The Routledge Handbook of Linguistics

Keith Allan

Corpus linguistics

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315718453.ch32

Andrew Hardie
Published online on: 20 Jul 2015

How to cite: Andrew Hardie. 20 Jul 2015, Corpus linguistics from: The Routledge Handbook of Linguistics Routledge
Accessed on: 27 Oct 2023
https://www.routledgehandbooks.com/doi/10.4324/9781315718453.ch32

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
32.1 Corpus linguistics: an overview

What is a corpus, and what is corpus linguistics? Answering either of these questions involves answering the other, as the notion of a ‘corpus’ in contemporary linguistics is inextricably tied up with the ideas and techniques of corpus linguistics. The term corpus (plural corpora) is simply the normal word in Latin for ‘body’, as in a body of text. Over time, the term came to be used for any natural-language dataset used by a linguist – so that, for example, a field linguist might refer to a set of sentences elicited from an informant as their ‘corpus’. Since the 1960s, a yet more specific meaning for the term has emerged: corpora in this sense are very large, machine-readable collections of spoken and written text that are analysed using computational methods. The set of methods required to approach the quantitative and qualitative analysis of a collection of language data on a scale far larger than any human being could hope to analyse by hand – together with related areas such as the compilation and annotation of these corpora – constitute the modern field of corpus linguistics.

Describing corpus linguistics as a ‘field’ or subdiscipline of linguistics is a statement which requires qualification, however. Most subfields of linguistics relate to the study of some aspect of the language system or its usage, such as phonology, morphology or sociolinguistics. Corpus linguistics, by contrast, is not concerned uniquely with any single facet of language, but rather is an approach which can be applied to many or all aspects of language. For this reason, many corpus linguists prefer to describe it as a ‘methodology’. As a methodology it has certainly been a great success in fields ranging from the history of English to lexicography to language teaching to discourse analysis. But for others in the field, the insights into language that arise from the use of the corpus are so radically different from the traditional understanding of how language works that they consider corpus linguistics to constitute an independent field of study or theory of language. This latter view is associated with the neo-Firthian school of researchers led by the late John Sinclair.

In its early days, corpus linguistics was something of a controversial approach. This is because it emerged at a time when the leading model of linguistic theory was the Chomskyan formalist approach. Without going into too much detail, the linguistics espoused by Noam
Chomsky and colleagues presume a rationalist rather than empiricist approach to knowledge: the linguist’s intuition and introspection are central as sources of information about language. In corpus linguistics, however, while the utility of the researcher’s linguistic (and metalinguistic) intuitions are not discounted, empirical data in the form of naturally-produced discourse is very much front and centre – both as a practical matter and as a matter of principle. Geoffrey Leech, for instance, has emphasised the notion of total accountability as critical within corpus linguistics (Leech 1992). For our analysis to be totally accountable to the corpus data, we must ensure that we take into account all the examples in our corpus relevant to what we are investigating – or, if that is impractical, that we examine random subsets of the data, not selected subsets. What we must not do is intentionally ignore relevant evidence. Within the neo-Firthian tradition, Sinclair has promulgated a maxim ‘trust the text!’ with generally similar import: it is the corpus data, and not theorising based on intuition, that is the ultimate arbiter of what generalisations and hypotheses we can come up with about language.

Corpus-based methodologies remained something of a niche approach through the 1960s and 1970s, partly because of the influence of the Chomskyan arguments against the use of corpus data, but perhaps more importantly because few linguists had access to the computer technology without which corpus linguistics is effectively impractical. Through the 1980s, and especially in the 1990s, all this changed. Advances in information technology put computer power sufficient for the analysis of larger and larger bodies of data within the reach of any researcher who cares to use a corpus. As a result, in the twenty-first century, corpus linguistics has become – in the words of Tony McEnery – a ‘killer method’ in linguistics, applied to a hugely diverse array of types of linguistic research. This is, in short, why corpus linguistics matters. It is not merely an improved method within lexicography, or a method for researching grammar, or an approach to political or media discourse. It is all of the above and more.

There is no room in this chapter to attempt a comprehensive review of the impact that corpus research has had across the field of linguistics. Instead, I will provide an overview of the fundamental ideas of corpus linguistics and some key methods and practices – starting with how we approach corpus design and construction, moving on to a summary of the most widely used corpus methods, and finishing with a very brief survey of some applications and advanced forms of analysis.

