Citation: Vaughan, E. and O'Keeffe, A. (2015) "Corpus Analysis". In K. Tracy, C. Ilie and T. Sandel (Eds) *The International Encyclopedia of Language and Social Interaction*, Denver, CO: John Wiley & Sons, Inc. pp. 252-268.

Corpus Analysis

Elaine Vaughan and Anne O'Keeffe

University of Limerick, Ireland and Mary Immaculate College, Ireland

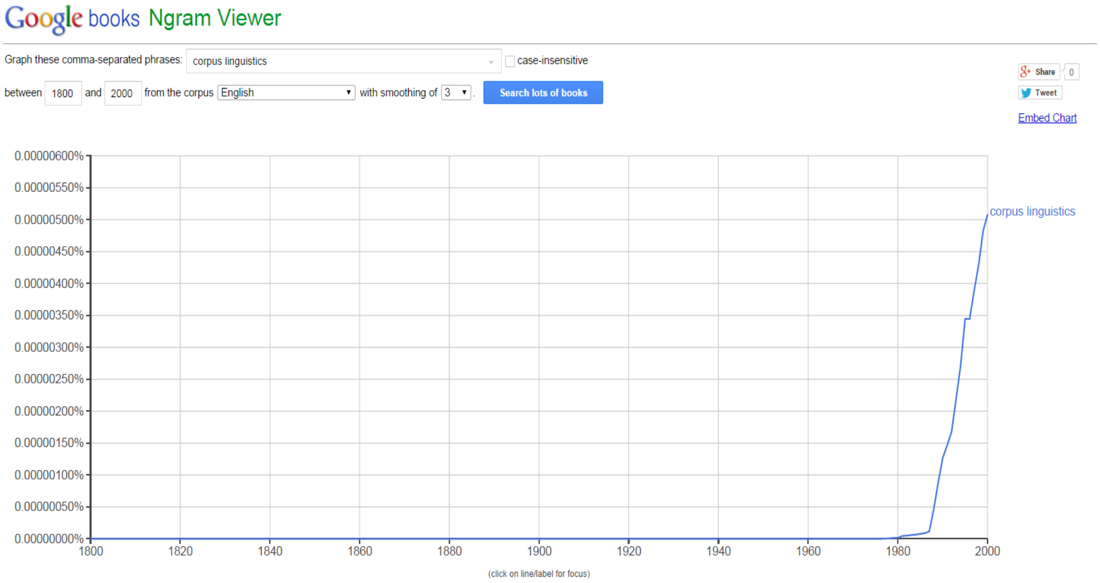
Large and small language text corpora have become quite ubiquitous in the broad fields that make up the study of language and social interaction. This entry provides an introduction to the concept of the "corpus" where language research is at issue and to the field of corpus linguistics. It reviews the main corpus analysis tools and the sort of perspectives they can open for language data. Finally, it gives a very broad overview of the ways in which corpus analysis has to date informed, or is beginning to permeate, different areas of language study.

Background

Corpus linguistics involves the use of computers to rapidly search and analyze databases of real language. These databases are called *corpora* (the plural of Latin *corpus*) and they can comprise any principled collection of written or transcribed spoken language. Examples of well-known corpora are the British National Corpus (BNC), which contains over a hundred million words of mostly written British English, collected between the 1980s and 1993; the American National Corpus, set up as a corpus comparable to the BNC and available as an online resource comprising a total of over 14.5 million words, 3.2 million of which are spoken data; the Corpus of Contemporary American English, a collection of 450 million words of spoken and written English, available and searchable online, out of which 85 million words are spoken data that include unscripted conversations from nearly 150 different TV and radio programs; the Michigan Corpus of Academic Spoken English, available online and comprising more than 150 transcripts of academic speech events recorded at the University of Michigan, USA that total 1.8 million words. English language corpora still tend to dominate, but corpora of many other languages now exist, including Spanish, French, German (both European and Canadian), Mandarin Chinese, Japanese Vietnamese, Egyptian Arabic, Farsi, Bulgarian, Greek (among others). Many of these are available from the Linguistic Data Consortium at the University of Pennsylvania (see www.ldc.upenn.edu). ELDA, the Evaluations and Language resources Distribution Agency in Europe, also makes available a number of corpus resources in different languages (see www.elda.org).

When the notion of using a computer to store and analyze real language first came into being in the late 1950s and punched-card technology was used for storage, the processing of 60,000 words took over 24 hours. The rise in popularity and use of corpora in the empirical study of language parallels the development of computer storage capacity and processing speed. At the time of writing, corpora of over a billion words can be searched in an instant. With the advent of cloud computing, storage is no longer an issue. McCarthy and O'Keeffe (2010) note that it was the 1980s and the 1990s that really saw the arrival of corpora as we know them now—as tools for the linguist or the applied linguist. McEnery, Xiao, and Tono (2006) point out that the more specific term *corpus linguistics* did not come into common usage until the early 1980s, when it was coined in Aarts and Meijs (1984). At this point scholars discussed the possible parameters of the field of "corpus linguistics" and wrestled with issues of how to define what could—or should, in fact—be referred to as a "corpus"; later on they debated the status of corpus linguistics and whether this new approach to considering language constituted a distinct *theory* of language (Tognini-Bonelli, 2001) or rather a *method* of language analysis.

Now, the number of language disciplines that utilize some or all of the tools of corpus analysis has grown exponentially; and the discipline of corpus linguistics, which has as its core a focus on language as a collection of data, has become firmly established in the academic sphere. Figure 1 illustrates the use of the term "corpus linguistics" over time via Google Ngram Viewer. This instrument searches for the term in the published materials available electronically on Google Books and shows quite clearly how its use has increased between the 1800s (when the terms "corpus" might have been used in quite a different way, e.g., to designate all the works of an author like Shakespeare) and its introduction by Aarts and Meijs in 1984.



Graph these comma-separated phrases: corpus linguistics

□ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3 .

Figure 1 The term "corpus linguistics" in published literature over time.
Generated using Google Ngram viewer: https://books.google.com/ngrams

The link between this visual representation of a term coined to describe the new field in the literature and the practice of corpus analysis is, as previously mentioned, technology. The possibilities that access to vast repositories of electronic language data now affords have meant that developments such as using the World Wide Web as a potential language corpus have now become a reality, though one to be taken with certain caveats (see Lee, 2010). This situation is in marked contrast with the early days of concentrated corpus development for the purpose of language analysis. Access to electronically available language samples via the World Wide Web means that the sheer scope of what corpus analysis techniques can be applied to is remarkable, though not without attendant issues of copyright and ethics (see McEnery & Hardie, 2012).

