24

Corpus linguistics

Studying language as part of the digital humanities

Gill Philip

Introduction

Corpus linguistics is a computer-aided approach to the study of language based on the search for regularities in word use. Such regularities may feature the habitual co-occurrence of two or more words (collocation), of a word with a grammatical or syntactic feature (colligation), a word's consistent use within a particular semantic field (semantic preference), or couched within contexts that suggest connotative or evaluative meanings (semantic prosody). The focus is firmly placed on the word, or more specifically, the word *form*, since even singular and plural forms of a noun are of preference viewed as different 'words'. At first glance this might seem an unnecessarily pedantic distinction to make, but its relevance can be appreciated from a simple example: 'eye' co-occurs (collocates) with the words 'blink', 'opener', 'cast', 'private', and 'witness', amongst others (Sinclair 2003: 168), while 'eyes' counts amongst its collocates colours, body parts, and vision verbs (ibid.). From this information it can be appreciated that the plural tends to refer to the physical feature or organ of sight, and that the singular form is, instead, more typically involved in compounds and idiomatic expressions, e.g. 'an eye-opener', 'cast an eye over something', 'an eye-witness account'.

Observations such as this lay bare typical features of language use which speakers often recognise only in retrospect – a phenomenon Louw (1993: 173) has described as 'twenty-twenty hindsight'. Declarative knowledge of the language is not necessarily unreliable, but it is selective: all users of language, linguists included, tend to focus on salient forms and meanings with the result that much is overlooked. Corpora make it possible to overcome this partial blindness, thus enriching descriptions of language with detail that is normally inaccessible to our intuitions (ibid.: 157). Since the salient meaning of 'eye(s)' is the organ of sight, it is reasonable to assume that both singular and plural refer to a person's facial features or to the act of seeing. The data reminds us that other meanings exist, and allows us to identify how they are realised via combination with other words. It is this sense of discovery that lies at the heart of corpus linguistics.

This chapter will focus on the main contributions of corpus linguistics to English Language studies. Part 2 looks at what a corpus is and how it is compiled, the main functions

found in concordancing software, and how these are used in corpus linguistics. Part 3 addresses current issues in the use of corpus linguistics in the study of English. It introduces the three broad approaches currently in use, namely corpus-driven, corpus-based, and corpus-assisted, and goes on to discuss the ways in which they have been applied to the study of English language, discourse and literature. Part 4 outlines the main areas of debate within corpus linguistics, in particular which approach is best for different kinds of research, and the integration of corpus linguistics with existing theoretical and methodological practices. Part 5, by way of conclusion, suggests likely future developments in terms of data (what can be studied), tools (how to study it) and theory (outstanding issues requiring attention).

Studying language through electronic text: corpora and concordancing software

What is a corpus?

The term 'corpus' is used in many branches of linguistics, as a general term meaning 'a collection of examples'. In corpus linguistics, a corpus is defined as a *principled* collection of *naturally-occurring texts*. The texts may be written or spoken (or, more recently, multimodal), of varying lengths, and represent any range of text types, genres and subject matter. What differentiates a 'corpus' from a 'text collection' is that its composition is based on specific guidelines, aimed at ensuring breadth of coverage and balance (no single text or author may be included disproportionately). A large archive of texts, such as www.gutenberg.org is a 'text collection' rather than a 'corpus', since it is not concerned with offering such breadth and balance: it makes available all it can. The Internet, used by many as a corpus, is for the same reason more accurately described as a text collection. However, a corpus can be *drawn* from a text collection or repository:

> The data were collected through an online interface of newspaper and periodicals (LexisNexis) by way of the following search query:
>
> refugee* OR asylum* OR deport* OR immigr* OR emigr* OR migrant* OR illegal alien* or illegal entry OR leave to remain AND NOT deportivo AND NOT deportment
>
> Data were collected from nineteen UK newspapers, including six daily tabloids (*Sun, Daily Star, People, Daily Mirror, Daily Express, Daily Mail*) and their Sunday editions (*Sunday Express, Mail on Sunday, Sunday Mirror, Sunday Star*), five daily broadsheets (*Business, Guardian, Herald, Independent, Telegraph*), two Sunday broadsheets (*Observer, Independent on Sunday*), and two regional newspapers (*Evening Standard, Liverpool Echo*). Data were obtained for most of the newspapers from January 1996 through October 2005, although in a few cases data were not available until 1999 (*Business*), 2000 (*Sun, Daily Star, Sunday Star*), or 2001 (*Liverpool Echo*).
>
> (Gabrielatos and Baker 2008: 9)

The corpus whose composition is described above is an illustration of a new generation of specialised corpora. Once known as small corpora – small both in size and in coverage – such specialised corpora can now be very large (e.g. '140 million words, containing 175,000 full newspaper articles spanning ten years', ibid. 6). The specialised focus is ensured by specifying which key words should appear in the texts; and the genre (news) is likewise

```
<w lemma="why" type="AVQ">Why </w>
<w lemma="can" type="VM0">ca</w>
<w lemma="not" type="XX0">n't </w>
<w lemma="bloody" type="AJ0"> <seg function="mrw" type="met"
vici:morph="n">bloody </seg> </w>
<w lemma="rabbit" type="NN2">rabbits </w>
<w lemma="come" type="VVB">come </w>
<w lemma="and" type="CJC">and </w>
<w lemma="eat" type="VVB">eat </w>
<w lemma="i" type="DPS">my </w>
<w lemma="lawn" type="NN2">lawns</w>
<c type="PUN">?</c>
```

Figure 24.1 POS-tagged, lemmatised, and metaphor-tagged data

restricted, making this corpus appropriate for sociolinguistic study but not for more general forms of language description, such as lexicography.

Specialised corpora need not be so large. It is a relatively easy matter for a researcher to compile a corpus for their own research, teaching or translation purposes, since 'a corpus' – once planned and the texts selected and located – is one or more files saved in .txt format. The crucial point is that the texts must be selected for a clear purpose so that they can be said to be a representative sample of whatever type of language the researcher wishes to examine. It is also possible to call the complete works of an author, or all extant texts of a period, a corpus.

Although some researchers prefer to compile their own corpus, many corpora are available, often free of charge for academic use (see Xiao 2008). These are usually enhanced with metadata, i.e. data about the data, including date, source, and author of each text. Very often they are annotated, or 'tagged', meaning that information is inserted for each word, e.g. to specify its part of speech (POS), or to indicate the lemma (the base form, or root, which all derived forms can be traced back to, e.g. BE is the lemma for *be, being, been, am, are, is, was,* and *were*). Tagging can additionally be used to manage variant spellings (ibid.), which is particularly relevant when dealing with Middle and Early Modern English texts, or social media texts where abbreviations, creative spellings and typos are common. Semantic tagging also exists, but this is far more complex than POS-tagging or lemmatisation, because the number of possible semantic categories is non-finite (Rayson et al. 2004; see also Rayson 2009). Some researchers tag their data manually, often to highlight language functions. For example, Semino and Short (2004) tagged speech and thought presentation in a range of different genres; and a portion of the British National Corpus (BNC) has been tagged for metaphor (Steen et al. 2010). Figure 24.1 illustrates the sentence 'Why can't bloody rabbits come and eat my lawns?' (BNC data) that has been lemmatised, POS (part-of-speech)-tagged, and metaphor-tagged. In the example, it can be seen that each word is coded separately, e.g. 'lawn' is the lemma (w lemma='lawn') to which 'lawns', a plural common noun (type='NN2') belongs.

