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 24 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 (2003), 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 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 is 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*, *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 systemically. 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 distribution 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).