One of the most significant developments for the modern era of corpus-building was the Survey of English Usage (SEU), a project led by Randolph Quirk at University College London, which started in the late 1950s. While it was not conceived of as a corpus in the sense in which the term is understood today (that is, as a searchable database in electronic form), it was revolutionary in that it collected examples of everyday spoken interaction as well as written data. This had not been a priority for linguistic analysis before that point. The SEU also contained written data; but the spoken data it contained were recorded on reel-to-reel tapes, laboriously transcribed, and then typed. Later projects built on this innovation and on the SEU were later computerized as part of the London–Lund corpus project.

A distinction is still made between spoken corpora and written corpora. Written corpora are more plentiful because they are easier, quicker, and cheaper to create in comparison to the protracted and costly process of building spoken corpora, which includes making recordings and painstakingly transcribing them, so that they are in computer-readable form. As noted, technological advances are closely linked to advances in corpus linguistics. Developments in digital recording devices, for example, greatly enhanced the quality of spoken corpora and, with advances in the use of digital media, spoken corpora can now be enriched modally; in this way, for example, video recordings can be aligned with transcripts, and speech, body language, and suprasegmental elements can all be annotated and made accessible to the researcher. These innovations are exciting, but the fact is that the sophisticated transcription of spoken data is still not entirely automated (and may never be). Therefore the cost, in terms of human labor, makes the development of very large corpora of spoken language realistic only for large enterprises conducted by well-established publishing houses.

Even the largest corpora available tend to display a 90 : 10 ratio of written to spoken language; the spoken language contained, for example, in the BNC (see Table 1) comes from media sources (e.g. radio programs) as well as from samples of government meetings (matters of public record) and from spontaneous, naturally occurring conversation. Some corpora are commercially available or freely available for online search, for example the Corpus of Contemporary American English (COCA). At over 450 million words, COCA is one of the largest corpora to be available and searchable in this way. Its website (http://corpus.byu.edu/coca) also incorporates a searchable version of the BNC. The Scottish Corpus of Speech and Texts (SCOTS) has approximately 4.6 million words available to search online at the time of writing (visit http://www.scottishcorpus.ac.uk). There are many more corpora and corpus resources available. One way of locating literature and identifying corpora that may inform and advance a research agenda is via a typology of corpus construction to date. Table 1 summarizes in a very basic way the main types of corpora available and gives some examples of each type. This summary is by no means exhaustive; readers will find more elaborate surveys of information in O’Keeffe, McCarthy, and Carter (2007, Appendix 1) and at a very detailed website maintained by David Lee (visit http://tiny.cc/corpora).

Type of corpus	Main purpose and characteristics	Examples of this type
Sample corpus also known as *general* or *reference corpus*	Usually monolingual corpora that aim to capture features of a language variety (e.g., American English, Irish English) in use in normal, everyday situations. They tend to be "snapshots" of a language, given that they are collected usually at a particular point in time, e.g., between 1980 and 1990.	The American National Corpus (ANC): http://www.americannationalcorpus.org The British National Corpus (BNC): http://www.natcorp.ox.ac.uk
Monitor corpus	A sample or general corpus that is consistently being added to in order to keep the language data it contains current.	The Corpus of Contemporary American English (COCA): http://corpus.byu.edu/coca
Parallel corpus	Two or more corpora of the same texts in different languages that have been translated and can be compared side by side, often line by line.	The English-Norwegian Parallel Corpus (ENPC): https://www.hf.uio.no/ilos/english/services/omc/enpc/
Historical corpus also known as *diachronic* corpus	Texts from different, specified periods of time, which can be used to identify features of language in use at that time, but also to track changes in language use over time. Often this involves digitizing texts that do not originally exist in electronic format.	A Representative Corpus of Historical English Registers (ARCHER): manchester.ac.uk/archer
Learner corpus	Texts gathered to represent the features of learner language, i.e. language used by non-native speakers of a foreign language. The goal of gathering a corpus like this is usually to inform teaching and learning processes and materials.	The International Corpus of Learner English (ICLE): http://www.uclouvain.be/en-cecl-icle.html
Specialized corpus	Specialized corpora that aim to capture a specific type of language use, in order to describe in highly contextualized terms language use in this domain.	The Bergen Corpus of London Teenage Talk (COLT): http://www.hd.uib.no/colt

Table 1 At a glance: main types of corpus and examples.

Analyzing corpora

It has been noted that access to these impressive databases of language has to date yielded a huge amount of detail about language in use, revealed surprising or unexpected patterns of use, and served to challenge received wisdom about "standard" language use. Certainly some of these aspects of use will have been intuited, and it is possible to perceive a pattern of language use without necessarily having recourse to empirical data. However, the particular view on language afforded by corpora has made these sorts of perspectives observable in ways that are not possible to achieve through manual investigation techniques. In fact, simply the ways in which language data can be displayed by corpus software tools and the sort of information that these can compute and present in milliseconds have not necessarily changed the nature of language, but rather the way we perceive it (Hunston, 2002). We review and present practically the basic functionalities of commercially and freely available corpus software below, and we discuss the type of insights they potentially provide on the corpora that we use them to search. The best known tools for linguistic research—both commercially available ones, like WordSmith Tools (Scott, 2008), and freeware ones, like AntConc (Anthony, 2014)—are used in our illustrating examples.

Word lists and clusters

As previously mentioned, once a corpus has been assembled or accessed for research, there are a number of corpus software applications for it, all of which essentially perform the same functions. The first, the generation of a word list, allows the user to load a corpus and investigate basic frequency patterns. This frequency view shows which words are occurring the most regularly in a text or collection of texts. Different concordancers will count "words" in different ways, but for word list purposes the most important distinction is between how many *tokens*—that is, how many individual strings of characters that the software recognizes as individual words—and how many distinct strings (*types*) there are in a text. The sentence *you put your right leg in, your right leg out…* contains 10 tokens and 7 distinct types:

