CHAPTER 23

Corpora in English language teaching

Hilde Hasselgård

Corpora, in the sense "databases that contain texts", are a widely used tool in language research. This tool can be a helpful resource for language teachers and learners as well, and many corpora are available online. In this chapter, Hilde Hasselgård discusses the use of corpora in the English classroom. She explains what a corpus is, and what answers a corpus can provide to questions about language use. Then she presents various possibilities for making use of corpora in language learning.

Finding out how language is used

Linguists use corpora to investigate how language is used in natural texts and to test their hypotheses about language. Such investigations sometimes concern the frequencies and the patterns of words. What words are used more or less frequently than others and in what types of texts? What lexical and grammatical patterns do words tend to occur in? What words and grammatical patterns tend to occur together?

Investigating language through corpora is not limited to researchers: a corpus can be a useful tool for language teachers and learners alike. In a language learning situation a corpus can act as a reference tool along with dictionaries and grammars. Everyone can search for particular words and expressions and get examples of how these are used in natural contexts.

This approach to language is descriptive rather than normative: it shows people how language is used in actual discourse but does not tell them how to speak or write. Unlike a dictionary, a corpus does not explain what a word means, but it can show how the word is used in sentences. It thus enables corpus users – after having interpreted the search results – to answer questions such as:

- What preposition is used after *critical*?
- Do people say *eggs and bacon* or *bacon and eggs*?
- Is the word *bloody* still used in the sense "full of blood"?
- Can I write *and so on* in a formal text, or is *etc*. more common?

Many corpora are freely available on the Internet. Some of these are listed at the end of this chapter. Corpus linguistics is a fast-developing field, so there may be good and useful corpora that are not listed. Nevertheless, the list should give plenty of material for studying speech and writing in the major varieties of English.

What is a corpus?

As explained above, a corpus is a digital collection of texts. A more precise definition is that a corpus is a structured database of natural texts prepared for use in linguistic research. That is, the texts are not produced in order to be included in a corpus, but in order to communicate in authentic settings. However, the selection of texts and the way in which they are organized have been planned with linguistic research in mind. This means that not all databases of texts are corpora, because they may have been created for other purposes. Examples are archives of newspapers and public documents. When linguists compile a corpus, they typically aim to make the corpus *representative*: that is, it should give a fair image of the language of a particular period, region or genre, for example. Users of a corpus need to be aware of what the corpus contains and thereby what kind of language it is meant to represent. For example, we cannot make claims about language in literature or in classroom settings based on a corpus that consists exclusively of newspaper text. Similarly, a corpus that contains only writ-

ten material cannot show us how people speak, and one that contains only conversations cannot reveal whether the singular *is* in *There's hundreds of tourists* is acceptable in writing.

The fact that a corpus has been prepared for use in linguistic research means, among other things, that the end user can see where the texts come from. In the *British National Corpus* (BNC), for example, each sentence of each text has a code that reveals its source. So in the case of the sentence below, the code (KS8 785) reveals – if we click on it in the corpus display – that the sentence comes from a concert programme.

> Tonight's programme features music from all these styles. (KS8 785)

In the BNC, as in many other corpora, the preparation also includes *tagging* each word for part of speech, which means that each word gets a tag attached to it with a code for the word class it belongs to. Such tags are thankfully not visible in a normal view of the corpus, because they clutter the text spectacularly. Here is what the above sentence looks like with visible tags:

> Tonight<AV0>'s<POS> programme<NN1> features<VVZ>
> music<NN1> from<PRP> all<DT0> these<DT0> styles<NN2>.

However, because the tags can be included in corpus queries, the searches can be made very precise. For example, we can search for *feature* as a verb (as opposed to a noun) or for an adjective (such as *clever*) followed by any preposition. The latter search in the BNC shows that the most frequent preposition to occur after the adjective *clever* is *of*, as in the first sentence below. Next in frequency we find *at*, *by*, *with* and *for*, as in the next sentences below. These sentences illustrate not only different expressions with *clever* and a preposition, but also different meanings the adjective.

> It was very *clever of* you to find it.
> He was *clever at* finding bargains.
> Oh, you're too *clever by* half.
> Still, I expect she's *clever with* her hands.
> According to her, he was too *clever for* his own good.

