5
Building small specialised corpora

Almut Koester

1 What's the point of a small corpus?

Over the years there have been two opposing trends in the compilation of corpora. On the one hand, corpora are getting ever larger, with "mega-corpora", such as the Bank of English, the Cambridge International Corpus (CIC) and the Corpus of Contemporary American English (COCA) having millions or billions of words. On the other hand, smaller, more specialised corpora are being compiled, focusing on specific registers and genres.

But what is the point of a small corpus? Surely, the point of a computer-based corpus is to allow the electronic storage and machine analysis of huge amounts of text which could not be handled manually. According to John Sinclair (2004), the "father" of corpus linguistics, 'small is not beautiful; it is simply a limitation' (p. 189). While he concedes that it may be possible to get valid results from a small corpus, he argues that these results will be limited. In a large corpus, on the other hand, 'underlying regularities have a better chance of showing through the superficial variations' (ibid.:189). To illustrate this, Sinclair looked for the phrase *fit into place* in a 2-million, 20-million and 200-million-word corpus and did not find any examples until he searched the largest 200-million-word corpus, and even then only half a dozen.

This anecdote illustrates the fact that small corpora are not suitable for certain types of analysis, particularly lexis and phraseology, but others have argued that a smaller corpus may be perfectly adequate for some purposes. Lexical items, except for the most common words, are relatively infrequent, and therefore a large corpus is necessary to carry out lexicographical research (see Chapter 27, this volume). However, grammatical items, such as pronouns, prepositions and auxiliary and modal verbs, are very frequent and can therefore be reliably studied using a relatively small corpus (Carter and McCarthy 1995). There may even be some disadvantages to working with a very large corpus. The sheer volume of data for high-frequency items may become unmanageable and result in analysts having to work with a smaller sub-sample (ibid.: 143). In a small corpus, on the other hand, *all* occurrences, and not just a random sample of high-frequency items, can be examined. Furthermore, in working with very large corpora,

DOI: 10.4324/9780367076399-5

where the samples examined come from many vastly different contexts, it is difficult, if not impossible, to say anything about the original context of use of the utterances (Flowerdew 2004). See also Sections 2 and 3 later and Chapter 4, this volume.

Smaller, more specialised corpora have a distinct advantage when it comes to contextual information: They allow a much closer link between the corpus and the contexts in which the texts in the corpus were produced. Where large corpora, through their de-contextualisation, give insights into lexico-grammatical patterns in the language as a whole, smaller, specialised corpora give insights into patterns of language use in particular settings. With a small corpus, the corpus compiler is often also the analyst and therefore usually has a high degree of familiarity with the context. This means the quantitative findings revealed by corpus analysis can be balanced and complemented with qualitative results (Flowerdew 2004). As we shall see, specialised corpora are also usually carefully targeted and set up to reflect contextual features, such as information about the setting, the participants and the purpose of communication. Therefore, analysis of such corpora can reveal connections between linguistic patterning and contexts of use.

This link between the corpus and the contexts of use is particularly relevant in the fields of English for Specific Purposes (ESP) and English for Academic Purposes (EAP), where small, specialised corpora have been compiled to inform pedagogy (see Chapter 28, this volume). Tribble (2002) argues that large corpora do not meet the needs of teachers and learners in ESP/EAP, as they provide 'either too much data across too large a spectrum, or too little focused data, to be directly helpful to learners with specific learning purposes' (p. 132). Smaller, more focused corpora, which have been set up for a specific research or pedagogical purpose, are much more likely to yield insights that are directly relevant for teaching and learning for specific purposes (ibid.).

Furthermore, from a practical point of view, any corpus an individual researcher or practitioner, such as an ESP/EAP teacher, can construct will necessarily be small, due to the limitation of collecting, and for a spoken corpus, transcribing the data. The aim of this chapter is therefore to provide some guidelines for building a small, specialised corpus and to discuss, with concrete examples, what can be learnt from such a corpus.

2 How small and how specialised?

But just how small and how specialised can a corpus be? The answer to this question depends crucially on what the corpus will be used for, that is the purpose of the research. But let's first define what we mean by a "small" corpus. Opinions diverge regarding what is considered "large" or "small" when it comes to corpora. First, it depends on whether the corpus is written or spoken. As it takes a long time to compile a spoken corpus (see Chapter 3, this volume), spoken corpora tend to be smaller than written ones. According to O'Keeffe *et al.* (2007: 4), spoken corpora containing over a million words of speech are considered large, whereas with written corpora, anything under 5 million words of text is quite small. But many small corpora, even written ones, are a great deal smaller than that, and Flowerdew (2004: 19) notes that there is general agreement that small corpora contain up to 250,000 words.

