* to de-emphasize the systematical nature of Simpson's attacks and imply that Simpson only lost control and beat his wife occasionally and that these events were unrelated. Nominalization not only de-emphasized Simpson's role by removing agency from a number of references to the attacks, it also turned a violent actional event into a non-violent verbal event.

In the fact-finding procedure of court trials, the coherence of the defendant's account is an important criterion which may be used to measure its reliability. Szakos and Wang (1999) presented a corpus-based study of coherence phenomena in the investigative dialogues between judges and criminals. Their study was based on the Taiwanese Courtroom Spoken Corpus, which includes 30 criminal cases with 17 different types of crimes. The authors demonstrated that word frequency patterns and concordancing of corpus data could assist judges in finding out the truth and arriving at fair judgments.

Another important issue in legal cases is to establish the authorship of a particular text, e.g. a confession statement, a blackmail note, a ransom note, or a suicide note. We have already discussed authorship attribution of literary texts (see unit 10.13). The techniques used in those contexts, such as Principal Component Analysis and cluster analysis, however, are rarely useful in forensic linguistics, because the texts in legal

cases are typically very short, sometimes only a few hundred words. The techniques used in forensic linguistics are quite different from those for authorship attribution of literary texts. Forensic linguists often rely on comparing an anonymous incriminated text with a suspect's writings and/or data from general corpora.

Baldauf (1999), for example, reported on the work undertaken at the 'linguistic text analysis' section of the Bundeskriminalamt (BKA) in Wiesbaden, Germany, which has been dealing with the linguistic analysis of written texts, mainly in serious cases of blackmail, for more than ten years. During this time a method has been established that consists partly of computer-assisted research on a steadily growing corpus of more than 1,500 authentic incriminated texts and partly of *ad-hoc*, case-specific linguistic analysis.

Perhaps the most famous example of authorship attribution in forensic linguistics is the case of Derek Bentley, who was hanged in the UK in 1953 for allegedly encouraging his young companion Chris Craig to shoot a policeman. The evidence that weighed against him was a confession statement which he signed in police custody but later claimed at the trial that the police had 'helped' him produce. Coulthard (see 1993, 1994) found that in Bentley's confession, the word *then* was unusually frequent: it occurred 10 times in his 582-word confession statement, ranking as the 8th most frequent word in the statement. In contrast, the word ranked 58th in a corpus of spoken English, and 83rd in the Bank of English (on average once every 500 words). Coulthard also examined six other statements, three made by other witnesses and three by police officers, including two involved in the Bentley case. The word *then* occurred just once in the witnesses' 930-word statements whereas it occurred 29 times – once in every 78 words in the police statements. Another anomaly Coulthard noticed was the position of *then*. The sequence subject + *then* (e.g. *I then*, *Chris then*) was unusually frequent in Bentley's confession. For example, *I then* occurred three times (once every 190 words) in his statement. In contrast, in a 1.5-million-word corpus of spoken English, the sequence occurred just nine times (once every 165,000 words). No instance of *I then* was found in ordinary witness statements, but nine occurrences were found in the police statement. The spoken data in the Bank of English showed *then I* was ten times as frequent as *I then*. It appeared that the sequence subject + *then* was characteristic of the police statement. Although the police denied Bentley's claim and said that the statement was a verbatim record of what Bentley had actually said, the unusual frequency of *then* and its abnormal position could be taken to be indicative of some intrusion of the policemen's register in the statement. The case was re-opened in 1993, 40 years after Derek was hanged. Malcolm Coulthard, a forensic linguist, was commissioned to examine the confession as part of an appeal to get a posthumous pardon for Derek Bentley by his family. The appeal was initially rejected by the Home Secretary; but in 1998, another court of appeal overthrew the original conviction and found Derek Bentley innocent. In 1999 the Home Secretary awarded compensation to the Bentley family.

An issue related to authorship attribution in forensic linguistics is plagiarism, which is sometimes subject to civil or criminal legal action, and in the context of education, subject to disciplinary action. Corpus analysis techniques have also been used in detecting plagiarism. For example, Johnson (1997) carried out a corpus-based study in which she compared lexical vocabulary and hapaxes (i.e. words that occur only once) in student essays suspected of plagiarism in order to determine whether those essays had been copied. Woolls and Coulthard (1998) demonstrated how a series of corpus-based computer programs could be used to analyze texts of doubtful or disputed authorship.

