Chapter 40 The corpus method

Vaclav Brezina and Dana Gablasova

40.1 What is corpus linguistics?

Corpus linguistics is a versatile methodology of computational analysis of language, which can be used for the study of grammar (e.g. Biber et al. 1999), discourse (e.g. Baker 2006), different social actors and processes (Baker et al. 2008) as well as in language learning and teaching (e.g. Aijmer 2009). Corpus linguistics is a scientific approach to language, which follows the tradition of empirical investigation, that is, investigation in which the focus is on the collection and interpretation of data (McEnery & Hardie 2011). Somewhat symbolically, the importance of data for corpus linguistics is evident from the fact that corpus linguistics has a term in its name that refers to data, namely **corpus**, in its name. A corpus (the plural form of this word is *corpora*) is a particular type of linguistic data, comprising thousands of pages of texts and/or transcripts of spoken language that we can search by a computer. A corpus is a sample of language, which allows us to observe language use in different situations. McEnery et al. (2006: 5) provide the following comprehensive definition of a corpus:

> A corpus is a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety.

Let us consider the individual aspects of this definition. First, a corpus is not just any collection of texts; the books (hard copies) that you have on a shelf at home are not a corpus, because they cannot be 'read' (processed) by a computer. Instead, a corpus is a collection of texts in an electronic format (e.g. ebooks), which allow us not only to read them page-by-page, but also and more importantly, to search them to find out, for example, the frequencies of different words and phrases in these texts. Second, what we include in a corpus are authentic texts and authentic transcripts of spoken language. This means that we don't produce texts specifically for language corpora but rather collect texts and samples of spoken language that are around us 'in the wild' such as books, newspaper articles, academic papers, blog posts, tweets, TV and radio broadcasts, lectures, informal conversations etc. Third, texts and transcripts of speech are carefully selected (sampled) to capture different aspects of language use, which we want to investigate. A corpus is not any pile of texts we happen to come across, but rather a carefully designed database, where we have as much information about each text as possible such as its source, genre, date etc. Fourth, we use corpora as samples which tell us something interesting about language as such or a particular variety of language, e.g. academic writing, newspaper language, informal speech etc. We say that corpora are representative of language (or a variety of language) by which we mean that corpora accurately reflect on a small scale how language is used every day on a large scale in a variety of situations.

In very practical terms, our typical encounter with a corpus is via a specific corpus analysis software package (see Section 40.2), which allows us to search the corpus for

linguistic purposes. For example, Figure 40.1 shows the search interface of a corpus tool called #LancsBox (Brezina et al. 2015), in which we searched for the word *love* in BE06, a one-million-word corpus of written British English (Baker 2009). We can see that the form *love*, which can be used either as a noun (*Love is all around us*) or a verb (*How can anyone love her as much as I do?*), occurs 288 times in one million words. The tool also gives us all examples of the word *love* in the corpus, which we can sort according to different criteria and look for patterns of use. In a similar way, we can compare the frequency of the word *love* with the frequencies of other words. We can, for instance, find out that *love* is less frequent than the definite article *the,* which occurs almost sixty thousand times in one million words but more frequent than the word *hate,* which occurs only 30 times in BE06.