Combining a word and a tag in a corpus search makes it possible to carry out sophisticated searches: for example, a search for 'minute', which can be either a noun (ˈmɪnɪt) or an adjective (maɪˈnjuːt), can be limited to just the adjective. If the data has been lemmatised, a search for the lemma BE will simultaneously retrieve all inflected forms of the verb, thus circumventing the need to carry out separate searches for each one. It is also possible to search by tag, thereby retrieving all instances of a tagged category, e.g. metaphors in the metaphor-tagged portion of the BNC data (Steen et al. 2010).

Concordancing software: main functions

The increasing availability of electronic text over the past two decades has made corpus compilation simple and accessible. However, the corpus is just one part of 'doing corpus linguistics' (Hunston 2013: 619), which involves combining a basic tool-box (the corpus data and specific software) with methodological know-how. Despite some claims to the contrary (e.g. McEnery and Hardie 2012) all corpus linguistics also makes use of theoretical frameworks, whether using traditional semantic and grammatical labels to describe the form and function of language items, or by referring to other branches of linguistics in order to interpret the data, e.g. by incorporating insights from cognitive linguistics, sociolinguistics, and so on.

Once the corpus data has been prepared, it is accessed using concordancing software designed to make the language data accessible via a range of on-screen viewing options. At its most basic, it allows the user to extract the searched-for word (node) together with a limited amount of surrounding context (co-text), displaying the output as a KWIC (key word in context) concordance such as that shown in Figure 24.2, which shows a KWIC concordance of 'minute' in a corpus compiled from three collections of *Sherlock Holmes* short stories (Conan Doyle 2007, 2008, 2011) hosted at Project Gutenberg www.gutenberg.org/. The user scans the output in a search for repetitions and regularities occurring to the right and left of the node, mainly reading vertically (from top to bottom) rather than horizontally (from left to right).

Although a corpus linguistics approach can be taken to texts in any language, English benefits particularly since it has a very simple morphology: even the most highly-inflected lemmas (e.g. BE) only have a handful of possible inflected forms. The corollary of this is that in English, words (more specifically, word forms) rarely have just one meaning. Working from the premise that meanings are associated with words in combination rather than in isolation, corpus linguistics allows researchers to find meanings by identifying the word patternings that characterise them. Even from the small amount of data shown in Figure 24.2, there are two distinct meanings of 'minute' in the data – the time period noun, and the size adjective. Looking more closely at the noun, in this data, 'minute' is not a precise measurement but instead means 'moment' ('a minute later', 'for a minute or so'); while the adjective 'minute' overwhelmingly refers to abstract nouns ('attention', 'examination') rather than physical objects, and can be paraphrased as 'extremely detailed', i.e. an indication of manner rather than size. While neither


```
1       in bed. Then they will not lose a minute, for the sooner they do their work the
2  Pall Mall, and then, leaving me for a minute, he came back with a companion whom I
3        after me. At first it was only a minute' s chat, but soon his visits lengthened,
4      to me." We waited in silence for a minute -one of those minutes which one can
5  the table, and began to study it with minute attention. My indignation at this calm
6     his losses or winnings at cards. A minute examination of the circumstances served
7  some 30 pounds, to say nothing of the minute knowledge which you have gained on every
8     there rose a thin spray of smoke. A minute later a carriage and engine could be
9       was leaning through the window. A minute later, however, when Hunter rushed out
10  There had been a ring at the bell. A minute later we heard steps upon the stairs,
11    wasting your time, sir, and every minute now is of importance,' cried the
12 his anger and resumed his seat. For a minute or more we all sat in silence. Then the
13  when your master left?" "Only for a minute or so. Then I locked the door and went
14  live. I sat frozen with horror for a minute or two. Then I seized the poker and went
15      had happened then? I stood for a minute or two to collect myself, for I was
16 had deduced from signs so subtle and minute that, even when he had pointed them out
17 room! Perhaps you will kindly wait a minute until I have examined the floor. No, I
```

Figure 24.2 KWIC concordance of 'minute' in *Sherlock Holmes*

of these meanings of 'minute' is unusual, they are sub-senses, not main senses of their respective noun and adjective forms, and the context allows us to identify them as such.

The interaction of words: collocations, wordlists, key words, clusters

Words tend to co-occur, or collocate, with a limited range of other words, and collocation makes it possible to distinguish between different meanings. Identifying collocations by their perceived frequency of occurrence in the language has always been possible (see especially Palmer 1933 and Firth 1957), but it does not follow that frequent collocations are also significant collocations. In corpus linguistics, collocation is tied to statistical measures of significance: a collocation, to be recognised as such, must occur at least twice in the corpus, and its statistical significance is determined using tests of standard deviation, e.g. *t-score*, or of observed/expected occurrence, e.g. *Mutual Information* (Church and Hanks 1990). Statistical tests offer validity and perspective, demonstrating the degree to which the co-occurrence of words in a collocation deviates from a hypothetically random distribution of words in the language (see Evert 2009). Computational retrieval and statistical calculation can also take account of collocates found not only in the immediate proximity of the node (contiguous collocations) but also, typically, up to five words before or after. Many concordancing software packages allow users to view the distribution of collocates, listing the number of times they occur in each position (Figure 24.3a) or by presenting their distribution in each position in descending frequency (Figure 24.3b).

Concordance patterns like the one in Figure 24.3b represent a first step in seeing recurrent phrasal chunks or multi-word expressions, but the researcher has to piece them together mentally. However, the identification of recurrent strings can also be computed in most software packages. Figure 24.4 shows the output for three-word clusters that incorporate the word 'minute' in the same *Sherlock Holmes* data. Three recurrent clusters are retrieved: 'for a minute' (9 occurrences), 'a minute later' (8) 'a minute or' (8), and the program used (WordSmith Tools v.7, Scott 2017) also offers suggestions as to how these relate to some longer clusters, e.g. 'for a minute or two'.