Readers can refer to Coulthard (1994), Heffer (1999) and Kredens (2000) for further discussion of the use of corpora in forensic linguistics. While forensic linguistics is a potentially promising area in which corpora can play a role, it may take some time to persuade members of the legal profession to accept forensic linguistic evidence. Yet in real life cases, Coulthard's testimony helped to bring a happy ending to the Bentley case. Other cases have been less successful, however. Stubbs's evidence against the judge's biased summing up was not accepted by the Lord Chief Justice who looked at the appeal. But whatever initial outcomes, forensic linguistics needs to demonstrate that it can indeed arrive at correct answers so that the discipline can gain more credibility. For this, more experimental tests need to be carried out where linguists are given problems to solve where the answer is already known by an independent judge.

10.15 What corpora cannot tell us

We have so far reviewed the use of corpora and corpus analysis techniques in a wide range of areas of language studies. This review might give the misleading impression that corpora are all-powerful and capable of solving all sorts of language problems. But in fact, they are not. This section will briefly discuss a number of limitations of the corpus-based approach to language studies. We will return to discuss the pros and cons of using corpora in unit 12. For the moment, let us review the problems with using corpora that we have noted so far.

First, corpora do not provide negative evidence. This means that they cannot tell us what is possible or not possible. Everything included in a corpus is what language users have actually produced. A corpus, however large or balanced, cannot be exhaustive except in a very limited range of cases. Nevertheless, a representative corpus can show what is central and typical in language.

Second, corpora can yield findings but rarely provide explanations for what is observed. These explanations must be developed using other methodologies, including intuition.

Third, the use of corpora as a methodology also defines the boundaries of any given study. As we have emphasized throughout the book, the usefulness of corpora in language studies depends upon the research question being investigated. As Hunston (2002: 20) argues, 'They are invaluable for doing what they do, and what they do not do must be done in another way.' It is also important, as will be seen in units 13-16 of Section B as well as in Section C of this book, that readers learn how to formulate research questions amenable to corpus-based investigation.

Finally, it is important to keep in mind that the findings based on a particular corpus only tell us what is true in that corpus, though a representative corpus allows us to make reasonable generalizations about the population from which the corpus was sampled. Nevertheless, unwarranted generalizations can be misleading.

The development of the corpus-based approach as a tool in language studies has been compared to the invention of telescopes in astronomy (Stubbs 1996: 231). If it is ridiculous to criticize a telescope for not being a microscope, it is equally pointless to criticize the corpus-based approach for not doing what it is not intended to do (Stubbs 1999).

10.16 Unit summary and looking ahead

This final unit of Section A reviewed the corpus-based approach to language studies, drawing examples from a wide range of studies. The first part of this unit (units 10.2 – 10.8) explored the major areas in linguistics which have used corpus data. They

include lexicographic and lexical studies (unit 10.2), grammatical studies (unit 10.3), register variation and genre analysis (unit 10.4), dialect studies and language varieties (unit 10.5), contrastive and translation studies (unit 10.6), diachronic studies and language change (unit 10.7), and language learning and teaching (unit 10.8). These areas will be further discussed in Section B and explored in Section C of this book. The second part of this unit (units 10.9 – 10.14) introduced other areas of linguistics in which the corpus-based approach has started to play a role. They include semantics (unit 10.9), pragmatics (unit 10.10), sociolinguistics (unit 10.11), discourse analysis (unit 10.12), stylistics and literary studies (unit 10.13), and forensic linguistics (unit 10.14). In unit 10.15, we also discussed a number of limitations of the corpus-based approach to language studies, which readers should keep in mind when reading the excerpts in Section B and exploring specific research questions in Section C of this book. These warnings are also useful when readers pursue their own corpus-based language studies.

Having introduced the key concepts in corpus linguistics and having considered the applications of corpora in language studies, we are now ready to move on to Section B of the book, which will further discuss, on the basis of excerpts from published works, major issues in corpus linguistics and the use of corpora in linguistics.