As already noted, when analysing high-frequency items, a relatively modest corpus may still yield robust and powerful findings, for example, the European Corpus of Academic Talk (EuroCoAT), a corpus of office hour consultations with just under 60,000 words (MacArthur *et al.* 2014), and Koester's (2006, 2010) 34,000-word Corpus of American and British Office Talk (ABOT). What is more important than the actual size of the corpus is how well it is designed and that it is "representative". There is no ideal size for a corpus; it all depends on what the corpus contains and what is being investigated (Flowerdew 2004). Nevertheless, it is possible to give some general guidelines regarding minimal sample size. These issues will be discussed in Section 3.

With regard to the degree of specialisation, a corpus may be more or less specialised, and it may be specialised in different ways, depending, again, on the purpose of the research (see also Hunston 2002: 14). Flowerdew (2004: 21) lists a number of parameters according to which a corpus can be specialised:

- Specific purpose for compilation, e.g. to investigate a particular grammatical or lexical item;
- Contextualisation: particular setting, participants and communicative purpose;
- Genre, e.g. promotional (grant proposals, sales letters);
- Type of text/discourse, e.g. biology textbooks, casual conversation;
- Subject matter/topic, e.g. economics;
- Variety of English, e.g. Learner English.

Many specialised corpora have been compiled to study a particular language variety, for example, the Limerick Corpus of Irish English (Farr *et al.* 2004) or the International Corpus of Learner English (ICLE website). Since the early 2000s, interest in *English as a lingua franca* (ELF) has led to the compilation of a number of ELF corpora, including the Vienna-Oxford International Corpus (VOICE 2013), the spoken academic English as a lingua franca (ELFA 2008) corpus and the Written ELF in Academic Settings (WrELFA 2015) corpus. Other corpora focus on specific academic or professional genres; for example, the Michigan Corpus of Spoken Academic English (MICASE, see Simpson *et al.* 2002) and the British Academic Spoken English (BASE) corpus both consist of spoken academic genres, primarily lectures and seminars. Examples of professional genre corpora (see also Cheng 2014) are the Cambridge and Nottingham Business English Corpus (CANBEC, see Handford 2010), a corpus of business meetings, the Enron email corpus (Kessler 2010) and the NHS Feedback Corpus (NHSFC), consisting of online patient comments and provider responses (Baker *et al.* 2019). The Construction Industry Corpus (CONIC) is an example of a corpus that is specialised both in terms of professional genre and variety (ELF) (Handford and Koester 2019). Specialised corpora can vary considerably in size, ranging from those with fewer than 100,000, such as EuroCoAT with 58,834 words (MacArthur *et al.* 2014), to a corpus like NHSFC, which contains approximately 40 million words (Baker *et al.* 2019).

The degree of specialisation also varies, for example, corpora representing a language variety such as Irish English are quite general in terms of the genres they comprise (see Chapter 6, this volume). However, such corpora can be set up to include more specialised sub-corpora, for example, the Hong Kong Corpus of Spoken English (HKCSE) has four sub-corpora: conversation, business discourse, academic discourse and public discourse (Warren 2004). Many ESP/EAP corpora are very specialised indeed, as they have been compiled for specific research or pedagogical purposes, for example, the 250,000-word Corpus of Environmental Impact Assessment (EIA) consisting of 60 summary reports commissioned by the Hong Kong Environmental Protection Department (Flowerdew 2008). An example of a corpus designed for a specific pedagogical purpose is the Indianapolis Business Learner Corpus (IBLC), which consists of 200 letters of application (Connor *et al.* 1997; Upton and Connor 2001). The letters were written by business communication students from three different countries as part of an international business writing course (see Section 3). Specialised corpora like these are designed to provide insights into the genres investigated, such as specific types of scientific (e.g. environmental impact statements) or academic genres (e.g. letters of application). They will obviously not be useful for predicting language patterns in other registers and genres or, for example, for teaching English for general purposes.

While specialised corpora may vary in size, an important point is that such corpora do not need to be as large as more general corpora to yield reliable results. The reason for this is that as specialised corpora are carefully targeted, they are more likely to reliably represent a register or genre than general corpora. Even with relatively small amounts of data, 'specialised lexis and structures are likely to occur with more regular patterning and distribution' than in a large, general corpus (O'Keeffe *et al.* 2007: 198).

The next two sections will provide practical guidelines for building a small, specialised corpus, but see also Chapter 2 for general guidelines for corpus design, Chapter 3 for compiling spoken corpora and Chapter 4 for written corpora.