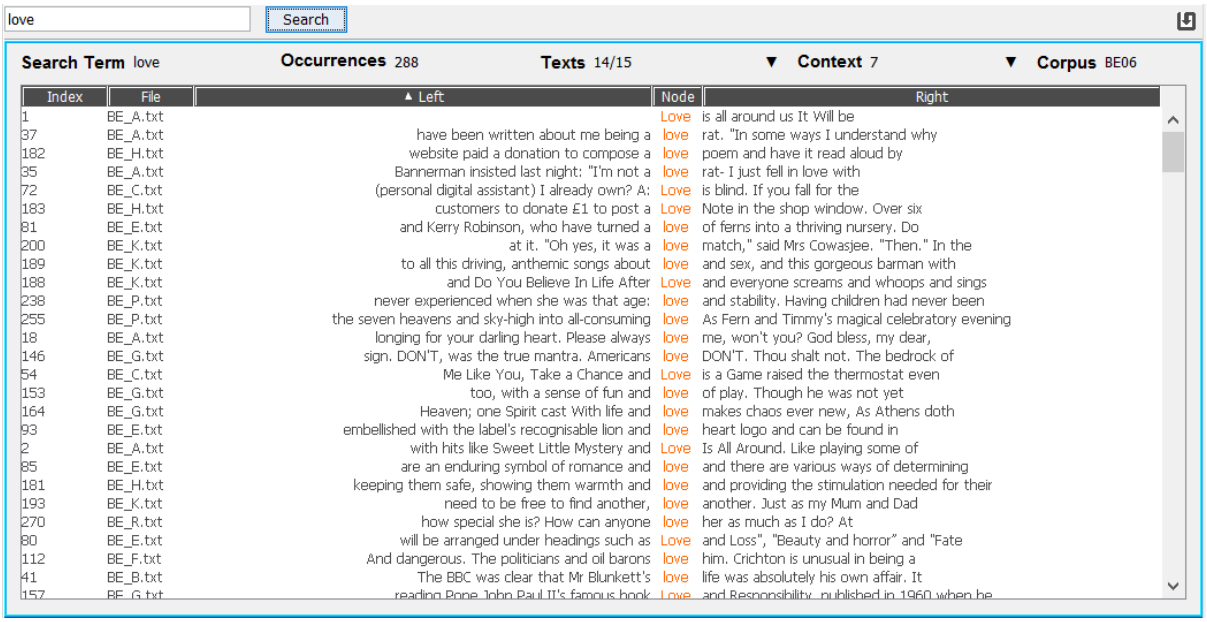

Figure 40.1. Concordance lines of 'love'

40.2 Corpus tools and techniques

There are different tools which we can use for corpus analysis. They can be divided into two broad groups: (i) desktop (offline) and (ii) web-based (online) tools. **Desktop tools** need to be downloaded to (and sometimes also installed on) users' own computers. These tools are suitable for processing small and mid-size corpora (up to several million words). Examples of desktop tools include #LancsBox (Brezina et al. 2015), MonoConc Pro (Barlow 2002), Word Smith Tools (Scott, 2016) and AntConc (Anthony 2014). **Web-based tools**, on the other hand, work inside a web browser and do not require any download or installation. They allow the users to access and search large corpora (hundreds of millions or even billions of words) that are provided as part of the tools. Some of these tools (see 'Advances box 40.1: Overview of tools') also allow the users to upload their own corpora and analyse them online. Examples of web-based tools include CQPweb (Hardie 2012), SketchEngine (Kilgarriff 2014), Wmatrix (Rayson 2008) and BYU corpora (Davies 2002-).

ADVANCES BOX 40.1

Overview of tools

Tool	Analysis of own data	Provides corpora	Brief description
Desktop (offline) tools			
#LancsBox	YES	YES	This flexible tool, which runs on all major operating systems, represents a new generation of corpus analysis software. It provides a simple interface, yet powerful analytical and visualization capabilities for corpus data. It allows easy comparing and contrasting of multiple corpora. The tool is freely available from http://corpora.lancs.ac.uk/lancsbox
MonoConc Pro	YES	NO	This Windows tool has a powerful search functionality and allows simple and advanced searches with easy navigation. It calculates and displays the distribution of linguistic features in individual text files. Paid license as well as a free simple version (MonoconcEsy) are available at http://www.monoconc.com
WordSmith	YES	NO	This Windows tool has a large number of analytical and data manipulation functionalities, with new features regularly added. The tool is recommended for more advanced users. Paid license as well as a free older version (v. 4) are available from http://www.lexically.net/wordsmith
AntConc	YES	NO	This tool is available in versions for different operating systems. AntConc is a toolbox which searches corpora and provides all core corpus analytical functionalities with easy connection between individual tools. The tool is freely available from http://www.laurenceanthony.net
Web-based (online) tools			
CQPweb	NO	YES	This tool offers a range of pre-loaded corpora for English (current and historical) and other languages including Arabic, Italian, Hindi and Chinese. It has a number of powerful analytical functionalities. The tool is freely available from https://cqpweb.lancs.ac.uk/
Wmatrix	YES	NO	This tool allows processing users' own data and adding part-of-speech and semantic annotation. Corpora can also be searched and compared with reference wordlists. Paid access as well as a free trial are available from http://ucrel.lancs.ac.uk/wmatrix/.

SketchEngine	YES	YES	This tool can be used for processing users' own data, collecting data from the web and exploring a very large number of pre-loaded corpora for all major languages. SketchEngine includes the TenTenTen family of web-based corpora, each of which consists of billions of words. Paid access as well as a free trial are available from https://www.sketchengine.co.uk/.
BYU corpora	NO	YES	This website offers large corpora of American and other varieties of English. It also contains NOW, a large monitor corpus of web-based newspapers and magazines updated daily. Non-English corpora include a corpus of Spanish and a corpus of Portuguese. The tool is freely available from http://corpus.byu.edu/. Its usability is somewhat limited by frequent requests for donations (after every 10-12 searches).