The type of collocation seen in Figure 24.4, consisting of an uninterrupted string of at least three words, is a multi-word sequence. Corpus linguistics recognises three distinct types of multi-word sequence: clusters, n-grams, and lexical bundles. Clusters incorporate the search word, as in Figure 24.4, while n-grams contain no specified word but are all repeated strings present in the data. Both clusters and n-grams are identified on the basis of their frequency, i.e. they are effectively multi-word word-lists. Lexical bundles are a specific class of clusters

Word	*Texts*	*Total*	*Total left*	*Total right*	*L5*	*L4*	*L3*	*L2*	*L2*	*centre*	*R1*	*R2*	*R3*	*R4*	*R5*
MINUTE	3	34								34					
A	3	31	26	5					26			2	1	1	1
THE	3	15	7	8	1	1	3		2			2	2	1	3
FOR	3	11	10	1			1	9			1				
LATER	3	8	0	8							8				
OR	3	9	1	8	1						8		1		
AND	3	8	3	5	1		1	1			2			2	1
WE	3	7	2	5	1							3	2		
OF	3	7	4	3		1	1	2				2	1		

Figure 24.3a Detailed collocate list for 'minute' in *Sherlock Holmes*

L5	*L4*	*L3*	*L2*	*L2*	*centre*	*R1*	*R2*	*R3*	*R4*	*R5*
TO	WILL	THE	FOR	THE	MINUTE	LATER	TWO	WE	ON	THE
YOU	TO		IN	EVERY		OR	WE	WERE	AND	WHICH
OVER			OF			AND	MORE	WHEN	ALL	
THEN						HE	THE	THE		
							OF	THEN		
							SO	TO		

Figure 24.3b Concordance pattern list for 'minute' in *Sherlock Holmes*

Cluster	*Freq.*	*Length*	*Related*
FOR A MINUTE	9	3	FOR A MINUTE OR (7), FOR A MINUTE OR TWO (3), FOR A MINUTE OR MORE (3)
A MINUTE LATER	8	3	A MINUTE LATER WE (3)
A MINUTE OR	8	3	FOR A MINUTE OR (7), FOR A MINUTE OR TWO (3), A MINUTE OR MORE (3), FOR A MINUTE OR MORE (3)

Figure 24.4 Three-word clusters involving 'minute' in *Sherlock Holmes*

or n-grams which occur *with statistically-significant frequency in particular registers or genres* (Biber et al. 1999: 989, emphasis added). In other words, they occur significantly more often than other clusters/n-grams do in the register or genre being studied.

The computation of collocates relies on calculating the probability of two words occurring together on the basis of their frequency in the data. In order to do this, the software initially compiles a word list, i.e. a list of all word forms (types) in the data and the number of times they occur (tokens). Word lists offer an initial entry point into corpus data by showing the words that are present and their actual frequencies. Some words are more frequent than others; the top of a frequency-determined word list of any English language corpus will always feature closed-class items – articles and determiners, prepositions, and pronouns. The middle cut is lexically rich, featuring recurrent content words, especially basic lexis, while the lowest-frequency words are hyponyms, specialised terminology and other infrequently used words such as proper nouns, foreign words, and non-standard spellings. Words occurring only once in a corpus normally account for well over half the total number of tokens. Viewed on a graph, a frequency-based word list ought to take the form of a sharply-dropping curve with a very long 'tail' comprising the non-repeated forms. Word lists can be applied to studies of authorship attribution, historical change, and regional varieties (see contributions in Archer 2009). However, they cannot reveal significant, communicatively-meaningful frequency: to ascertain whether a frequent word is also a statistically significant (key) word, the word list for the data (the *focus corpus*, Kilgarriff 2009) is compared with another word list, prepared from a reference corpus (ibid.), to generate a key word list.

Key words reveal what a text is about, although precisely how they do so depends on the choice of reference corpus: the key words change when the reference corpus changes. By way of illustration, Figure 24.5 shows two different key word lists for the same *Sherlock Holmes* short story, 'A Scandal in Bohemia', calculated against the BNC and against the small – but focused – *Sherlock Holmes* corpus already described in this chapter. The most obvious difference between the lists is their length: the comparison with the focused corpus yields very

few key words. All of those key words are also found on the list generated with reference to the BNC, but they occupy different ranks and their keyness score is very different. Of the top ten key words in the BNC comparison, *Holmes* and the pronouns 'I', 'my', 'his' and 'you' tell us about the *Sherlock Holmes* stories in general (Watson's first-person narrative, direct speech, predominance of male characters) rather than 'A Scandal in Bohemia' in particular. Less predictable to those who already know the stories are 'cried' and 'street' (ranks 19 and 21 respectively in the BNC comparison) and 'she', which is the only personal pronoun to appear in the comparison with the other *Sherlock Holmes* stories: these are items that a researcher might want to investigate in order to understand why they are key words.

Generating key words by comparing like with like (a specific story with the corpus of stories featuring the same protagonists, in the same genre, by the same author) allows us to identify the 'aboutness' (Phillips 1989) of the short story: the protagonists (King of Bohemia, Irene Adler) and significant objects and places (photograph, Briony Lodge) where the action takes place. Both of the comparisons shown in Figure 24.5 are valid, depending on what the researcher intends to us the key words for; but the differences in the results reminds us that keyness is not an absolute measure, and highlights how important it is to choose a reference corpus with care.

Although collocations, word lists, clusters and key words have been dealt with separately here, it is normal in corpus linguistics research to make use of all of these tools in combination, i.e. to navigate between key words, collocates, clusters, and KWIC concordances. In this way the abstracted, decontextualised view of the data presented in the various data lists

	A Scandal in Bohemia vs BNC			*A Scandal in Bohemia* vs *Sherlock Holmes*		
Rank	*Key word*	*Frequency*	*Keyness*	*Key word*	*Frequency*	*Keyness*
1	HOLMES	48	526.28	PHOTOGRAPH	21	71.99
2	I	261	364.58	MAJESTY	16	71.83
3	BRIONY	11	182.71	KING	17	61.83
4	ADLER	13	161.60	ADLER	13	54.68
5	MY	78	158.53	IRENE	13	54.68
6	MAJESTY	16	156.00	BRIONY	11	49.38
7	PHOTOGRAPH	21	150.65	SHE	71	42.50
8	IRENE	13	122.81	LODGE	11	33.82
9	SHERLOCK	11	120.08	NORTON	7	31.42
10	HIS	105	95.27	BOHEMIA	8	31.39
11	YOU	129	91.05			
12	UPON	25	82.38			
13	BOHEMIA	8	81.67			
14	LODGE	11	77.20			
15	AM	23	66.82			
16	KING	17	62.41			
17	ME	46	62.09			
18	IT	154	60.40			
19	CRIED	10	56.62			
20	NORTON	7	54.54			
21	STREET	18	54.37			
22	HE	109	54.15			
23	SERPENTINE	5	53.23			
24	SHE	71	49.54			
25	REMARKED	8	49.41			

Figure 24.5 Top 25 key words for *A Scandal in Bohemia* calculated against the BNC and *Sherlock Holmes*

is brought back into relation with the original (co)-text. How researchers do this is covered in the next section.

Corpus linguistics and the study of English: current issues

Corpus-driven, *corpus*-based, *or corpus*-assisted?

Although all corpus linguistics makes use of corpora, three distinct strands of research currently coexist: corpus-driven (Tognini-Bonelli 2001), corpus-based (ibid.), and corpus-assisted (Partington 2006). Corpus-driven research aims to cut through the heterogeneity and richness of natural language in order to uncover underlying regularities. Its central tenet is to 'trust the text' (Sinclair 2004), and so it is the patterns attested in the corpus data that determine the direction of the research. This is essentially an exploratory approach, in which word use is mapped out on the basis of co-occurrence features – collocation, colligation, semantic preference, and semantic prosody (Sinclair 1996). All frequently-occurring patterns are be treated on an equal footing, irrespective of their perceived salience. One of the major beneficiaries of the corpus-driven approach has been lexicography, particularly dictionaries for learners of English as a foreign language, where the documentation of non-salient but common uses of language is a high priority.

