10

What are the basics of analysing a corpus?

Christian Jones

1 What are the basics?

Although corpora are not new and have been used to analyse language for many years, we are fortunate to live in an era where many are freely available and increasingly, the software we can use to analyse our own data is also free (for example, Anthony 2019; Davies 2019). However, Evison (2010) is right to suggest that on their own, corpora cannot tell us much about language (see also Chapter 2, this volume). The insights come from our own analysis of the data. In order to analyse a corpus, we need to know three main things: why we are using a corpus, what we aim to find out and what the limitations of our analysis are. I will briefly discuss each area in turn.

Why am I using a corpus?

A corpus is simply a searchable collection of texts (written or spoken) stored in an electronic form. Sinclair (1991) suggests that this collection should be principled so that it represents the types of language we wish to better understand. Corpora are often mono-modal (using one medium, normally text) but can be multi-modal (using more than one medium such as text and video). Due to costs and ethical considerations (we cannot anonymise video easily), most corpora are mono-modal, though this situation is changing and increasingly, multi-modal corpora are appearing (Adolphs and Carter 2013; Ishikawa 2019). Such corpora are particularly useful when analysing speech and sign languages (see also Chapter 7, this volume).

We can use a corpus to uncover patterns of usage and to test our intuitions about how language is used by particular groups of users, in particular texts (Jones and Waller 2015). It can tell us which form or forms occur with greater frequency, how words are used together in discourse and allow us to compare this usage across different types of data (commonly, spoken vs. written), historical periods and genres. What a corpus cannot do is tell us why certain forms are used, why speakers tend to use one item with or before another or why a certain form is rarely used. Such findings must come from our own interpretations of the data. As long as we keep this in mind, even a basic search

 DOI: 10.4324/9780367076399-10

of a corpus can reveal something to us and will often provide a useful starting point for further investigations.

What am I trying to find out?

Before we start any search, it will help if the purpose is clear. This will influence how we search, what corpora we search and what kinds of analysis we undertake. Jones and Waller (2015: 158) describe a principled research cycle as one which starts from our beliefs and intuition about a language area, moves on to examining the literature to check what others have told us about this, sets clear research questions, decides on which corpora to search and then how best to analyse them. To give an example, a search can start from a question which an English as a second language (ESL) student might ask a teacher, such as "Do people say X?" We can give an intuitive answer and then check in a corpus. In this case, our "research question" is "how frequently do people say x?" As the student uses the word "say", we will want to look in a spoken corpus and preferably one which contains the kind of speech they wish to use. It will then be helpful to make comparisons between the frequency of this item in different corpora, perhaps comparing its frequency with written corpora. Finally, we may wish to undertake qualitative analysis of concordance lines to explore how speakers use this item in context. Carter (2016) is a good example regarding how quantitative and qualitative exploration can be combined to answer questions about how language is used; in this case to analyse creativity in everyday conversations.

What are the limitations of my analysis?

It is important to be aware of the limitations of any analysis. Any search will be limited by the design and size of a corpus. A spoken corpus, for example, might only contain examples of monologic talk, which will mean we cannot use it to explore how items are used in interactive dialogue. Also, there can sometimes be a tendency to think that the larger the corpus, the more authoritative it is. This is not automatically true, and even a corpus of many millions of words is still only a partial view of the language (Cook 1998). Instead, it is better to consider whether the size of the corpus allows you to make observations about language in use which match the questions you are seeking answers for.

The sections in this chapter show some basics of analysing corpora, looking at both quantitative frequency analysis and qualitative analysis of concordance lines and extended texts, often in combination. In each section, I have tried to show how we can work with open-access corpora or use freely available analysis tools to work with our own data. I have also tried to show that analysing corpus data does not need to be limited to searching for single words. For obvious reasons of space and brevity, the examples given are not accompanied by extensive explanations of the research project they could be part of and are simply included to be illustrative.

