42

Corpus linguistics

Svenja Adolphs and Phoebe M. S. Lin

Introduction

Corpus linguistics most commonly refers to the study of machine-readable spoken and written language samples that have been assembled in a principled way for the purpose of linguistics research. At the heart of empirically based linguistics and data-driven description of language, corpus linguistics is concerned with language use in real contexts. Therefore, it is often contrasted with Chomskyan linguistics, which emphasises language competence and often involves made-up examples as the basis of its exploration of language. Access to ever larger spoken and written corpora has already revolutionised the description of language in use; however, the impact of corpus linguistics has reached far beyond the disciplines that are purely concerned with linguistic descriptions of language. As an approach, corpus linguistics continues to gain recognition and popularity, with an increasing number of researchers across different disciplines exploring innovative ways of using corpus-based research as part of their methods toolkit.

This chapter provides a brief overview of some of the different types of corpora available and some of the methods used within the area of corpus linguistics, including the generation of frequency lists, concordance outputs and keyword analyses. It then moves on to a discussion of selected current issues in corpus linguistics. We focus here on three issues which we believe are marked by the persistent attention they have received in the field, as well as by their prominent status among researchers and end-users. The issues we will introduce include an area of language description (phraseology and corpus research), an area of application (English language teaching and corpus research), and an area of resource development (the Web as corpus). The chapter will conclude with a discussion of the impact which technological developments may have on the discipline. All the corpus resources mentioned in this chapter can be found after the 'Further reading' section.

Corpus as data

Corpora are designed to represent a particular language variety. Common distinctions are made between *specialised* and *general* corpora, where the former includes texts that belong to a particular type, e.g. academic prose, while the latter includes many different types of texts, often assembled with the aim to serve as reference resources for linguistic research or to produce reference materials such as dictionaries. Other types of corpora include *historical* and *monitor* corpora, *parallel* corpora and *learner* corpora. Historical corpora include texts from different periods of time and allow for the study of language change when compared with corpora from other periods. Monitor corpora can be used for a similar purpose, but tend to focus on current changes in the language. New texts from the same variety are added to the existing corpus at regular intervals, thus contributing to a constantly growing text database. Parallel corpora include texts in at least two languages that have either been directly translated, or produced in different languages for the same purpose. Such corpora are often used in translation studies. Learner corpora contain collections of texts produced by learners of a language. They allow the researcher to identify patterns in a particular variety of learner English, and to compare the language of the learner to that of other users of a language.

In terms of the history of corpus design, a distinction is often made between the early corpora developed in the 1950s, 1960s and 1970s and the larger corpora developed from the late 1980s onwards. Early corpora include the London-Lund Corpus of Spoken English (LLC), the Brown Corpus based on American written English, and the Lancaster-Oslo/Bergen Corpus based on written British English. The parallel design of the latter two corpora allowed for a corpus-based comparison between British and American English. Early corpora were often limited in size to a one million word threshold, which is partly a reflection of the technological possibilities at the time.

Two of the most substantial corpus projects developed in the 1980s and 1990s are the Collins and Birmingham University International Language Database (COBUILD) and the British National Corpus. Both offer a valuable resource for the study of everyday spoken and written English. The COBUILD corpus, which is also referred to as the Bank of English, was developed in the 1990s as a monitor corpus. This means that new texts are constantly added to this database: the size of the corpus stood at 450 million words. One of the main aims of this project has been to provide a textual database for the compilation of dictionaries and lexicography research. The corpus contains samples of mainly British written language, as well as transcribed speech from interviews, broadcast, and conversation. The British National Corpus (BNC) was compiled in the late 1980s and early 1990s, and is a 100 million word corpus of modern British English, consisting of 90 per cent written and 10 per cent spoken texts (including speeches, meetings, lectures, and some casual conversation). Apart from these two major corpora, many publishing houses have developed their own corpora which serve as a resource for authors, mainly in the area of lexicography. Examples are the Cambridge International Corpus (CIC), the Longman Corpus Network and the Oxford English Corpus. Another large corpus project is the International Corpus of English (ICE), which was initiated in 1990 as a resource for comparing different varieties of English. At the time of writing, the ICE consists of 22 one-million-word corpora, each representing a regional variety of English. More recently, two substantial American English corpora have been developed: the American National Corpus (ANC) and the Corpus of Contemporary American English (COCA). By 2009 the ANC contained 22 million words of written and spoken texts in American English produced since 1990. The COCA consists of more than 400 million words of American English, with 21 per cent spoken and 79 per cent written material. With 20 million words added each year, the COCA can also be used as a monitor corpus to capture language change.

The corpora above mainly focus on the collection of general English in use. As such they contrast with specialised corpora which range from those that represent the language of a particular group of people, such as the Bergen Corpus of London Teenage Language (COLT), to those that represent a particular mode of discourse. Some of the major developments of specialised corpora have taken place in the domain of academic discourse and include, for example, the Michigan Corpus of Academic Spoken English (MICASE), and its British counterpart, the British Academic Spoken English corpus (BASE).