The corpus-based approach treats the corpus as a source of data on which the researcher can draw 'in order to explore a theory or hypothesis, typically one established in the current literature, in order to validate it, refute it or refine it' (McEnery and Hardie 2012: 6). More specifically, 'corpus-based' is used to describe research which makes use of corpora for purposes that go beyond the description of the data held in the corpus, and is typically used in conjunction with discourse analysis, stylistics, text linguistics, or other methods. The corpora used may be general reference corpora (to check general norms), and/or specialised corpora (to verify norms in a particular genre, register, or discourse); where their use differs from the corpus-driven approach is that only the central, most typical features are typically taken into consideration, leaving aside peripheral uses.

Corpus-assisted research combines the above approaches. It uses corpora to verify researchers' intuitions and thus lend validity to their interpretation, but very often it is to compare norms in the corpus with an 'oddity' in a text. As Partington explains, 'if texts are not compared to other bodies or corpora of texts it is not possible to know or to prove what is normal and only against a known background of what is normal and expected can we detect the unusual and meaningful' (2006: 6-7).

Corpora and language documentation

As mentioned already, the corpus-driven approach is closely associated with lexicography. Foremost in this area is COBUILD, a project which resulted in the publishing of a range of language reference books and the first ever dictionary to be compiled from a corpus, using the data to identify word senses, syntactic preferences, and frequency information. The early years of the project are documented in a collection of papers edited by Sinclair (1987), in which the enthusiasm and excitement of the research team leap off the pages. It was anticipated that corpus data would require language descriptions to be refined, but not that new ways of describing language would be necessary. The most important discovery, now taken for granted within corpus linguistics but still largely overlooked in theoretical linguistics, is that lexis and grammar tend not to operate on separate planes but are instead intertwined. Natural language is largely idiomatic

(in the broad sense of the term), and meaning emerges from words in combination. Collocation is just one of the phrasal types to have been documented in detail. Others include collocational frameworks, lexicogrammatical frames, and semi-prepackaged phrases (see Philip 2011: 35–58).

Although corpora are now deemed essential in lexicography, their direct use in language teaching is rare. Data-driven learning (DDL, see Johns 1991) and other pedagogical applications of corpora are mainly restricted to researcher-teachers working in higher education (Leńko-Szymańska and Boulton 2015), and although learner corpora such as the International Corpus of Learner English (Granger et al. 2002) have shed light on various features of learner English, very little learner corpus research has filtered down directly into mainstream language pedagogy. The indirect use of corpora in English language teaching has, however, been widespread and is particularly evident in the intensified focus on collocation in general English as a Foreign Language textbooks, and the use of large amounts of authentic, albeit usually adapted, text (Meunier and Gouverneur 2009). Collocation is now viewed as an essential part of vocabulary building, necessary for the production of proficient, fluent speech and writing, and few language teaching textbooks fail to treat it systematically. Authentic text lies at the base use of increasingly text-heavy teaching materials: even though such texts are almost always adapted rather than reproduced wholesale (see Clavel-Arroitia and Fuster-Márquez 2014), they support the inductive approach to language learning and increase learners' exposure to phraseology (Meunier and Gouverneur 2007). Exposing learners to authentic texts mirrors, to some extent, the factors that contribute to mother-tongue acquisition, and has favourable repercussions on the acquisition of collocations, lexical bundles, and other native-like features (De Bot et al. 2005).

Two further areas of corpus-driven research have been influential in (academic) English language study. One is the compilation of the Academic Word List (Coxhead 2000), a list of around 3000 words found in general academic English, as attested in a 3.5 million-word corpus of academic texts. The list is widely available on the Internet, and many print and web-based teaching materials have been developed around it. Less well-known outside corpus linguistics is the lexico-grammatical approach to genre research. Biber's (1988) seminal study into spoken and written varieties of English demonstrated how registers can be differentiated by comparing the distribution of phraseological features across a range of dimensions. Such dimensions include, for example 'narrative versus nonnarrative discourse'; here, it is reasonable to expect narrative to feature e.g. more past tense forms, more verbs for speech and thought representation, fewer text-organising features and fewer instances of hedging strategies (amongst other variables) compared to non-narrative discourse. Later work (Biber 2006) demonstrated that related registers can be further differentiated on the basis of lexical bundle distribution. What is important to note is the contribution of corpus-driven analysis, which is intuitively characterised by lexical choices and the presence (or absence) of particular grammatical structures. Biber and other scholars have highlighted the phraseological nature of genre using frequency and co-occurrence measures that can only be revealed via the processing of large amounts of data.

While the use of corpora for lexicographical purposes is almost exclusively corpus-driven, this does not mean that corpora must always be used within such large-scale projects. Individual scholars routinely conduct small-scale studies into lexis, grammar, and phraseology using the same approach. Additionally, it must be stressed that it is not only the standard version of the language that is investigated: there is a vast body of research into all kinds of non-standard Englishes based on corpora of historical, geographical, age-specific and emerging varieties of the language. All these contribute to its documentation – for native speakers, for learners, and for philologists and scholars of the literature of past ages. However,

corpus data is not only useful for lexicology; it is also used to analyse the communicative functions and effects of texts, as outlined in Parts 3.3 and 3.4.

Corpora and text analysis

In discourse analysis, the purpose of the research is not simply to document linguistic features, but to connect them with how language is used to construct meaning and a vision of the world. The use of corpora in this context allows comparisons to be drawn between forms located in the chosen text with general features of the discourse (or of the language) as a whole. In this area of research, specialised corpora are often compiled using texts that characterise the discourse being studied. One such specialised corpus, used to investigate discourses of migration (Gabrielatos and Baker 2008; Baker et al. 2008) was described in Part 2.1. Using a specialised, discourse- or topic-specific corpus allows researchers to move from the micro-analysis of a single text to an intermediate level focusing on corpora of specialised discourse, to the macro-analysis of large amounts of data, so that generic features can be described (Bednarek 2009). It may also be informative to 'downsample' one or more texts from a corpus so that detailed examination can be carried out on a small component of the larger data set, using corpus methods or other linguistic approaches. Baker et al. (2008) do precisely this: they downsampled a selection of texts belonging to a precise time frame with the intention of subjecting them to a CDA (critical discourse analysis), after a quantitative analysis of the corpus suggested this time frame would be worth investigating in more detail.

