CORPUS METHODS IN LANGUAGE STUDIES

Chapter · August 2018

CITATIONS
0

READS
4,250

1 author:

Richmond Ngula
University of Cape Coast
4 PUBLICATIONS **8** CITATIONS

SEE PROFILE

CORPUS METHODS IN LANGUAGE STUDIES

Richmond Sadick Ngula
Department of English, University of Cape Coast, Cape Coast, Ghana
rngula@ucc.edu.gh

INTRODUCTION

For decades, beginning perhaps from the grammatical studies of Danish scholar, Otto Jespersen, in the 1920s, much of the empirical studies into language has been based on a manual analysis of a few texts. In these early studies, the analysts had – by necessity – been constrained by the small amounts of text they could gather, handle and analyse effectively. Corpus linguistics, in the last two decades especially, has brought a massive boost, and a significant turn-around, to the empirical investigation of language. Owing to corpus linguistics, not only are analysts now able to explore, with relative ease, texts running into millions of words, they have also become aware of the fascinating insights that can be derived from the application of corpus methods to textual analyses: insights which were missed in a human-only analysis. As Hunston (2002, p. 1) notes, it is not an exaggeration to say that corpus linguistics has "revolutionised the study of language". What then is corpus linguistics? And what is its primary resource? According to Baker (2010, p. 93), "[c]orpus linguistics is an increasingly popular field of linguistics which involves the analysis of (usually) very large collections of electronically stored texts, aided by computer software". McEnery and Hardie (2012, p. 1) say that corpus linguistics deals with "some set of machine-readable texts which is deemed an appropriate basis to study a specific set of research questions". Corpus linguistics is therefore a methodology or an approach used to investigate linguistic phenomena rather than a sub field within linguistics, comparable to such areas as semantics, syntax, sociolinguistics, forensic linguistics, etc. Its primary resource is a corpus, whose definition can be safely gleaned from the corpus linguistics definitions offered by Baker (2010) and McEnery and Hardie (2012) above: a corpus is a large 'body' of texts stored electronically. For

any specific investigation, an analyst invariably works with a corpus or a set of corpora by first uploading it onto a corpus software, and then applying specific methods on the software, such as running frequency counts or concordance lists to derive results. It is clear, then, that the development of corpus linguistics has been facilitated by the role of computers. Indeed, the major advances in corpus linguistics are inextricably linked to advances in computer technology.

English is doubtless the language that has benefitted the most from the application of corpora for language studies and the reason is not farfetched: the field itself started and developed in English-speaking countries, primarily the United Kingdom and the United States of America (cf. McEnery & Hardie, 2012). Hence, tools were originally designed with English studies in mind. But there are now compilations and analyses of corpora in several other languages such as Chinese, French, Dutch, Danish, Maltese, Arabic, Urdu, Hindi, etc. – a fact which has totally diversified the field and given its practice a truly global outlook. The focus of this chapter is to examine the role corpus methods and corpora play in language studies. In the remainder of this chapter, some of the key theoretical issues around corpus linguistics are discussed, how to design and build a corpus is explained, and how to analyse a corpus (using one or two examples of corpus research to explain the process) is highlighted. The chapter ends with a concluding remark that makes a statement on the field's prospects. As I have a background in English linguistics, my discussion will be based on English corpus linguistics, hoping that readers interested in working with other languages can still benefit from the issues discussed.

Key Theoretical Issues in Corpus Linguistics

Empiricism and rationalism are long-standing, yet opposing, philosophical positions that seek to explain how new knowledge is acquired by humans. With the former, knowledge (or reality) is evidence-based and thrives on direct observation, experience and experimentation. By contrast, rationalism takes a mentalistic and an innate view, and suggests that knowledge is acquired intuitively through reason. Corpus linguistics is hinged on empiricism, and as an approach, its strength lies in the evidence derived from what corpus data may help us