Another category of corpora captures the language use of language learners. The analysis of learner corpora makes it possible to track developmental aspects of learner language, as well as to highlight particular areas of difficulty for the learner. At the same time, learner corpora can be used as a basis for better descriptions of different varieties that emerge from communication between speakers who communicate in a language other than their first language. The design criteria for learner corpora have a slightly different focus to native speaker corpora in that particular emphasis has to be placed on the level of consistency of the resource in terms of the language background of the speakers, including their level of proficiency and first language. Examples of learner corpora include the Cambridge Learner Corpus the Longman Learners' Corpus and the International Corpus of Learner English (ICLE). Examples of corpora which are used as the basis for exploring the use of English as a lingua franca include the Vienna-Oxford International Corpus of English (VOICE) and the English as a Lingua Franca in Academic Settings (ELFA) corpus.

Many of the corpora outlined above come with their own concordancing interface, often available via the Internet. The next section will consider in more detail the various tools and methods which may be used to explore the language captured in spoken and written corpora.

Metadata

Apart from the process of assembling written and spoken language samples in a principled way into a corpus, it is also important to collect and document further information about the collected discourse itself. Metadata, or 'data about data', is the conventional method used to do this. Burnard (2005) states that 'without metadata the investigator has nothing but disconnected words of unknowable provenance or authenticity'. Thus, metadata are critical to a corpus to help achieve the standards for *representativeness*, and of *balance* and *homogeneity* (see Sinclair 2005).

Burnard (2005) uses the term *metadata* as an umbrella term which includes editorial, analytic, descriptive and administrative categories:

- Editorial metadata: providing information about the relationship between corpus components and their original source.
- Analytic metadata: providing information about the way in which corpus components have been interpreted and analysed.
- Descriptive metadata: providing classificatory information derived from internal or external properties of the corpus components.
- Administrative metadata: providing documentary information about the corpus itself, such as its title, its availability, its revision status, etc.

Metadata are particularly important when the corpus is shared and reused by others in a research community, and they also assist in the preservation of electronic texts. Metadata can be kept in a separate database or included as a 'header' at the start of each document (usually encoded though mark-up language). A separate database with this information makes it easier to compare different types of documents and has the distinct advantage that it can be further extended by other users of the same data. The documentation of the design rationale, as well as the various editorial processes that an individual text has been subjected to during the collection and archiving stages, facilitates replicability of research and validation of results.

Corpus linguistics: tools and methods

A number of user-friendly software packages are available which facilitate the manipulation and analysis of corpus data. Common functionalities include the generation of frequency counts according to specified criteria, comparisons of frequency information in different texts, different formats of concordance outputs, including the Key Word In Context (KWIC) concordance, and the extraction of multiword units or clusters of items in a text. Many of these programs can be downloaded from the Internet or used directly via an interactive Website (see, for example, the Compleat Lexical Tutor, BNCWeb and BYU-BNC). Other programs are distributed commercially, often by publishing houses, and can be purchased for a fee.

Word lists

The frequency of a word or a phrase in different contexts is an important part of its description. Various word lists that are based to some degree on word frequency in a corpus exist especially in the English language teaching (ELT) context, such as the Academic Word List (Coxhead 2000) and the Academic Formulas List (Simpson-Vlach and Ellis 2010). Word lists are a good starting point for subsequent searches of individual items at concordance level and can be useful in the comparison of different corpora. Word lists can be generated to account for individual items or for recurrent sequences of two or more items. Lemmatised frequency lists group together words from the same lemma. For example, the words ‘say’, ‘said’, ‘saying’, ‘says’ are all part of the lemma SAY. Lemmatisation can be done manually using an alphabetical frequency list, or in an automated way which is often based to some degree on lists of predefined lemmas. Different forms of the same lemma tend to vary significantly in terms of their overall frequency, with one particular form tending to be more frequent than others in the lemma. Previous research has shown that there often are variations in meaning between different variants of the lemma (Stubbs 1996; Tognini-Bonelli 2001). Lemmatised lists also have a place in more applied contexts, including ELT where it can be beneficial to teach all forms of one lemma together while giving priority to the most frequently used form.

The kind of basic information that can be gathered from a frequency list is illustrated with reference to Table 42.1, which shows the ten most frequent items in the spoken Limerick Corpus of Irish English (LCIE) and in the written component of the British National Corpus (O’Keeffe *et al*. 2011). LCIE is a corpus of naturally occurring contemporary spoken Irish

Table 42.1 Ten most frequent words in the BNC (written) and LCIE (spoken)

	BNC	*LCIE*
1	*the*	*the*
2	*of*	*I*
3	*and*	*and*
4	*a*	*you*
5	*in*	*to*
6	*to*	*it*
7	*is*	*a*
8	*was*	*that*
9	*it*	*of*
10	*for*	*yeah*

English (for more details see Farr *et al*. 2004). A comparison of the ten most frequent words in a spoken and a written corpus highlights some of the key differences between the two modes of discourse. Both contain mainly grammatical items, which is expected in terms of the general distribution of different items in the English language. However, the spoken corpus list also includes the personal pronouns 'I' and 'You' which shows the interactive nature of the discourse that makes up this corpus. In addition, the vocalisation 'Yeah' occurs amongst the most frequent items in the spoken data reflecting the pervasive occurrence of listener response tokens in conversation. These three items are at the heart of spoken interaction and the frequency list helps to identify those defining items.

