Unit 10 Corpora and language studies

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 John's *Oxford English Dictionary*), 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 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 been 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

[see my posting “Corpora and Language education”]