In the rest of this section, we review four core corpus linguistic techniques. The techniques discussed will be (1) Frequency lists, (2) Concordances, (3) Collocations and (4) Keywords.

40.2.1 Frequency lists

Have you ever thought about which words we use most often? This question can be answered with the use of language corpora. The technique that helps us count words is called the **frequency list** or **wordlist** technique. Because computers, unlike humans, are very good at counting words, we can find the answer to the question asked above in a matter of seconds. For example, the top ten most frequent words in the British National Corpus (BNC), a dataset that contains one hundred million words representing spoken and written British English, are listed in the left panel of Table 40.1. Next to these (still in the left panel) are the top ten words from the Poetry subcorpus (i.e. component of a corpus) of the BNC. Top ten words in the BNC and the Poetry subcorpus are very similar and consist of grammatical words such as articles (*the, a*), prepositions (*of, to, in, on, with*), pronouns (*it, I*) etc. This is because in English, grammatical words are very frequent and are used in any text or genre as requirement of the English grammar.

Table 40.1 Most frequent words in general British English and British poetry.

Top ten words		
	BNC	**Poetry**
1.	the	the
2.	of	a
3.	and	and
4.	to	of
5.	a	to
6.	in	in
7.	that	I
8.	it	it
9.	is	on
10.	was	with

Top ten content words		
	BNC	**Poetry**
1.	said	time
2.	time	see
3.	like	eyes
4.	now	old
5.	new	day
6.	people	love
7.	know	man
8.	see	light
9.	get	come
10.	way	life

In contrast to grammatical words, content words (words that carry non-grammatical meaning) displayed in the right panel of Table 40.1 differ between general English (BNC) and poetry. For example, while *time* is the most frequent noun in both lists, the poetry wordlist generally favours more specific nouns such as *eyes*, *day, love, light* and *life.* Interestingly, the BNC list includes *new* as the most frequent adjective, while the poetry wordlist favours *old*.

ILLUSTRATION BOX 40.1
Which are the most frequent words in English?

The Frequency list technique was used in research, which asked the question in the title of this illustration box (Brezina & Gablasova, 2015). The purpose of the research was to identify core English vocabulary, words that are used in a large number of different contexts. In order to find these words, the research looked at spoken, written and online communication across a variety of genres, which are included in large general corpora. While previous research in this area considered only one source of data, Brezina & Gablasova (2015) used and compared four corpora of English: the British National Corpus, British English 2006, the Lancaster-Oslo-Bergen corpus and EnTenTen12. Altogether these corpora contained over twelve billion words.

Comparing multiple wordlists, the research found that a group of merely 2,500 high-frequency English words ('lexical core') represents over 80% of all English text regardless of the topic or genre. A combined wordlist, the New General Service List (new-GSL), was produced to capture the results of the research. The new-GSL can be used for teaching of English and creating syllabi, teaching materials and dictionaries. The full list is available at http://corpora.lancs.ac.uk/vocab/

40.2.2 Concordances

A **concordance** is a list of all instances of a word, phrase, grammatical structure etc. in the corpus usually displayed in a special format called 'KWIC' ('key word in context'). KWIC display places the search term, called the 'node', in the middle and shows a few words to the left and a few words to the right of the node (see Table 40.2). The point of placing the node in the middle is to allow efficient skim-reading through many examples to identify typical patterns of use of the node. The examples in the concordance can be sorted, randomised or filtered by different criteria. For instance, Table 40.2 shows instances of the verb *hate* used with the first person pronoun *I* (a filter was applied to display only lines with *I*), which are sorted alphabetically according to the words that immediately follow the node (*hate*). We can see that the words following the node in Table 40.2 start with B (*bloody*) and move down the alphabet (*changing, everything, it*…). Looking at the use of the verb *hate*, we can see that it is followed by either a noun (*Christmas, Ravel, house, dog*), pronoun (*everything*, *it, that*), an *-ing (changing)* or a *to* (*to drink, to think*) construction. By using the verb *hate,* speakers express a strong dispreference towards people (*Ravel*), things (*house*), events (*Christmas*) and actions (*changing nappies*). The concordance of the word *hate* gives us an insight into the typical contexts in which the word appears and is essential for understanding the meaning of the word. Corpus linguistics often defines the meaning of a word in terms of its use in context. All modern English dictionaries therefore use corpora (and the concordance technique) to describe the meanings of words.