32.2 Corpus design and construction

32.2.1 General considerations

More linguists work with corpora than actually collect corpus data or work on corpus design. A very large set of questions in corpus linguistics can be answered without going beyond standard, widely available corpora. A good example of such a corpus is the British National Corpus (BNC). The BNC is a 100 million word collection of British English, compiled in the early 1990s, and including samples of a wide range of spoken and written genres. As a broad sample of the English language in general, it is suited to many different research aims. In other cases, a more specialised collection of data may be required. A wide range of such specialised corpora have been constructed and likewise made generally available, covering historical and dialectal data, specific genres, or the spoken language of specific types of people.

But there are other types of research question for which no standard corpus is available. In this case, the first step for the researcher is to build the corpus on which the analysis will
be based. For instance, many corpus-based analyses of media discourse are based on examining media texts that address some particular topic. It is unlikely that such a highly specialised dataset would be available in advance. If I am interested in how local newspapers in Britain discussed the issue of crime in the years 2010 to 2013, I will almost certainly need to build a corpus myself for this specific purpose. Such goal-oriented corpora are informally known as ‘do-it-yourself’ or DIY corpora and very often are not made available to researchers other than the original compiler (often, for copyright reasons, they cannot be).

The virtue of working with a DIY corpus is that the design of the data can be tailored directly to the specific research question at hand. The virtue of working with a standard published corpus is that the corpus serves as a common basis between different researchers – allowing results to be tested and replicated by other scholars, for instance. Both approaches have their place in different kinds of corpus-based study.

However, corpus researchers need an understanding of the issues that arise in corpus construction even if they never build any corpora of their own. This is because we have to understand the contents, design and structure of the corpora that we use, in order to select a corpus appropriate to what we want to research, in order to make appropriate use of that corpus, and in order to treat our results critically in terms of how far they may be extrapolated beyond the specific corpus at hand. The key concepts here are the notions of the corpus as a sample and of balance and representativeness.

### 32.2.2 The corpus as a sample

Any corpus is fundamentally a sample. That is, there is some large phenomenon that we want to know about – a language, or some specified variety of language, as a whole – and since we cannot look at all the possible text within that language, we must select a sample. Borrowing terms from statistics, we can talk about the whole of the language variety we are approaching as the population, and the corpus we are using to look at that language variety as the sample. Our interest is typically on what we can say about the population on the basis of the sample – and how confidently we can say it. The very fact that the corpus is a sample means that we need to think carefully about how our sample relates to the population, which means we need a detailed and critical awareness of the process of sampling that has been used in corpus construction.

It is rarely possible to avoid sampling. As Chomsky pointed out, language is in principle non-finite (it is always possible for a speaker of a language to create a new sentence in that language that has never been used before) and it is – of course – impossible to collect a dataset of infinite size! Only in the case of certain historical data, where the amount of text that has survived from a particular period is finite, can we have a complete corpus of some language or language variety. For example, it would be possible to compile a complete corpus of Old English, because only a limited number of documents in Old English have survived. But even in this case, we are effectively letting the forces of history – the preservation of some documents, and the loss of others – do the sampling for us.

An important difference between sampling in corpus linguistics and in other sciences is that a corpus is never a random sample of the population. In research such as opinion polling or medical trials, where the population is a literal population of people, the sample must be as close as possible to a randomly-selected subset of the population, such that every person has an equal likelihood of being selected to be in the sample. Corpus collection is very different. It is possible to select the texts of a corpus randomly from a population of texts of interest. But we rarely use the whole text as the unit of analysis. Most of the analyses we will
want to undertake with our corpus will be centred on a much smaller linguistic unit, such as the sentence, the clause or, perhaps most commonly, the word. However, even if the texts of the corpus have been selected randomly, the sentences and words are not random. For instance, consider the word elephant. It is not especially common. Many texts contain no examples at all of elephant. However, a text that does use the word elephant once is then rather likely to go on and use it several more times (for example, if it is a text about elephants). This means that the probability of a given word occurring at some point in a text is influenced by the words that have occurred before and thus the words in a corpus are not random. Similar logic can be applied to sentences.

What does this mean for corpus analysis? Three things. First, we have to be very careful in applying statistics to corpus data that are unproblematic in other fields, precisely because of this issue. Second, when we analyse a corpus we must always be aware of dispersion – looking out for the difference, for instance, between a word that occurs 100 times in our corpus where all 100 instances are bunched close together in one text, and a word with the same 100 occurrences that are spread out across the whole corpus. Third, in the process of corpus construction we have to think about the relationship between the population and the sample in a somewhat different way to how this is done in fields where sampling can be done on a random basis. The concepts that we normally apply for this purpose are representativeness and balance.