Since corpus linguistics is centrally concerned with word use, in text and discourse analysis the researcher's focus alternates between the word, the text, and the corpus. If the research is corpus-driven, the ideal procedure would be to start with the corpus data as a whole, extract keywords, collocations and recurrent clusters, and use this information to formulate the initial research questions (e.g. 'Why does 'group(s)' collocate with 'ethnic' and with 'tribal', but not with 'racial'?', cf. Krishnamurthy 1996). In a corpus-based perspective, researchers select in advance which data to focus on, using corpus-external criteria, e.g. the desire to focus on a selection of near-synonyms, to uncover their similarities and differences in a given discourse. For example, Baker et al.'s (2008) study focused on 'refugee(s)', 'asylum seeker(s)', 'immigrant(s)' and 'migrant(s)', in a 140 million-word corpus of news articles dealing with the topic. Their research not only studies these terms: they were used as 'seed terms' for the initial corpus compilation (Gabrielatos and Baker 2008: 9; see 2.1 above) to ensure that their corpus would be centrally relevant to the topic, rather than just a generic corpus of newspaper texts. Another way into the data is to perform a close reading of a sample of the corpus in order to identify words, semantic fields or structures of potential interest, and then to search for them in the full corpus. This approach is favoured by many metaphor scholars and is outlined in detail in Charteris-Black (2004).

A different use of corpus data is found in corpus-assisted (Partington et al. 2013) or corpus-informed (O'Halloran 2007) studies, in which the text being studied is generally not part of the larger corpus nor is it studied on the basis of findings derived from the corpus. Instead, corpus data is used to check how words in the text are typically used in the language as a whole. This allows researchers to verify their intuitions, and can support hypotheses regarding how readers are likely to interpret meanings (ibid.). Bolstering one's claims with corpus data ensures that the treatment of the text does indeed amount to analysis rather than interpretation, and, at the same time, counteracts any tendency to overinterpretation (O'Halloran and Coffin 2004).

At its most detailed, research fusing corpus linguistics and discourse analysis is capable of revealing not only the main topics of a discourse, but also lexicogrammatical patterns that

communicate stance, positive/negative evaluation, and connotation, all of which operate above the level of word meaning. Of particular relevance in this context is semantic prosody (Louw 1993), the attitudinal, evaluative and pragmatic force that is encoded in the 'extended unit of meaning' (Sinclair 1996; see also Philip 2011: 38–82). In other words, a type of meaning usually considered intangible can actually be pinned down to specific patterns and structures in the language. Semantic prosody is subtly revealing and elusive in equal measure. Not identifiable in a single instance of language, and not inherent in word-level semantics, it is in the KWIC concordance that it surfaces.

Finding and defining semantic prosody requires careful analysis. In Louw (1993), KWIC concordances reveal how 'bent on' collocates with negative (destructive, disruptive) activities, including 'destroying', 'harrying', 'mayhem'. The expectation is, therefore, that this expression will normally be used to convey annoyance (the semantic prosody) in relation to individuals who are *bent on* such activities. Louw argues that intentional deviation from this expectation results in irony (1993: 171). Unintentional deviation, on the other hand, reveals insincerity (ibid). By way of example, Philip (2017) discusses an unfortunate slip of the tongue in an informal statement made on TV news by the former UK Prime Minister, David Cameron, who described the influx of refugees from war-torn Syria as 'a swarm of immigrants'. Corpus data confirms that 'swarm' is typically used to refer to large quantities of insects, particularly those that bite or sting, and as a result its semantic prosody conveys annoyance that such insignificant creatures should be so bothersome. To find 'immigrants' in the syntactic slot normally occupied by insects transfers this prosody onto the people arriving. Contemporary news reports confirmed the public reception of the phrase as being negative, dehumanising, and – importantly – revealing the speaker's private sentiments; statements to the press attest that Cameron had not used the expression deliberately (ibid.). O'Halloran's (2007) analysis of metaphorically-used words such as 'erupt' and 'simmer' offers a reminder of the subtlety of semantic prosody: different inflected forms (present participle, past simple) appear to harbour distinct evaluative meanings; and these may additionally be associated with particular registers, e.g. journalism.

Corpora and stylistics

Corpus stylistics is normally defined as the analysis of literary texts using corpus linguistic techniques, although some argue that its scope is wider, involving 'the application of theories, models and frameworks from stylistics in corpus analysis' (McIntyre 2015: 60–61). However, it is on literary texts than most corpus stylistic research currently concentrates. Using corpora changes the reading paradigm from the chronological, horizontal unfolding of the text, to a synchronic, vertical reading (Tognini-Bonelli 2001). While this can be said of all corpus analysis, this change in viewpoint is particularly marked in the study of narrative prose and drama: the narrative progression is fragmented, attention is drawn away from the gradual development of plot and character. Concentrated into the space of an on-screen concordance, subtle aspects of characterisation, plot, and style that are distributed throughout the text come to the fore.

Corpus-driven stylistics is illustrated in Toolan's (2009) work on narrative progression in short stories. He argues that if a form or feature is prominent enough to be noticed by the reader, it must also be identifiable formally, via corpus analytic techniques. Toolan identifies eight textual features in short stories that are central to prospection (clues that allow the reader to anticipate what is to befall the characters), all eight of which can be identified using corpus linguistics tools, including searching for particular word forms or classes (e.g. reporting verbs, modal verbs), extracting key words and their collocates, and investigating clustering of

repetition. Toolan does not limit himself to studying key word lists nor the clusters of which they form a part; instead, he argues that the sentences in which a character key word appears can be extracted and recompiled to provide a potted version of the story. He also observes that key words have a tendency to cluster at particular points of the narrative, and suggests that such clustering indicates plot intensification. Since corpus linguistics analyses work best with repeated word forms, the only phenomenon that Toolan finds problematic is 'para-repetition' (ibid: 103), i.e. repetition of meaning which does not involve the reiteration of the same lexis: corpus software is primarily designed to retrieve recurrent word forms, not recurrent meanings.

Toolan's approach is original within corpus stylistics, where it is more common to use key words to identify the main topic, participants, and stylistic features of a text. For example, Culpeper (2009) uses them to investigate the links between style markers and keyness in Shakespeare's *Romeo and Juliet*, while Scott (2006) is interested in how they collocate with one another, thus revealing how characters, places, events and language interact. Fischer-Starcke (2010) proposes an analysis of Austen novels, starting with key words and then analysing the (four-word) clusters associated with them. Mahlberg (2013), investigates longer (five-word) clusters in Dickens' novels, in an attempt to identify general phraseological features of this author's style, particularly his 'authorial habit' (ibid. 60) of using repeated clusters both within one work and across different texts. Mahlberg has also focused on recurrent clusters within suspensions – the narrator comments that interrupt direct speech – using a purpose-built corpus tool, CLiC (Mahlberg et al. 2016). Mahlberg's (2013) analysis of body-part lexis in suspension clusters demonstrates how body language contributes to the depiction of characters' personalities, appearance, and typical behaviour.

In his early investigations of semantic prosody, Louw (1993) also discusses ways of interpreting poetry by comparing collocations with those found in a general reference corpus. He argues that the 'inescapable feeling of melancholia' (p. 162) experienced by readers of Larkin's 'Days' can be explained by referring to the habitual patterns found in the corpus. Larkin writes 'Days are where we live', which appears to be a neutral statement; yet an examination of 'days are . . . ' in the corpus reveals that it carries a semantic prosody of regret for time past which cannot be relived, not for present time and the actions carried out therein. In a similar vein, Semino (2010) makes use of corpora to validate conventional use of words and their collocates, this time within a metaphor-related study of Elizabeth Jennings' 'Answers'. Another way in which a literary text can be compared to language norms is illustrated by McIntyre (2015). He compares the relative frequencies of direct speech, indirect speech and narrator comment in Mark Haddon's (2005) *The Curious Incident of the Dog in the Night-time* with those found in a corpus of contemporary fiction, in order to confirm his hypothesis that the novel uses a disproportionately large amount of direct speech. McIntyre also compares the dialogues of the TV series *Deadwood* with a contemporary (19th-century) corpus of American English, to investigate the perceived authenticity of the characters' discourse (ibid.), finding both similarities and divergences in the use of anachronisms such as 'an honor and a pleasure'.