understand about real occurrences of language use. Interestingly, a major opposition to corpus work came from a linguistic theory that foregrounded rationalism: Noam Chomsky's *Generative Grammar*. Chomsky's *Syntactic Structures* (1957), and later *Aspects of the Theory of Syntax* (1965), virtually revolutionarised the study of language as it succeeded in shifting the focus of linguistic inquiry from an external description of authentic language use (*Performance*) towards an abstract cognitive model that stressed a speaker's tacit, internalised knowledge of their language (*Competence*). Chomsky argued that a linguist's primary goal should be to build a model of linguistic competence, and, in his view, performance data could not be relied upon to achieve this goal. Once the generative movement was embraced by the linguistic fraternity, it implied also that corpus work was to become unpopular. In the words of McEnery and Wilson (2001, p. 6), "[a] corpus is by its very nature a collection of externalised utterances and, as such, it must of necessity be a poor guide to modelling linguistic competence". Unsurprisingly, by the 1960s and early 1970s especially, corpus linguistics had virtually been subdued. The impact was so great that "many early corpus linguists almost felt as if they had to work in secret cells" (Lindquist, 2009, p. 9).

But it did not take too long for corpus linguistics to regain popularity among linguists. From the 1980s, the Chomskyan criticism of performance data was not only shown to be overstated, the idea of *communicative competence* from Hymes (1972), which highlighted the role of context for any successful communication, had reinforced the value of authentic (corpus) data in the study of language. The renewed interest of linguists in corpora is stated by Meyer (2002, p. 1) who observes that "[l]inguists of all persuasion are now far more open to the idea of using linguistic corpora for descriptive and theoretical studies of language". If for nothing at all, corpora – more than methods based on introspection – offer objective and speedy analysis of linguistic items, give reliable frequency information, and allow researchers to be able to verify and replicate studies.

In contemporary corpus linguistics, one issue that has generated considerable debate within the field is the *corpus-based* vs. *corpus-driven* divide, first argued for by Tognini-Bonelli (2001). It is with regards to the theoretical contribution of corpus

linguistics that this divide has gained prominence. While many linguists agree that corpus linguistics is a methodology, some others working from the Firthian (and Sinclairian) framework of linguistics think that corpus linguistics is more than a methodology: it has a strong theoretical status, they would argue. This is the basis of the *corpus-based* and *corpus-driven* distinction. A corpus-based analysis explores a corpus with the primary aim of testing existing linguistic hypotheses or theories, especially if these were based more on introspection rather than on corpus evidence, to ascertain, based on corpus data, whether such hypotheses or theories can be supported, or may have to be modified or refuted. On the other hand, a corpus-driven analysis approaches a corpus with a more open mind, without an eye on existing hypotheses or theories. It aims to allow the corpus itself to drive the research and for the analyst to observe what is salient to explore in the corpus. Approaching a corpus this way helps to arrive at much stronger, and sometimes entirely new, theoretical conclusions.

Not everyone supports the distinction between corpus-based and corpus-driven linguistics. McEnery and his team (e.g. McEnery & Wilson, 2001; McEnery & Hardie, 2010; McEnery & Hardie, 2012), for instance, think that all corpus linguistic work should be characterised as corpus-based. Generally, it seems the linguists in favour of this distinction, and who particularly have a stronger inclination towards corpus-driven analysis, have been the followers of John Sinclair's work – including Tognini-Bonelli herself, Stubbs, Hunston, Hoey, Krishnamurthy, Teubert, among others – at the University of Birmingham in the UK (see McEnery & Hardie, 2010). But perhaps the important point to note is that linguists on both sides of the divide have, over many years, healthily co-existed and worked together, sharing ideas at the same conferences and publishing research findings in the same journals. Indeed, recent corpus studies tend to apply key ideas from the two camps (e.g. Baker, 2014), where in some of Baker's chapters the analysis is corpus-based, while in others he follows a corpus-driven approach.

Corpus Design and Construction

Compiling a corpus when existing ones are not suitable

To start studying any linguistic item in a corpus, there must first be a corpus. Corpora (by their very nature as texts processed and stored in digital form) once compiled can be used by many other researchers under certain conditions. To work with an already-existing corpus, you should be sure that the corpus you have in mind is available and can be accessed for your study. Besides, the corpus should be one that is suitable enough to address your specific research questions. But these two conditions may not always be met, as not all existing corpora are publicly accessible, and not every corpus might usefully be able to address your research questions. So, you might necessarily be required to design and construct your own corpus for some specific research goal. What does designing and constructing a corpus entail?