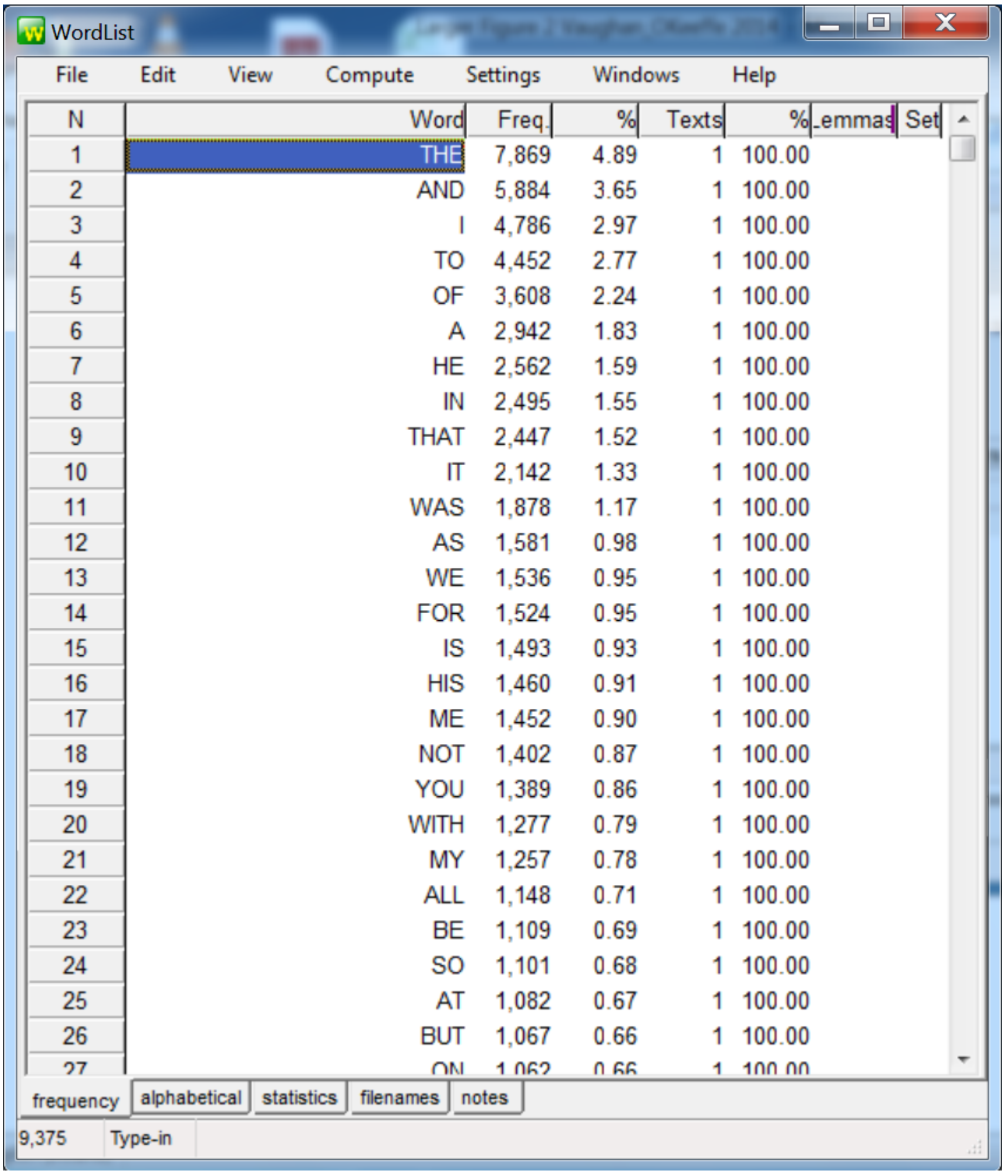
Tokens

You	put	your	right	leg	in,	your	right	leg	out
1	2	3	4	5	6	7	8	9	10

Types

You	put	your	right	leg	in,	your	right	leg	out
1	2	3	4	5	6	3	4	5	7

Figure 2 shows an example of a word list created by using WordSmith Tools version 5 (Scott, 2008) and a small corpus compiled by copying and pasting the text of the novel *Dracula* by Bram Stoker (1897), sourced from Project Gutenberg (http://www.gutenberg.org), into a Word document and then saving it as a "plain text" file—the only type of file that most concordancers will "read."



N	Word	Freq.	%	Texts	%	Lemmas	Set
1	THE	7,869	4.89	1	100.00		
2	AND	5,884	3.65	1	100.00		
3	I	4,786	2.97	1	100.00		
4	TO	4,452	2.77	1	100.00		
5	OF	3,608	2.24	1	100.00		
6	A	2,942	1.83	1	100.00		
7	HE	2,562	1.59	1	100.00		
8	IN	2,495	1.55	1	100.00		
9	THAT	2,447	1.52	1	100.00		
10	IT	2,142	1.33	1	100.00		
11	WAS	1,878	1.17	1	100.00		
12	AS	1,581	0.98	1	100.00		
13	WE	1,536	0.95	1	100.00		
14	FOR	1,524	0.95	1	100.00		
15	IS	1,493	0.93	1	100.00		
16	HIS	1,460	0.91	1	100.00		
17	ME	1,452	0.90	1	100.00		
18	NOT	1,402	0.87	1	100.00		
19	YOU	1,389	0.86	1	100.00		
20	WITH	1,277	0.79	1	100.00		
21	MY	1,257	0.78	1	100.00		
22	ALL	1,148	0.71	1	100.00		
23	BE	1,109	0.69	1	100.00		
24	SO	1,101	0.68	1	100.00		
25	AT	1,082	0.67	1	100.00		
26	BUT	1,067	0.66	1	100.00		
27	ON	1,062	0.66	1	100.00		

Figure 2 Sample word-list output: Screenshot of WordList using WordSmith Tools 5 and Dracula corpus.

As can be seen, most of the single-word items are "functional," or grammatical operators like determiners (*the*: 7,869 occurrences), conjunctions (*but*: (1,067 occurrences), and prepositions (*at*: 1,082). Pronouns such as *I*, *he*, and *me* are also high-frequency items (4,786, 2,562 and 1,452, respectively). This characteristic of word lists—to contain primarily "small" items with higher frequencies—is fairly consistent across corpora. Frequency is a central concept in corpus analysis (Baker, 2006), and a criticism of corpus linguistics in the past has been that it is primarily concerned with quantification. However, what is not frequent can also be interesting to the researcher—a corpus of teacher talk that did not contain the word *student*, for example, would perhaps be an anomaly and the reasons behind it would be a line of further investigation. There are at least two more observations we could make in this regard: (1) the frequency list for a specific corpus may be interesting in and of itself, but becomes more so in comparison to a frequency list for another corpus; and (2) the potential of these “small” items to do important grammatical, syntactic, and pragmatic work should not be underestimated (Vaughan & Clancy, 2013). Even with this small *Dracula* corpus (approximately 161,000 words in total), hypotheses could be formulated about the frequent items, such as *he* (could this be related to the vampire?); but in order to bring the interesting trends in the corpus into clearer view, we can compare it with another corpus. Corpus analysis is inherently comparative, particularly as the trend toward smaller, more specialized corpora gathers pace.

For comparative purposes, then, it is possible to compare high-frequency items—say, the top 20—with items from another, larger corpus. Table 2 presents the top 20 items in the *Dracula* corpus and the top 20 words in a 1-million-word sample from the BNC (British National Corpus, 1999).