Concordance

A corpus may come with a so-called search interface. That means that it has its own search tool which is tailored to searching in the corpus and displaying the output in a way that is useful for language study. This leads us to another central concept in corpus linguistics: *concordance*. A concordance is a list of the occurrences of the word or phrase that was searched for. Each occurrence has a context, which can be a sentence or a specific number of words or characters on either side of the search word(s). Figure 23.1 shows what a concordance may look like. The search word, *hopeful*, is highlighted and appears in the middle of the line. The lines have been sorted so that they appear in the alphabetical order of the word to the right of *hopeful*. This function of the search interface makes it easier to see what patterns the word occurs in. For example, the concordance in Figure 23.1 shows that *hopeful* can describe both people and things, with the difference that a "hopeful person" is hoping for something, but a "hopeful thing", as in *a hopeful sign*, is one that gives people hope. The concordance also gives an example of *hopeful* used as a noun. We can compare *hopeful* with the similar adjective *wishful*, which turns out to have a much less varied pattern of use in the same corpus: it occurs only in the expressions *wishful thinking* and the related *wishful thinker*.

	Details	Left context	KWIC	Right context ↓
1	Christian Unity in …	churches . </s><s> Many people seem	hopeful	, yet it is difficult to predict whether or
2	Land of the Silver …	istance , while the men were stern but	hopeful	. </s><s> All , of course , except the D
3	various	tion of the seven principal Presidential	hopefuls	: five Democrats -- Senator Hubert H.
4	Deadlier Than the …	s><s> Jeb cautioned him not to be too	hopeful	and then , ignoring his own advice , s
5	My Hero.	k Adam Herberet is guilty of being too	hopeful	and better informed on defense financ
6	The Heartless Light.	else to tell them : no assurances , no	hopeful	hints at great discoveries that day . </
7	Values and Moder…	rtising agency ; ; </s><s> and many a	hopeful	incipient business executive decides i
8	various	g March 4 . </s><s> Mr. Hodges is so	hopeful	over the outlook that he does <g/> n't
9	Peace Corps. Fact…	s new peaceful program , this will be a	hopeful	sign to the world . </s><s> Congress
10	A Passion in Rome.	it aside for good . </s><s> But it was a	hopeful	sign , he told himself . </s><s> She n

Figure 23.1 Concordance for *hopeful* (Brown Corpus, via Sketch Engine).

Frequency

Corpus linguists are often interested in *frequency*; that is, how often a word or expression occurs in the corpus, particularly in comparison with other words and expressions. Information on frequency is useful in choosing among alternative wordings. For example, the word *different* can occur with various prepositions: *from, to* and *than*. A search in the BNC for *different* + preposition shows that *different from* occurs 3243 times, *different to* 429 times, and *different than* 50 times. Thus the safest option is *different from*. It is of course possible that the prepositions signal different meanings, as we saw with *clever* above. This is also the case with *sorry for/about*, as indicated by concordance lines with those combinations in the BNC (in Table 23.1).

Table 23.1 *Sorry for* and *sorry about* in the British National Corpus.

H8T	But when nobody else feels sorry for you, you tend to feel **sorry for** yourself, don't you?
KBL	No I'm not being rude to Brian! I feel **sorry for** Brian! What! I'm really hurt by that! Well that's
HWL	Swedish number plates in order to follow him. I felt a wee bit **sorry for** the driver—reindeer probably don't drive like that—but
CDM	. I'd liked her until the money lending began, and I was **sorry for** being unkind to her. Then Frankie saw me. " No use
G06	horrible but who is too polite to draw attention to the fact by seeming **sorry for** her. I didn't yet know Lili.
CDM	went out dolled up. Everybody seemed to shun her but I felt **sorry for** her and we became close friends. However, she didn't
HHC	, I know you've said it doesn't matter, but I am **sorry about** what happened up at Handley Farm. " Sorry for what,
CCE	make do with animal furs and leather where we are going. (**Sorry about** that.) Let's pause for a while in order to breathe in
HGM	this time by Ace. " Dara, great to see you! **Sorry about** last night! Do you two girls know each other? " Kate, on
B0U	which comes from solitary confinement was already maturing inside me. " **Sorry about** the bread. It was the best I could do.
FP7	; if it comes to that, so do you. " I'm **sorry about** the phone call. I got the instruction but I didn't really take

In order to compare frequencies from two different corpora, we need to know how big the corpora are. In the case of *different from*, for instance, we find 35 examples in the Brown Corpus, which was used for Figure 23.1, while the BNC has 3243. But since the Brown Corpus contains one million words and the BNC has 100 million, *different from* is actually slightly more frequent in Brown, with 35 vs. 32.4 occurrences per million words.