3 Important considerations in the design of a small, specialised corpus

As with any corpus, the most important consideration in designing a small, specialised corpus is that it should be representative (see Chapters 2, 3, 4 and 6, this volume, for more on representativeness). Biber (1993: 243) defines representativeness as 'the extent to which a sample includes the full range of variability in a population'. Biber (ibid.) identifies two types of variability: *situational* and *linguistic*. Situational variability refers to the range of registers and genres in the target "population", i.e. in the text types or speech situation to be included in the corpus. Linguistic variability refers to the range of linguistic distributions found in the population.

If a very specific type of genre is being investigated, then it may be straightforward to establish situational representativeness, as all the samples collected will accurately represent that genre. However, in most cases, there is some degree of variability even within a given genre, and it is therefore important to ensure that the corpus is "balanced" so that the corpus reflects the full range of variability found in the genre (see Chapter 6, this volume). For example, there may be different sub-genres, or perhaps the genre is used in different types of organisations or by different people. If all the samples come from just one organisation, then the corpus will be representative of the genre as used in that organisation, but not of the genre as a whole. Of course, the aim of the research may simply be to study the genre in that particular organisation, but generally the purpose of a corpus is to yield insights not only into itself but also into typical language use in the genre, register or variety from which it was taken (Tognini-Bonelli 2001: 53–4). The NHS Feedback Corpus (NHSFC), for example, comprises a complete set of genre exemplars within an organisation in a given time period, as it includes all patient comments and provider replies over two and a half years (Baker *et al.* 2019). While this genre set comes from just one organisation, the findings and implications are potentially far-reaching, as the NHS is such a large organisation with a unique position in the UK.

Good sampling is therefore essential for designing a representative corpus, but there are, of course, practical limitations to sampling. It will never be possible, particularly for a small corpus, to collect samples from *all* the situations in which a fairly widespread genre is used. What is important is to ensure that the samples are collected from a range of typical situations. For example, data for the ABOT Corpus, which was designed to investigate the most frequently occurring genres in spontaneous face-to-face office interactions, were collected from offices in a range of organisations and business sectors (Koester 2006). Only those genres which occurred across various office settings were selected for inclusion in the corpus, thereby ensuring that the corpus was not biased towards any one setting.

For specialised corpora, linguistic representativeness (at least at the lexical level) can be measured by the degree of "closure" or "saturation" (see McEnery and Wilson 2001: 148–67; McEnery *et al.* 2006: 15–16). A corpus is considered to be saturated when the addition of new data (i.e. word tokens) does not yield new lexical items (i.e. word types). Moreover, linguistic representativeness also depends on the number of words per text sample and number of samples per register or genre included in the corpus. According to Biber (1993), the most common linguistic features (e.g. personal pronouns, contractions, past and present tense and prepositions) are relatively stable in their occurrence across 1,000-word samples. To adequately represent a register or a genre in a corpus, Biber (1990) found that the linguistic tendencies are quite stable, with ten (and to some extent even five) text samples per genre or register.

Biber's studies indicate that it is not necessary to have millions of words or a huge number of texts in a corpus to get reliable results (at least for high-frequency items). But even these relatively modest criteria cannot always be met, for example, having text samples that contain at least 1,000 words. This is especially the case for a spoken corpus, as many spoken interactions (for example, service encounters) are relatively short and do not amount to 1,000 words; and even some written texts, especially in workplace contexts (e.g. emails), may contain fewer than 1,000 words. It is more important to collect complete texts or interactions, rather than artificially controlled samples of a certain length, to adequately represent the genre or text type (Flowerdew 2004). One can still try to ensure that any sub-corpus within the corpus (for example, a particular genre or subgenre) is represented by at least 1,000 words (even if these are spread across different texts or conversations) and that every sub-corpus contains at least five, if possible ten, different samples.

The ABOT Corpus will again be used to illustrate how a small corpus can be designed, as well as to indicate some of the problems and pitfalls. A 'corpus-driven' approach (Tognini-Bonelli 2001: 84) was used to establish the genres in the ABOT Corpus; this meant that it was not possible from the outset to gather a minimum number of exemplars per genre. Some genres, such as decision-making, were much more frequent than others, such as reporting, which meant that some sub-corpora contain more generic episodes or "texts" than others. About half of the sub-corpora contain between seven and eleven text samples, but others contain fewer than five. Clearly for those genres represented by fewer than five exemplars, the results of corpus analysis will be less reliable than for the genres with more text samples. But again, this reflects the reality of the target situation in that certain genres are typically more frequent in office interactions than others. To ensure that results from analysing the ABOT Corpus were reliable, comparisons were often made between "macro-genres" (similar genres grouped together), rather than between individual genres.

It was not possible to achieve lexical saturation or closure for the ABOT Corpus, as adding data from a different workplace setting would most likely have resulted in the addition of new lexical items. However, as the aim was to study generic features rather than lexical items, this limitation was acceptable. A related issue is that of 'local densities' (Moon 1998: 68): With genres that are under-represented, certain items may appear to be frequent in a genre simply because they occur frequently in one particular encounter. However, such local densities are usually easy to spot (e.g. if most examples of a lexical item come from one encounter), and this should be taken into account when interpreting the results.