Dracula corpus (161k)			BNC sampler (1m)		
N	Word	Freq.	N	Word	Freq.
1	THE	7,869	1	THE	68,856
2	AND	5,884	2	OF	33,798
3	I	4,786	3	AND	29,077
4	TO	4,452	4	TO	27,137
5	OF	3,608	5	A	22,092
6	A	2,942	6	IN	21,545
7	HE	2,562	7	FOR	10,275
8	IN	2,495	8	IS	10,230
9	THAT	2,447	9	THAT	8,555
10	IT	2,142	10	WAS	8,167
11	WAS	1,878	11	IT	7,951
12	AS	1,581	12	ON	7,280
13	WE	1,536	13	BE	7,171
14	FOR	1,524	14	WITH	7,152
15	IS	1,493	15	AS	6,676
16	HIS	1,460	16	BY	6,320
17	ME	1,452	17	I	6,218
18	NOT	1,402	18	AT	5,693
19	YOU	1,389	19	ARE	5,346
20	WITH	1,277	20	HE	4,889

Table 2 Top 20 words for Dracula corpus and 1-million-word sample of BNC.

Looking at the frequency list in this way shows us the rank order for the frequency information (the column N); but in order to compare the two corpora systematically the fact that they are of such different sizes needs to be addressed. One method for doing this is to *normalize* the frequency figures. This is a basic but illustrative method; in essence, we assume that, if we had 1 million words in the *Dracula* corpus, the rank order and frequency pattern would hold. A normalized frequency (*nf*) is based on the following calculation (McEnery & Hardie, 2012: 49–50):

$$nf = \text{(number of examples of the word in the whole corpus} \div \text{size of the corpus)} \times \text{(base of normalization)}$$

A simple calculation to establish if *he* is frequent in any marked way (our hypothesis) tells us that it is: normalized per 1 million words (the base of normalization), *he* gives a frequency of 15,884, or over 10,000 more occurrences than in the BNC sample corpus. So this is, potentially, an area for further investigation.

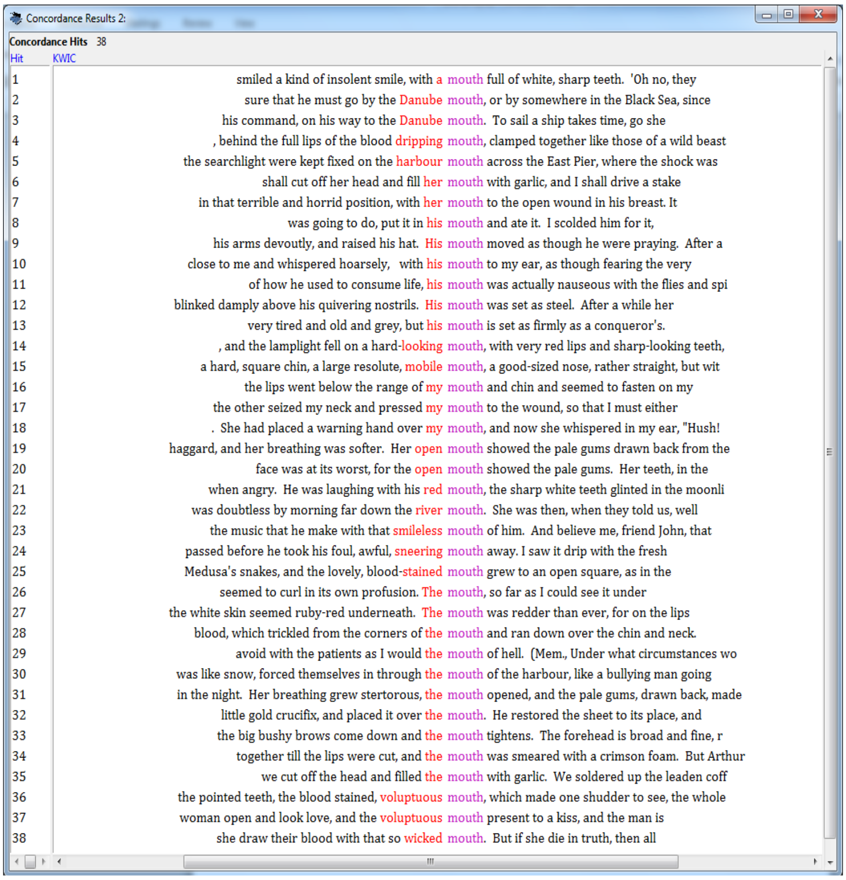
While the word-list function of the concordancer, in this case WordSmith Tools 5, can give a view of single-word item frequency, it is also possible to generate a list that presents clusters of items and their frequency. In the literature, clusters are also referred to as *chunks*, *n-grams*, or *lexical bundles*. The concept of *collocation*, discussed in the next section, is implicated here, but the crucial point is that the corpus perspective allows a view of language that acknowledges extended units of meaning beyond the single word—units that constitute a phrase or a particular pattern of meaning (Greaves & Warren, 2010). Comparing 2-, 3- and 4-word clusters across different registers of spoken and written discourse has provided useful insights for areas such as teaching English for academic purposes (EAP), and the description of characteristics of spoken and written grammar (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Carter & McCarthy, 2006).

Concordances

<pf>The potential of computationally based analysis to help linguists establish empirically which words tend to co-occur and how this co-occurrence affects their meaning has had a significant impact on lexical studies. *Collocation*, one of the pivotal concepts in lexical studies today, had been discussed in an early paper by Firth (1935), but it was the advent of the corpus method that allowed researchers to really

flesh out the concept. The study of collocation is particularly associated today with the work of John Sinclair (e.g., Sinclair, 1991) and explains why it feels more natural for speakers to say "*blatantly* obvious," for example, than to choose a different, perhaps equally possible adverb (*clearly*, *unashamedly*). This natural preference is reinforced each time the selection is made, and the meaning of "*blatantly* obvious" is entailed in both of its components: it has attained a unitary meaning.

The software tool that has allowed this perspective on corpus data is the concordance view. All of the tools of corpus analysis require human interaction with the information that the software tools can automatically generate, and arguably none more so than the concordance view. This output view presents a particular, preselected search word in its immediate linguistic context—usually five to eight words to its left and right, though it is possible to expand this view. Figure 3 shows a concordance view for the word *mouth* in the *Dracula* corpus (*mouth* is considered an apposite search term there given the source material for the corpus—a novel about a vampire): the search (or *node*) word *mouth* appears in the middle of the lines, and there are approximately eight or nine words on either side.