Key areas of dispute and debate

How is corpus linguistics to be done?

The most persistent area of dispute and debate within corpus linguistics is that of corpus-driven vs. corpus-based analysis (Tognini-Bonelli 2001). The debate is often polarised: at one extreme is an absolutist empirical stance, whereby the corpus data is reified, requiring the analyst to investigate all (and only) the language it contains; at the other extreme, the corpus is

considered as nothing more than a convenient repository of examples which can be 'cherry-picked' at will. An alternative view of the same debate is to argue whether corpus linguistics is a theory or a method (McEnery and Hardie 2012; Viana et al. 2011). What emerges clearly from Viana et al.'s interviews with notable corpus linguists is that is it both theory *and* method, albeit in varying proportions depending on one's academic background and research interests, and on the scope and slant of each individual study. Corpus linguistics has made considerable contributions to linguistic theory, refining existing models and also proposing new ones. Corpus linguistics is also a method; it is one of many tools that any linguist can make use of, without necessarily embracing (or rejecting) existing theoretical stances, corpus-derived or otherwise. If two broad approaches are to be identified, therefore, it would be more accurate to speak of a lexicologically-oriented stance in which the data serves to make generalisations about the structures and functions of language, and a discourse-oriented stance in which the data is used to enhance the interpretation of texts.

The lexicologically-oriented stance can be seen in studies which focus on a limited set of words, rarely (if ever) examining the context that lies beyond the boundary of the KWIC concordance on-screen. This local focus allows for detailed examination of language items in their immediate textual environment and extrapolates to the corpus as a whole, making it possible to explain how meanings arise in context. Notable advances in linguistic theory have been made using this approach, particularly the idiom principle (Sinclair 1991), linear unit grammar (Sinclair and Mauranen 2006) and the theory of norms and exploitations (Hanks 2013). The idiom principle proposes that spoken and written text is normally composed of intersecting or superimposed phrasal fragments, rather than by alternating starkly between grammatical and lexical choices. Corpus data supports this notion: on-screen KWIC concordances as well as lists of collocates and n-grams demonstrate repeatedly that words co-occur in recurring patterns, that they 'prefer' to be used as part of a limited set of collocates, phrasal structures, syntactic positions, and so on. Linear unit grammar extends this observation, focusing on the way in which each consecutive word in an utterance or text reduces the potential for other words to appear. This theory of language can be seen in predictive text applications and Internet search engines, where the user is frequently prompted for ways of completing phrases after entering just one or two words. Linear unit grammar does not stop at three- or four-word sequences, however, but suggests that *in extremis*, entire texts may be built up by a process of prediction and exclusion based on the words that have already appeared. The theory of norms and exploitations is also a phraseological theory of language, but it aims to explain creative or unusual word combinations in relation to formulaic language including the collocations and n-grams that typify the idiom principle. Each of these theories may be described independently, but they share two principles: first, a rejection of the traditional opposition between grammar and lexis, and second, a view of language that is based on habit and familiarity rather than on a constant renewal and re-composition of word combinations. These theoretical notions are still viewed with scepticism outside corpus linguistics, where traditional views of grammar, syntax and semantics still hold sway.

Within the discourse-oriented stance, the analysis is primarily text-oriented rather than word- or language-oriented. In other words, the analysis of the lexical items is not an end unto itself (lexicological) nor is it intended to be automatically extrapolated to the language (or language variety under study) as a whole, but rather aims to contribute a quantitative aspect to the (predominantly qualitative) interpretation of texts. Ideally, several levels of interpretation should be included: 'macro- (large-scale quantitative analysis), meso- (small-scale quantitative analysis), and micro- (individual text analysis) levels' (Bednarek 2009), in the spirit of mixed-methods research (ibid.). The study by Baker et al. (2008), discussed in Part 3.3,

is an example of this approach: a large specialised corpus was compiled from a list of pre-selected terms which were then examined in the large corpus; successively, a smaller section of the corpus was isolated; finally, individual texts appearing in that sub-corpus were subjected to qualitative analysis using a CDA approach.

The two broad approaches just outlined tend not to overlap: individual corpus linguists favour one or the other, depending on where their initial training was conducted and the type of language research that interests them. Each will invariably claim that their preferred approach is 'better' than the other, but in truth there is no 'best' way of doing corpus linguistics research. Instead, researchers need to use the approach that is best suited to addressing their research objectives, combining as many features as necessary in order to reach this aim.

Integrating corpora with other linguistic analyses

Corpora introduce a degree of quantitative analysis in areas of language study that are predominantly qualitative, particularly discourse analysis, literary text analysis, and stylistics. Using corpus tools, the researcher can focus on isolated forms in a way that close (and chronological) reading of texts does not allow, picking out details that might otherwise go unnoticed. However, caution may be advised. McIntyre (2015: 60) remarks that some corpus linguists overstate their case by implying that text analysis carried out without corpus tools somehow lacks rigour (ibid.). Corpus linguistics provides a level of quantitative analysis and replicability that is difficult to attain without the aid of computers, but literary scholarship conducted along traditional lines is by no means inferior if it fails to include statistical scores, collocates listings and the like: it is, after all, primarily qualitative. Corpus methods can enhance such research by adding a rigorous quantitative dimension, but they cannot replace it entirely. Another potential pitfall awaiting corpus linguists working with literary text is to overlook the importance of the existing scholarship in the field, comprising both the detailed textual analyses just mentioned, and important insights into the context of production and reception of the texts under examination. Meaningful, insightful interpretation does not come about only as a result of empirical analysis: even the most impeccable analysis can appear naïve if not supported by adequate and appropriate background knowledge.

Criticisms of corpus linguistics, notably from (critical) discourse analysis and systemic functional linguistics, take issue with the validity of focusing on words in a restricted context. Although corpus linguistics techniques offer new insights, they cannot be seen as substituting the analysis of complete texts (Hunston 2013). Indeed, the close attention to words that is typical of corpus approaches can distort the interpretation of a text as a whole, in that it overlooks the ways in which the specific lexical items contribute to text-level meanings, including cohesion, coherence and linguistic function. For this reason – as in literary stylistics – corpus linguistics is often used to add a quantitative dimension to text linguistics, but it is predominantly viewed as an additional, complementary means of analysis, subordinate to existing qualitative methods. It offers one of several computer-based ways of 'reading' texts which allow scholars to enhance their analyses: it is indeed one of the earliest manifestations of what is now referred to as digital humanities.