Another challenge to representativity can arise with historical corpora, such as the Bolton/Worktown Corpus (BWC) compiled from manually recorded conversations of working-class people in Bolton (1937–40) totalling approximately 80,000 words (Timmis 2018). As Timmis points out, historical corpora can only include those documents that happen to have survived. To increase the representativeness of a corpus faced with the problem of data scarcity, researchers should try to include data from multiple sources in the corpus (ibid.).

These examples regarding the compilation of relatively small spoken corpora illustrate how the principles of corpus design interact with practical considerations relating to the nature of the data collected and how limitations regarding sampling can be dealt with. While every effort should be made to make the corpus as representative as possible, optimum representative sampling may not always be possible, particularly when compiling a spoken corpus, due to restrictions in relation to access and obtaining permissions, or simply limitations on the researcher-cum-corpus compiler's time (particularly as transcription is time consuming).

The most important consideration regarding corpus design is that the corpus should be set up in a way that is suitable for the purpose of the research. While many larger corpora were compiled for research into general linguistic phenomena, specialised corpora are often designed to answer specific research questions. For example, the aim of the IBLC was 'to study language use, accommodation across cultures, and genre acquisition of native and non-native speaking students in an undergraduate business communication class' (Upton and Connor 2001: 316). The data collected for the corpus consisted of letters of application written by business communication students in institutions in different countries as part of a writing project. The project involved students at each university reading and evaluating letters of students from other countries (see Connor *et al.* 1997). The corpus thus included data from both native and non-native students, had a cross-cultural element and involved a specific genre (letters of application) and could therefore be said to be well-designed to answer the research question.

4 Compiling and transcribing a small, specialised spoken corpus

Many of the limitations of a small corpus can be counterbalanced by reference to the context. Indeed, for specialised corpora, gathering contextual data about the setting from which the texts or discourses were collected can be essential, as it is often not possible to make sense of such specialised discourse without some background knowledge. For the 500,000-word sub-corpus of business discourse collected as part of the HKCSE, data collection was preceded by a period of observation in the organisations, which enabled the research assistant to choose sites for recording that would reflect a cross-section of the organisations' functions (Warren 2004). Warren notes that this period of observation and orientation was found to be essential at a later stage to interpret the data.

Although methods of ethnographic observation, note-taking and interview are not usually associated with corpus studies, there is no reason these methods cannot also inform and complement corpus analysis (ibid.: 137). In the case of small, specialised corpora, such contextual information is extremely valuable: It is often essential for interpreting the data, and it can be drawn on in qualitatively analysing the corpus results. For example, Harrington (2018) used both ethnographic and corpus linguistic methods in a study of interactions in an asylum reception centre. In addition to field notes and interview data, untranscribed data, which is not part of the corpus, can also inform the analysis in various ways. In some cases, it may be necessary to consult discourse participants or other representatives of the organisation to aid with transcription or corpus compilation. Warren (2004) gives the example of an encounter which the compilers of the business sub-corpus of HKCSE were not able to assign to any of the team's list of genres. After consulting an employee of the organisation, a new genre category was created for the sub-corpus.

Background information is useful not only in interpreting the data but can also be an integral part of corpus design. Having detailed information about the speakers or writers, the goals of the interactions or texts and the setting in which they were produced as part of the corpus database means that linguistic practices can easily be linked to specific contextual variables. This can be done by having each contributor complete a speaker/writer information sheet and obtaining as much information as possible about the text samples (e.g. through participant observation and interviews with people in the organisation). This information can be included as a header at the top of each text file or transcript and/or stored in a database, which can be drawn on in carrying out the analysis. Speaker information collected for HKCSE includes place of birth, gender, occupation, educational background, time spent living or studying abroad and mother tongue (Warren 2004). In designing the Cambridge and Nottingham Business English Corpus (CANBEC), a spoken meetings corpus of 1 million words, Handford (2010) collected information in the following main categories:

1. Relationship between the speakers, e.g. peer, manager-subordinate, colleagues from the same or different departments;
2. Topic, e.g. sales, marketing, production;
3. Purpose of the meeting, e.g. internal/external, reviewing, planning;
4. Speaker information, e.g. age, title, department, level in the company;
5. Company type and size.

In Section 5, we will see how this information was used in the corpus analysis.