Table 40.2 Concordance lines for *hate* in BE06 filtered for the occurrence of *I* and sorted 1R

to worry about now. 'Christmas. Christmas. I	**hate**	bloody Christmas,' she said, rolling away from
cut out for this motherhood stuff. I	**hate**	changing nappies, and...' The dragon paused, peered
an anti-connoisseur of these events – I	**hate**	everything about them – but not even
marks either side of it. Because I	**hate**	it and it's crap and I JUST
said loftily. "And, for your information, I	**hate**	Ravel. Try Rachmaninov's third piano concerto. About
Alison Findlay examines Cavendish's drama in '"I	**hate**	such an old-fashioned House": Margaret Cavendish and
to get something for nothing, and I	**hate**	that. You know, as though you are
half an hour to kill and I	**hate**	to drink alone.' Maybe it was the
You need a properly balanced diet. I	**hate**	to think what sort of state your
baffling graffiti: "Love is for Suckers', "I	**Hate**	Your Dog' and "Sit on it' are

40.2.3 Collocations

Collocation is systematic co-occurrence of words in text and discourse that we identify statistically. In practice, the corpus tool will produce a list of collocates, i.e. words that systematically co-occur with our word of interest (node). For example, the node *love* often co-occurs with collocates such as *affair, fall, fell, I'd, I'm* etc. Collocations are useful because they show us important meaning connections between words and help us identify multi-word units as basic 'building blocks' of language (see Chapter 9 on the "meaning connections"

between words). To illustrate this phenomenon, when we form a sentence such as *Peter fell in love with Jane* we do not merely put individual words together; instead, we express the meaning in a set format by using the collocation *fall in love with* as one unit. Collocates can be displayed in a table (Table 40.3) or a graph (Figure 40.2).

Table 40.3 Collocates of *love*: Tabular display

Collocate	MI score	Frequency
affair	8.9	5
fell	8.5	14
falling	8.5	5
fallen	8.4	5
i'd	6.4	6
i'm	5.7	8
me	5.5	22
life	5.1	8

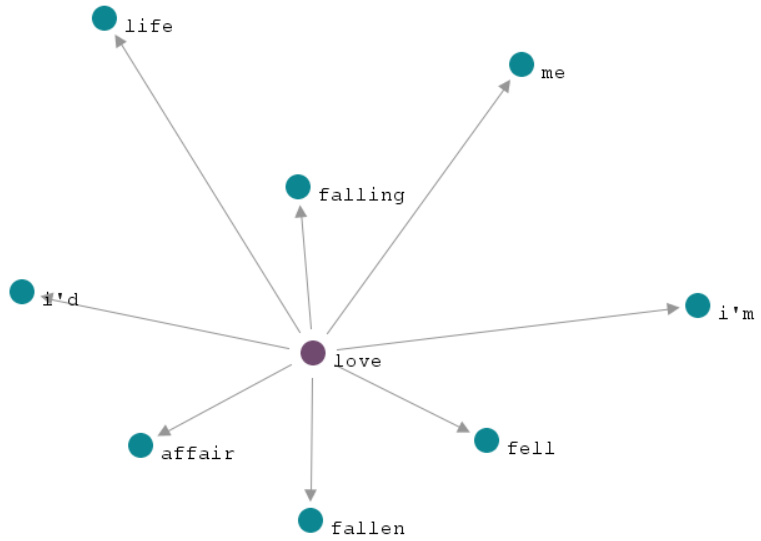

Figure 40.2 Collocates of *love*: Graphical display in #LancsBox

Table 40.3 shows the collocates of *love* ordered according to their strength (attraction between words evidenced by their systematic co-occurrence in text) measured by a statistical association measure (see the Advances Box 40.2) called the Mutual Information (MI) score. Figure 40.2 visually displays the information from Table 40.2. The strength of the attraction between *love* and its collocates is indicated by the length of the arrow. The closer the collocate is to the node the stronger the attraction.

ADVANCES BOX 40.2

Association measures

Association measures (AMs) are statistics used by corpus tools to identify collocations. We can select a particular statistic according to the type of collocations we wish to identify. The simplest option is to look at raw frequency of co-occurrence (see column 1 in Table 4). With this measure we typically highlight grammatical words because these are very frequent and thus co-occur with other words by virtue of being present in every context. More sophisticated AMs take into account not only how frequently words appear in each other's company, but also how frequently words appear generally in the corpus.

Table 4. Top ten collocates of *love* according to different AMs.

Raw frequency	MI score	T-score	MI2	Log Dice
and	affair	and	fell	fell
the	fell	i	i	love
in	falling	in	affair	affair
i	fallen	the	you	falling
of	martin	of	falling	can't
to	march	to	love	fallen
a	letters	with	fallen	me
with	love	you	me	martin
you	can't	a	in	march
was	i'd	was	and	letters

The most popular AMs in corpus linguistics are the MI score and the t-score, although t-score has been criticised for its lack of transparency (Evert, 2004). The MI-score uses a logarithmic scale to express the ratio between the frequency of the collocation and the frequency of random co-occurrence of the two words in the combination. MI score highlights collocates which are relatively rare and which are exclusively associated with the node. T-score is calculated as an adjusted value of collocation frequency based on the raw frequency from which random co-occurrence frequency is subtracted. This is then divided by the square root of the raw frequency. T-score, like raw frequency, highlights frequent combinations of words (Gablasova et al. forthcoming).