Variation

Corpora can also reveal *variation* in language use. The variation can be between different patterns of the same (or similar) words, as in the examples with *different from/than/to* and *hopeful* vs. *wishful*. But it can also be between regional varieties of English or different text types. The corpus search interface at https://www.english-corpora.org/ lets the user compare text types. For example, a search for the word *stuff* in the *Corpus of Contemporary American English* (COCA) shows that the word is clearly most common in speech and rather rare in academic writing. Using the GLOWBE corpus (short for *Corpus of Global Web-Based English*) we can find, for example, that the modal auxiliary *ought to* is more frequent in Nigerian English than in any of the other regional varieties included in the corpus, and least frequent in Irish English. This information may not be of much interest in itself, but could inspire further investigations of modal expressions in different varieties of English.

Another aspect of the linguistic variation in a corpus is that not all corpus examples will be universally recognized as good or correct usage. This is an inevitable consequence of the fact that the corpus consists of authentic texts. Speakers/writers simply do not always express themselves in ways that language learners should copy. Therefore it is a good idea to use frequency as a pointer to what expressions are preferred in the corpus, since we can assume that acceptable uses will outnumber erroneous ones. When in doubt about the acceptability of a corpus example, we can also look up the same expression in a dictionary.

How can we use corpora in English language teaching?

The answer to this question depends on a number of factors, especially the proficiency level of the learners, their learning goals, and the practical issue of whether computers are available in the classroom. Corpora can be helpful to a language teacher regardless of how these factors play out, but the degree of involvement with the corpus will vary. Various uses of corpora in language teaching and learning are visualized in Figure 23.2.

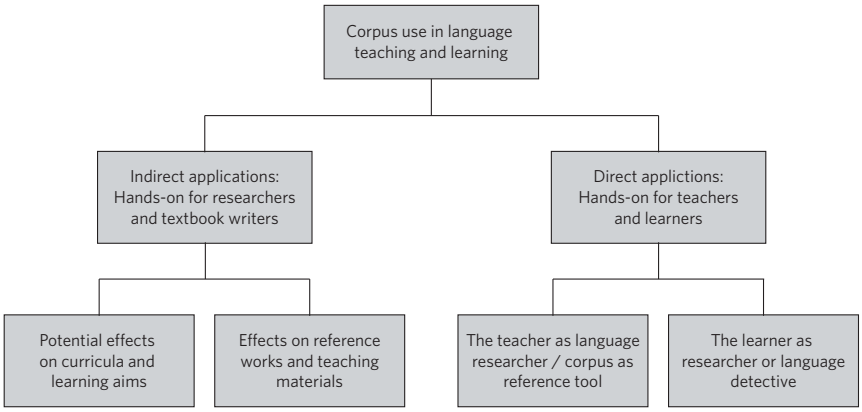

Figure 23.2 Corpus use in language teaching and learning (based on Römer 2011, p. 207).

Indirect applications of corpora

Even without using corpora ourselves, we can benefit from teaching materials and reference tools that are based on corpus investigations. For example, most major English-language dictionaries are corpus-based. As a result, the dictionaries can give reliable information on the kinds of contexts a word or phrase occurs in; for example whether it is academic or colloquial, rare or frequent. Besides being an excellent source of authentic examples that can show the usage patterns of words and phrases, corpora can be used to define a core vocabulary. Studies have shown that the 2000 most frequent words in a ten-million-word corpus of spoken and written English account for 83% of the text (O'Keefe et al., 2007). This means that a learner who has

acquired these words will get by in most situations. In addition, it may be helpful to work with specialized corpora to identify and teach slightly less frequent words that are nevertheless useful in the kind of situations and genres that the learners are likely to encounter. One use of specialized corpora is to extract word lists to tailor the vocabulary teaching to specific learner groups and learning purposes. For example, young learners may focus on vocabulary that is frequent in corpora of children's books and everyday conversations in family settings, while those learning for a particular profession (such as cooking, carpentry, engineering, business) could learn vocabulary drawn from corpora of books, magazines, journals and transcribed lessons within their field, or, say, from recorded service encounters if the goal of the learning is to be able to serve customers in English.