Future directions

The increasing availability of electronic text has not only increased the scope for corpus linguistics studies: the appeal of corpus linguistics methods is extending beyond the language sciences and into the humanities at large. Corpus linguistics has much to offer the many areas

of study which make use of language data, since the tools and analytical techniques developed for linguistic description are equally suited to the analysis of text-based source materials in history, geography, sociology, and elsewhere. While the primary purpose of digitisation is often simply to make precious documents available without exposing them to unnecessary deterioration, one important corollary is that a gold mine of previously under-documented language varieties is opened up for linguistic research. Not only are corpus linguists exploiting the greater availability of text data to linguistic ends; they are increasingly participating in cross-disciplinary research too, contributing their expertise in the processing and analysis of text data to other fields of study. Corpus linguistics, once seen as a niche area, is therefore poised to take on a pivotal role within digital humanities research.

Within corpus linguistics, the digitisation of library holdings has had a major impact on diachronic language study in particular. Although English corpus linguistics has encompassed historical varieties from the outset, the quantity and range of data available has always been limited in comparison with contemporary varieties. Drawing on data that is now available on digital format, thanks to the systematic digitisation of library holdings, historical corpus linguistics is now able to refine existing studies of grammar and lexis and is also able to address other linguistic phenomena within a historical perspective, such as pragmatics, which was previously beyond its reach (Kytö 2011).

Over the past decade the widespread use of social media has led to new ways of using language, and there is much to be investigated in the emerging varieties of English that ensue (see Spilioti, this volume). A hybrid genre is evolving that is part-spoken, part-written (Knight et al. 2014). 'Conversations' with others are increasingly asynchronous, characterised by short turns, non-standard spellings, and alternation between text, emoticons and images. Conversation threads may in fact comprise several interlaced conversations – as well as irrelevant intrusions (e.g. trolling) – meaning that the very notion of what 'a text' is needs to be refined or indeed redefined.

Corpus analysis tools continue to be developed in response to the specific needs of researchers, and researchers working with English have a distinct advantage since it is the best-resourced language both in terms of data availability (corpora, electronic text) and software. It is usually the test bed for new applications, with the result that innovations in English corpus linguistics tend to precede those in other languages. Data visualisation is currently a growth area. Once limited to the kind of output visible in Figures 24.2–24.5, there is a growing interest in graphic interfaces such as word clouds and dynamic models of word use. Word cloud technology is widespread and easily accessible, but output is normally static. Recent tools allow users to navigate between the word cloud and the texts from which it has been generated *WordWanderer* (Dörk and Knight 2015). Word relationships are made visible in *GraphColl* (Brezina et al. 2015), which shows each word as a node, and draws lines of differing lengths to connect closer (significant) and more distant (less significant) collocates; the output is dynamic and can be manipulated on-screen using the mouse. These developments represent a growing trend for more flexible, interactive and meaningful forms of data visualisation, appealing to those outside the corpus linguistics community.

Corpus linguistics software can process and display data in such detail that some might be tempted to believe that the software is carrying out the linguistic analysis. However, the problem of repeated meanings which do not manifest in repeated word forms remains, as does the question of how to capture nuances of meaning which analysis of frequent collocations does not capture. Semantic tagging is not consistently reliable and improving automatic semantic-class assignation is a challenge: when dealing with hundreds of semantic classes, the error rate inevitably rises; and determining which class fits the word, given the context

in which it occurs, introduces yet more complexity. Unlike grammatical classes, which are unequivocal, a word can belong to several semantic fields at the same time, meaning that even output which is 'correct' may not be 'complete'. Connected to the matter of semantic tagging is what is sometimes called 'long tail' semantics, in reference to the tailing off of repeated word forms in a corpus when viewed on a typical distribution curve (see 2.2 above). Corpus linguistics allows researchers to investigate frequent forms, but has yet to offer reliable means of retrieving frequent meanings, often realised as low-frequency or non-repeated forms in the long tail, and often clustering in single texts within the corpus (Serrano et al. 2009). Research into this area is only just beginning, and is driven by market concerns rather than purely academic interest: Internet search engines need to incorporate lesser-used lexicalisations of popular search-strings into their algorithms, so that the sites retrieved satisfy users' requirements. This technology will inevitably be made available within corpus linguistics, thus refining current semantic tagging systems and allowing users to search for meanings, not just words.

Further reading

Baker, P. and J. Egbert (eds) (2016) *Triangulating Methodological Approaches in Corpus Linguistic Research*. London: Routledge. In this volume, ten scholars tackle the same research question using the same 400,000-word corpus, each from their preferred methodological-analytical angle.

McEnery, T. and A. Hardie (2012) *Corpus Linguistics. Method, Theory and Practice*. Cambridge: Cambridge University Press. A good, all-round introduction to (English) corpus linguistics, with tasks and study questions.

Scott, M. (2017) *WordSmith Tools Version 7*. Stroud: Lexical Analysis Software. One of the most widely-used and functionally complete concordancing packages available; the demo version is free.

Sinclair, J. M. (2004) *Trust the Text*. London: Routledge. A collection of Sinclair's major writings on corpus linguistics, focusing particularly on analytical procedures, descriptive techniques, and theoretical implications.

Viana, V., S. Zyngier and G. Barnbrook (2011) *Perspectives on Corpus Linguistics*. Amsterdam: John Benjamins. Fourteen well-known corpus linguists give their answers to a set of general questions about the discipline, plus others that are specific to their field of expertise.

Related topics

- Stylistics: studying literary and everyday style in English
- Sociolinguistics: studying English and its social relations
- Discourse analysis: studying and critiquing language in use.

References

Archer, D. (ed.) (2009) *What's in a Word-List?* Farnham: Ashgate.

Baker, P., C. Gabrielatos, M. Khosravinik, M. Krzyżanowski, T. McEnery and R. Wodak (2008) 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', *Discourse & Society* 19 (3): 273–306.

Bednarek, M. (2009) 'Corpora and discourse: A three-pronged approach to analyzing linguistic data', in M. Haugh (ed.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*. Somerville, MA: Cascadilla Proceedings Project, 19–24.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (2006) *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