Other association measures, which each highlight different types of collocates, include MI2, log Dice, delta P etc. (Brezina et al. 2015, Evert 2008).

40.2.3 Keywords

Keywords are words typical of the corpus we are interested in (C) when compared with another corpus, which we call a reference corpus (R). Keywords are produced automatically by using a statistical procedure (Brezina, forthcoming: Chapter 3), which takes into account the frequencies of individual words in C and compares them with the frequencies of the same words in R thus identifying words that are more frequent in C than in R. Keywords are used to show the characteristics of a particular variety of language, genre or discourse etc. For example, the following keywords (Table 40.5) are typical of academic writing when compared with general English. The corpus used in this analysis is BAWE (British Academic Written English Corpus); it is compared to the BNC as a reference corpus representing general English.

Table 40.5 Top 15 keywords typical of academic writing.

Keyword	**Frequency in BAWE** (per million)	**Frequency in BNC** (per million)
this	2323.1	980.9
however	942.5	245.1
is	13281.9	8672.8
press	467.6	21.2
et	493.4	46.1
due	594.3	133.1
also	1873.9	1059
figure	459.3	47.4
order	755.6	282.3
therefore	617.1	183.7
different	920.9	417.5
data	542.6	140.7
p.	511.9	125
social	760.5	311.4
model	485.6	106.7

We can see that the top academic keywords (i.e. words that are much more frequent in academic writing than general English) include academic content words occurring across different disciplines such as *Figure, different data, social* and *model* and also grammatical words such as *this*, *however, is, due, also, order* and *therefore*. The grammatical words are typically used to explicitly state facts (*is*), indicate the relations between concepts and ideas (*however, due to, also, in order to, therefore*) and point or refer to (*this*) events and phenomena. The remaining keywords (*Press, et, p.*) are indicative of the references to literature used in academic writing, e.g. a book published by Cambridge University *Press*, with multiple authors listed as *et al*. with reference to a particular page *p. 3*.

40.3 Research design: How do we search corpora?

Corpora, as large collections of data, are an important source of information about language use. So how do we go about searching corpora? First, we have to operationalize our questions or topics we are interested in; that means we need to express the questions/topics in specific terms, which can be searched in the corpus. For example, if we are interested in finding expressions of anger in language, we have to come up with a list of words, phrases (e.g. swearwords) and non-linguistic clues such as hesitations and pauses in speech that signal anger. Second, once the question/topic is operationalized, we can start searching the corpus, count the instances of the occurrence of our target words/phrases etc. and decide whether a particular text or speaker displays linguistic evidence of what we were initially interested in (e.g. anger). Third, we often use comparison as one of the main methods of corpus linguistic inquiry. For example, we compare language use in different modalities (e.g. speech vs. writing), genres (e.g. academic writing vs. newspaper language) and speaker groups (e.g. men vs. women, younger vs. older speakers).

Corpora can be searched at different levels. The simplest way is to search for specific words (*love*, *hate* etc.). We can also search for multiple words and phrases such as *I love* or *I hate.* However, if our corpus includes additional information, which is called **annotation** and which can be added when building or processing the corpus, we can also search at a more abstract level, i.e. for morphological, syntactic or semantic (meaning-related) patterns. For example, we can ask how many nouns, verbs, adjectives etc. there are in a text/corpus, how many noun phrases, complex sentences etc. or how many words in the text/corpus express an emotion. When we annotate a corpus, we linguistically analyse it, often using automatic or semi-automatic methods, according to the target categories and add this information to the corpus where it is attached to each word or phrase as a 'tag'. The process is therefore sometimes referred to as 'tagging' a corpus.

Different types of annotation encode linguistic and meta-linguistic information at various levels. A starting point for any type of annotation is plain text files, where nothing but text is included. As an example of how annotation works, the following plain-text utterance from the Trinity Lancaster Corpus of spoken learner language (Gablasova et al. 2015) will be used:

[40.1] The politic men er think er want only earn money but they doesn't doesn't help us to live.