32.2.3 Representativeness and balance

If the words and sentences of a corpus are not a random sample, then how can we have any confidence that findings we arrive at using a corpus are applicable to the language or variety as whole? We must consider the issue of how representative the corpus is of that language or variety, based on the decisions that were made about what to include and what to exclude in that corpus. As Leech (2007: 135) points out, ‘Without representativeness, whatever is found to be true of a corpus, is simply true of that corpus – and cannot be extended to anything else.’ We consider a corpus to be representative if we have adequately captured the nature of the whole population within our sample. That means that the full variety of what is found in the language must be represented in the corpus, if we aspire to extrapolate from the corpus to the language. If we are trying to represent written English, for instance, our corpus cannot contain just newspaper language. It must also contain examples of magazines, books, reports, letters, and other media both published and unpublished. Within each medium, we would also need to make sure that the full range of genres (functional varieties) and the full range of domains (subject matters) were represented adequately. Clearly, building a corpus big enough to contain an example of everything from this massive landscape of possibilities would be a gargantuan undertaking. Even a very large standard corpus such as the BNC will not sample every possible type of text. Smaller corpora, such as the Brown Corpus of written American English (one million words in extent), will struggle even more to represent the full range of what is ‘out there’. Perfect representativeness is, therefore, something that almost certainly cannot be achieved. Rather, we must strive towards better and better representativeness, even though complete representativeness is unattainable. Moreover – and critically – we must always be aware of the limitations in representativeness inherent in the design of the corpus we are using, and restrict the scope and strength of the claims we make on the basis of the corpus data accordingly.

Balance is a related, and equally problematic, issue. It relates to the relative proportions of different types of data within a corpus. A corpus of general English, for instance, must
A. Hardie

sample both spoken and written data (of various kinds) in order to be representative; however, what should the relative proportions of speech versus writing be? A corpus is considered balanced if its subsections are correctly sized relative to one another. However, there is no single answer to the question of what the correct proportions are. For instance, in the case of speech versus writing, we might propose a corpus that is 50 per cent spoken and 50 per cent written data (this is, in fact, the design of the International Corpus of English, ICE). However, this assumes that speech and writing are equally important. Is that the case? How do we even determine the parameters of what is ‘important’ in this context? Most people probably hear more language than they read; does this mean the spoken language should make up more than half the corpus? On the other hand, speech is generally much more ephemeral than writing; does that mean written texts should dominate? As with representativeness, the issues around balance are fraught with difficulty, but the critical thing for the user of any corpus is to be actively aware of how that corpus is balanced, to think carefully about what this means for the generalisability of results based on that corpus.

32.2.4 Metadata, markup and annotation

A corpus always contains the actual words of one or more texts that have been collected as discussed above. These will typically be collected in the form of plain text files – that is, computer files containing just the actual characters of the text, without any kind of formatting, as found in word-processor files, for instance. The plain-text words of the corpus are sometimes called the raw data of the corpus. However, raw data is not necessarily all that the corpus contains. There are three types of additional information that can be added to a corpus, namely metadata, markup and annotation.

The definitions of (especially) the terms markup and annotation in the context of corpus design and encoding vary quite a lot in the literature; the term tagging is also sometimes used as a synonym for either or both of these terms. However, for present purposes we will treat them as distinct concepts. Markup is information added to a corpus to represent features of the original texts other than the running words themselves – for instance, in the case of a written text, features such as the beginning and end points of sentences or paragraphs, or the position of page breaks, or the position and content of elements such as illustrations which would be omitted in a plain text corpus. Annotation, by contrast, encodes into the corpus the results of some linguistic analysis which we have undertaken on the corpus text.

The third thing which can be added to the raw text of a corpus is metadata. Metadata is data about data, that is, information about the texts that have been included in the corpus – such as whether they are written or spoken, information about genre or domain, bibliographic information on original publication of a written text, or demographic information about the speakers in a spoken text. There are several ways to store metadata for a corpus, but one popular way is to insert a chunk of descriptive information at the beginning of each corpus text in a fixed, specified format: this metadata chunk is called the header of the text.

Critically, markup, metadata and annotation can be processed by computer, just like the actual words of the corpus texts, and therefore they can be exploited in automated corpus analysis. We will return to annotation in the next section; let us deal briefly here with metadata and markup by discussing some examples of how they may be used to enhance a linguistic analysis – and, thus, the reason corpus builders invest effort into adding them into the corpus in the first place!