First, it must be decided what type of corpus is to be constructed. There are a variety of corpus types which, due to the scope to be covered in this chapter, I am unable to discuss fully. Luckily, there are excellent introductory textbooks on corpus linguistics that offer detailed explanations on the various types of corpora (see, for example, Hunston, 2002; McEnery, Xiao & Tono, 2006). Here, I will simply explain the two most common types; namely, *general* and *specialised* corpora, and then focus on how specialised corpora are designed and constructed since they are the type individual researchers utilise quite often and can easily compile on their own.

A general corpus is one that includes a variety of text types in its compilation. It may contain written texts, spoken texts, or both, and very often it represents a national, regional or sub variety of a language. There are several general corpora of approximately a million words, such as the Lancaster-Oslo-Bergen (LOB) written corpus, and others of a much bigger size that include both written and spoken texts, such as the over 450 million-word Contemporary Corpus of American English (COCA). Constructing a general corpus can be quite a task and therefore it very often requires a collaboration of researchers and/or institutions.

Designing and compiling a specialised corpus

A specialised corpus, in contrast to a general one, targets one text type (or genre), say, political speeches, newspaper editorials, master's theses, or business letters. Because of its narrowed text focus, a specialised corpus is usually smaller in size compared to a general one, yet some specialised corpora are quite large and have been compiled by a team of researchers as well (e.g. the 1.8 million Michigan Corpus of Academic Spoken English (MICASE), or the 5 million Cambridge and Nottingham Corpus of Discourse in English (CANCODE)). Depending on an analyst's research aim and questions, a specialised corpus can be much smaller. For example, Handford and Matous (2011) compiled a 13, 000 -word (preliminary) corpus of construction industry discourse to study interaction features in that context; Stubbs (2005) compiled a specialised corpus of less than 40, 000 words (i.e. Joseph Conrad's *Heart of Darkness*) for a corpus stylistic study; and Baker (2006) compiled a 130, 000-word corpus of debates in the House of Commons to study discourses around foxhunting. At first sight, it might seem unworthy to build (very) small, specialised corpora, such as the above examples, given that corpora are now becoming even larger. But as Koester (2010, p. 67) writes, the point of such small corpora is that 'they allow a much closer link between the corpus and the contexts in which the texts in the corpus were produced', noting further that '[w]ith a small corpus, the corpus compiler is often also the analyst, and therefore usually has a high degree of familiarity with the context'. So, what are the main issues and/or the stages in constructing a (specialised) corpus? The process usually involves designing, gathering and processing relevant texts, and possibly annotating the corpus.

Designing a corpus

For a start, you have to design the corpus by planning, deciding, and generally putting up a framework to guide the gathering and processing of texts for the corpus. Designing a corpus is thus much like designing a plan for a building construction. Design procedures for a specialised corpus trigger a few relevant questions to ask and to attempt giving tentative answers. What text type (or genre) and which author is involved? What would be the size of the corpus? For a

specialised corpus, you probably would be able to decide easily the text type your corpus would contain. But would the text collection be based on text excerpts or full texts? How many text samples would be included? Which publication dates for texts qualify to enter the corpus? Would the texts to be collected require any permissions? A corpus compiler should think of the answers to give to such questions before proceeding to collect the relevant texts for inclusion in the corpus. When I wrote my own PhD thesis (Ngula, 2015), it was based on a specialised corpus of research articles (RAs) I constructed, and I had to answer such design questions. In my PhD research, I set out to explore epistemic rhetorical resources of argumentation in RAs written in English by Ghanaian scholars based in Ghana in the disciplines of Sociology, Economics and Law, and to compare these RAs with similar RAs written by international scholars who are native speakers of English. When I discovered that already-existing RA corpora would not usefully address my research questions, I decided to construct my own corpus. It was obvious I needed a corpus of RAs in the three disciplines for the two groups of authors. I decided to build an overall RA corpus of approximately 1 million words to have sub parts for the three disciplines of the two groups of authors. Given that the linguistic items I wanted to study (i.e. epistemic modality devices) occur quite frequently in RAs, I thought that 1 million words would be sufficient to reveal useful tendencies and findings to address my research questions. I planned that I would include 20 articles for each of the three disciplines on the Ghana side and on the international scholars' side, so that I estimated 120 RAs in total.