Grammarians have drawn on corpora to improve the description of language structure. Here, too, the corpus is an invaluable source of authentic examples of grammar in use. But even more importantly, the corpus can provide accurate information on the occurrence of particular grammatical constructions in different contexts. An example of a corpus-based English grammar is the *Longman Student Grammar of Spoken and Written English* (Biber, Conrad & Leech, 2002). This book contains information that would have been impossible to discover and document without the corpus, for example that the passive construction with *get* is most common with the verbs *(get) married, hit, involved, left* and *stuck*. Other verbs seldom make use of their potential to occur in the passive, for example, *exclaim, guess, hate, joke, love, try* and *want* (pp. 171–172). The corpus that this grammar is based on contains four genres: news articles, fiction, academic prose and conversation. We can thus also learn that the passive is most common in academic English and least common in conversation (p. 178). Less obviously, the passive is more common in news than in fiction.

Corpora can furthermore be the basis of exercise materials for both vocabulary and grammar learning. It is possible to find ready-made exercises of this kind both online and in printed form, often under the heading "data-driven learning". However, it is also relatively simple for teachers to design their own exercise materials from a corpus. For example, we can create exercises on "confusable words" (for example *teach/learn, take/bring*) by searching for these words, deleting the "keywords" from the concordance

lines so that the students can fill them back in, as shown in the sample exercise. It may be useful to omit concordance lines from the exercise that might confuse the learner, for example because of difficult vocabulary. If students have corpora available to them, they can create such exercises for each other.

Sample exercise

This exercise is based on a selection of sentences retrieved from searches for *teach* and *learn* in the British National Corpus. In the exercise, these words have been removed. The same technique can be used for other word pairs or for different grammatical forms (such as *walk/walks, walks / is walking*).

Instruction for the learners: Fill in *teach* or *learn* in the sentences below.

1. And our spinners too have had to _____ how to bowl on these pitches.
2. He goes on to argue that we can _____ to cope with the anxiety associated with an anticipated event.
3. It may not be possible to get rid of the mite completely but we can _____ to live with it.
4. It will also help _____ your child computer interaction at an early age.
5. Keep an eye on how employees develop, and _____ from their mistakes.
6. She had the joint X-rayed on Wednesday night and was somewhat relieved to _____ that it was merely a bad case of bruising.
7. That will _____ your father to allow your "admirers" to visit the Black Lion.
8. We have to _____ from the past.

Direct applications of corpora: teachers and learners as corpus users

Regardless of the proficiency level of the students, a language teacher can use corpora to check words and expressions in student texts, to create teaching materials, and to find good illustrations to explain words, phrases and grammatical constructions. A big corpus will help a non-native English teacher answer the question *Can you say this in English*? It can also be

used to check a suspicion that the style level is inappropriate. For example, learners commonly use a lot of informal and spoken-like expressions in their written English. A corpus can show us that certain expressions belong mainly to conversational style, such as *you know, a little bit, really great*. Similarly, we can search for expressions that sound odd. If they cannot be found in the corpus, or there are very few of them compared to other expressions with similar meaning, we can use this to advise learners to choose different wordings. However, it is important to remember that no corpus can contain everything. The fact that an expression is not found in the corpus is not proof that it does not exist in the language. A word or expression may, for instance, be more recent than the corpus texts, or it may belong to a genre that the corpus does not contain.

An exciting feature of corpus use is that students can use the same methods as language researchers. Even at relatively low levels of proficiency, learners can come to the corpus with questions about the language and discover something new. And as Stig Johansson (2011) observes, corpus users often make interesting discoveries in addition to what they were originally looking for. In order for learners to benefit from corpus use, however, it is important that the material is suitable for their level. Most language teachers have experienced that learners at low levels of proficiency are overwhelmed when faced with authentic texts. It is therefore vital that students are given manageable tasks. As they become more proficient, they can handle increasingly more advanced and independent tasks.

When exposing young learners to corpus material, it may be useful to give them a simple concordance focusing on a word or phrase that is central to what the class is working on. The sentences below came out of a search for *cat* in an interface to the BNC where it is possible to specify the target readership of the texts, in this case children and teenagers (unfortunately this interface is not freely available, but see the section on "Do-it-yourself corpora"). Such sentences can be the starting point for work with words and expressions for describing a cat and its activities.