Metadata has two main functions. First, it allows us to trace the corpus evidence we see back to its source. If we are examining a series of examples extracted by searching a corpus,
for instance, it is often very useful to be able to see, via the metadata, precisely what text each example comes from: knowing, for example, whether a particular instance comes from a novel or from a newspaper report can be a valuable aid to interpretation. The second function is that metadata allows us to isolate and compare different sections of a corpus. If, for instance, we have a corpus containing both writing and speech, we may on occasion wish to search just within the spoken texts, or to compare the results of some analysis in the written data versus the spoken data.

Markup has similar applications in that it allows analyses to take account of the structure of the corpus texts. For instance, in a spoken corpus, who said what is often critical to an analysis: only the markup of utterance boundaries and speaker identities allows us to know this. But the same can apply to written data – we might be interested in differences in how a word is used paragraph-initially versus medially or finally, for instance, and this can only be automated if the corpus has paragraph breaks marked up.

32.2.5 Corpus annotation

To annotate a corpus is to insert codes into the running text to represent a linguistic analysis. In many cases, the analysis is done at the word level, so a single analytic label or tag is assigned to each word of the corpus. However, analyses at higher or lower levels of linguistic structure can also be represented as corpus annotation. Much research has been done to automate the process of adding the most common forms of analysis, so that it is not necessary for a human being to read through the whole corpus and manually insert the tags. Let us consider four of the most widely used kinds of corpus annotation – those often pre-encoded in general-purpose corpora prior to their being distributed.

Part-of-speech (POS) tagging is the assignment to each word of a single label indicating that word’s grammatical category membership. It is the longest-established form of corpus annotation, with the first efforts in the direction of automatic taggers going back as far as the early 1960s, and the first practical, high-accuracy tagging software emerging in the 1980s – although even the best POS taggers have a residual error rate of around 3–5 per cent which can only be removed by painstaking manual post-editing. The utility of POS tagging, especially for languages like English, is that many words are ambiguous in terms of their part-of-speech. For instance, walk can be both a noun and a verb, and the only way to know which is to look at it in context. Having the computer evaluate the contexts for us and on that basis disambiguate each instance of the word allows us to treat walk the noun and walk the verb separately, according to what makes most sense for a given research question.

Different POS tagger systems use different tagsets, that is, systems of grammatical (specifically, morphosyntactic) categories. Some simple tagsets only distinguish major word classes such as noun, verb, adjective, preposition and determiner. However, there are many more morphosyntactic distinctions that can be useful to linguistic analysis, and more fine-grained tagsets will try to encode these distinctions. For example, it is common for POS tagsets to distinguish common nouns from proper nouns, or lexical verbs from auxiliary verbs. Inflectional distinctions will also usually be represented: singular versus plural, or past tense versus present tense, for instance.

The underlying principle of semantic tagging is the same as POS tagging, namely, the assignment to every word in a text or corpus of labels representing categories within a scheme of analysis. The difference is that in the case of semantic tagging, the tags represent semantic fields, i.e. categories of meaning – and the tagset as a whole thus represents some ontology, or way of dividing up all possible meanings into various domains or concepts.
Semantic tagging is a harder task for a computer than POS tagging. This is because, while a word can be ambiguous for both grammatical category and semantic field, disambiguating the semantics involves actually understanding the meaning of the context, which computers cannot do. For this reason, we must be prepared to work with a higher error rate when analysing the output of a semantic tagger than when working with POS tags. Nevertheless, semantic tags can be extremely useful, especially for detecting patterns in a corpus that affect groups of semantically-related words which are individually very rare.

A *lemma* is a group of word-forms that are judged to constitute a single entry in the lexicon of the language – different inflectional forms of a single dictionary headword. Lemmatisation is conceptually the most straightforward of the word-level annotations; every word is, simply, annotated with the lemma to which it belongs (where each lemma is represented by its morphological base form, such as *go* for *go*, *goes* and *went*). Thus, computer searches of a corpus can be based on grouping together all the separate inflectional forms of a single word. However, even this relatively simple operation can raise issues regarding what is, and what is not, considered to be part of the same lemma. The English word *broken*, for instance, can be tagged either as a participle or as an adjective. Many lemmatisers will treat it differently depending on how it is tagged. If it is tagged as a participle, it is lemmatised as *break* – since a participle is considered an inflectional form of a verb base. On the other hand, if it is tagged as an adjective, *broken* is lemmatised as *broken* – since the change of category to adjective is deemed to create a new lemma. So the lemmatisation does not always group elements together; in some cases it can draw distinctions between elements that are superficially identical.