Data collected for a spoken corpus will need to be transcribed and decisions made as to how detailed or "close" the transcription should be. If the corpus is quite small, it may be possible to transcribe the recording in more detail than for a large corpus, where there may be so much data to transcribe that time-consuming close transcription is not possible. As with decisions about corpus size and sampling, the level of detail required for the transcription depends on the aim of the project (see Chapters 2 and 3, this volume). For example, there is no need to transcribe prosodic features (intonation) if these features will not be analysed. However, it is worth remembering that the more detailed the transcription is, the more faithfully it represents the original interaction and the more features are available for later analysis. The transcription conventions used also need to be computer-readable, and as most corpus software requires texts to be stored as plain text files, any codes used should be available in plain text format.

Small, spoken corpora are often used to examine interactive features, and for such analysis, pauses, overlaps, interruptions and unfinished words or utterances, as well as non-linguistic features of interaction, such as laughter, should be indicated. An even closer transcription would also code for certain features of intonation, showing, for example, any syllables that are emphasised or whether an utterance ends in a rising or falling tone, as is done for the VOICE corpus of spoken ELF interactions (see VOICE Project). A very detailed prosodic transcription was pioneered with an early general corpus, the Survey of English Usage corpus (Svartvik and Quirk 1980), part of the London-Lund Corpus of Spoken English. An example of a specialised corpus showing detailed prosodic information is the HKCSE, which uses Brazil's (1997) discourse intonation system, where the utterances are transcribed as tone units and prominence, tone and key (pitch) are shown (see Cheng *et al.* 2008).

5 What can be learnt from a small, specialised corpus?

Having covered the issues involved in designing and compiling a small, specialised corpus, this section will discuss the advantages of small corpora in terms of what can be learnt from them. As already mentioned at the beginning of this chapter, one of the main advantages of a small, specialised corpus is that, unlike with a large corpus, the language is far less de-contextualised. Indeed, as Handford (2010: 7) points out, such contextual information may be essential in interpreting the data.

According to Flowerdew (2008), context is relevant for corpus analysis in two ways:

1. The context can inform the corpus-based analysis, for example, when the compiler-cum-analyst of a small, specialised corpus has access to background information to aid in the interpretation of the data;
2. The linguistic patterns identified through corpus analysis can tell us something about the social and cultural context from which the data were taken.

For both types of contextual links between corpus and context, small, specialised corpora have a clear advantage over large corpora. The first type was discussed in Section 4 on corpus compilation; here we will examine the second more closely.

Corpus analysis using word or keyword lists and concordances (see Chapter 9, this volume) can reveal insights into contexts of use. Patterns identified can be linked to specific contexts, for instance, showing in which genre(s) certain words or expressions occur and who uses them most. For example, a study of hotel interactions from the business sub-corpus of HKCSE (Cheng 2004) showed that the word "minibar" occurred in all checking-out interactions and was used exclusively by hotel staff, not by guests. Further investigation of frequent or key words and phrases may lead to the discovery of pragmatically specialised uses within a professional context. A good example of this is from a study of a 60,000-word corpus of telephone calls to NHS Direct (Adolphs *et al.* 2004). Starting with a keyword list and following up with a qualitative examination of these words revealed that many of the words and expressions performed interpersonal functions aimed at eliciting symptoms from callers. One such key expression is the vague category marker "or anything", which health advisors frequently used in eliciting symptoms from callers, for example:

(1)
N: ‘And so there’s no swelling anywhere to your face or anything?’
(Adolphs et al. 2004: 19)

The use of this vague expression invites callers to add their own description to the proposed symptoms, and thus performs a pragmatically specialised function within this professional genre.

Finally, let’s consider some specific examples of how factors, such as genre, topic or the relationship between the participants, can influence local contexts of use. As shown in Section 4, CANBEC was designed to enable searches according to topic and purpose of the meetings and relationship between the speakers. Quantitative findings, such as frequency counts, can therefore be linked to such factors. The use of the lexical items *issue* and *problem* in CANBEC provides an interesting illustration of the role such factors can play in local contexts. These words appear to be synonyms, and basic corpus searches do indeed reveal similar patterns: They both have a high frequency and enter into similar collocations. However, the frequency of these two lexical items varies considerably when looking at the topics discussed in the meetings and the relationship between the speakers (Handford 2010: 188–95). *Issue*, for example, is more frequent in human resources and marketing meetings, whereas *problem* occurs most in procedural and technical meetings. In terms of speaker relationships, *issue* occurs more in interactions between managers and subordinates, whereas *problem* is used more in peer discussions. The following example from a meeting between peers, in which both *issue* and *problem* are used, illustrates how these two words in fact perform slightly different functions:

(2)
Well I- I thi- think that’s another **issue**. And the other the and and another **issue** which comes on- onto that is that erm I'm still waiting … apparently one of the **problems** with getting some of the information off the computer is …
(adapted from Handford 2010: 193)