2 Exploring frequency

One of the simplest and most common means of analysing a corpus is to look at frequency. The most basic search which can be undertaken is to check on the frequency of a word or longer sequence. All freely available corpora will enable you to do this via the "search" function, and this can give you the frequency of items in one corpus or when

Table 10.1 The frequency of *I mean* per million words in informal TV shows across time using the TV corpus

1950s	*1960s*	*1970s*	*1980s*	*1990s*	*2000s*	*2010s*
339.34	379.66	701.38	756.35	713.81	761.18	803.66

comparing two corpora or sub-corpora (see Chapter 9, this volume). Increasingly, comparisons can also be made across different times to track the historical change of an item, if the corpus is large enough to do so. This is achieved by simply searching for the item you wish to check and typing it into the "search" box. The example in Table 10.1 (see later) shows the frequency of *I mean* per million words across the TV corpus (Davies 2019), a corpus of 325 million words of informal TV shows from the 1950s to the 2010s from the UK and Ireland, the United States, Australia and New Zealand. These data are produced using the "chart" button available with this corpus.

This shows a steady increase across time so that there are now more than double the number of occurrences than in the 1950s in these data. A finding such as this does not tell us a great deal about the language, but it allows us to say, as suggested by analyses of other spoken corpora (e.g. O'Keeffe *et al.* 2007), that *I mean* is very frequent in much spoken language, particularly the language of informal conversations. A simple search also enables us to compare frequency across sub-corpora. We can do this by clicking "sections" and choosing which two sub-corpora we wish to compare. For example, we may wish to check the frequency of *I mean* in the UK/Ireland shows in this corpus compared to US shows in this corpus. Selecting both tells us that the item is somewhat more frequent in the American shows (US = 822.01 per million words, UK/Ireland = 508.1 per million words). This basic information can provide a starting point for further analysis. We could, for example, look at some of the data in context and try to work out if there are any differences in the way this item functions in the US and UK/Ireland data.

Commonly, frequency is also analysed by means of producing a frequency list, which can tell us the words or larger sequences which occur most often in any given corpus. Many corpora will now produce frequency lists, and in some cases, (e.g. Spoken BNC2014, Love *et al.* 2017) several different types are available, including all word forms or only lemmas (a base form such as "run" from which all inflections such as "runs" can be derived). A list based on all word forms will provide separate frequency counts for items such as "runs", "running" and "run", while one for lemmas would give one count only for "run".

We can also use corpus analysis software to produce frequency lists from our own data. Table 10.2 gives an example of three such frequency lists, two derived from the Spoken BNC2014 (11 million words of informal conversations of British English) and one produced using data from the UCLan Speaking Test Corpus (USTC) (Jones *et al.* 2018: 46), a corpus of 91,000 words of successful (at pass level) speaking tests at Common European Framework of Referenced for Languages (CEFR) B1–C1 levels. It shows the top ten most frequent items in each corpus. The data in this example are taken from the B1 section of the USTC corpus, and the figures show the total number of occurrences in each corpus.

Table 10.2 Sample frequency lists

N	*Spoken BNC2014 (all word forms)*	*Spoken BNC2014 (lemmas)*	*USTC B1 sub-corpus (all word forms)*
1.	I (435,613)	Be (743,461)	Er (1,391)
2.	It (354,259)	I (467,272)	I (1,126)
3.	You (311,450)	It (354,267)	The (688)
4.	The (297,665)	You (311,515)	And (523)
5.	's (292,876)	The (297,666)	To (461)
6.	And (275,495)	And (275,496)	Erm (320)
7.	Yeah (260,026)	Yeah (260,026)	In (296)
8.	That (230,781)	Not (240,180)	Is (288)
9.	A (215,473)	Do (236,550)	So (288)
10.	? (213,437)	A (232,816)	A (269)

As mentioned previously, such frequency lists on their own tell us little about the language used in any given corpus, and there is obviously a need to interpret them ourselves. It is also important to remember that most frequency lists will contain a high proportion of grammatical items (such as the determiner *the*) as their most common item. Frequency lists will also be heavily influenced by the type of data in any corpus. This results here in common items, such as the contracted *s,* which reflect the general nature of unrehearsed speech. The register of each corpus also has a large influence on the most frequent item in such lists (Biber 2012). Here, the Spoken BNC2014 features conversations between native speakers on topics they have chosen, while the USTC features spoken test data. Alongside the fact that these are second-language speakers at a pre-intermediate level, the frequency of items such as *er* and *so* in the learner data reflects the register of these tests. Learners do not choose the topics, have to speak in real time and will be aware that this is a test where they may need to give extended answers and reasons. All of these factors may account for the frequent use of items such as *er* and *so* in these data.