The other form of corpus annotation that is often applied automatically is *parsing*, the annotation of syntactic structures and relations. Unlike the forms of annotation mentioned so far, parsing does not operate strictly at the word level; rather, what is annotated are grammatical phenomena at the phrase, clause and sentence level. As such, parsing represents an implementation within a computerised text of the traditional tree-diagram style of syntactic analysis; for that reason parsed corpora are often called *treebanks*. There are two major forms of parsing, each corresponding to a different kind of syntactic analysis. Constituency parsing is about identifying the phrases (syntactic constituents) and how they are nested within one another. This is done by introducing codes into the text to represent the beginning and end points of all the phrases. Dependency parsing, by contrast, labels the relationships between words: each word is linked to the grammatical structure by depending on another word in a particular way. For instance, a noun may relate to a verb by being its subject, or its object; an adjective may relate to a noun by being its modifier. Different tags are used to represent different types of dependency relationship.

Like semantic tagging, automated parsing has a much higher error rate than POS tagging, because the task is inherently more difficult. In both cases, the benefits of being able to access the semantic or grammatical analysis make it worth the trouble of working with data that has a known error rate. That said, many of the generally-available treebanks have been parsed manually, or at least have had their automated parsing manually corrected.

There are many other kinds of annotation which a researcher might apply for the specific purposes of their own research questions, such as different kinds of discourse-pragmatic or stylistic annotation. But these are often very different from research project to research project, and therefore must often be applied manually rather than automatically.

When a corpus has been tagged, the fundamental techniques of corpus analysis – frequency counts and concordance searches, which will be introduced at length in the next section – can be undertaken at levels of analysis higher than the orthographic
word-form. For instance, lemmatisation allows frequency lists to be compiled at the level of the dictionary headword, which may subsume many distinct word-forms. Conversely, tagging can disambiguate the words: break is both a noun and a verb, and broken is both a participle and an adjective – POS tags allow a computer to distinguish the different grammatical functions.

As I mentioned above, there exists recurring confusion regarding the distinction between annotation and markup. Part of the reason for this is that, as a practical matter, both can be inserted into the corpus using the same encoding system of tags inserted into the running text. The current standard method for adding tags to a corpus is by means of XML (the eXtensible Markup Language). In an XML document, anything inside <angled brackets> is part of a tag; anything outside angled brackets is part of the text itself. The special meaning of the angled brackets < and > means that actual less-than or greater-than signs in the text itself must be represented by the special codes &lt; and &gt;. Different XML tags are used for markup, metadata and annotation. For instance paragraph boundaries might be shown in XML using <p> tags; a <title> tag in the header might contain the original title of a text; and a POS tag on a word could be represented as <w pos='NN1'> (where NN1 is a common tag for a singular common noun).

32.3 Analytic techniques in corpus linguistics

32.3.1 Corpus analysis software

The two most basic forms of data which we can extract from a corpus are the concordance and the frequency list. It is, in theory, possible to generate either of these through hand-and-eye analysis of a stack of paper documents. Indeed, in the pre-computer age this was sometimes done; the earliest concordances were compiled manually for the study of the language of the Bible, and in the early twentieth century word frequency lists were often compiled manually to help inform foreign language teaching. However, in practice, all techniques for corpus analysis – these two most basic methods, and all the more complex methods built upon them – are nowadays supported by the use of various pieces of corpus analysis software. Thus, what possibilities we have for the analysis of a corpus is inevitably a function of the affordances of the available software tools.

Nearly all corpus analysis software permits the generation of a concordance – a listing of all the instances of a word or phrase in the corpus, together with some preceding co-text and some following co-text (usually anything from a handful of words to a couple of sentences). As this functionality is so common, the term concordancer is an often-used synonym for ‘corpus software tool’. Concordancers vary greatly in terms of what analyses, other than basic concordance searches, they allow. Since the late 1970s and early 1980s, a very wide range of such programs have been created – some very general and some highly specialised. For our present purposes, it suffices to distinguish two broad varieties of concordancer that are used today. Some programs (e.g. WordSmith, AntConc) run from the researcher’s desktop computer; this means that the user can analyse any corpus they have available locally using these programs. Other tools operate as client-server systems over the internet: the user accesses a remote computer via a client program, typically a web browser, and uses search software that actually runs on a remote server machine where some set of corpus resources and analysis systems has been made available. Client-server tools currently in wide use include SketchEngine, Wmatrix and CQPweb; some of these allow users to upload their own data to the corpus server, while others restrict users to a static set of available corpora.
In some cases, an analyst may not have much choice in whether to use a desktop concordancer or a client-server system: the choice is determined by what data they wish to use. Many large, standard corpora are only available via the web (such as the Corpus of Contemporary American English (COCA) – see Davies 2009). This is for reasons of intellectual property: unless permission to redistribute has been obtained from the copyright holder of every text in a corpus, then the corpus as a whole cannot be made available to anyone other than the original compiler. However, it is generally assumed that making isolated snippets of text from the corpus available via a web-based concordancer does not breach copyright, due to the consideration of fair use.