Handford (ibid.) notes that, while both words have the ‘prosody of difficulty’, *problem* seems to indicate more of a concrete obstacle, something that should be solved, whereas *issue* is somewhat more nebulous, and perhaps indicates that further discussion is needed. This fits with the nature of the meeting topics, where each of these words occurs most frequently: In technical and procedural meetings, more concrete *problems* are raised, whereas in human resources and marketing meetings, wider discussions “around” *issues* seem to be required. Considering speaker relationship, Handford argues that *problem* comes across as more categorical and could therefore potentially be face-threatening. This explains its higher frequency in peer meetings, where threats to face are less likely, thanks to the equal relationship between participants. In meetings between unequal participants (managers and subordinates), *issue* may be a useful euphemistic alternative to *problem*, serving to mitigate a potentially face-threatening act (ibid.: 192–4).

In the ABOT Corpus of workplace interactions (see Section 3), we can also observe the influence of local contexts on the frequency and use of various words and patterns. Both CANBEC and ABOT show that modals of obligation (*have to*, *need to*, *should*) are very frequent in workplace interactions (Koester 2006, 2010; Handford 2010). However, in both corpora, these modals, as well as their collocational patterns, are differentially distributed according to local contexts, such as genre and speaker relationship. The

Table 5.1 Total number of occurrences of modals of obligation in each macro-genre (per thousand words)

	Collaborative		*Unidirectional*	
	Raw freq.	*Density PTW*	*Raw freq.*	*Density PTW*
have (got) to	64	4.4	42	2.9
need (to)	32	2.2	22	1.5
should	28	1.9	22	1.5

genres in ABOT are grouped into two "macro-genres": unidirectional and collaborative (Koester 2010: 24–5). In unidirectional genres, one of the speakers clearly plays a dominant role by imparting information or giving instructions. In collaborative genres, such as decision-making and planning, participants contribute more or less equally towards accomplishing the goal of the encounter. In the ABOT Corpus, all the modals of obligation are more frequent in collaborative genres than in unidirectional genres, as shown in Table 5.1 (Koester 2006: 85–8).

Table 5.1, which shows both raw frequency and frequency per thousand words (or "density"), also shows that the difference in frequency is greater the stronger the modal: i.e. *have to*, which is the most forceful, occurs much more frequently, whereas *should*, the least forceful, is only marginally more frequent in collaborative genres.

Moreover, collocational patterns of modals and pronoun combinations also vary systematically with genre. Thus, in collaborative genres, *we* and *you* are the most frequent pronouns used with the modals noted earlier, whereas in unidirectional genres, *I* occurs most frequently in combination with all three modals. In unidirectional genres, *you have to* does not occur at all: there is just one example of *you'll have to* and a few instances of *you don't have to*.

Both the lower frequency of the more forceful modals and the infrequent use of the pronoun *you* in combination with all three modals of obligation can be linked to the feature that all unidirectional genres have in common, namely the fact that one speaker plays a dominant role. This imbalance in the speakers' roles generally means more care is taken to avoid face-threatening acts, even if the actual social or institutional relationship between the speakers varies. This results in more indirect and hedged language, as illustrated in the following example, where a speaker makes a request using *I need you to* instead of *you need to*:

(3)
I need you to sign off on this pack too.

[Author's data]

Another reason for the frequency of the first-person pronoun *I* is that in procedural discourse or instruction-giving (the most frequent unidirectional genre), the person receiving instructions frequently "invites" directives by saying *should I*, e.g.:

(4)
What should I do. Just - get the estimate…

[Author's data]

In collaborative genres, on the other hand, participants play a more equal role, and therefore more direct forms, such as *you have to* or *you should* are unproblematic, e.g.:

(5)
You have to make sure you can get access to that.

[Author's data]

Also, most collaborative genres are action-orientated, meaning people are trying to get things done (decisions, plans, arrangements), which results in the frequent use of modals of obligation with the first-person pronoun *we*, e.g.:

(6)
We need to get it moving.

[Author's data]

Corpus analysis can also reveal specific pragmatic meanings of collocational patterns, so-called "semantic prosodies" (see Chapter 27, this volume), within a specialised genre. Flowerdew (2008) found that the collocation *associated with* was very frequent in a corpus of environmental reports. Not only did it occur 139 times in the 250,000-word corpus, but it was found across all 23 companies from which the reports were drawn, indicating that this is a typical phrase for the genre and not a result of "local prosody" (see Section 3). In 135 of these instances, the phrase seemed to have a negative semantic prosody, for example:

(7)
difficulties associated with hydraulic dredging

(Flowerdew 2008: 121)