Table 42.2 shows the ten most frequent two-word, three-word and four-word n-grams in LCIE. This type of frequency output highlights the phrasal nature of language. Although the kinds of sequences generated in this way do not necessarily reflect the underlying phraseology of language fully, the output is strongly suggestive of common phrases of which the sequences in Table 42.2 form a part. However, a mere frequency-based list of continuous sequences is limited in its explanatory power when it comes to the study of phraseology. Research in the area of computational linguistics has introduced new techniques for extracting meaningful units from corpora, both on the basis of frequency information (see, for example, Danielsson 2003) and on the basis of part-of-speech tagged corpora which include further annotation of semantic fields (Rayson 2003).

Keywords and key sequences

Keywords are as words which occur either with a significantly higher frequency (positive keywords) or with a significantly lower frequency (negative keywords) in a text or collection of texts, when they are compared to a reference corpus (Scott 1997). Keywords are identified on the basis of statistical comparisons of word frequency lists derived from the target corpus and the reference corpus. Each item in the target corpus is compared with its equivalent in the reference corpus, and the statistical significance of difference is calculated using chi-square or log-likelihood statistics (see Dunning 1993). Both of these statistics compare actual observed frequencies between two items with their expected frequencies, assuming random distribution. If the difference between observed and expected frequency is large then it is likely that the relationship between the two items is not random. The procedure thus generates words that are characteristic, as well as those that are uncharacteristic in a given target corpus. The choice of the

Table 42.2 Ten most frequent 2-word, 3-word and 4-word units in LCIE results per million words

Frequency rank	*2-word units*		*3-word units*		*4-word units*	
1	*you know*	4406	*I don't know*	1212	*you know what I*	230
2	*in the*	3435	*do you know*	769	*know what I mean*	215
3	*of the*	2354	*a lot of*	522	*do you know what*	208
4	*do you*	2332	*you know what*	379	*I don't know what*	134
5	*I don't*	2200	*do you want*	373	*do you want to*	121
6	*I think*	2003	*I don't think*	338	*are you going to*	103
7	*It was*	1939	*you know the*	323	*you know the way*	103
8	*I was*	1891	*you have to*	308	*I don't know I*	91
9	*going to*	1849	*going to be*	307	*thank you very much*	91
10	*on the*	1801	*yeah yeah yeah*	297	*the end of the*	85

reference corpus used as the basis for such a comparison is crucial in this context, as it affects the output of keywords. For example, in a comparison of a transcript of medical consultation with a reference corpus that consists solely of written texts, the characteristics of spoken versus written language may interfere with the analysis of keywords in the medical consultation genre.

We can generate keyword lists as well as lists of key sequences. The list below contains key sequences resulting from a comparison of health communication used as part of a UK telephone health advice service with the CANCODE, a five-million-word corpus of casual conversation in British English. The sequences are the ten most significant positive key sequences in the Nottingham Health Communication Corpus featuring health communication in the British context (see Adolphs *et al*. 2004; Adolphs 2006).

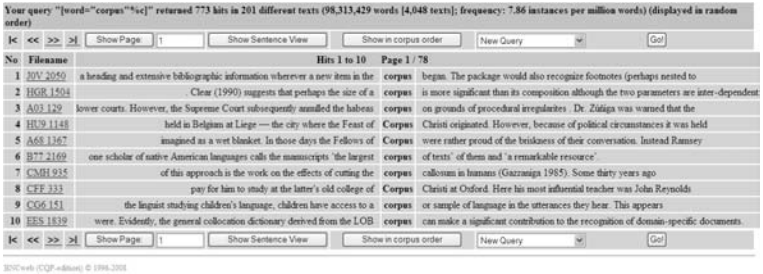
1 NHS Direct
2 NHS
3 Just bear with
4 Call you back
5 Bear with me
6 Date of birth
7 Your date of
8 You're calling from
9 Manage their services
10 However anybody with

As can be expected, quite a few of the recurrent sequences in this list form part of the responses that typically mark the beginning of telephone interactions with the health advice service NHS Direct. Other sequences relate to the gathering of basic information about the caller. The most significant negative key sequence, i.e. the one that occurs with a significantly lower frequency in the Health Communication Corpus, in comparison to the corpus of casual conversation, is 'I don't know'. This highlights the professional nature of this encounter where the emphasis is on providing knowledge and advice. 'I don't know' is a common hedge and politeness marker in casual conversation and does not fit with the more asymmetrical medical exchanges by telephone.

The concordance output

An example of a Key Word in Context (KWIC) concordance of the word 'corpus' using the BNCWeb is shown in Figure 42.1. Corpus users can normally specify the number of words to the left and to the right of the search word that are displayed as part of the output. If a corpus is tagged for part-of-speech (POS), then users may also carry out a concordance search based on word class or grammatical structure.

There are many ways of examining and interpreting concordance data. A concordance output can be useful in providing a representation of language data which allows the user to notice patterns relating to the way in which a lexical item or a sequence is used in context. Sinclair (1996) argues that a new unit of meaning emerges from the analysis of concordance data that extends beyond the single word and takes into account the properties and patterns that are revealed by concordance analysis. Such units, as Sinclair points out, are going to be 'largely phrasal' (1996: 82). In order to describe the nature of individual units of meaning, Sinclair (1996) suggests four parameters: collocation, colligation, semantic preference and semantic prosody. Collocation refers to the habitual co-occurrence of words and will be discussed in more detail below. Colligation is the co-occurrence of grammatical choices. Grammatical



Your query "[word="corpus"%c]" returned 773 hits in 201 different texts (98,313,429 words [4,048 texts]; frequency: 7.86 instances per million words) (displayed in random order)