The most common type of annotation at the level of words is automatic **part-of-speech (POS) tagging**. This procedure uses a computer program called a POS-tagger to process large amounts of text and add to each word the information about the word class (noun, pronoun, adjective, verb etc.). CLAWS, a system developed at Lancaster University, is an example of a POS-tagger for the English language. When we process our model utterance by CLAWS (http://ucrel.lancs.ac.uk/claws), we get:

[40.2] The_AT0 politic_AJ0 men_NN2 er_UNC think_VVB er_UNC want_VVB only_AV0 earn_VVB money_NN1 but_CJC they_PNP does_VDZ nt_XX0 does_VDZ nt_XX0 help_VVI us_PNP to_TO0 live_VVI ._PUN

Each of the tags encodes the information about the word class of a word such as the definite article (AT0), adjective (AJ0) or noun in plural (NN2). This type of annotation helps us distinguish different uses of the same form, e.g. *love* as a noun (*Love*_NN1 *is all around us*.) and *love* as a verb (*How can anyone love*_VVI *her as much as I do?*). We can thus, for example, ask the corpus tool to search only for those instances in which *love* occurs as a verb.

Semantic annotation, another type of annotation, adds information about the meaning categories of words. USAS, a system developed at Lancaster University, provides an automatic semantic annotation of texts. When we process our model utterance by USAS (http://ucrel.lancs.ac.uk/usas), we get:

[40.3] The_Z5 politic_G1.2 men_S2.2m er_Z4 think_X2.1 er_Z4 want_X7+ only_A14 earn_A9+/I1 money_I1 but_Z5 they_Z8mfn does_A1.1.1 nt_Z6 does_Z5 nt_Z6 help_S8+ us_Z8 to_Z5 live_H4 ._PUNC

Each of the tags encodes the information about the semantic (meaning-related) category of a word. For instance, G1.2 category subsumes all words related to 'Politics', while S2.2 includes words related to 'People: Male'. We can thus use the USAS tags to search for broad areas of meaning in the corpus.

A type of annotation common in corpus-based research on language learning is **error annotation,** in which different types of errors in learner language are identified and coded. This is largely a manual process. Below is an example of simple error annotation of the model sentence. It uses a different type of notation with an error code in angle brackets (< >) showing the beginning (e.g. <lex>) and the end (</lex>) of an error.

[40.4] The <lex>politic men</lex> er think er want only earn money but they <grammar>doesn't </grammar><grammar>doesn't</grammar> help us to live.

Two types of errors were identified: lexical choice (<lex>) and grammatical errors (<grammar>). Instead of *politicians* the speaker, whose mother tongue is not English, opted for *politic men*. The grammatical error *doesn't* is a subject verb agreement error with the target (correct or expected) variant *don't.* Researchers can than quantify the frequency and type of each type of error and use this to evaluate the effectiveness of different teaching methods or common traits in the language of speakers from a specific linguistic background. These are merely examples of the most common types of annotation that can be added to a corpus; corpora, however, can be annotated for a many other features such as syntactic structures or pragmatic functions depending on the research question we want to investigate.

Finally, we need to be aware that corpus research is a process of constant engagement with the data at various levels of abstraction, which moves from close reading of examples to abstract statistical analyses and visualizations and back to examples. Doing corpus research means going through a 'circle of interpretation'. In this process, we need to ask not only what we see in the data but also think about linguistic and discourse functions of the observed patterns and possible interpretations (see Illustration box 40.2).

ILLUSTRATION BOX 40.2
How to analyse corpus data?

When analysing corpus data, we usually follow three basic steps: (i) observe, (ii) interpret and (iii) contextualise. The example below uses BE06 to demonstrate this three-step process:

OBSERVE
Table 40.6 shows different written genres in BE06 and the frequency of pronoun use. From the table we can see that pronouns are much more common in fiction (107,664 per million words) than any other written genres; on the other hand, they are especially infrequent in academic writing (17,062 per million words). Note that we use so-called **relative** (or **normalised**) **frequencies** to compare different genre-based parts of the corpus, which are of unequal sizes. Relative frequency is somewhat similar to percentages; however, instead of calculating the proportions from one hundred we use a larger number (basis of normalisation) such as ten thousand or one million. This can be seen in the example below where, for fair comparison, the frequencies were normalised to one million words.

Table 40.6 Use of pronouns in BE06

Genre	**Freq. per million**
Fiction	107,664.09
General prose (non-fiction)	54,174.12
Academic writing	17,062.09
Newspapers	58,050.73

INTERPRET
The concordance that displays specific examples of pronoun use in different genres can help us understand the functions pronouns have in various genres. In the concordance, we can observe that fiction often uses pronouns to express a subjective style as in example [40.5].

[40.5] This is my first meeting, and it's something I'm only just coming to terms with (BE06, K01).