Flowerdew (2008: 121) concludes that this phrase is 'most likely an attenuated form of "caused by"' which is used by scientists to 'avoid claiming a direct causal effect, thereby forestalling any challenges from their peers', and therefore forms part of the discourse practices of the genre of environmental reports. In order to determine whether this finding is generalisable to other types of scientific writing, Flowerdew searched for the phrase *associated with* in the much larger 7-million-word Applied Science domain of the British National Corpus (BNC) and found that in 40 per cent of the samples examined, the phrase also has a negative semantic prosody. Such comparisons with a larger corpus, covering a similar variety or genre as the smaller, specialised corpus, are useful in testing the validity of findings from a smaller corpus and reinforcing the robustness of any generalisations made (see also Flowerdew 2003). By comparing a small corpus against a larger "benchmark" corpus, "keywords" can also be identified (e.g. using Wordsmith Tools [Scott 2019], Antconc or other applications, as described in Chapter 9, this volume): These are words that are unusually frequent in the small corpus compared to their normal frequency in the language (see Chapters 9 and 10, this volume).

This chapter has shown that while small corpora are not suitable for all types of analysis, a small, specialised corpus can nevertheless provide valuable insights into specific areas of language use and can even have certain advantages over large corpora. The main advantage is in the close link that exists between language patterns and contexts of use, as illustrated throughout this chapter from corpus design, through compilation and transcription to corpus analysis and findings. This interplay of language and context in corpus studies can be followed up in other chapters in this volume which deal with special areas of language use. Chapter 6 (this volume) looks at building a corpus to represent a language variety, and Chapter 7 at building a specialised audio-visual corpus. Other chapters focus on specific registers and genres; for example, Chapter 28 examines English for Academic Purposes, Chapter 41 looks at forensic linguistics and Chapter 43 explores health communication.

Transcription conventions used in data extracts:

.	falling intonation at end of tone unit
…	ellipted utterance
-	sound abruptly cut off, e.g. false start

Further reading

Flowerdew, L. (2004) 'The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings', in U. Connor and T. Upton (eds) *Discourse in the Professions*, Amsterdam: John Benjamins, pp. 11–33. (This chapter is useful for anyone wanting to build a specialised corpus. As well as presenting a rationale for using specialised corpora, it provides useful guidelines for defining a specialised corpus and for corpus design.)

Handford, M. (2010) *The Language of Business Meetings*, Cambridge: Cambridge University Press. (This book provides a complete description of all the steps involved in building and exploiting a corpus of one professional genre (the business meeting) from data collection and corpus compilation to corpus analysis and interpretation.)

Harrington, K. (2018) *The Role of Corpus Linguistics in the Ethnography of a Closed Community*, London: Routledge. (This book provides a good representative study of a small, specialised corpus which focuses on the spoken interaction of a specific community: the residents of an asylum reception centre. It shows how corpus linguistics can be combined with other methods, in this case ethnography and conversation analysis.)

O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) *From Corpus to Classroom: Language Use and Language Teaching*, Cambridge: Cambridge University Press. (This book provides an accessible introduction to the most important topics in corpus research. The role of qualitative as well as quantitative analysis is a theme throughout the book, and many chapters address the topic of what can be learned from small, specialised corpora, in particular Chapters 8 and 10.)

References

Adolphs, S., Brown, B., Carter, R., Crawford, C. and Sahota, O. (2004) 'Applying Corpus Linguistics in a Health Care Context', *Journal of Applied Linguistics* 1(1): 9–28.

Baker, P., Brookes, G. and Evans, C. (2019) *The Language of Patient Feedback*, London: Routledge.

BASE (*British Academic Spoken English*) and BASE Plus Collections, https://warwick.ac.uk/fac/soc/al/research/collections/base/ [Accessed 9 August 2020].

Biber, D. (1990) 'Methodological Issues Regarding Corpus-Based Analyses of Linguistic Variation', *Literary and Linguistic Computing* 5(4): 257–69.

Biber, D. (1993) 'Representativeness in Corpus Design', *Literary and Linguistic Computing* 8(4): 243–57.

Brazil, D. (1997) *The Communicative Role of Intonation in English*, Cambridge: Cambridge University Press.

Carter, R. A. and McCarthy, M. J. (1995) ‘Grammar and the Spoken Language’, *Applied Linguistics* 16 (2): 141–58.

Cheng, W. (2004) ‘//→ did you TOOK//↗ from the miniBAR//: What is the Practical Relevance of a Corpus-driven Language Study to Practitioners in Hong Kong’s Hotel Industry?’, in U. Connor and T. A. Upton (eds) *Discourse in the Professions*, Amsterdam: John Benjamins, pp. 141–66.

Cheng, W. (2014) ‘Corpus Analyses of Professional Discourse’, in V. Bhatia and S. Bremner (eds) *The Routledge Handbook of Language and Professional Communication*, London: Routledge, pp. 13–25.