No	Filename	Hits 1 to 10	Page 1 / 78	
1	J0V 2050	a heading and extensive bibliographic information wherever a new item in the	corpus	began. The package would also recognize footnotes (perhaps nested to
2	HGR 1504	. Clear (1990) suggests that perhaps the size of a	corpus	is more significant than its composition although the two parameters are inter-dependent:
3	A03 129	lower courts. However, the Supreme Court subsequently annulled the habeas	corpus	on grounds of procedural irregularites . Dr. Zúñiga was warned that the
4	HU9 1148	held in Belgium at Liege — the city where the Feast of	Corpus	Christi originated. However, because of political circumstances it was held
5	A68 1367	imagined as a wet blanket. In those days the Fellows of	Corpus	were rather proud of the briskness of their conversation. Instead Ramsey
6	B77 2169	one scholar of native American languages calls the manuscripts 'the largest	corpus	of texts' of them and 'a remarkable resource'.
7	CMH 935	of this approach is the work on the effects of cutting the	corpus	callosum in humans (Gazzaniga 1985). Some thirty years ago
8	CFF 333	pay for him to study at the latter's old college of	Corpus	Christi at Oxford. Here his most influential teacher was John Reynolds
9	CG6 151	the linguist studying children's language, children have access to a	corpus	or sample of language in the utterances they hear. This appears
10	EES 1839	were. Evidently, the general collocation dictionary derived from the LOB	corpus	can make a significant contribution to the recognition of domain-specific documents.

Figure 42.1 A KWIC concordance of the word 'corpus' using the BNCWeb
Source: Extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium.

patterning around a particular word accounts for the 'variation' of a phrase, which 'gives the phrase its essential flexibility, so that it can fit into the surrounding co-text' (Sinclair 1996: 83). Many of the so-called 'fixed phrases' are therefore only fixed if we consider the lexico-grammatical 'core'. If we extend the units of meaning, however, to patterns in the co-text, the expressions become more variable. One of the examples Sinclair provides is the phrase 'true feelings' which, in the Bank of English, exhibits the following patterns:

> At N-3 position and beyond: 'will never reveal', 'prevents me from expressing', 'careful about expressing', 'less open about showing', 'guilty about expressing'
>
> At N-2 position: 'communicate', 'show', 'reveal', 'share', 'pour out', 'give vent to', 'indicate' and 'make public'
>
> At N-1 position: possessives such as 'our'

The collocates of 'true feelings' show clear patterns in terms of semantic prosody, semantic preference and colligation. The semantic preference of a lexical item or expression is a semantic abstraction of its prominent collocates. In his discussion of the expression 'the naked eye', Sinclair (1996) finds that most of the verbs and adjectives preceding this expression are related to the concept of 'vision'. The verbs 'see' and 'seen' together occur 25 times within four words to the left of the expression in a sample of 151 examples of 'the naked eye' that he studies. Sinclair (1996) uses, as his fourth parameter in the description of the units of meaning, the concept of *semantic prosody*. First discussed by Sinclair (1987) and Louw (1993), semantic prosodies are associations that arise from the collocates of a lexical item and are not easily detected using introspection. Semantic prosodies have mainly been described in terms of their positive or negative polarity (Sinclair 1991a; Stubbs 1995) but also in terms of their association with 'tentativeness/indirectness/face saving' (McCarthy 1998: 22). Carter and McCarthy (1999) find, for example, that there is a consistently negative semantic prosody associated with the *get*-passive in the corpus data they examine (e.g. 'get arrested', 'get sued' and 'get nicked').

Current issues in corpus linguistics

Phraseology

John Sinclair's (1991a) contributions relating to lexical patterning have been highly influential in the field (see Stubbs 2009). One of the most notable aspects of his work is his research on the prevalence of phraseology or multiword units. The idea that everyday language is full of highly recurrent sequences of words challenges the traditional perception of language processing in the brain and the belief that language production (and reception) relies on a completely rule-based system. Corpus evidence, which highlights the prevalence of highly recurrent chunks or multiword units, suggests that we have a large store of conventional, preferred expressions that we use regularly to express meaning. Therefore, Sinclair (1991a) suggests that highly recurrent chunks are fundamental to the organisation and the production of language, and proposes that language production is the result of the alternation between the idiom principle and the open-choice principle. The idiom principle posits that 'a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments' (1991a: 110). The open-choice principle is also known as the slot-and-filler model, for it sees a text as constituted by a series of slots to be filled by lexis which satisfies certain grammatical constraints. According to Sinclair (1991a), for normal texts the idiom principle is the first (i.e. the default) mode to be applied. The switch to the open-choice principle occurs only when there is good reason for the switch to occur. This explains why multiword units are so prevalent. According to a study by Erman and Warren (2000), multiword units account for slightly over half of English speech and writing. Altenberg (1998: 102), who adopts an alternative definition of phraseology, estimates that 'over 80 percent of the words in the [London-Lund] corpus form part of a recurrent word-combination in one way or another'.

A related major contribution to the field relates to the study of word sense and relations where J. R. Firth's (1957: 11) idea that 'you shall know a word by the company it keeps' has found support from corpus evidence. A simple example to illustrate the dynamics of word meanings as a result of collocation is shown in the concordance of the word 'stand' (Table 42.3 below), taken from the BNCWeb. We know that the basic meaning of the word 'stand' is to rise to or to maintain an upright position. But when 'stand' collocates with other words, it takes up a metaphorical meaning.