Brezina, V., T. McEnery and S. Watten (2015) 'Collocations in context: A new perspective on collocation networks', *International Journal of Corpus Linguistics* 20 (2): 139–173.
Charteris-Black, J. (2004) *Corpus Approaches to Critical Metaphor Analysis*. Basingstoke: Palgrave Macmillan.
Church, K. and P. Hanks (1990) 'Word association norms: Mutual information and lexicography', *Computational Linguistics* 16 (1): 22–29.
Clavel-Arroitia, B. and M. Fuster-Márquez (2014) 'The authenticity of real texts in advanced English language textbooks', *ELT Journal* 68 (2): 124–134.
Conan Doyle, A. (2011) *The Adventures of Sherlock Holmes*. Urbana, Illinois: Project Gutenberg. Retrieved June 13, 2017, from www.gutenberg.org/ebooks/1661
Conan Doyle, A. (2008) *Memoirs of Sherlock Holmes*. Urbana, Illinois: Project Gutenberg. Retrieved June 13, 2017, from www.gutenberg.org/ebooks/834
Conan Doyle, A. (2007) *The Return of Sherlock Holmes*. Urbana, Illinois: Project Gutenberg. Retrieved June 13, 2017, from www.gutenberg.org/ebooks/108
Coxhead, A. (2000) 'A new academic word list', *TESOL Quarterly* 34 (2): 213–238.
Culpeper, J. (2009) 'Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet', *International Journal of Corpus Linguistics* 14 (1): 29–59.
De Bot, K., W. Lowie and M. Verspoor (2005) *Second Language Acquisition. An Advanced Resource Book*. London: Routledge.
Dörk, M. and D. Knight (2015) 'WordWanderer: A navigational approach to text visualisation', *Corpora* 10 (1): 83–94.
Evert, S. (2009) 'Corpora and collocations', in A. Lüdeling and M. Kytö (eds), *Corpus Linguistics: An International Handbook*. Vol. 2. Berlin: Mouton De Gruyter, 1212–1248.
Firth, J. R. (1957) *Papers in Linguistics 1934–1951*. London: Oxford University Press.
Fischer-Starcke, B. (2010) *Corpus Linguistics and the Study of Literature*. London: Continuum.
Gabrielatos, C. and P. Baker (2008) 'Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press, 1996–2005', *Journal of English Linguistics* 36 (1): 5–38.
Granger, S., E. Dagneaux and F. Meunier (eds) (2002) *The International Corpus of Learner English*. Louvain-la-Neuve: Presses universitaires de Louvain.
Hanks, P. (2013) *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.
Hunston, S. (2013) 'Systemic functional linguistics, corpus linguistics, and the ideology of science', *Text & Talk* 33 (4–5): 617–640.
Johns, T. (1991) 'Should you be persuaded – two examples of data-driven learning materials', *English Language Research Journal* 4: 1–16.
Kilgarriff, A. (2009) 'Simple maths for keywords', in M. Mahlberg, V. González-Díaz and C. Smith (eds), *Proceedings of Corpus Linguistics CL2009*. Liverpool: University of Liverpool. Available at www.ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc
Knight, D., S. Adolphs and R. Carter (2014) 'CANELC: Constructing an e-language corpus', *Corpora* 9 (1): 29–56.
Krishnamurthy, R. (1996) 'Ethnic, racial and tribal: The language of racism?', in C. R. Caldas-Coulthard and M. Coulthard (eds), *Texts and Practices: Readings in Critical Discourse Analysis*. London: Routledge, 129–149.
Kytö, M. (2011) 'Corpora and historical linguistics', *Belo Horizonte* 11 (2): 417–457.
Leńko-Szymańska, A. and A. Boulton (eds) (2015) *Multiple Affordances of Language Corpora for Data-Driven Learning*. Amsterdam: John Benjamins.
Louw, W. E. (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, 157–176.
Mahlberg, M. (2013) *Corpus Stylistics and Dickens's Fiction*. London: Routledge.
Mahlberg, M., P. Stockwell, J. de Joode, C. Smith and M. O'Donnell (2016) 'CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics,' *Corpora* 11 (3): 433–463.
McEnery, T. and A. Hardie (2012) *Corpus Linguistics. Method, Theory and Practice*. Cambridge: Cambridge University Press.
McIntyre, D. (2015) Towards an integrated corpus stylistics. *Topics in Linguistics* 16: 59–68.

Meunier, F. and C. Gouverneur (2007) 'The treatment of phraseology in ELT Textbooks', in E. Hidalgo, L. Querada and J. Santana (eds), *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi, 119–139.
Meunier, F. and C. Gouverneur (2009) 'New types of corpora for new educational challenges Collecting, annotating and exploiting a corpus of textbook material', in K. Aijmer (ed.), *Corpora and Language Teaching*. Amsterdam: John Benjamins, 180–201.
O'Halloran, K. (2007) 'Critical discourse analysis and the corpus-informed interpretation of metaphor at the register level', *Applied Linguistics* 28 (1): 1–24.
O'Halloran, K. and C. Coffin (2004) 'Checking overinterpretation and underinterpretation: Help from corpora in critical linguistics', in C. Coffin, A. Hewings and K. O'Halloran (eds), *Applying English Grammar: Functional and Corpus Approaches*. London: Hodder Arnold.
Palmer, H. E. (1933) *Second Interim Report on English Collocations*. Tokyo: Kaitakusha.
Partington, A. (2006) *The Linguistics of Laughter.* London: Routledge.
Partington, A., A. Duguid and C. Taylor (2013) *Patterns and Meanings in Discourse*. Amsterdam: John Benjamins.
Philip, G. (2011) *Colouring Meaning*. Amsterdam: John Benjamins.
Philip, G. (2017) 'Conventional and novel metaphors in language', in E. Semino and Z. Demjén (eds), *Routledge Handbook of Metaphor and Language*. London: Routledge, 219–232.
Phillips, M. (1989) *Lexical Structure of Text*. Birmingham: Birmingham ELR.
Rayson, P. (2009) *Wmatrix Corpus Analysis and Comparison Tool*. Computing Department, Lancaster University. Available at www.ucrel.lancs.ac.uk/wmatrix
Rayson, P., D. Archer, S. L. Piao and T. McEnery (2004) 'The UCREL Semantic Analysis System', *LREC 2004 Proceedings* 7–12.
Scott, M. (2006) 'Key words of individual texts: Aboutness and style', in M. Scott and C. Tribble (eds), *Textual Patterns: Key words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins, 55–72.
Scott, M. (2017) *WordSmith Tools* (version 7). [computer software]. Stroud: Lexical Analysis Software.
Semino, E. (2010) *Metaphor in Discourse*. Cambridge: Cambridge University Press.
Semino, E. and M. Short (2004) *Corpus Stylistics*. London: Routledge.
Serrano, M. Á., A. Flammini and F. Menczer (2009) 'Modeling statistical properties of written text', *PLoS ONE* 4 (4): e5372. DOI 10.1371/journal.pone.0005372.
Sinclair, J. M. (ed.) (1987) *Looking Up*. Glasgow: Collins ELT.
Sinclair, J. M. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, J. M. (1996) 'The search for units of meaning', *Textus* 9 (1): 75–106.
Sinclair, J. M. (2003) *Reading Concordances*. London: Longman.
Sinclair, J. M. (2004) *Trust the Text*. London: Routledge.
Sinclair, J. M. and A. Mauranen (2006) *Linear Unit Grammar.* Amsterdam: John Benjamins.
Steen, G., A. Dorst, B. Herrmann, A. Kaal, T. Krennmayr and T. Pasma (2010) *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Amsterdam: John Benjamins.
Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
Toolan, M. (2009) *Narrative Progression in the Short Story.* Amsterdam: John Benjamins.
Viana, V., S. Zyngier and G. Barnbrook (2011) *Perspectives on Corpus Linguistics*. Amsterdam: John Benjamins.
Xiao, R. (2008) 'Well-known and influential corpora', in A. Lüdeling and M. Kytö (eds), *Corpus Linguistics. An International Handbook*. Vol. 1. Berlin: Mouton De Gruyter, 383–457.