Despite the earlier caveats, as with basic frequency searches, lists such as these give useful initial indications about corpus data and provide starting points which we can then follow up on. The frequency lists provided earlier, for example, seem to indicate some of the typical linguistic features which characterise the interactive nature of conversations: the need for speakers to express viewpoints, to address each other (*I, you)*, to respond to each other (*yeah*) and to link turns and ideas together (*it, that*). We can choose areas such as these for further investigation. A closer look at the USTC data, for example (Jones *et al.* 2018: 46), shows how the use of *it's* is the eighteenth most frequent word in this corpus at the B1 level (152 occurrences). This significantly increases as levels progress from B1 to C1 (at the C1 level there are 276 occurrences), as learners become better able to link the ideas not only within their own turn but also across turns, adding to what other speakers are saying and referring back to things said by others or themselves. Jones *et al.* (2018: 119) give the following as an example of this. The students are discussing tourism as part of an interactive task.

(1)
<$2 M > I personally think the same about the location the hotel is really important.

<$1 F > Yeah.
<$2 M > I said if **it's** near to the attraction point then that's good but if **it's** far away then you know it will be costly to travel up.
<$1 F > Yeah spend more

[USTC]

Normalisation

As suggested, raw frequency is a fairly basic measure, and it is also important to note that an item may display more frequency simply because the size of the corpus is bigger. *The*, as an example, will obviously occur more times in a larger corpus than in a smaller one. For this reason, we also need to consider normalisation, which simply means how many times a word or larger pattern occurs per million words, or per thousand words in smaller corpora. Many corpora will now produce these data for you (see, for example, the data in Table 10.1), but if not, you can calculate this figure for yourself.

To take a simple example, we may wish to look at the use of the pronoun *I* from the frequency lists in Table 10.2. Such comparison is simple in the Corpus of Contemporary American English (COCA) (Davies 2008), a written and spoken corpus totalling more than a billion words. When we look at the "chart" view in search, we can see the overall frequency is a little higher overall in the spoken data (2,283,884 occurrences in 126.1 million words) compared to the fiction sub-corpus (2,191,345 occurrences in 118.3 million words), but the occurrences per million words are a little higher in the fiction corpus (18,106.58 spoken vs. 18,520.17 fiction). This simply shows us that looking only at raw frequency is not always the most helpful form of analysis.

You may also wish to calculate this with your own data and make comparisons with other smaller corpora. To produce such a figure in your own data, Evison (2010: 126) gives a simple formula. Divide the total amount of occurrences by the number of tokens in the corpus and multiply by 1,000 to get occurrences per 1,000, or 10,000 or 1 million to obtain those figures. Your choice of which calculation to use will be dictated by the size of your corpus, and it is important to be consistent in your calculations so that you compare data from different corpora in the same way. With smaller corpora, using a per 1,000 figure is generally considered to be most useful. For example, in the USTC data (Jones *et al.* 2018) at the C1 level, *yeah* occurs 442 times in this sub-corpus of 23,083 tokens (this excludes the examiner utterances), giving a normalised frequency of 19.14 occurrences per 1,000 words, while at the B1 level it occurs 250 times in 17,171 tokens (excluding the examiner utterances), giving a normalised frequency of 14.55 per 1,000 words. This shows us that the use of *yeah* increases as learners move from the B1 to C1 level, and as Jones *et al.* (2018: 127) note, is one linguistic marker which shows the increased ability to interact across turns and to co-construct conversations, which increases as levels progress.

3 Exploring keyness

Another way we can compare frequency is by looking at keyness. This allows us to look at the frequency of items in one corpus in comparison to another, normallya larger, reference corpus, and check which words occur significantly more (positive keywords) or significantly less (negative keywords) in our corpus compared to a reference corpus. The

Table 10.3 Positive keywords in USTC (B2) compared with LINDSEI as a reference corpus

N	*Keyword*	*Freq.*	*%*	*Texts*	*RC. Freq.*	*RC. %*	*Keyness*	*P-value*
1	Er	1,312	6.16	17	29,325	3.02	536.42.	p<.001
2	Culture	88	0.41	6	174	0.02	316.25	p<.001
3	Agree	55	0.26	15	61		243.59	p<.001
4	Can	204	0.96	17	2,224	0.28	204.39	p<.001
5	Technology	31	0.15	4	8		186.75	p<.001
6	You	518	2.43	17	9,842	1.24	185.92	p<.001
7	Will	137	0.64	16	1,165	0.15	184.46	p<.001
8	Preston	24	0.11	10	0		174.89	p<.001
9	UK	19	0.09	10	0		138.45	p<.001
10	Weather	42	0.2	5	101	0.01	138.28	p<.001