One other approach exists to corpus software: that is, for the researcher to write ad hoc programs specifically to carry out the particular analyses they wish to undertake. The argument for this practice – which necessarily requires that corpus linguists learn computer programming – is primarily one of flexibility; without the ability to create specialised analysis programs, the researcher is limited to solely those procedures that their concordancer happens to make available (see Gries 2010). However, given the advanced level of programming skills that this approach requires, it is difficult to see it being adopted by more than a minority of corpus analysts.

### 32.3.2 Corpus frequencies and their interpretation

One basic output that nearly all concordancers can produce is a frequency list: that is, a listing of all the words in the corpus, together with their frequency. In discussing the corpus frequency of words, it is useful to introduce the distinction between word types and word tokens. A *token* is a single instance of any word at a particular place in the corpus; a *type* is an individual word-form, which can occur many times in the corpus: thus one type is usually instantiated by multiple tokens. (The exception, types which occur only once in a corpus, are called *hapax legomena* – Greek for ‘(they were) said once’. ) So for instance, if we see on a frequency list for an English corpus that *the* is the most frequent word, occurring 60,000 times, then we could say that the single type *the* is represented in this corpus by 60,000 tokens. Frequency lists can also be compiled for any annotation that is present in the corpus, and we can talk about types and tokens of annotation tags in the same way.

Any given type will be more frequent in a bigger corpus; for that reason, corpus frequencies are often given in relative terms, as frequencies per thousand words or per million words. Relative frequencies of types are calculated by taking the raw frequency, dividing it by the number of tokens in the corpus overall, and multiplying by the normalisation factor (one thousand, one million or whatever). So, if *the* occurs 60,000 times in a 1.5 million word corpus, its relative frequency is forty per thousand words, or 40,000 per million words.

Speaking meaningfully about corpus frequencies is not straightforward. Simply stating that *the* occurs forty times per thousand words in a corpus tells us nothing of linguistic interest, whether about the word *the* or about the corpus in question. It is only when we compare frequencies that we arrive at interesting linguistic generalisations. We could compare the frequency of *the* to that of other words on the frequency list, and observe whether it is more or less frequently used than those other words. We could, moreover, compare the frequency lists of two corpora or sub-corpora, to ascertain whether the contents and ordering of their frequency lists are similar, as a means of contrasting the types of language those corpora represent. We would find, for instance, that in a written English corpus *the* will probably be the most frequent word, but in conversational English the personal pronouns *I* and *you* are likely to top the list.
A frequency list is inherently quantitative in nature. To be interpreted meaningfully, then, it must therefore nearly always be combined with some form of qualitative analysis – that is, an analysis that involves the linguist interacting with the actual discourse within the corpus and its structure and/or meaning. This is most straightforwardly accomplished within the context of a concordance analysis.

32.3.3 Concordance analysis

As we saw above, a concordance is the result of a corpus search: all the examples in a corpus matching some specified search pattern, together with some preceding and following context. In raw or unannotated corpora, we can only search for specified words or phrases. However, if tags are present, they can be used instead of or as well as a word pattern – for instance, to search for can as a noun versus can as a verb, or to search for all adverbs in the corpus. Most concordancers allow the use of special characters called wildcards that indicate a position in the pattern where ‘any letter’ or ‘any sequence of letters’ may be present, and thus allow searches to be underspecified; for instance, if the * character is a wildcard, then searching for act* will find not just act but also actor, action, active and so on. Some corpus software allows the use of the extremely sophisticated system of wildcards called regular expressions. The use of tags and wildcards together allows wholly abstract grammatical structures to be specified in a search pattern, for example, phrase-level constructions such as the English passive, perfect and progressive.