> So the striped **cat** made her way to the garden of the tower. (FUB 611)
> Slowly the **Cat's** eyes, then its ears, and then the rest of its head appeared. (FNS 448)

The **cat** was black and white: half its face was black and half was white; half its body was black and half was white. (FSL 251)
Tom **Cat** jumped down. (B2N 45)
She stroked the **cat**. (CHo 1898)
… the man waited, like a **cat** waiting for a mouse. (FSJ 75)

A more advanced task, though of the same type, might stimulate language awareness and (cross-)linguistic reflection by targeting words or expression that have multiple meanings and/or lack a direct counterpart in Norwegian. One such a word is *mind*, illustrated in Table 23.2. *Mind* can be a noun or a verb, and it enters into a number of (relatively) fixed expressions in both functions. Besides, from a Norwegian point of view, *mind* is a hard word because there is no Norwegian word with exactly the same meaning; rather, we use a number of different words to express the meanings of *mind*, depending on the context. Here are some suggested activities using a concordance of *mind* (see Table 23.2):

- Identify the expressions with *mind* that the learners are familiar with. This can be followed up by, for example, finding equivalent expressions in Norwegian (or other languages that the learners speak) or synonymous expressions in English.
- Identify the expressions with *mind* that the learners are unfamiliar with. Try to work out what they mean from the context. Search for more examples to check if the assumptions are correct.
- The students may search in a corpus for more examples of the expressions in the concordance, to see if there is any variation in the wording. For example, searching for "on * mind" from line 3 (where the * can stand for any word), we find *his, my, her, your, the* and *its* between *on* and *mind*. *On his mind* occurs in expressions with *be* (something is on his mind) or *have*, as in line 3. Since *the* is different from the rest, it is worth exploring. *On the mind* occurs mainly in formal contexts, for example *Other ideas seem to operate on the mind with great force*. It combines with other verbs than *have* and *be*, and seems to be less of a fixed expression.
- If the corpus contains different genres, it is also possible to find out where the expression is typically used. For example, *never mind* and

I don't mind both occur mainly in conversation and fiction, and not so much in the other genres in the BNC.

- Focus on complex constructions with *mind*, such as the pattern in lines (4) and (9) and produce similar sentences with this pattern.

Table 23.2 Concordance of *mind* from the British National Corpus.

1	ABX 3249	I don't **mind**.'
2	ACB 670	'**Mind** you,' Gazzer continued, 'I'll remember what you said about them pill-boxes.
3	ACV 838	Mungo had so much on his **mind** that he was unable to concentrate on Mary Ann's stories.
4	ACV 2293	'Don't you **mind** him stealing your father's eggs?'
5	AEB 1717	'Give me time to make up my **mind**.
6	AT4 3311	'Well, Deirdre's a determined girl once she's set her **mind** on anything, I'll give her that.
7	B0B 2029	The thought of Joe came at once into her **mind**, but where could Joe be?
8	BMS 2486	Can't you **mind** your own business?'
9	BMU 1507	If people don't **mind** my having no degrees, I could give a few music lessons!
10	BMU 2671	Oh, never **mind** — that's beside the point.
11	C85 3475	I wouldn't **mind** taking up with her … '
12	C87 1798	Snooker's a lot more fun when you don't have to wear a tie and waistcoat — who in their right **mind** would want to do that?
13	CE0 331	But we didn't **mind**, we soon forgave him
14	CEJ 216	He wrote with a particular audience in **mind** and therefore emphasised the points of interest most suited to that audience.
15	CFJ 265	'She's made up her **mind**.'

Another task type, which can be easily adjusted to the learners' needs and proficiency level, is to search for a word or phrase with multiple meanings and identify instances of the different meanings from the concordance. For example, the verb *appreciate* can mean "value" or "realize/understand", as illustrated by the following sentences from the BNC.

What a nice thought! I'm sure they'd **appreciate** it.

First, you must **appreciate** that a helicopter produces two different types of lift.

An area in which the corpus is unbeatable is *collocations*. A collocation is a combination of two or more words that routinely occur together. It is not necessarily an idiom, since collocations may have literal meaning and need not constitute complete phrases. But collocations represent the linguistic habits of native speakers of a language. We can find collocations in the corpus by searching for component parts of them, for example *a piece of* *, where the asterisk represents an unspecified word. The ten most frequent nouns to follow *a piece of* in the Corpus of Contemporary American English are *paper, cake, music, land, wood, bread, meat, furniture, fruit, legislation*. This information can be used in working with fixed expressions with *a piece of*, but also in a more cognitively advanced task where learners are asked to consider the meaning of *a piece of* in each collocation. They will then realize that *piece* can sometimes be translated by the Norwegian *stykke*, for example in *a piece of cake*, but not in *a piece of furniture*. Furthermore, if *a piece of cake* is used with metaphorical meaning, the Norwegian *et kakestykke* does not work as a translation. The concordance of *a piece of cake* will reveal whether the literal or the metaphorical meaning is more common.