RC = reference corpus; % = the percentage of the total corpus which this item's total occurrences represent.

corpora must be comparable to make keyness searches worthwhile, and it is important to be clear about what we are comparing and why. For example, if we compare the USTC data with another corpus of spoken learner data, we need to ensure it contains data similar to interactive spoken exams and not monologues.

Many open-access corpora now have keyness built in as a feature. For example, the CLiC corpus (Mahlberg *et al.* 2016) features the works of Dickens and allows us to compare these with a mixed reference corpus of nineteenth-century novels using the "keywords" function. We can also calculate this with our own data. The example in Table 10.3, taken from Jones *et al.* (2018: 56), shows the positive keywords in comparison between the USTC learner spoken test data (B2 level) and the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin *et al.* 2010) corpus, which has over a million words of learner speech from learner of different L1s, levels and ages, and used interview data which had clear similarities to the USTC data. For example, there were discussion tasks in both datasets. To undertake the analysis, a wordlist was made from learner turns in the LINDSEI data with the Wordlist function in WordSmith (Scott 2015) and then the keyword function was used to make comparisons with our data. Other freely available software will also calculate key words in the same way by comparing word lists (see Chapter 9, this volume).

The keyness score here is based on log likelihood, which measures the statistical significance of keywords. Significance measures help us to be sure results are important in terms of keyness, and commonly anything below 0.05 ($p < 0.05$) is considered significant; that is, there is a less than 5 per cent chance that the results are due to chance. The p-values displayed in Table 10.3 simply show that each keyword is significantly more frequent in the USTC corpus in comparison to LINDSEI.

Although some of the words here (such as *UK* and *weather*) clearly reflect the topics discussed in the USTC data, others such as *can* and *will* are of interest to us as researchers and are something we could follow up on with further study. We could examine uses of *can*, for example, by looking at this in context and deciding how form(s) and functions compare with those in the LINDSEI data (see Jones *et al.* 2018, chapter 2 for an extended discussion of keywords in USTC).

Quick and simple comparisons to determine more or less frequent use of an item in comparisons between one corpus and another can also be made using an online log-likelihood calculator (Log-Likelihood and Effect Size Calculator 2020; Rayson and Garside 2000). This calculator allows you to input the frequency of words or phrases in corpora of different sizes and make a comparison using log likelihood as a measure, determining whether each item is significantly more (+) or significantly less (-) frequent in one corpus compared with another. The calculator produces a score, and the online site provides a detailed breakdown of how different scores relate to different levels of significance. For example, if the score is 6.63, it will be significant at the level $p < 0.05$ (see also Rayson and Garside 2000 and the calculator website for more details).

An example of such a calculation can be made by looking at *think* from the Spoken BNC2014 and comparing its frequency with the spoken section of COCA (Davies 2008), which consists of 126 million words of US TV and radio show data of an unscripted nature. When inputting these data, you need the raw frequency (not the normalised frequency) of the item plus the number of tokens in each corpus. Checking this via the calculator shows that *think* is used significantly more (+4040.81, $p < 0.0001$) in the Spoken BNC2014 and indicates that it may be an item worthy of further investigation. We may wish to look at this item in context and see what forms and functions exist with it (which other words or phrases are used in conjunction with it, for example) and then try to understand what this can tell us about each set of data. One obvious comparison here might be differences in usage between UK and US English and why these might occur in each corpus.