The term *multiword unit* is often used in this context as an umbrella term for sequences of interrelated words which are retrieved from memory as single lexical units. They occur with varying degrees of fixedness, including formulae (e.g. *have a nice day*), metaphors (e.g. *kick the bucket*) and collocations (e.g. *rancid butter*) (Moon 1998; Wray 2002). The description and conceptualisation of multiword units are a key concern in many different areas of language study ranging from psycholinguistics to Natural Language Processing (NLP). There are many different ways of identifying multiword units. These include intuitive identification, the use of discourse analytical techniques, and automatic extraction from electronic texts (see

Table 42.3 Concordance of the word 'stand', taken from the BNCWeb

up for huge amounts of emotional strain, knowing that their clients	*stand*	to lose more than their liberty. Since almost all those on
, all modes of thought and life. Certainly Lady Lisa might	*stand*	as the embodiment of the old fancy, the symbol of the
, or obeyed it only in so far as he took a	*stand*	against conservative taste. He did not take sides against Abstraction,
any view. Reynolds was a notable conversationalist, well able to	*stand*	up to his friends, who included Samuel Johnson and Edmund Burke
approach, it must be admitted. In Rome a visitor can	*stand*	in front of a Baroque church, but a few minutes later
the progress of this mad love – to which he did not	*stand*	all that close at the time, brother as he was –
to revel in his own discomfiture here, in his desperate last	*stand*	for freedom – he clearly enjoys the sensation of falling in love,
of them were particularly concerned with 'glamour'. What does	*stand*	out is that everyone believes that the profession, its standards and
number of higher clergy in Ulster and many of its lay intellectuals	*stand*	a long way off from protestant–loyalist politics and are in
will do now, he wrote. It must be impossible to	*stand*	up against it, he wrote, impossible to draw breath before

Wray 2002, 2008). The classification into different types of multiword units tends to be linked to particular characteristics. Formulae, for example, are marked by their pragmatic function (see Aijmer 1996) while collocations are marked by their frequency of co-occurrence in discourse. Multiword units are closely linked to the particular genre in which they occur (see, for example, Biber and Conrad 1999; Biber *et al*. 2003; Oakey 2002; Schmitt 2004; Simpson 2004; Biber 2006, 2009 for discussions of multiword units in academic speech and writing).

Corpora and English language teaching

While corpus linguistics has enabled better descriptions of language in use, its real impact lies in the enhancement of applications based on those descriptions. A key area to highlight in this context is that of English language teaching, where the latest findings from corpus research have led to real innovations in material design and classroom practice. There are two main areas in which corpora can benefit language teaching and learning: first, by incorporating the latest corpus-based findings into language syllabuses, teaching materials and dictionaries; second, by encouraging teachers and learners to examine language patterns in corpus as part of their (independent) learning activities in and outside classrooms (see Gavioli and Aston 2001).

Corpus linguists and language teaching researchers are often found collaborating in these two areas and there are now publications on the subject. Some of these (e.g. Meunier and Granger 2008) provide further corpus-based descriptions of aspects of language which target the needs of specific groups of language learners, e.g. ESP/EAP learners or learners of the same L1 background. Others (e.g. Hunston 2002; Sinclair 2003) aim to equip teachers and learners with the skills of concordancing and extracting useful information from concordance lines. Other publications (e.g. Tribble and Jones 1997; O'Keeffe *et al*. 2007) include practical suggestions on the various ways in which corpus research can be introduced into the language classroom to enrich the experience of language learners.

Despite the growing interest in the pedagogical applications of corpus linguistics, there have been a number of debates relating to the place of corpus linguistics in language teaching (see Sinclair 1991b; Widdowson 1991; Seidlhofer 2003). Widdowson (1991), for instance, argues that the fact that a language pattern is particularly frequent in a corpus does not necessarily mean that it should take priority in the language teaching syllabus. Further discussion centres around the issue of authenticity and whether it is beneficial to present learners with authentic, real language in use (see McCarthy and Carter 1995; Carter and McCarthy 1996; Prodromou 1996a, 1996b, 1998). According to Prodromou (1996b), it is a 'fallacy' to assume that real language is spontaneously interesting and useful to foreign language learners. He argues that train timetables, advertisements, letters published in British newspapers and consumer leaflets are only real to members of the speech community that these texts target. When such data are used as teaching material in a foreign language classroom, they mean very little to the language learners because they lack the same reality for this specific audience. Prodromou (1996a) suggests that an 'authentic' discourse has its 'here and nowness', and when the discourse is presented in a context that is detached from the 'here and now' it automatically loses its authenticity. Similarly, Widdowson (2000) argues that the language presented in a corpus is decontextualised and only partially real. If the decontextualised language in a corpus is to be presented to learners as language in use, it has to be recontextualised. Yet, the reconstituted context is not always the same as the original context of the texts (see Prodromou 2008).

Despite these arguments, corpus data are increasingly becoming an accepted and desirable basis for the development of English language teaching materials, and most major dictionaries and grammars now advertise the fact that they are based on 'real' language from a corpus.

The Web as corpus

Today, corpus size has long exceeded the one million word standard set by the Brown Corpus in the 1960s. The Cambridge International Corpus (CIC), which collects spoken and written texts of American English, British English and learner English, is currently one of the biggest corpora of English, with over a billion words. However, with the advent of the world-wide-Web we now have access to language data which far exceeds even the most substantial corpus.