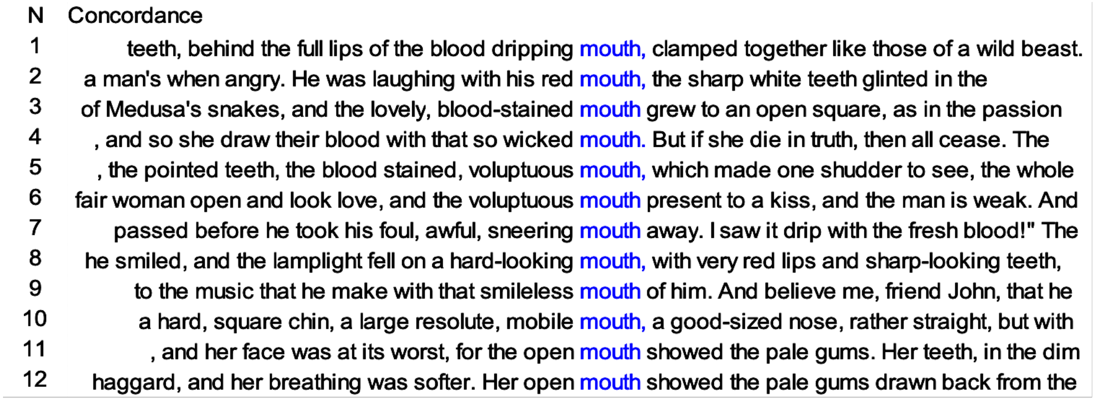
Concordance Results 2:

Concordance Hits 38

Hit	KWIC
1	smiled a kind of insolent smile, with a mouth full of white, sharp teeth. 'Oh no, they
2	sure that he must go by the Danube mouth, or by somewhere in the Black Sea, since
3	his command, on his way to the Danube mouth. To sail a ship takes time, go she
4	, behind the full lips of the blood dripping mouth, clamped together like those of a wild beast
5	the searchlight were kept fixed on the harbour mouth across the East Pier, where the shock was
6	shall cut off her head and fill her mouth with garlic, and I shall drive a stake
7	in that terrible and horrid position, with her mouth to the open wound in his breast. It
8	was going to do, put it in his mouth and ate it. I scolded him for it,
9	his arms devoutly, and raised his hat. His mouth moved as though he were praying. After a
10	close to me and whispered hoarsely, with his mouth to my ear, as though fearing the very
11	of how he used to consume life, his mouth was actually nauseous with the flies and spi
12	blinked damply above his quivering nostrils. His mouth was set as steel. After a while her
13	very tired and old and grey, but his mouth is set as firmly as a conqueror's.
14	, and the lamplight fell on a hard-looking mouth, with very red lips and sharp-looking teeth,
15	a hard, square chin, a large resolute, mobile mouth, a good-sized nose, rather straight, but wit
16	the lips went below the range of my mouth and chin and seemed to fasten on my
17	the other seized my neck and pressed my mouth to the wound, so that I must either
18	. She had placed a warning hand over my mouth, and now she whispered in my ear, "Hush!
19	haggard, and her breathing was softer. Her open mouth showed the pale gums drawn back from the
20	face was at its worst, for the open mouth showed the pale gums. Her teeth, in the
21	when angry. He was laughing with his red mouth, the sharp white teeth glinted in the moonli
22	was doubtless by morning far down the river mouth. She was then, when they told us, well
23	the music that he make with that smileless mouth of him. And believe me, friend John, that
24	passed before he took his foul, awful, sneering mouth away. I saw it drip with the fresh
25	Medusa's snakes, and the lovely, blood-stained mouth grew to an open square, as in the
26	seemed to curl in its own profusion. The mouth, so far as I could see it under
27	the white skin seemed ruby-red underneath. The mouth was redder than ever, for on the lips
28	blood, which trickled from the corners of the mouth and ran down over the chin and neck.
29	avoid with the patients as I would the mouth of hell. (Mem., Under what circumstances wo
30	was like snow, forced themselves in through the mouth of the harbour, like a bullying man going
31	in the night. Her breathing grew stertorous, the mouth opened, and the pale gums, drawn back, made
32	little gold crucifix, and placed it over the mouth. He restored the sheet to its place, and
33	the big bushy brows come down and the mouth tightens. The forehead is broad and fine, r
34	together till the lips were cut, and the mouth was smeared with a crimson foam. But Arthur
35	we cut off the head and filled the mouth with garlic. We soldered up the leaden coff
36	the pointed teeth, the blood stained, voluptuous mouth, which made one shudder to see, the whole
37	woman open and look love, and the voluptuous mouth present to a kiss, and the man is
38	she draw their blood with that so wicked mouth. But if she die in truth, then all

Figure 3 Concordance output view for *mouth* using *Dracula* corpus and AntConc 3.4.1w.

This view of language data will yield nuanced information about their patterns only with a significant amount of view manipulation. For illustration purposes, the number of concordance lines generated by searching for *mouth* in the *Dracula* corpus is eminently manageable: there are only 38 occurrences of this word altogether, of which six refer to the mouth of a river or the mouth of a harbor or represent a figurative use (the *mouth of hell*). These can be deleted, since the focus of the search is the human mouth. It is now also possible to re-sort the data so that the patterning in words that occur to the left and right of this meaning of *mouth* becomes more visible. If, for example, the possessive pronouns (*his mouth*, *my mouth*) and the definite and indefinite articles (*a mouth*, *the mouth*) are also deleted, so that what remains is the attributive adjectives that occur before *mouth*, the view changes again, as Figure 4 shows.



```
 N  Concordance
 1    teeth, behind the full lips of the blood dripping mouth, clamped together like those of a wild beast.
 2     a man's when angry. He was laughing with his red mouth, the sharp white teeth glinted in the
 3    of Medusa's snakes, and the lovely, blood-stained mouth grew to an open square, as in the passion
 4    , and so she draw their blood with that so wicked mouth. But if she die in truth, then all cease. The
 5   , the pointed teeth, the blood stained, voluptuous mouth, which made one shudder to see, the whole
 6    fair woman open and look love, and the voluptuous mouth present to a kiss, and the man is weak. And
 7      passed before he took his foul, awful, sneering mouth away. I saw it drip with the fresh blood!" The
 8  he smiled, and the lamplight fell on a hard-looking mouth, with very red lips and sharp-looking teeth,
 9        to the music that he make with that smileless mouth of him. And believe me, friend John, that he
10        a hard, square chin, a large resolute, mobile mouth, a good-sized nose, rather straight, but with
11        , and her face was at its worst, for the open mouth showed the pale gums. Her teeth, in the dim
12      haggard, and her breathing was softer. Her open mouth showed the pale gums drawn back from the
```

Figure 4 Concordance lines of attributive adjectives occurring before mouth in the Dracula corpus (concordance lines edited using WordSmith Tools 5).