4 Exploring larger patterns

Analysing keywords and wordlists may suggest that we can only use frequency measures to look at single words. However, in most data, it is common to look at larger patterns of language use. One simple way to do this is to search data for collocates of frequent words or keywords, an option which most open-access corpora and all commonly available software will give us. We can do this in a corpus such as the 100-million-word American soap opera corpus (Davies 2011) by clicking the "collocates" section, then inputting our search term and then choosing the span of collocates (the number of words to the left or right) which we are checking collocate with the node word (the target word). Commonly, researchers wish to look at collocates which occur up to four words before or after the target word, but it is perfectly possible to simply check what comes directly after or before a target word. In this corpus (and associated corpora available from the same site), if we click "relevance" below the search bar, we can produce a mutual information score. This shows us how high the chance of co-occurrence is in any one corpus, and the higher the score is from zero, the higher chance there is of co-occurrence, suggesting a strong association (McEnery and Wilson 2001; Oakes 2004). Note that this score does not necessarily show us the words which most frequently come before or after a target item, but those which are most likely to co-occur. To give an example, we can search for a word such as *paper* in the American soap opera corpus and look for collocates occurring up to four words after this. The data show us that the top five items (in terms of MI scores) are *clips* (11.93), *shredder* (10.67), *airplanes* (10.42), *scissors* (10.13) and *trail* (10.02), which all differ in frequency (*clips* = 24, *shredder* = 7, *airplanes* = 11, *scissors* = 38, *trail* = 112). This tells us that in these data, we are most likely to find *paper* in co-occurrence with *clips*. Such searches are useful, as they start to

Table 10.4 Four-word n-grams in the CLiC Dickens corpus

N	*Four-word n-grams*	*Frequency*
1	As if he were	405
2	As if he had	266
3	At the same time	255
4	In the course of	215
5	In the midst of	206
6	As if it were	204
7	I beg your pardon	200
8	I don't know what	200
9	On the part of	200
10	What do you mean	199
11	With an air of	197
12	It would have been	194
13	Said the old man	194
14	In a state of	192
15	For the first time	190

show how words are patterned together and can give us insights to usage in a particular corpus, which we can then compare to usage in other corpora. For more about collocates, see Chapters 9, 14 and 15, this volume.

We can also look at larger patterns of co-occurrence by searching for what are termed "n-grams", also commonly referred to as clusters, chunks or formulaic sequences in the literature, albeit with slight differences in the definitions (see Chapter 15, this volume). N-grams have been defined as 'two, three, four (or more) sequence[s] of words that combine in data' (Jones and Waller 2015: 194). Online corpora and corpus tools allow us to specify the length of n-gram we wish to search. Although longer sequences are possible, there is a large drop-off in frequency after four-word sequences (O'Keeffe *et al.* 2007), and so most searches will focus on items from two to four words in length. A search in the CLiC fiction corpus (Mahlberg *et al.* 2016) under "clusters", for example, produces the list of four-word n-grams from the whole Dickens corpus displayed in Table 10.4.

Such n-grams can be a useful basis for further investigation of corpora. We could, for example, simply click on an n-gram to see how it is used in context in sets of concordance lines and how it contributes to meaning in extended text. Extract 2 shows some samples of *with an air of* from the CliC corpus, looking at the texts from Dickens. These are displayed in a list here, but in the actual corpus, the keyword in context (in this case "with an air of") is displayed in the middle of the concordance line and will be highlighted.

(2)

1. carrying a reticule came curtsying and smiling up to us **with an air of** great ceremony. "Oh!" said she. "The wards in Jarndyce!
2. Yes." He folded his arms and sat looking at me **with an air of** the profoundest astonishment

3. friend," pursued Miss Flite, advancing her lips to my ear **with an air of** equal patronage and mystery, "I must tell you a secret
4. one knee, and gently smoothed the calves of his legs, **with an air of** humble admiration. 'That I had but eyes!' he cried
5. you see.' 'What is this!' said Gashford, turning it over **with an air of** perfectly natural surprise. 'Where did you get it from
6. replied Barnaby, finishing his task, and putting his hat on **with an air of** pride. 'I shall be there directly.' 'Say "my lord,"
7. the more impressive, Mr. Micawber drank a glass of punch **with an air of** great enjoyment
8. 'Trotwood,' said Mr. Dick, **with an air of** mystery, after imparting this confidence

(CliC [online] 2020)

We can see that in terms of form, it tends to be followed by abstract nouns such as *mystery*. In terms of function, it serves to tell us about how certain characters behave, allowing the author to build a picture of their character. It is therefore a device by which the author can speak to us about characters and help to formulate our view (the view the authors hope we will take) of them. A more detailed picture of this n-gram could be developed by looking at how it relates to particular characters in particular Dickens books. It is possible to use the "search for types" function in CLiC to produce all examples of *pride,* for example, and then look at which texts they are used in and by which characters. Such investigations allow us to interrogate the data in some depth, moving from the general to the specific and away from single words.