Kilgarriff and Grefenstette (2003) suggest that checking spelling and usage of a word by typing it into an Internet search engine is a practical example of how the World Wide Web is already being used as a language corpus on a daily basis by a large number of people. They give the example of 'speculater' and 'speculator'. A search engine reveals that these two spellings generate, respectively, 67 hits and 82,000 hits on the Web. Therefore, based on the higher frequency of occurrence of 'speculator', one may conclude that this is the preferred spelling.

However, for the Web to provide more than free, instant suggestions on spellings, corpus linguists have developed Web-based interfaces that allow researchers to use the Web as a compatible resource for linguistic research. WebCorp, for example, allows users greater control over the type of texts to be searched. They can specify the register, textual domain, topic range, date of modification and so on. These facilities support investigations into both synchronic and diachronic changes in language (see Renouf 2003; Renouf *et al*. 2007). Another advantage of using WebCorp over general Internet search engines in lexical research is that the former offers basic statistical information, including the collocational profile of search items and the option to disambiguate polysemous items (Renouf 1993). The WebCorp interface can also be used to generate frequency lists of Websites specified by the user. It is clearly a valuable resource to use in its own right, but it can also be used to complement research on finite corpora in terms of the up-to-date evidence of language in use that it offers.

While the World Wide Web is a very large repository of naturally occurring language, further research is needed as to the type of language that is being used on the Web, what it represents, and how balanced it is in the context of a particular research question. Given the ubiquity of Internet-based and Internet-stored discourse, this endeavour becomes particularly urgent.

The impact of new technologies on corpus linguistics: an example

One of the main impacts of new technology on the area of corpus linguistics is no doubt the use of the Web as a corpus. In addition, there have been significant advances in spoken corpus linguistics which have been afforded by the alignment of different modalities with a transcript. This development started with the alignment of audio recordings with transcripts, and has recently been extended to include video data as well. It has long been pointed out by corpus linguists working with spoken data that the lack of audio and video leads to problems in the analysis of this kind of corpus data. De Cock (1998), for example, in a discussion of the sequence 'you know', argues that it is virtually impossible to decide whether 'you know' has a literal or a formulaic meaning on the basis of the orthographic transcript alone. Similarly, Lin and Adolphs (2009) observe that it is not possible to determine the functions of some instances of 'I don't know why' in context unless one can refer to their prosody. Similar concerns arise from a corpus-based analysis of multimodal written texts, i.e. those containing images and graphics.

Gesture, prosody and kinesics all add meaning to utterances and discourse as a whole, and recent research in the area of spoken corpus analysis has started to explore the potential impact of drawing on multimodal corpus resources for our descriptions of spoken language (see, for example, Knight *et al*. 2009). In addition to offering a more comprehensive resource for describing discourse, multimodal corpora also allow us to reflect on and evaluate some of the methods for analysing textual renderings of spoken discourse established so far. The representation and analysis of 'textual' concordance data thus becomes limited and limiting in a way that can now be avoided by using one of the tools and interfaces developed for aligning and searching text, audio and video data, such as ELAN or Transana.

Summary

In this chapter we have provided a brief overview of some of the key methods and current issues in corpus linguistics. This has included an overview of different types of corpora, as well as an introduction to analytical methods. Key issues that have been highlighted in this chapter relate to the use of corpus linguistics in phraseology research, English language teaching and the use of the Web as corpus. We have also discussed the role of new technologies in the development of multimodal corpus resources. As a discipline, corpus linguistics is gathering pace with the development of ever larger data-sets and with an increasingly sophisticated suite of tools that can be used to analyse these data and represent the outputs. One of the main challenges for the future will be to fully explore the implications of these advances, not only for language description in its own right, but also crucially for other disciplines, and the impact that this work may have in applied contexts.

Related topics

critical discourse analysis; discourse; English for academic purposes; ESP and business communication; language learning and language education; lexicography; lexis; medical communication; SLA; technology and language learning; the media; translation and interpreting; world Englishes

Further reading

Anderson, W. and Corbett, J. (2009) *Exploring English with Online Corpora*, Basingstoke: Palgrave Macmillan. (This volume offers an introduction to how online corpora can be used in the teaching and learning of English.)

Baker, P., Hardie, A. and McEnery, A. (2006) *A Glossary of Corpus Linguistics*, Edinburgh: Edinburgh University Press. (This book provides a comprehensive overview of key concepts and relevant references in corpus linguistics.)

Hoffmann, S., Evert, S., Smith, N., Lee, D. and Prytz, Y. B. (2008) *Corpus Linguistics with BNCWeb: A Practical Guide*, Frankfurt: Peter Lang. (This book offers a practical, hands-on introduction to corpus linguistic methods using the BNCWeb corpus interface.)

Sinclair, J. (2004) *Trust the Text: Language, Corpus and Discourse*, London: Routledge. (This volume is a collection of some of Sinclair's most influential papers in the area of corpus linguistics and lexico-grammar.)