Whatever form of search is used, the result is the same: a list of examples in context, typically presented so that the word or phrase matching the search term is centred on the screen, and often sorted alphabetically (by the match expression, or by a preceding or following word). How do we handle the analysis of this concordance? There are several different approaches. At one level, analysis can be more or less impressionistic – based on scanning the eye up and down the concordance lines in an attempt to observe features of note that recur in the concordance, or to identify different functions of the word or phrase that was originally searched for. A more careful analysis will often attempt to quantify the number of concordance lines that exemplify a given function or contextual feature, and to make sure that every single concordance line has been inspected individually, to identify exhaustively all possible categories and patterns of usage. The most rigorous form of concordance analysis will systematise the aspects of each example that are considered, building a matrix of different features – grammatical, semantic or pragmatic – and the values these features have for each concordance line. All this can make a full concordance analysis a major undertaking; in compliance with the principle of total accountability, concordances too long to analyse in full should be reduced (or thinned) randomly to a manageable size. It is especially important not to look at just the beginning of a concordance, as these early concordance lines may present examples from the early part of the corpus, which is not necessarily representative of the corpus as a whole.

32.3.4 Key items analysis

A key items analysis is based on contrasting frequency lists. As we saw above, frequency lists often become most interesting when we compare lists from different corpora or different texts. However, simply comparing two or more lists by eye is only a very approximate way to do this. A more rigorous way to compare a pair of frequency lists is to use a statistical procedure to identify on a quantitative basis the most important differences between the two
lists. Several difference statistics can be used, but the most common is the log-likelihood test of statistical significance, a procedure similar in principle to the better-known chi-squared test. A key items analysis is performed as follows: first, frequency lists are created for two texts or corpora – either two corpora which we wish to contrast, or a single text of interest and a generic reference corpus. The frequency list may be of word types, lemmas or any kind of tag – thus, we often talk about keywords and key tags. Second, the frequencies of each item on the two lists are compared by calculating a log-likelihood score from the two frequencies and the total sizes of the two corpora. Third, the list is sorted in descending order of log-likelihood score; those items whose frequencies have the most significant differences between the two corpora under comparison will appear at the top of the list. Key items can be classed as positive (more frequent in the first corpus) or negative (more frequent in the second or reference corpus); both can be of interest, but studies based on keywords tend to focus on the positive items.

The interpretation of a keyword or key tag list, done properly, is an onerous task, as it will typically involve undertaking at least a partial concordance or collocation analysis of a large part of the highest-scoring key items. In fact, for many analysts, a primary virtue of a key items list is that it provides a way in to identifying which words or categories in a corpus will be of interest for more detailed study. Other analytic approaches to the key items list might involve grouping the items according to either an ad hoc or pre-established scheme of classification, to identify patterns or trends of usage across some large fraction of the list.

### 32.3.5 Collocation analysis

In the same way that key items analysis is based on, and abstracts away from, two frequency lists, so a collocation analysis is based on, and abstracts away from, a concordance search. Collocation in the broadest sense means simply those aspects of a word’s meaning which subsist in its relationship with other words alongside which it tends to occur. The idea that collocation is of high or indeed primary importance to understanding the meaning of a word or other linguistic form originates with J.R. Firth and is very important in the neo-Firthian school of corpus linguistics; but collocation analysis itself is a commonplace of most or all approaches to corpus linguistics. There are many different ways to operationalise the notion of co-occurrence. One is to look at sequences of words which recur in a fixed order with high frequency in the corpus – these sequences are variously referred to as n-grams, clusters or lexical bundles. A more common approach is to begin with a single word (or lemma) of interest, called the node, for which a concordance search has been performed; the surrounding text is then scrutinised to establish what elements occur frequently in the vicinity of the node. Depending on one’s methodological preferences and theoretical orientation, there are then a number of ways in which one might proceed. One way is manual and largely qualitative: to undertake a hand-and-eye analysis of the concordance, as explained above, where the focus of analysis is to identify co-occurring elements, the collocates of the node, in the concordance lines. However, given how labour-intensive this approach is, it is very common to automate the process of extracting collocates from the local context around the node. The software compiles a list of words that occur in the context (either all words, or words which stand in some particular relationship to the node) and then applies statistical analysis to identify words which are more common in the vicinity of the node than elsewhere in the corpus. These are then presented as a list of statistical collocates for the researcher to interpret. It should be noted that this is fundamentally rather similar to the procedure of a key items analysis, except that we are comparing the vicinity of the node to the rest of the
corpus, rather than comparing two different corpora, and some of the same considerations apply. First, as with key items, a number of different statistical procedures can be used; log-likelihood is one of the more common, but mutual information, z-score, and t-score are also common – among others – and the choice of statistic can result in almost totally different results. Second, as with key items, a rigorous collocation analysis will always go beyond the raw list of statistical collocates to look at actual instances of those collocates in usage alongside the node.