The internet as corpus?

The internet is by far the largest existing digital collection of texts, and it can be used in language studies in much the same way as a corpus. Besides finding information on specific topics, a search engine such as Google can also show how an expression is used (especially if we use quotation marks around the expression we search for). So if we are wondering, for example, whether it is more common to say *texted him* or *texted to him*, Google will show up more than 70 times as many hits for *texted him*. A clear advantage of Google is its easy access to huge amounts of material in most languages. It is a goldmine for information on the use of new words and expressions, which may not have found their way into dictionaries or established corpora yet. At the same time one must be critical when using the internet as corpus. For example, many English texts on the web have been written by people who are not native speakers of the language. There is thus a greater risk of finding unidiomatic expressions on the web than in a corpus. This risk can be reduced by using advanced search mode, which enables us to limit the

search to websites in countries where English is the (major) first language. Another disadvantage is that, unlike a concordance, the output of a Google search is not displayed in a way that makes it easy to see patterns of use.

There are, however, web interfaces that show search output in the form of concordances, such as WebCorp (see Figure 23.3). The concordance format makes it easier to focus on the linguistic expression, and a bit harder to be distracted by the content. WebCorp sorts the hits according to the websites they come from. This is useful because the hits may come from dictionaries, which is not the type of material one is primarily after in corpus linguistics. WebCorp gives a choice of three alternative search engines, of which FAROO and Guardian Open seem to give the most relevant results for those who are interested in studying natural language use rather than definitions and examples from dictionaries.

1) http://www.adweek.com/digital/facebook-is-working-on-technology-that-lets-you-type-and-control-vr-devices-with-your-mind/

Text, Wordlist, text/html, UTF8 (Content-type), 2017-01-01 (Copyright footer)
1: Lets You Type and Control VR Devices With Your Mind Share About / AdvertisingMedia Kit Sponsor
2: Lets You Type and Control VR Devices With Your Mind Could a 'brain mouse' be coming? By Marty Swant
3: is creating a way to control VR and AR with the mind. Getty Images Share
---
4) http://dailycaller.com/2017/04/21/trudeau-says-trump-is-willing-to-change-his-mind-unlike-many-politicians/
Text, Wordlist, text/html, UTF8 (Content-type), 2017-01-01 (Copyright footer)
11: Trudeau Says Trump Is Willing To Change his Mind 'Unlike Many Politicians'
12: ideology makes him open to changing his mind. "I've learned that he listens," Trudeau told

Figure 23.3 Concordance of *mind* from WebCorp.

Alternative solutions are offered by the NOW corpus (*News-On-the-Web*) and *Querying Internet Corpora*, where existing webpages are organized as a corpus. The NOW corpus is updated every day. Table 23.3 shows some concordance lines from Querying Internet Corpora. The clickable abbreviations in the leftmost column give the sources of the examples.

Table 23.3 Concordance of *funny* from Querying Internet Corpora

nrs	Well, my God, he' s lost it. This is n't	**funny**	! " And then the more they read,
gou	"I just thought the word 'goat' was	**funny**	, " the court then pressed her to explain
ofw	the mayor controlling the system. It's	**funny**	, Joel has a sense of grass roots
nik	Gil Cates, said that, " He is smart, quick,	**funny**	, and loves movies.
pnq	fun and wrong, but it ended up being very	**funny**	, and the Nick/Phyllis affair.

"Do-it-yourself" corpora

Corpus methods can be used with all kinds of texts. There are several corpus tools (concordancers) for this purpose, one of which is called AntConc. It is freely downloadable and designed to be used with files in plain text format; that is, with the extension .txt at the end of the filename. If we have a file in a different format, such as .docx or .odt, we can convert it to plain text in the word processor by means of the function "Save as". The use of AntConc is explained on the website of its originator, Laurence Anthony (see the list of corpus resources below). Concordancers such as AntConc make it possible to use corpus methods creatively. For classroom purposes it may be interesting, for instance, to use literary texts or texts produced by students and explore them as one would a corpus.