Cheng, W., Greaves, C. and Warren, M. (2008) *A Corpus-Driven Study of Discourse Intonation*, Amsterdam/Philadelphia: John Benjamins.

Connor, U., Davis, K., De Rycker, T., Phillips, E. M. and Verckens, J. P. (1997) ‘An International Course in International Business Writing: Belgium, Finland, the United States’, *Business Communication Quarterly* 60(4): 63–74.

ELFA (2008) *The Corpus of English as a Lingua Franca in Academic Settings*, Director: Anna Mauranen, http://www.helsinki.fi/elfa [Accessed 4 August 2020].

Farr, F., Murphy, B. and O’Keeffe, A. (2004) ‘The Limerick Corpus of Irish English: Design, Description and Application’, *Teanga: The Irish Yearbook of Applied Linguistics* 21: 5–29.

Flowerdew, L. (2003) ‘A Combined Corpus and Systemic-Functional Analysis of the Problem-Solution Pattern in a Student and Professional Corpus of Technical Writing’, *TESOL Quarterly* 37(3): 489–511.

Flowerdew, L. (2004) ‘The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings’, in U. Connor and T. A. Upton (eds) *Discourse in the Professions*, Amsterdam: John Benjamins, pp. 11–33.

Flowerdew, L. (2008) ‘Corpora and Context in Professional Writing’, in V. K. Bhatia, J. Flowerdew and R. H. Jones (eds) *Advances in Discourse Studies*, London: Routledge, pp. 115–31.

Handford, M. (2010) *The Language of Business Meetings*, Cambridge: Cambridge University Press.

Handford, M. and Koester, A. (2019) ‘The Construction of Conflict Talk across Workplace Contexts: Towards a Theory of Conflictual Compact’, *Language Awareness* 28(2):186–206.

Harrington, K. (2018) *The Role of Corpus Linguistics in the Ethnography of a Closed Community*, London: Routledge.

Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.

ICLE website, https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html [Accessed 4 August 2020].

Kessler, G. (2010) ‘Virtual Business: An Enron Email Corpus Study’, *Journal of Pragmatics* 42(1): 262–70.

Koester, A. (2006) *Investigating Workplace Discourse*, London: Routledge.

Koester, A. (2010) *Workplace Discourse*, London: Continuum.

MacArthur, F., Alejo, R., Piquer-Piriz, A., Amador-Moreno, C., Littlemore, J., Ädel, A., Krennmayr, T. and Vaughn, E. (2014) *EuroCoAT, The European Corpus of Academic Talk*, http://www.eurocoat.es.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics*, Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies*, London: Routledge.

Moon, R. (1998) *Fixed Expressions and Idioms in English: A Corpus-based Approach*, Oxford: Clarendon 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.

Scott, M. (2019) *Wordsmith Tools, Version 7 (corpus analytical software suite)*, Oxford: Oxford University Press.

Simpson, R. C., Briggs, S. L., Ovens, J. and Swales, J. M. (2002) ‘The Michigan Corpus of Academic Spoken English’, Ann Arbor, MI: The Regents of the University of Michigan, https://lsa.umich.edu/eli/language-resources/micase-micusp.html [Accessed 7 August 2020].

Sinclair, J. (2004) ‘Trust the Text: Language’, *Corpus and Discourse*, London: Routledge.

Svartvik, J. and Quirk, R. (1980) *A Corpus of English Conversation*, Lund: Liberläromodel.

Timmis, I. (2018) *Historical Spoken Language Research*, London: Routledge.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*, Amsterdam: John Benjamins.

Tribble, C. (2002) ‘Corpora and Corpus Analysis: New Windows on Academic Writing’, in J. Flowerdew (ed.) *Academic Discourse*, London: Longman, pp. 131–49.

Upton, T. A. and Connor, U. (2001) ‘Using Computerized Corpus Analysis to Investigate the Textlinguistic Discourse Moves of a Genre’, *English for Specific Purposes* 20: 313–29.

VOICE. (2013) *The Vienna-Oxford International Corpus of English* (version 2.0 online).

VOICE. (2007) VOICE Transcription Conventions [2.1], available at http://www.univie.ac.at/voice/voice.php?page=transcription_general_information [Accessed 5 March 2020].

Warren, M. (2004) ‘//so what have YOU been WORKing on REcently//: Compiling a Specialised Corpus of Spoken Business English’, in U. Connor and T. A. Upton (eds) *Discourse in the Professions*, Amsterdam: John Benjamins, pp. 115–40.

WrELFA. (2015) *The Corpus of Written English as a Lingua Franca in Academic Settings*, Director: Anna Mauranen, Compilation manager: Ray Carey, http://www.helsinki.fi/elfa [Accessed 8 October 2021].