Although we can use collocation as a general covering term for different kinds of co-occurrence phenomena observed in corpora, we can also use a set of more specific terminology, developed largely by John Sinclair and colleagues, to distinguish co-occurrence patterns at different linguistic levels. In this more specific sense, collocation refers to the co-occurrence of particular word-forms with the node, and three other terms are used to refer to grammatical, semantic or discourse-pragmatic or affective co-occurrence pattern: colligation, semantic preference and semantic prosody.

Colligation refers to a recurrent co-occurrence of some node with a particular grammatical category or structure. For example, if we observe that a node word almost always occurs after an adjective – even if no single adjective co-occurs frequently enough with the node to be considered a collocate – we could say that the category of adjective is a colligate of that node. A colligate can also be a syntactic structure that the node tends to occur within or alongside – such as a sentence position, or a verb complementation pattern. Semantic preference or semantic association refers to a consistent co-occurrence with a set of words which – again, while perhaps not individually significant collocates – are drawn from a recurring semantic field. Finally, semantic prosody or discourse prosody refers to a broad function or meaning which tends to co-occur with the node but which may be variously realised. For instance, the verb cause tends to occur in contexts where the thing that is caused is negatively evaluated – but the linguistic realisation of that negative evaluation can take many different forms both in terms of the words and the grammatical structures used.

32.4 Conclusion: applications and frontiers of corpus linguistics

In this chapter, we have briefly considered a number of central topics of corpus linguistics. It bears emphasis that any analysis conducted using these techniques will be applied in addressing some specific research question, and the nature of that research question will to some extent dictate the selection of data and method. Unfortunately, there is simply no space here to give a full account of all the areas of linguistics, and of other fields of study, where the methods of corpus linguistics have been productively applied. As previously noted, corpus techniques have acquired the status of a key methodology applicable to nearly all subfields of language study (perhaps the sole exception being those Chomskyan formalist approaches whose opposition to corpus data is a matter of principle). It is not possible to do more here than provide a non-exhaustive list of applications without further commentary. The two initial applications of corpus data were improved grammatical description (see, for instance, Biber et al. 1999) and improved lexicography (as pioneered by corpus-based dictionaries such as the Collins COBUILD series). Other areas where corpus methods have been gainfully deployed include: theoretical linguistics, especially from a functional or cognitive perspective; linguistic typology; field linguistics; contrastive linguistics; genre and register analysis; dialectology; sociolinguistics; psycholinguistics, including the study of first and second language acquisition; stylistics; discourse analysis, including critical discourse analysis; applied linguistics topics including the study of learner language and
language teaching and testing; forensic linguistics; and both computational linguistics and natural language processing. In short, while there are many linguists who themselves do not use and do not need to use corpora, there are increasingly few linguists working in areas where no one uses corpora.

The current frontiers of corpus linguistics are twofold. The first is to extend yet further the range of applications in the field – beyond the concerns of linguistics, to serve the research interests of scholars across the humanities and social sciences. Many such disciplines are focused on one form or another of close analysis of text, for instance history and literary criticism. As more and more texts become available in large electronic archives – especially texts from before the late twentieth century – so there will arise an increasing requirement in humanities and social sciences research for approaches that allow such large amounts of data to be handled. This is exactly what the methodology of corpus linguistics offers to researchers in disciplines beyond linguistics. Not entirely unrelated to this is the second front of current progress in the field – namely, the continuing development of more refined techniques and tools for corpus analysis. In this chapter we introduced four basic analytic methods, but a number of more complex approaches, often based on advanced corpus statistics, have been introduced over the years, such as collocation network analysis, advanced multivariate statistics applied to concordance analysis and corpus frequencies, and so on. Further extension of these methods, and the development of more such complex techniques, is something we can expect to see over the next several years.

Further reading

Recommended introductory texts on corpus linguistics include Biber et al. (1998) and Hunston (2002), Sinclair (1991) and McEnery and Hardie (2012) are, in different ways, slightly more advanced texts. Key collections of research papers in the field of corpus linguistics include Aijmer and Altenburg (1991), Short and Thomas (1996), and Sampson and McCarthy (2004).

References