A small study of J. D. Salinger's novel *The Catcher in the Rye* can illustrate how corpus methods can be applied to a literary text, after uploading it in AntConc in txt format. The WordList function of AntConc can provide an alphabetical list of the words in the text, but, more interestingly, it also produces a *frequency word list*, in which the most frequent words are listed first and the least frequent ones last, and which shows the number of times each word occurs. In *The Catcher in the Rye*, the most frequent word is *I*, followed by *and* and *the*. This is not unexpected, because all texts contain

a lot more so-called function words (such as pronouns, articles and prepositions) than content words (nouns, verbs, adjectives and adverbs). For literary studies we may be more interested in the content words. One of the most frequent ones in *The Catcher in the Rye* is *old*. The concordance lines reveal that *old* often appears in front of the name of a person who is objectively quite young, so in those contexts the word means "well known to the main character" rather than "aged". Furthermore, the word *goddam* is fairly frequent, which will not surprise anyone who has read the novel. Sorting the concordance by the word to the right of *goddam* we discover that all kinds of things are characterized by this adjective. And the concordance of *pretty* shows that this word is used as an adverb (for instance in *pretty good, pretty nervous* and *pretty ugly*) more often than as an adjective meaning "good-looking".

Concordances of central words in a literary text can say something about both style and content in the text. So can searches for repeated combinations of words. If we search in *The Catcher in the Rye* for repeated sequences of at least four words, we find *I don't know, all of a sudden, for God's sake* and *but I didn't* at the top of the list (AntConc counts *'s* and *n't* as words). The phrase *to know the truth* is also among the ten most frequent ones. The list of repeated sequences highlights a good number of negative sentences with *I* as subject (such as *I don't know, I didn't feel, I didn't say anything*), which may suggest a negative attitude on the part of the main character. *To know the truth* is part of the expression *if you want to know the truth*, which the main character uses often, perhaps to emphasize his truthfulness.

Using corpus tools with student texts can also be a useful exercise. Merging texts by all the students in a single text file should normally secure enough anonymity for the file to be used in class (but check whether any personal/confidential information should be removed first). Some issues that can be worth investigating include:

- What words are frequent in the students' texts compared to texts by native speakers of English on a similar topic? Are there words that are very frequent in the student texts that could be omitted or replaced by other words and expressions?

- How are words combined? For example, what words occur around *man* and *woman*?
- How do the students use particular words that they should master, for example linking adverbs (such as *however, in fact*) or technical terminology in their field?

It may be noted that the study of learner corpora is a research field of its own. See for example Hilde Hasselgård and Stig Johansson (2011) for an overview. In a Norwegian context, this research has deepened our insight into particular features of the language produced by Norwegian learners of English, including problems, successes, and features that set this learner group apart from other learner groups and/or native speakers of English.

Free online corpus resources

Below follows an overview of some of the corpus resources that are available online. Some of these resources require registration, but no subscription fee. No matter which corpus we choose to work with, it is useful to remember the following advice about sensible corpus use:

Common corpus sense

- Know your corpus! You need to know what kinds of texts the corpus contains to tell what you can use it for.
- The corpus cannot show you what is right and wrong, only what is common usage.
- Different corpora are suitable for different purposes.
- The fact that something is not found in a corpus is not proof that it does not exist in the language.
- The corpus itself cannot answer a linguistic question. The corpus user must interpret the concordances and make sense of the data.

The corpora referred to in this chapter

- British National Corpus (BNC): Contains 100 million words of written and spoken British English, from the early 1990s. https://www.english-corpora.org/bnc/
- Brown Corpus: Contains a million words of written American English from the 1960s. Searchable from Sketch Engine (Open corpora) at https://app.sketchengine.eu/#open
- Corpus of Contemporary American English (COCA): one billion words of written and spoken American English from the 1990s up to present (the corpus is continually updated). https://www.english-corpora.org/coca/
- GLOWBE Corpus: A large collection of texts drawn from the internet, representing first- and second-language varieties of English from around the world. https://www.english-corpora.org/glowbe/

Other corpora with search interfaces

- A range of large, English-language corpora are available from https://www.english-corpora.org/. Apart from the BNC and COCA, these include the NOW corpus, a corpus of American soap operas, and more.
- Some corpora are freely available from Sketch Engine (Open corpora) at https://app.sketchengine.eu/#open. Note particularly the British of Academic Written English Corpus, which contains student essays written at British universities. More corpora are available by subscription from the same provider.
- Michigan Corpus of Upper-Level Student Papers (MICUSP) contains highly rated American university student assignments in a range of disciplines. Available at http://micusp.elicorpora.info/. The interface may not work in all browsers (it works in Chrome and Firefox). The related Michigan Corpus of Academic Spoken English is found at https://quod.lib.umich.edu/m/micase/.