We can also look at n-grams within our own data and produce frequency lists, which we can normalise and also explore for keyness in some of the ways described in previous sections. In the freely available corpus analysis tool Antconc (Anthony 2019), this function is under the "clusters/n-grams" tab, and as with the CLiC corpus, you can specify the length of n-gram you wish to search for. Table 10.5 shows the top 10 four-word n-grams from the USTC data, as described in Jones et al. (2018: 67), at the B2 level. Note that in this data, a contracted form is counted as two words

When we view n-grams in this way, it gives a clearer indication about the prefabricated nature of much language use and how the most frequent items of language are often formulaic in nature, something which has been described for many years in corpus research (e.g. Sinclair 1991). As mentioned in earlier discussion of word frequency, such data also reflect the specific register of the corpus. In these USTC data, learners are answering and discussing questions and tasks set by an examiner. The nature of such questions and tasks often requires learners to express their own view and the views of their fellow test takers. This is one reason why there are several sequences with *think.* It is also notable that there are no examples of sequences many learners will be taught to express a viewpoint around words such as *opinion* or *view*. Jones *et al.* (2018) suggest that this is because learners in these data favoured items which are multi-functional, and *I think it's* can be used both to buy time and express an opinion, for example.

It is also worth noting that some n-grams may seem fragmentary (such as *They don't have* in Table 10.5) and need further investigation and interpretation in order to understand how the language is being used. These fragments can be part of larger, more meaningful frames such as e.g. *They don't have + noun* and can serve important functions in discourse.

Table 10.5 Four-word n-grams from USTC at the B2 level

N	*Four-word n-grams*	*Frequency*
1.	I agree with you	21
2.	I think it's	20
3.	A lot of time	19
4.	What do you think	16
5.	Spend a lot of	15
6.	I don't like	12
7.	I don't know	10
8.	They don't have	10
9.	Do you think about	9
10.	I don't think	9

Lexico-grammar

Investigating larger patterns of language beyond the single word can also enables us to explore lexico-grammar. Halliday and Matthiessen (2004: 45) define this term as 'patterns which lie somewhere between structures and collocations having some of the properties of both'. They give the example of *take pride/delight + in + -ing,* where we can see the collocates of *take* are also commonly associated with *in* and *ing*. In other words, words go together with other words, but are also likely to be found in the company of certain grammatical patterns. Many modern corpora make it relatively simple to search for these patterns (see Chapter 11, this volume, for more on patterns). We can start from an example such as *I think* and search for what most commonly comes after it in the COCA spoken corpus (Davies 2008), which consists of unscripted TV and radio shows. Using a wildcard search (search for *I think* *), we can see that one very common item which comes directly after it is *there's*. If we then search for *I think there's* * we see the most common pattern is *I think there's a* and what follows this is most often *lot*. We can then search the data further using the parts of speech (POS) next to the search bar or further wildcard searches to find out the most common patterns. In this case, the most common pattern is *a lot + of + noun phrase* (e.g. *people/blame/truth*). This allows us to build a picture of the lexico-grammar of *I think* in these data. We can see from this simple analysis that in this spoken section of the COCA corpus, *I think* is often patterned in a particular way, which we could describe as *I think + there's + a lot of + noun phrase* and we can also list the most frequent noun phrases. We might also wish to check the frequency of *there's* with countable nouns such as *people* in contrast to the use of *I think there are + countable noun phrase.* Viewing concordance lines also allows us to investigate how this pattern functions. Some sample concordance lines are shown in extract 3.

(3)

1. **I think there's a lot of people** that are pulling for her. Mr-JACKSON: Yeah. STORM: Yeah
2. .Well **I think there's a lot of people** who are committing on a plan that they haven't read.

3. my ideals or some of my views. But **I think there's a lot of people** similar to me who maybe don't get labeled the same way
4. like in "The Wizard of Oz," and **I think there's a lot of people** now stepping back and saying, "Oh, my God.
5. **I think there's a lot of people** that would love to attack us in the United States.
6. Well, I think there is there -- **I think there's a lot of people** who believe it was a suicide, but there's still a lot
7. **I think there's a lot of people** that have ideas against the government or against whatever philosophies or groups,
8. **I think there's a lot of - I think there's a lot of people** over there that's been around too long. But I mean…

[COCA spoken sub-corpus]

One obvious function here (due to the nature of the discourse) is to use this sequence to refer back to the topic of interest and add support to your view. Saying *a lot of people* adds authority to an opinion and, of course, means you do not need to say exactly who those people are and therefore your viewpoint is harder to challenge! It is also clear that such a sequence contributes to the cohesion of this discourse, allowing speakers to link ideas within their own turns in relation to the topic under discussion.