It is now possible to see that the adjectives that occur before *mouth* could be interpreted as being largely sinister and negative, *the blood dripping mouth*, *sneering mouth*, *wicked mouth*, while even an adjective such as *voluptuous*, which could be positive, reveals on closer examination a negative meaning.

This practical illustration raises two important and connected points. First, the analysis of concordance lines that present the search word cannot proceed without constant recourse to its immediate co-text—the five, six, or more words occurring in the vicinity of the search or node word. It is also possible, in the concord view of most software, to click on a line and to see the search word as part of the text of the corpus itself, a view that may be crucial to interpreting its meaning. Second, this view of corpus data was critical to the empirical development of the concept of *semantic* *prosody*: the way in which words collocate can have a direct impact on their connotational meaning. Some words are quite obviously positive or negative, though rarely neutral—*slim*, *thin*, and *skinny* being good examples of broadly synonymous items that illustrate this point. Corpus analysis has brought into view many words with positive and negative semantic prosodies that would not be immediately obvious out of context. Hunston (2002) gives the example of *sit through*: there is no core meaning of either component of this phrasal verb that suggests something negative, and yet *sit through* tends to collocate with experiences that are protracted and boring. In other words, it has a negative semantic prosody.

Keywords

The concordance view therefore requires the analyst to have identified, possibly via frequency lists, which item (s)he wishes to investigate; and the frequency list is often the first entry point into a data set (Baker, 2006). While frequency lists give a general overview of frequency in a corpus, it is also possible to compare two different word lists and generate a list of keywords. Keywords are not the most frequent words in any corpus or text, but rather the most unusually frequent (or infrequent) relatively to some comparative baseline. In other words, while word lists present *frequency* information, the calculation of which words are key measures the *saliency* of words in a text (Baker, 2006). Corpus software will relate the word-frequency list of one corpus to that of a larger "reference corpus" by comparing frequencies of words in the target corpus, the number of running words in this and the reference corpus, and the cross-tabulation of these figures, using a statistical test (either Chi-square or log-likelihood) to ascertain which items occur with unusually high or low frequency.

A view of keywords in a document gives the analyst a sense of what items characterize the data set—or the "aboutness" of a given text or set of texts—as well as a sense of the style of the text. The "aboutness" information tends to come from the lexical items that appear in the keyword list, and the style information from what has been referred to above as the "small," but high-frequency grammatical words. The reference corpus used for comparative purposes can have a significant impact on what items emerge as key. Table 3 shows two different keyword lists, created by comparing a small, domain-specific corpus of teacher talk in meetings (c. 40,000 words; Vaughan, 2008) with two much larger corpora. The first is the 1-million-word spoken Limerick corpus of Irish English (Farr, Murphy, & O'Keeffe, 2004); the second is a 1-million-word written sample from the BNC (BNC, 1999). The differences between the spoken and the written keyword lists are certainly observable, and fairly immediately so; however, in this case, over 50% of the words calculated as key occur in both lists. What emerges most strongly is the distinction between spoken and written styles coming to the fore, as the jargon of the teachers' workplace is salient in comparison with that of the spoken corpus—*students* (3), *semester/s* (4) and (15), *certificate* (11), *elementary* (12), and so on—while the "spoken-ness" of the original data returns as a key element by comparison with the written corpus (*yeah* (1), *okay* (4), *em* (19).

Reference: LCIE (1m)		Reference: BNC written (1m)	
1	KET	1	YEAH
2	PET	2	I
3	STUDENTS	3	THINK
4	SEMESTER	4	OKAY
5	CLASS	5	YOU
6	EXAM	6	KNOW
7	WE	7	WE
8	ENGLISH	8	SO
9	CLASSES	9	DO
10	THINK	10	KET
11	CERTIFICATE	11	IT'S
12	ELEMENTARY	12	THEY
13	BOOK	13	PET
14	PASS	14	THAT'S
15	SEMESTERS	15	THEY'RE
16	THEY	16	MEAN
17	INTERMEDIATE	17	JUST
18	OKAY	18	SEMESTER
19	PRE	19	EM
20	TOEFL	20	LAUGHTER
21	SO	21	KIND
22	KIND	22	EXAM
23	LAUGHTER	23	STUDENTS
24	MEAN	24	CLASS
25	MAYBE	25	DON'T

Table 3 Keywords in CMELT using spoken and written reference corpus.

Exploring this notion further, O’Keeffe (2012) uses the internationally recognizable BBC 1 *Panorama* television interview of Diana, Princess of Wales, taken by Martin Bashir (broadcast November 1995) to illustrate how using a different references corpus can alter the range of words that are identified as “key”; and this in itself can be insightful. As Barlow (2004) argues, a heightened awareness of the role of the tools in analyses is essential to understanding how these tools impact on the interpretation of the data. This is critical to evaluating what corpus analysis can bring to the investigation of texts, and it is explored in depth below.

The transcript of the *Panorama* interview is readily available on the Internet, as is the actual television interview (visit http://www.bbc.co.uk/politics97/diana/panorama.html). First, the interview’s keywords are identified by using a corpus of media interviews as a reference corpus (O’Keeffe, 2006). This reference corpus comprises 271,553 words: 93,180 words from 29 political interviews; 89,225 words from 46 interviews on TV chat shows and radio involving known or public personae; and 89,148 words from 17 interviews from radio phone-ins involving unknown or private personae. Data are drawn from international English-speaking media sources, for instance from the UK, USA, Canada, Australia, and Ireland. The keywords generated from this analysis are relatively few: 24 in all (the interviewer’s and the interviewee’s names have been removed, as these were only part of the transcript rather than part of the discourse). Table 4 illustrates all of the keywords generated, in order of keyness, vertically.

did	husband	difficult	queen	your
was	had	William	were	children
Wales	uh	royal	yourself	media
prince	monarchy	my	because	depression
marriage	bulimia	role	relationship	husband’s

Table 4 Keywords of Bashir–Diana Panorama interview with Media corpus.
(O’Keeffe 2006; 2012)

Notice that the vocalization *uh* is cited as a keyword. This is most likely a function of the variation in the transcription of vocalizations in the reference corpus, which comprises many media transcripts. Some transcribe the same or similar vocalization as *uhm*, *erm*, *ah*, among other versions. This is the first point regarding analysis of corpora: if similar words or vocalizations are transcribed differently, this will have a bearing on keyword calculations.