On the other hand, when used in academic writing, pronouns often refer to the authors (*we*) or are a part of an impersonal constructions (*it should be noted…*) as in example [40.6]. The first person pronouns *I* and *my* are notably infrequent. Academic writing also uses strategies to avoid pronoun use such as passivation (using the passive voice as in *this was done* instead of *I did it*).

[40.6] It should be noted that we suggest that the more serious risk cases in this scheme be allocated to specialist social workers (BE06, J04).

CONTEXTUALISE

On a more general level, we can relate the use of pronouns in fiction and academic writing to two different linguistic functions of these genres identified in the literature (e.g. Biber et al. 1999). While fiction focuses on the narrative and expressive aspect of language, academic writing emphasises accurate description of complex phenomena.

40.4 Types of corpora: How to choose a corpus?

Corpora are developed with different aims in mind. It is always important to choose a corpus that best suits the research question we want to answer and that most accurately represents the language we are interested in. Corpora can be divided into several categories. In this section, we review some of the major types of corpora and their areas of application to help you make an informed choice when looking for a corpus for your own research.

First, corpora can be categorised according to the mode of communication they represent. The majority of current corpora represent written language. **Written corpora** can include works of fiction, academic articles, student essays, newspaper articles but also language from the web such as emails, blogs, tweets and Facebook posts. The newly emerging genres and registers of internet language open a new dimension of exploration of English writing (e.g. Biber et al 2015, Huang, 2015). Another group of corpora are those representing spoken language. **Spoken corpora** are usually smaller than written corpora and are generally few and far between because they are more difficult (and expensive) to produce. Spoken language needs to be recorded and then painstakingly transcribed in order for it to be searchable by a computer. When transcribing speech, researchers have to make many decisions about how to capture the complexity of the spoken code including pauses, hesitations, false starts, overlaps etc. Example [40.7] comes from the Spoken BNC2014 (Love et al. 2017), a corpus of informal British speech.

[40.7] S1: when you're out and about?
S2: mm
S1: might just wanna smartphone do you know what I mean?
S2: no I bloody don't er sorry to swear I don't want a sm- if I want a smartphone I get one as well
S1: all right okay

You may have noticed some of the typical features of speech such as the hesitation (*er*), an unfinished word (*sm-*), and also the fact that in spoken language there are no sentences which we could mark by a full stop. Instead, we can identify different intonation patterns (e.g. questions), communicative units (utterances) and speaker turns (S1, S2). Apart from written and spoken, we sometime also build **multimodal corpora,** which capture language production in different modalities. In addition to the written electronic form (transcript) they typically include video recordings that are aligned with the transcript and can be analysed together with the transcript. These types of corpora are especially useful when the additional information from the audio/video is essential for the type of interaction we wish to study. For example, this would be the case when analysing a sign language (Fenlon et al. 2014) or the role played by gestures in communication (Adolphs & Carter 2013).

Another classification of corpora takes into account their scope and variety of language they represent. A basic distinction is made between general and specialised corpora. **General corpora** are usually very large (hundreds of millions or billions of words) and contain language from different areas of use and different situations. Their aim is to capture the diversity of communication across a language and thus to represent the language as a whole. General corpora are used, for example, when compiling a dictionary or a grammar book. **Specialised corpora,** on the other hand, represent language from a specific language use or a specific group of language users. For example, academic English corpora, healthcare communication corpora, aviation English corpora, classroom language corpora belong to this group.

Finally, a third major categorisation of corpora is related to the group of users of language that are included in the corpora. Here, we distinguish between **native-speaker** and **non-native speaker** (or **learner**) **corpora.** Native speaker corpora capture the target language of speakers and writers who grew up with the language. Non-native speaker corpora represent L2 (second language) production, that is speakers of English (or another language) from different L1 (native language) backgrounds and proficiency levels. The non-native user corpora are often compared with native user corpora for pedagogical reasons, that is to determine where the main areas of difference are between learners and native speakers. These areas can then be addressed in textbooks and teaching of foreign languages.