Another decision I made was that I would collect full RA texts rather than RA excerpts. If I wanted to study specific sections of the RA – such as the introduction section or the discussion section – then it would have been ideal to collect only those sections of the RA for the corpus. In my case, I wanted to explore occurrences of epistemic devices in the entire RA, and to study disciplinary and discourse community variation. So, collecting the full texts was the way to proceed as I was not looking at the feature in specific sections of the RA. Furthermore, I planned that the dates of publication for all the RAs (of both groups) to enter my corpus would be from the year 2000 to 2010. Two reasons

informed this decision: the first was that I wanted my RA corpus to reflect contemporary usage and I thought this year range worked quite well; and secondly, my pre-design checks suggested that, for both groups of scholars, I could relatively easily obtain sufficient RA texts published in this period. And since published RAs are already in the public domain, I did not have to obtain any permissions to gather the relevant RAs for my corpus.

Certain kinds of texts cannot be obtained easily without (written) permission, and sometimes organisations and individuals are reluctant to release confidential textual material even when corpus compilers seek permission to use such texts. A good example is Handford (2010) who recounts how challenging it was to record business meetings for his corpus. He notes that companies were not easily "persuaded … to allow recording, with roughly 95 per cent of companies who were approached refusing permission. Companies were especially concerned about confidentiality" (Handford, 2010, p. 4).

The design decisions just discussed above are important to the extent that the texts a corpus builder gathers for his corpus should be *representative*, *balanced* and *sampled*. Representativeness, balance and sampling are related features to consider in constructing a corpus. They are especially unavoidable principles in the construction of general and other types of corpora. But they may also be applied when constructing very specialised corpora. According to Biber (1993, p. 243), *representativeness* is to do with "the extent to which a sample includes the full range of variability in a population". *Balance* is achieved if the full range of genres or text types is included in the corpus. The way each text excerpt or full text is selected for inclusion in the corpus is called *sampling*. Baker (2010, p. 96) is of the view that "[b]ecause a corpus ought to be representative of a particular language, language variety, or topic, the texts within it must be [sampled] and balanced carefully in order to ensure that some texts do not skew the corpus as a whole". Representativeness, balance and sampling – in the building of general corpora – are not a simple and straightforward matter, and Biber (1993) and McEnery et al. (2006) offer detailed discussions on their nuances.

These principles can also be applied, more confidently, when constructing a specialised corpus. It is even possible to achieve total representativeness in some cases. If, for example, we wanted to construct a corpus of novels written by Chinua Achebe or Ayi Kwei Armah, we would simply have to include all the novels of these authors in wholes to achieve 100 per cent representativeness. In such a situation, there can be no skewing, and issues of sampling and balance will not even arise as the entire 'population', as it were, has been included in this case. However, if we take my specialised corpus of RAs once again, sampling and balance had to be carried out to ensure that the corpus was representative. As I designed my corpus, I knew it was not possible to include all RAs that met the collection criteria (i.e., the 'population'). Hence, I decided I would choose equal samples of RAs in the three disciplines of Sociology, Economics and Law for the two groups of scholars I wanted to study. This meant applying the concepts of sampling and balance, in the hope of maximising representativeness. Overall, design decisions help the text collection process to proceed smoothly when it starts. But it has to be mentioned that when text collection starts, following the design put in place, practicalities on the field could lead to slight or even major changes to the design. These on-the-field realities and possible changes in original design reflect the fact that "corpus building is of necessity a marriage of perfection and pragmatism" (McEnery et al., 2006, p. 73).

Text collection and processing

I turn now to the collection and processing of authentic texts for storage in digital (or machine-readable) form. This is where the real corpus compilation starts. I will still focus on specialised written corpora and draw on my own RA corpus compilation to explain the process. But let me begin with a general point of note. Collecting and processing spoken texts is more arduous a task than doing same for written texts mainly because of the extra work of recording and orthographically transcribing speech, which sometimes may require detailed extra linguistic mark-ups. This explains why in most cases a corpus which includes both written and spoken texts has the written component being larger in size. A clear example is the approximately 100-million-word British National Corpus (BNC) which is made up of 90 per cent written and only 10 per cent spoken.