The web as corpus and web-based corpora

- WebCorp: http://www.webcorp.org.uk/live/
- Querying Internet Corpora: http://corpus.leeds.ac.uk/internet.html

- The NOW corpus (News-On-the-Web). From 2010 until "yesterday", updated every day. https://www.english-corpora.org/now/

Databases of downloadable texts

- The Gutenberg Project: https://www.gutenberg.org/ offers thousands of downloadable books. If in plain text, the downloaded file(s) can be used with AntConc. Recent literature is not included for copyright reasons.
- Oxford Text Archive develops, collects and preserves electronic literary and linguistic resources for use in Higher Education, in research, teaching and learning. See https://ota.ox.ac.uk/.

Corpus tool (to use text files as a corpus)

AntConc is a corpus tool that can be easily installed at no cost on a pc or a mac. The tool can be used with texts in plain text (*.txt) format. AntConc, and instructions for its use, can be found at http://www.laurenceanthony.net/software/antconc/.

Learner corpora

Many of the available corpora of learner language (English and other languages) are listed at https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html. Note that relatively few of these corpora are freely available, mainly for copyright reasons.

YouTube

A YouTube search for "Corpus linguistics" gives a number of hits. The following presentations are recommended:

- Tony McEnery: Corpus linguistics: method, analysis, interpretation. https://youtu.be/YJTM3i5HxsQ
- Randi Reppen: Using corpora in the language classroom. https://www.youtube.com/watch?v=Qf46lOnMCfs
- Michaela Mahlberg on corpus linguistics and literature: https://www.youtube.com/watch?v=kvbrp5PqNxw

Corpus exercises for secondary school students

Dypedahl, Magne and Hilde Hasselgård. 2006. *Exploring English*. Website. http://exploringenglish.cappelendamm.no/c302684/artikkel/vis.html?tid=329736.

Concluding remarks

The linguist M.A.K. Halliday (1991) argues that "the immense scope of a modern corpus, and the range of computing resources that are available for exploiting it, make up a powerful force for deepening our awareness and understanding of language" (p. 41). Both teachers and learners of English can benefit greatly from learning how to use a corpus: they will have practically unlimited sources of information on English language use at their disposal. The corpus can offer invaluable assistance where our intuitions (even in our first language) are insufficient, for example when it comes to studying collocations, making sense of ambiguous words and expressions, and assessing how frequent an expression is compared to another. In fact, research has shown that people's intuitions about frequency are unreliable because we notice what stands out rather than what seems normal. But in order to benefit from the corpus, we must know how to use it. That means knowing how to ask good questions, how to find appropriate ways of searching in the corpus and, most importantly, how to interpret the search results. Anyone who has acquired these skills will soon discover that corpus use is addictive. And unlike many other addictions, corpus linguistics is purely beneficial.

Reflection questions

1. Do you agree that corpus work represents a descriptive approach to language and language learning? What are the reasons for your agreement or disagreement?
2. What do you see as the benefits and problems of using authentic (corpus) material in English language teaching?

3. Think of ways in which learners can be guided into making discoveries about language, with or without the aid of corpora.
4. Discuss whether (and/or how) you could use corpus methods with literary texts as a bridge between active language learning and literary interpretation.
5. How would you design corpus tasks that are relevant to particular groups of learners, taking account of age, proficiency level and so-called "specific purposes" (for example English in vocational training)?

References

Biber, D., Conrad, S, & Leech, G. (2002). *Longman student grammar of spoken and written English*. Longman.

Halliday, M.A.K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 30–82). Longman.

Hasselgård, H., & Johansson, S. (2011). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora. In honour of Sylviane Granger* (pp. 33–82). Benjamins.

Hidalgo, E., Quereda, L., & Santana, J. (Eds.). (2007). *Corpora in the foreign language classroom*. Rodopi.

Johansson, S. (2011). A multilingual outlook of corpora studies. In V. Viana, S. Zyngier, & G. Barnbrook (Eds.) *Perspectives on corpus linguistics*. Benjamins.

Mukherjee, J. (2002). *Korpuslinguistik und Englischunterricht. Eine Einführung*. Peter Lang.

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

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics, 31*, 205–25.

Sinclair, J. M. (Ed.) (2004). *How to use corpora in language teaching*. Benjamins.