ADVANCES BOX 40.3

Overview of some available corpora

Corpus	Modality	Variety	Size (words)	Availability
General				
BNC	spoken (10%) and written (90 %)	current British English	100 million	CQPweb
COCA	spoken (20%) and written (80 %)	current American English	530 million	BYU corpora
BE06	written	current British writing	1 million	CQPweb
AM06	written	current British writing	1 million	CQPweb
English Web (EnTenTen13)	written	current international English online	20 billion	SketchEngine
Specialised				
BASE	spoken	British academic speech	1.2 million	SketchEngine (free)
BAWE	written	British academic writing	7 million	SketchEngine (free)
EEBO-TCP v.3	written	historical British English (1400-1800)	1.2 billion	CQPWeb
COHA	written	historical American	400 million	BYU corpora

		English (1810s-2000s)		

40.5 Conclusion: Applications of the corpus method

Many areas to date have benefited from the use of the corpus method to study the English language. Corpora and corpus techniques have become one of the key tools in the modern dictionary-making with all major English dictionaries today (e.g. *Oxford dictionary of English, Cambridge English dictionary, Macmillan English dictionary*) being based on large general corpora. Corpus analysis also plays an increasingly more important role in understanding how our society functions. For example, the techniques mentioned in this chapter have been used by Semino et al. (2016) to better understand the language of healthcare communication, focusing in particular on areas such as the interactions between cancer patients and healthcare providers. The corpus method also lies at the core of many studies which investigate the media discourse and its impact on society. A study by Baker et al. (2013) examined how the British newspaper discourse referencing 'Muslim' and 'Muslims' creates and propagates a particular type of identity; Hardaker & McGlashan (2016) used corpus techniques in their study of abuse and misogyny on twitter. Corpora are also gaining prominence in applied linguistics (Gablasova et al. 2015) and foreign language pedagogy (Conrad 2000), where they assist teachers, learners, material developers and language testers with gaining insight into the use of English. Through its application in many areas of linguistic and social interest, the corpus method has become one of major approaches to the analysis of linguistic data.

Recommended readings

An excellent overview of the field, its history as well as corpus approaches and techniques is offered by McEnery & Hardie (2011). A practical introduction to corpus linguistics with readings and exercises can be found in McEnery et al. (2006). Adolphs (2006) offers an accessible entry-level textbook suitable for beginners in corpus linguistics. Baker et al. (2006) provide a very useful glossary of corpus terms and techniques with clear explanations and examples; this volume is an essential reference guide for anyone exploring the field. There are four main journals dedicated to current research in corpus linguistics: *Corpora*, *Corpus Linguistics and Linguistic Theory, International Journal of Corpus Linguistics* and *International Journal of Learner Corpus Research*. Lancaster University also offers a free online course (MOOC), *Corpus Linguistics: Method, Analysis, Interpretation,* which runs yearly.

References:

Adolphs, S. (2006). *Introducing electronic text analysis: A practical guide for language and literary studies.* New York: Routledge.

Adolphs, S., & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. New York: Routledge.

Aijmer, K. (ed.). (2009). *Corpora and language teaching*. Amsterdam: John Benjamins.

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/ [Accessed 10/10/2016].

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

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, *14*(3), 312-337.

Baker, P., Gabrielatos, C., & McEnery, T. (2013). Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim' in the British press 1998–2009. *Applied Linguistics*, *34*(3), 255-278.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, 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.

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press

Barlow, M. (2003). MonoConc Pro 2.2: A Professional Concordance Program. *Computer software]*. Houston, TX: Athelstan.

Biber, D., Egbert, J., & Davies, M. (2015). Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora*, *10*(1), 11-45.

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

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics, 36* (1):1-22.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139-173.

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century?. *TESOL Quarterly*, *34*(3), 548-560.

Davies, Mark. (2002-) BYU corpora. Available online at http://corpus.byu.edu [Accessed 10/10/2016].

Evert, S. (2008). Corpora and collocations. In Ludeling, Anke, & Merja Kyto. *Corpus linguistics. An international handbook*, *2*, Berlin: Mouton de Gruyter, 223-233.

Fenlon, J., Schembri, A., Rentelis, R., Vinson, D., & Cormier, K. (2014). Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua*, *143*, 187-202.

Gablasova, D., Brezina, V & McEnery, T. (forthcoming) Collocations in SLA research: identifying, comparing and interpreting the evidence. *Language Learning.*

Gablasova, D., Brezina, V., Mcenery, T., & Boyd, E. (2015). Epistemic stance in spoken L2 English: the effect of task and speaker style. *Applied Linguistics*, *36*(1), 1-22.

Hardaker, C., & McGlashan, M. (2016). "Real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, 91, 80-93.

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*(3), 380-409.

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59, 244-255.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*,*1*(1), 7-36.

Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. (forthcoming, 2017). The Spoken BNC2014 - designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

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

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*. *13* (4), 519-549.

Scott, M., 2016, WordSmith Tools version 7, Stroud: Lexical Analysis Software.

Semino, E., Demjen, Z., & Demmen, J. (2016, Advance Access). An Integrated Approach to Metaphor and Framing in Cognition, Discourse, and Practice, with an Application to Metaphors for Cancer. *Applied Linguistics*.