Resources mentioned in the chapter

Language corpora

American National Corpus: www.americannationalcorpus.org/

Bank of English: www.titania.bham.ac.uk/

Bergen Corpus of London Teenage Language (COLT, now a constituent of the BNC): www.hit.uib.no/colt/

British Academic Spoken English (BASE) corpus: http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/; www.reading.ac.uk/AcaDepts/ll/base_corpus/
British National Corpus (BNC): www.natcorp.ox.ac.uk/
Brown Corpus: http://khnt.hit.uib.no/icame/manuals/brown/index.htm
Cambridge and Nottingham Corpus of Discourses in English (CANCODE, now a constituent of the CIC): see McCarthy (1998)
Cambridge International Corpus (CIC): www.cambridge.org/elt/corpus/
Corpus of Contemporary American English: www.americancorpus.org/
English as a Lingua Franca in Academic Settings (ELFA) Corpus: www.uta.fi/laitokset/kielet/engf/research/elfa/
Health Communication Corpus: see Adolphs *et al*. (2004)
International Corpus of English (ICE): http://ice-corpora.net/ice/
International Corpus of Learner English (ICLE): see Granger *et al*. (2009)
Lancaster-Oslo-Bergen Corpus (LOB): see Johansson *et al*. (1978; 1986)
Limerick Corpus of Irish-English (LCIE) http://www.ul.ie/~lcie/homepage.htm
London Lund Corpus of Spoken English (LLC): http://khnt.hit.uib.no/icame/manuals/londlund/index.htm
Longman Corpus Network: www.pearsonlongman.com/dictionaries/corpus/index.html
Michigan Corpus of Academic Spoken English (MICASE): http://micase.elicorpora.info/
Oxford English Corpus: www.askoxford.com/oec/
Vienna-Oxford International Corpus of English (VOICE): www.univie.ac.at/voice/page/index.php

Corpus tools/interfaces

BYU-BNC: http://corpus.byu.edu/bnc/
BNCWeb: http://corpora.lancs.ac.uk/BNCWeb/home.html
Compleat Lexical Tutor: www.lextutor.ca/
ELAN: www.lat-mpi.eu/tools/elan
Transana: www.transana.org/
WebCorp: www.Webcorp.org.uk/

References

Adolphs, S. (2006) *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*, Abingdon and New York: Routledge.
Adolphs, S., Brown, B., Carter, R., Crawford, P. and Sahota, O. (2004) ‘Applying corpus linguistics in a health care context’, *Journal of Applied Linguistics* 1(1): 9–28.
Aijmer, K. (1996) *Conversational Routines in English*, London and New York: Longman.
Altenberg, B. (1998) ‘On the phraseology of spoken English: the evidence of recurrent word-combinations’, in A. P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*, Oxford: Clarendon Press.
Biber, D. (2006) *University Language: A Corpus-based Study of Spoken and Written Registers*, Amsterdam: John Benjamins.
——(2009) ‘A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing’, *International Journal of Corpus Linguistics* 14(3): 275–311.
Biber, D. and Conrad, S. (1999) ‘Lexical bundles in conversation and academic prose’, in H. Hasselgard and S. Oksefjell (eds) *Out of Corpora: Studies in Honour of Stig Johansson*, Amsterdam: Rodopi.
Biber, D., Conrad, S. and Cortes, V. (2003) ‘Lexical bundles in speech and writing: an initial taxonomy’, in A. Wilson, P. Rayson and T. McEnery (eds) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, Frankfurt: Peter Lang.
Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*, Harlow: Pearson Longman.
Burnard, L. (2005) ‘Metadata for corpus work’, in M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford: Oxbow Books.
Carter, R. and McCarthy, M. (1996) ‘Correspondence’, *ELT Journal* 50(4): 369–71.