5 Exploring language in context

Every time we look in a corpus, we are, of course, exploring language within a context or contexts. However, it is also possible to use a corpus to explore how language forms and functions are used in extended discourse. One way we can do this is by comparing language used in different corpora to explore how lexical items are "primed" (Hoey 2005) in terms of forms and function in different types of discourse. A definition of lexical priming is as follows:

> the theory of lexical priming suggests that each time a word or phrase is heard or read, it occurs along with other words (its collocates). This leads you to expect it to appear in a similar context or with the same grammar in the future, and this "priming" influences the way you use the word or phrase in your own speech and writing.
>
> *(Macmillan Online Dictionary 2020)*

Hoey (2005) further suggests that we are also primed to expect items to occur in particular positions in a sentence or spoken turn and with particular meanings and that these will vary according to the genre of the texts we are looking at. Jones and Waller (2015:31) give the example of *married,* which they show tends to colligate (go together with a particular grammatical form) with *is* as part of a non-defining relative clause in newspaper texts and functions to add information when people are being described in news reports. For example, 'The woman, who is married and lives in Sefton, was attacked when…'. They contrast this with its use in a spoken corpus (from COCA) where it is primed to co-occur with *get* and colligate with *going to* and allows TV show guests (particularly on chat shows) to describe future or future in the past plans such as *I was going to get married.*

Table 10.6 The priming of *divorced* in academic and fictional texts

	Academic corpus	*Fictional corpus*
Words which most frequently follow *divorced*	*from* (294, 2.5), *or* (37, 0.3), *families* (12, 0.1)	*him* (67, 0.6) *when* (40, 0.3) *from* (40, 0.3).
Common patterns	X+ is divorced from reality	She divorced him / …and divorced him
Common meaning/use	Used to form part of a discursive argument Figurative meaning.	Used to form part of a narrative description. Literal meaning.
Examples	*Yet the mantra is not entirely divorced from reality*	*A few years later, when his cocaine habit had bankrupted them, she divorced him.*
Position	End of clause	Complete clause, often following the conjunction *and* at the end of sentences.

Searching the COCA corpus and starting from a word or sequence allows us to investigate language in this way. To use a similar example to *married*, we may be interested in how an item such as *divorced* is used in contrasting corpora. We can investigate this by searching for items to get an overall picture of how they are used by using the "word" function in this corpus. This gives us collocates, clusters, texts, topics associated with a word and concordance lines at one click. In this case, common collocates listed are *parent* and *get,* while *divorced mother of two* and *parents are getting divorced* are common clusters. We can then explore these patterns by searching for them in contrasting corpora and checking how they are primed in different texts. In the search section of COCA, this can be achieved by a search for *divorced* and by simply checking what comes after it with a wildcard search (*divorced* *). If we also compare usage in different datasets (in this case the academic and fiction sub-corpora), we can then examine concordance lines to see how the item is primed in these different texts. Some summative results are displayed for *divorced* in Table 10.6. The academic corpus in this case is 119.8 million words of academic texts and the fiction corpus of 118.3 million words. Frequency figures are displayed here overall in total and then per million words in brackets.

This kind of description shows how the item *divorced* is primed in different ways within different corpora and shows how we can use a corpus to explore language in context. This helps to build descriptions that take account of how different genres of texts within corpora will affect the usage and meaning of particular items. Using corpora in this way can help us to move beyond basic frequency searches and to combine both qualitative and quantitative analysis of language in context.

Further reading

Adolphs, S. and Carter, R. A. (2013) *Spoken Corpus Linguistics: From Monomodal to Multimodal*, London: Routledge. (This provides a useful description and analysis, showing some of the possibilities available when designing and working with spoken corpora.)

Collins, L.C. (2019) *Corpus Linguistics for Online Communication: A Guide for Research*, London: Routledge. (This is a useful and highly practical introduction to using corpora to investigate forms of online communication such as the use of social media.)

Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*, London: Routledge. (This is an influential book, which shows how corpora can be used to investigate and further our understanding of language in use.)

Timmis, I. (2015) *Corpus Linguistics for ELT: Research and Practice*, London: Routledge. (Aimed at those teaching English as a second or foreign language, this is a useful and practical introduction to corpus linguistics which will benefit anybody interested in working with corpora.)

References

Adolphs, S. and Carter, R. A. (2013) *Spoken Corpus Linguistics: From Monomodal to Multimodal*, London: Routledge.

Anthony, L. (2019) *AntConc (Version 3.5.8) [Computer Software]*, Available from: https://www.laurenceanthony.net/software/antconc/ [Accessed 15 February 2020].

Biber, D. (2012) 'Register as a Prediction of Linguistics Variation', *Corpus Linguistics and Linguistics Theory* 3(2): 9–37.

Carter, R. A. (2016) *Language and Creativity: The Art of Common Talk*, 2nd edn, London: Routledge.

CLiC [Online] [1 February 2020], Available from click: http:\\lic.bham.ac.uk

Cook, G. (1998) 'The Uses of Reality: A Reply to Ronald Carter', *ELT Journal* 52(1): 57–63.

Davies, M. (2008-) *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present*. [Online], [6 February 2020], Available from: https://www.english-corpora.org/coca/.

Davies, M. (2011-) *Corpus of American Soap Operas: 100 million words*. [Online], [20 February 2020]. Available from: https://www.english-corpora.org/soap/.

Davies, M. (2019-) *The TV Corpus: 325 million words, 1950-2018*. [Online], [20 February 2020] Available from: https://www.english-corpora.org/tv/.

Evison, J. M. (2010) 'What Are the Basics of Analysing a Corpus?', in A. O'Keeffe and M. J. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge, pp. 122–35.

Gilquin, G., De Cock, S. and Granger, S. (2010) *LINDSEI: Louvain International Database of Spoken English Interlanguage*. [CD-ROM], Louvian: Presses Universitaires de Louvain.

Halliday, M. A. K. and Matthiessen, C. (2004) *An Introduction to Functional Grammar*, 3rd, London: Routledge.

Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*, London: Routledge.

Ishikawa, S. (2019) 'The ICNALE Spoken Dialogue: A New Dataset for the Study of Asian Learners' Performance in L2 English Interviews', *English Teaching* (The Korea Association of Teachers of English) 74(4): 153–77.

Jones, C. and Waller, D. (2015) *Corpus Linguistics for Grammar: A Guide for Research*, London: Routledge.

Jones, C., Byrne, S. and Halenko, N. (2018) *Successful Spoken English: Findings from Learner Corpora*, London: Routledge.

Log-Likelihood and Effect Size Calculator (2020) [Online], [2 February 2020]. Available from: http://ucrel.lancs.ac.uk/llwizard.html.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017) 'The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations', *International Journal of Corpus Linguistics* 22(3): 319–44.

Macmillan Online Dictionary (2020) *Definition of lexical priming*. [Online], [5 February 2020]. Available from: https://www.macmillandictionary.com/dictionary/british/lexical-priming.

Mahlberg, M., Stockwell, P., de Joode, J., Smith, C. and O'Donnell, M. B. (2016) 'CLiC Dickens: Novel Uses of Concordances for the Integration of Corpus Stylistics and Cognitive Poetics', *Corpora* 11(3): 433–63.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics*, 2nd edn, Edinburgh: Edinburgh University Press.
Oakes, M. P. (2004) *Statistics for Corpus Linguistics*, Edinburgh: Edinburgh University Press.
O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) *From Corpus to Classroom: Language Use and Language Teaching*, Cambridge: Cambridge University Press.
Rayson, P. and Garside, R. (2000) 'Comparing Corpora Using Frequency Profiling', in *Proceedings of the workshop on Comparing Corpora held in conjunction with the 38th annual meeting of the Association for Computational Linguistics 1-8 October 2000*, Hong Kong, pp. 1–6.
Scott, M. (2015) *WordSmith Tools. v.6.0.0.252. [Online]*, Stroud: Lexical Analysis Software Ltd, Available from: http://lexically.net/wordsmith/ [Accessed 29 December 2015].
Sinclair, J. (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.