When the same interview was compared with a second reference corpus, more and different results were generated. Whereas Table 4 illustrates the keywords generated when the *Panorama* interview was compared with a corpus of media interviews, Table 5 shows keywords from a reference corpus that is distinctly unrelated to media interviews, namely an academic corpus of English. The Limerick–Belfast Corpus of Academic Spoken English (LIBEL) is a 500k-word collection of academic spoken English. This is made up of lectures, tutorials, seminars, and presentations. Ninety-two keywords were generated (again, excluding the interviewer's and interviewee's names). Table 5 shows a sample of those 92 items (again, in order of keyness, vertically).

was	I've	I'd	because	people	think
I	it's	marriage	monarchy	difficult	never
don't	me	people's	myself	public	wasn't
husband	uh	bulimia	role	there's	Mr.
my	yes	you're	husband's	yourself	princess
I'm	didn't	queen	couldn't	relationship	royal
did	had	William	divorce	feel	pressures
Wales	prince	were	that's	loved	albeit

Table 5 Sample of top keywords of Bashir–Diana Panorama interview with LIBEL.

The results in Table 4 and Table 5 above show stark differences in both numbers of the keywords generated (25 when compared with a media reference corpus and 92 when compared with a very different reference corpus of academic discourse). By comparison with a corpus that is distant in terms of genre, we find keywords that point to these genre differences, for instance common first- and second-person pronouns (*I*, *I'm*, *my*, *myself*, *yourself*, *me*), which are not high-frequency words in academic lectures. We see high-frequency verbs and verb forms (*was*, *did*, *didn't*, *wasn't loved*, *think*), pronoun–verb combinations (*I've*, *you're*), everyday seeming nouns (*divorce*, *husband*, *marriage*, *people's*, *husband's*), and so on. These all reference the more private sphere domains of reference "you—I," relationships, marriage, problems like bulimia, marriage breakdowns, all of which would not normally be talked about in the more referential world of academia. The following extract from the *Panorama* interview illustrates this:

(1) From an interview between Martin Bashir and Diana, Princess of Wales, broadcast November 1995

```
<trd>BASHIR:        What effect did the depression have on your
                marriage?
<trd>Diana:         Well, it gave everybody a wonderful new label -
                Diana’s unstable and Diana’s mentally unbalanced.
                And unfortunately that seems to have stuck on and
                off over the years.
<trd>BASHIR:        Are you saying that that label stuck within your
                marriage?
<trd>Diana:         I think people used it and it stuck, yes.
```

If the *Panorama* Bashir–Diana interview is compared with two further data sets—both of which have in common with the interview the fact that they involve more reference within the “I–you” domain and that they refer more to everyday worlds of relationships and so on—further interesting analytical and methodological insights come to light. These two data sets are a corpus of the sitcom *Friends* and the Limerick corpus of Irish English (LCIE) (Farr et al., 2004).

was	media	because	relationship	yourself	not
very	and	that	The	interest	look
people	were	William	In	being	is
husband	prince	obviously	did	felt	know
had	role	royal	people’s	think	you
of	difficult	country	bulimia	depression	just
public	children	monarchy	queen	get	no
Wales	marriage	as	effect	can	oh

Table 6 Keywords Bashir Panorama interview with Friends sitcom corpus (48 in total).

husband	bulimia	had	divorce	effect	attention
marriage	role	that	depression	book	engagements
Wales	difficult	uh	yourself	children	daunted
prince	royal	my	albeit	obviously	were
relationship	William	people’s	feel	did	future
media	because	very	your	being	enormous
monarchy	public	yes	princess	pressures	knowledge
people	queen	husband’s	was	separation	duties

Table 7 Sample of keywords Bashir Panorama interview with LCIE as reference corpus (total 87 keywords).

From Tables 6 and 7 it becomes obvious that using the *Friends* corpus (Table 6) and the LCIE (Table 7) as reference corpora returns a broad spread of keywords in the same way as when the academic spoken corpus data from LIBEL is used (Table 5). All three sets of results—Tables 5, 6, and 7—have in common that they used reference corpora that were different in genre, whereas the results in Table 4 were generated by comparison with a reference corpus of a similar genre (media interviews).

Importantly, this tells us that, if a keyword analysis is carried out using a reference corpus that is very similar to the test corpus, it is likely that more concentrated (and fewer) keyword forms will be generated. Conversely, if a reference corpus that is very different from the test corpus is used, a very diverse range of keywords will be generated, including some that may be unexpected (e.g., all of the first- and second-person pronouns and high-frequency verbs generated when the academic corpus was used as a reference). A general corpus used for comparative purposes, representing how English is generally used (in this case, LCIE), will return a large number of keywords, but they will be less disparate. The results in Table 6 (*Friends*) and Table 7 (LCIE) have a lot in common with Tables 4 and 5, but they have a broader spread. Many of the noun forms are common to all four tables, for example *husband*, *bulimia*, and *monarchy*. Interestingly, *divorce* is in Tables 5, 6, and 7, but not in Table 4. As the word *divorce* is showing as a keyword when reference corpora outside of media discourse are used and not showing as a keyword when the *Panorama* interview is compared with a more similar corpus, it can be deduced that *divorce* is not an uncommon reference in media interviews.

Corpus analysis in language study

These practical illustrations of patterns of language use—on which it is possible to get specific perspectives by using corpus analysis methods—have deliberately used quite different types of corpora, which represent very different genres of language: a 19th-century gothic novel (written), a small corpus of workplace talk in meetings (spoken), and the transcript of an iconic interview (spoken). These sites of language use have in the past tended to be allied to particular areas of study, with very specific analytical frameworks. The final points that can be made in relation to the broad topic of corpus analysis connect to this emerging characteristic of the corpus method: that it facilitates complementary analyzes in fields where language data are key.

There are many traditions within the broad field of discourse analysis, for example, where this is the case. Critical discourse analysis is a research tradition that addresses theoretical concepts such as “ideology” and “power” with the help of many different qualitative techniques. It has in fact been criticized for being “too” qualitative, or for basing some of its conclusions on what may have been considered to be “insufficient” evidence. While this may or may not be a fair assessment in general, work by Baker et al. (2008) discusses the possibilities afforded by a “methodological cross-pollination” (p. 274) between corpus linguistics and critical discourse analysis.

This relates to a very real challenge in the field of corpus linguistics, especially as the use of corpora and corpus linguistic methods spreads to other fields, as a tool for exploring discourse in context: the challenge is whether corpus linguistics can offer a complete framework for analysis. Many argue that corpus linguistics is solely a powerful methodological tool that aids in the analysis of large text-based data sets. It can generate reliable, automatic, virtually instantaneous information about word frequencies in the data set, its keywords, its syntactic and semantic patterns, as well as aiding qualitative analysis by interactive access to the source file. Thematic and other types of tagging can be developed depending on the research question, and the tools allow for quick retrieval of, and hence access to, patterns in the phenomena selected for tagging. However, many would argue that empiricism without nuanced interpretation does not do justice to spoken or written language data, hence the trend toward blending corpus linguistic analysis methods and well-established theoretical frameworks.