——(1999) 'The English get-passive in spoken discourse: description and implications for an interpersonal grammar', *English Language and Linguistics* 3(1): 41–58.
Coxhead, A. (2000) 'A new academic word list', *TESOL Quarterly* 34(2): 213–38.
Danielsson, P. (2003) 'Automatic extraction of meaningful units from corpora: a corpus-driven approach using the word stroke', *International Journal of Corpus Linguistics* 8(1): 109–27.
De Cock, S. (1998) 'A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English', *International Journal of Corpus Linguistics* 3(1): 59–80.
Dunning, T. (1993) 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics* 19(1): 61–74.
Erman, B. and Warren, B. (2000) 'The idiom principle and the open choice principle', *Text* 20(1): 29–62.
Farr, F., Murphy, B. and O'Keeffe, A. (2004) 'The Limerick corpus of Irish English: design, description and application, *Teanga* 21: 5–30.
Firth, J. R. (1957) 'A synopsis of linguistic theory, 1930–55', in *Studies in Linguistic Analysis*, Oxford: Blackwell.
Gavioli, L. and Aston, G. (2001) 'Enriching reality: language corpora in language pedagogy', *ELT Journal* 55(3): 238–46.
Granger, S., Dagneaux, E., Meunier, F. and Paquot, M. (2009) *International Corpus of Learner English, version 2*, Louvain: Presses Universitaires de Louvain.
Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.
Johansson, S., Leech, G. and Goodluck, H. (1978) *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*, Oslo: University of Oslo.
Johansson, S., Atwell, E., Garside, R. and Leech, G. (1986) *The Tagged LOB Corpus: User's Manual*, Bergen: Norwegian Computing Centre for the Humanities.
Kilgarriff, A. and Grefenstette, G. (2003) 'Introduction to the special issue on the Web as corpus', *Computational Linguistics* 29(3): 333–48.
Knight, D., Evans, D., Carter, R., and Adolphs, S. (2009) 'HeadTalk, HandTalk and the corpus: towards a framework for multi-modal, multi-media corpus development', *Corpora* 4(1): 1–32.
Lin, P. M. S. and Adolphs, S. (2009), 'Sound evidence: phraseological units in spoken corpora', in A. Barfield and H. Gyllstad (eds) *Researching Collocations in Another Language: Multiple Interpretations*, Basingstoke: Palgrave Macmillan.
Louw, B. (1993) 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in M. Baker, G. Francis and E. Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair*, Amsterdam: John Benjamins.
McCarthy, M. (1998) *Spoken Language and Applied Linguistics*, Cambridge: Cambridge University Press.
McCarthy, M. and Carter, R. (1995) 'Spoken grammar: what is it and how can we teach it?', *ELT Journal* 49(3): 207–18.
Meunier, F. and Granger, S. (eds) (2008) *Phraseology in Foreign Language Learning and Teaching*, Amsterdam: John Benjamins.
Moon, R. (1998) 'Frequencies and forms of phrasal lexemes in English', in A. P. Cowie (ed.) *Phraseology: Theory, Analysis and Applications*, Oxford: Clarendon Press.
O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom: Language Use and Language Teaching*, Cambridge: Cambridge University Press.
O'Keeffe, A., Clancy, B., and Adolphs, S. (2011) *Introducing Pragmatics in Use*, London: Routledge.
Oakey, D. (2002) 'Formulaic language in English academic writing', in R. Reppen, S. Fitzmaurice and D. Biber (ed.) *Using Corpora to Explore Linguistic Variation*, Philadelphia, PA: John Benjamins.
Prodromou, L. (1996a) 'Correspondence', *ELT Journal* 50(4): 371–3.
——(1996b) 'Correspondence', *ELT Journal* 50(1): 88–9.
——(1998) 'Correspondence', *ELT Journal* 52(3): 266–7.
——(2008) *English as a Lingua Franca: A Corpus-based Analysis*, London: Continuum.
Rayson, P. (2003) 'Matrix: A Statistical Method and Software Tool for Linguistic Analysis Through Corpus Comparison', unpublished thesis, Lancaster University.
Renouf, A. (1993) *Making Sense of Text: Automated Approaches to Meaning Extraction*, Proceedings of the 17th International Online Information Meeting, 7–9 December 1993.

——(2003) 'WebCorp: providing a renewable data source for corpus linguists', in S. Granger and S. Petch-Tyson (eds) *Extending the Scope of Corpus-based Research: New Applications, New Challenges*, Amsterdam: Rodopi.

Renouf, A., Kehoe, A. and Banerjee, J. (2007) 'WebCorp: an integrated system for Web text search', in M. Hundt, N. Nesselhauf and C. Biewer (eds) *Corpus Linguistics and the Web*, Amsterdam: Rodopi.

Schmitt, N. (ed.) (2004) *Formulaic Sequences*, Amsterdam: John Benjamins.

Scott, M. (1997) 'PC analysis of key words – and key key words', *System* 25(2): 233–45.

Seidlhofer, B. (ed.) (2003) *Controversies in Applied Linguistics*, Oxford: Oxford University Press.

Simpson, R. (2004) 'Stylistic features of academic speech: the role of formulaic expressions', in U. Connor and T. A. Upton (eds) *Discourse in the Professions: Perspectives from Corpus Linguistics*, Amsterdam: John Benjamins.

Simpson-Vlach, R. and Ellis, N. C. (2010) 'An academic formulas list: new methods in phraseology research', *Applied Linguistics* [advance access published online 12 January 2010]: 1–26.

Sinclair, J. McH. (1991a) *Corpus, Concordance and Collocation*, Oxford: Oxford University Press.

——(1991b) 'Shared knowledge', in J. E. Alatis (ed.) *Linguistics and Language Pedagogy: The State of the Art*, Washington, DC: Georgetown University Press.

——(1996) 'The search for units of meaning', *Textus* 9(1): 75–106.

——(2003) *Reading Concordances*, Harlow: Longman.

——(2004) *Trust the Text: Language, Corpus and Discourse*, London: Routledge.

——(2005) 'Corpus and text: basic principles', in M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford: Oxbow Books.

Sinclair, J. McH. (ed.) (1987) *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*, London: Harper-Collins.

Stubbs, M. (1995) 'Collocations and semantic profiles: on the cause of the trouble with quantitative methods', *Functions of Language* 2(1): 1–33.

——(1996) *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture*, Oxford: Blackwell.

——(2009) 'Memorial article: John Sinclair (1933–2007): the search for units of meaning: Sinclair on empirical semantics', *Applied Linguistics* 30(1): 115–37.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*, Amsterdam: John Benjamins.

Tribble, C. and Jones, G. (1997) *Concordances in the Classroom: A Resource Guide for Teachers*, 2nd edn, Harlow: Longman.

Widdowson, H. G. (1991) 'The description and prescription of language', in J. E. Alatis (ed.) *Linguistics and Language Pedagogy: the State of the Art*, Washington, DC: Georgetown University Press.

——(2000) 'On the limitations of linguistics applied', *Applied Linguistics* 21: 3–25.

Wray, A. (2002) *Formulaic Language and the Lexicon*, Cambridge: Cambridge University Press.

——(2008) *Formulaic Language: Pushing the Boundaries*, Oxford: Oxford University Press.