An area of enquiry that has a long-established focus on spoken language is conversation analysis, which, despite its name, is concerned with language use in all contexts. One of the primary units of analysis within conversation analysis is the “turn,” in the sense that one speaker speaks first and then the next and so on. Research in conversation analysis has been driven primarily by questions such as who gets to talk and when, and what constraints exist in terms of how they do that. Conversation analysts use as data highly detailed transcripts of face-to-face talk, and much of the canonical work in the field focused on the sequential organization of telephone conversations.

The concept of the turn has been shown to be a useful complement to the corpus linguistic method in general, as a bridge between the bottom-up evidence provided by corpus searches, and it has been explored in detail in studies such as O’Keeffe and Walsh (2012). They illustrate how, as the analytical focus moves up from word to multiword unit to grammatical patterning to the level of “turn,” corpus linguistic methods can provide a view on turns in discourse—but only up to a point. They can certainly help quantify what happens *within* a turn, for example, providing evidence on the most common patterns of language used in a call opening or closing, but they do not provide a framework for analysis at turn level. For example, if what happens in a series of turns is the research focus, concordancing software can find those turns; however, a framework such as conversation analysis provides a lens for analysis in terms of turns and exchanges, turn sequentiality, topic management, power relationships, or interactional patterning.

Analyzing language by using corpus methodologies is no longer a niche undertaking, and the proliferation of language text corpora continues apace. Where once the corpus analysis tools themselves, or concordancers, were mainly only available commercially and used by the few, now there are a number of concordancers freely available to the many, for example, AntConc (Anthony, 2014). These are user-friendly to varying degrees, but, combined with the potentially unlimited language data available for example via the World Wide Web or with the increasing ease (with some caveats) of recording and transcribing spoken language events through voice recognition software, corpora and corpus methodologies have never been so widely referenced and used. However, there are still large questions surrounding the extent to which corpora are consulted in an informed way, how corpora are built, what language situations they can claim to be representative of, what perspectives on language the analysis of corpora can provide, and what positions they challenge. The power of corpus linguistics is its potential to analyze large quantities of data; but the merging of this potential with other existing analytical frameworks is conceptually underdeveloped at the time of writing.

SEE ALSO: Conversation Analysis, Overview; Critical Discourse Analysis; Genre Analysis

References

Aarts, J., & Meijs, W. (Eds.). (1984). *Corpus linguistics: Recent developments in the use of computer corpora in English Language research*. Amsterdam, Netherlands: Rodopi.

Anthony, L. (2014). *AntConc 3.4.1.* Retrieved from http://www.antlab.sci.waseda.ac.jp/index.html

Baker, P. (2006). *Using corpora in discourse analysis*. London, UK: Continuum.

Baker, P., Gabrielatos, C., Khosravinik, M., Kryżanowski, M., McEnery, T., & Wodak, R. (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.

Barlow, M. (2004). Software for corpus access and analysis. In J. McH. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 205–221). Amsterdam, Netherlands: John Benjamins.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. London, UK: Longman.

British National Corpus (BNC). (1999). *The BNC sampler corpus*. Oxford, UK: Oxford University Computing Services.

Carter, R., & McCarthy, M. (2006). *The Cambridge grammar of English*. Cambridge, UK: Cambridge University Press.

Farr, F., Murphy, B., & O’Keeffe, A. (2004). The Limerick corpus of Irish English: Design, description and application. *Teanga*, 21: 25–29.

Firth, J. R. (1935). The techniques of semantics. *Transactions of the Philological Society*, 7: 36–72.

Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 212–226). London, UK: Routledge.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, UK Cambridge University Press.

Lee, D. (2010). What corpora are available? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 107–121). London, UK: Routledge.

McCarthy, M., & O’Keeffe, A. (2010). Historical perspective: What are corpora and how have they evolved? In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 3–13). London, UK: Routledge.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge, UK Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London, UK: Routledge.

O’Keeffe, A. (2006). *Investigating media discourse*. London, UK: Routledge.

O'Keeffe, A. (2012). Corpora and media studies. In K. Hyland, M. H. Chau, & M. Handford (Eds.), *Corpus applications in applied linguistics* (pp. 441-454). London, UK: Continuum.

O'Keeffe, A., & Walsh, S. (2012). Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. *Corpus Linguistics and Linguistic Theory*, *8*(1): 159–181.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge, UK: Cambridge University Press.

Scott, M. (2008). *WordSmith Tools 5*. Liverpool, UK: Lexical Analysis Software.

Sinclair, J. McH. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam, Netherlands: John Benjamins.

Vaughan, E. (2008). "Got a date or something?" An analysis of the role of humour and laughter in the workplace meetings of English language teachers. In A. Ädel & R. Reppen (Eds.), *Corpora and discourse: The challenges of different settings* (pp. 95–115). Amsterdam, Netherlands: John Benjamins.

Vaughan, E., & Clancy, B. (2013). Small corpora and pragmatics. In J. Roméro-Trillo (Ed.), *The yearbook of corpus linguistics and pragmatics 2013: New domains and epistemologies* (vol. 1, pp. 53–73). Dordrecht, Netherlands: Springer.

Further Reading

Authors

Elaine Vaughan is lecturer in TESOL & applied linguistics at the School of Modern Languages and Applied Linguistics, University of Limerick. Her research interests include corpus linguistics and corpus-based discourse analysis, Irish English, including media representations of Irish English, and the discourse of teaching and learning. Her published work concerns community and identity in language, humor, and professional discourse. Her most recent publication in corpus linguistics is "Small Corpora and Pragmatics" in the *Yearbook of Corpus Linguistics and Pragmatics* (2013; with B. Clancy).

Anne O'Keeffe is senior lecturer in the Department of English Language and Literature, Mary Immaculate College, Ireland. She has published extensively in corpus linguistics and applied linguistics, and her most recent publications include *English Grammar Today* (2013; with R. Carter, M. McCarthy & G. Mark), *Introducing Pragmatics in Use* (2011; with B. Clancy & S. Adolphs), and *The Routledge Handbook of Corpus Linguistics* (2010; with M. McCarthy). She is principal investigator on the CUP/Cambridge English Language Assessment/University of Cambridge project *English Profile*. Her research interests include corpus linguistics and applied linguistics, media discourse, pragmatics, and Irish English.