Text collection starts with capturing and having every relevant text in electronic form, if it is not already in that format. Indeed, a lot of texts are in print or non-print (handwritten) format only and for those you have to either scan the texts or simply key in the relevant texts by typing them out directly onto the computer. While this latter process of keyboarding can be time consuming and very tedious, it is recommended when especially you are collecting non-printed texts, such as student handwritten essays. This is because OCRs hardly capture handwritten texts when scanned. But a general problem with keyboarding texts (printed or non-printed) is the potential for typing errors to occur, especially when large amounts of texts are being typed. Post-typing editing is therefore often needed to ensure that typed texts mirror exactly the original texts. Scanning texts is a lot faster and preferred where large amounts of printed texts not already in electronic form are involved. Normally a page is scanned at a time, and depending on the effectiveness of the OCR being used, nearly every character is captured when a scan is completed. Thus, with OCR scanning too there may be the need for minor editing after the texts are scanned.

In recent years, however, a much more easy and effective approach to compiling texts for a corpus is made possible because of the existence of many text storage sites on the internet. If the texts to be collected are already in electronic form, it would be unnecessary altogether to scan or type texts. As Baker (2006, p. 31) has pointed out, "due to the proliferation of internet use, many texts which originally began life in written form can be found on websites or internet archives". So, for example, when I constructed my RA corpus, all the relevant RA texts produced by the international scholars were already digitalized, as they were in E-journals that I could easily access online. I therefore downloaded the texts as *pdf* files for further processing. For RA texts produced by the Ghanaian scholars in Ghana, some of the texts were not available in digital form and for those I used a scanner and an OCR software to digitalise them.

The next important thing to do after text capture in digital form relates to the appropriate text file format and text filing name to use to store texts. With regards to file format, it is most suitable to save corpus text files as *plain text* on the

computer by clicking on the drop-down 'Save As' menu to select this option. This is because, as Reppen (2010) says, most corpus analysis software at present work best with this format, although other options like *Rich text* and *XML* can be used. It should be stressed that corpus analysis tools will not read digital texts in *pdf* or *word* format, and so when texts in these formats are downloaded from the internet, they ought to be converted to (or saved as) *plain text*. On text file naming, the compiler should decide on, and use, an appropriate naming system that considers the main features of the text, e.g., text number, genre, author name, speaker sex, etc. The importance of file naming is underscored by Reppen (2010, p. 33) who notes that file names should "clearly relate to the content of the file to allow users to sort and group files into sub categories or to create sub corpora more easily". In my RA corpus, for example, I named the Sociology, Economics and Law texts produced by the international scholars who are native authors using file names as follows: *SOC NA01*, *SOC NA02*, etc.; *ECO NA01*, *ECO NA02*, etc.; *LAW NA01*, *LAW NA02*, etc. to respectively reflect the discipline, the fact that writers are native authors and the text number. Once proper file names are given to the texts of the corpus, the texts can be stored in a folder on your computer and used for the intended analysis.

3.5 *Raw or annotated corpus*?

A corpus is in its raw state once the texts it contains have been processed as plain texts and given appropriate file names. A raw corpus is in a good enough shape to be used for many kinds of corpus analysis. However, one further step can be added to the processing of a corpus before it is used for analysis. This is referred to as corpus annotation. Annotating a corpus means adding valuable linguistic information to the corpus. Leech (1997, p. 2) defines corpus annotation as "the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data". Different types of corpus annotation are possible, including syntactic annotation, pragmatic annotation and prosodic annotation (see Garside, Leech & McEnery, 1997 for a detailed discussion of the various types and importance of corpus annotation). But perhaps the most basic and utilised type is grammatical word annotation (also known as *part-of- speech*

tagging or *POS* tagging). Grammatical word annotation involves assigning tags or labels to every word in the corpus to indicate its part of speech or grammatical function. This can be done manually if the corpus is of a small, manageable size. However, there are now automatic tagging programs that do the job relatively easily although they may not produce a 100 per cent tagging accuracy. A good example of automatic taggers, which have been used to tag several (English) corpora including the BNC, is the Constituent Likelihood Automatic Word-tagging System (CLAWS) (Garside, Leech & Sampson, 1987). CLAWS has been consistent in achieving a 97 per cent (or more) accuracy rate. This means that when the tagging is completed, post-tagging editing may be helpful to obtain total accuracy. Table 1 is a sample of the CLAWS tag set (version 5).

Table 1. Sample part-of-speech (POS) tagset (CLAWS, v.5)

Tag	Description	Examples
AJ0	Adjective (general or positive)	good, light
AJC	Adjective (comparative)	better, lighter
AJS	Adjective (superlative)	best, lightest
NN0	Common noun (neutral for number)	sheep, fish
NN1	Singular common noun	cat, light
NN2	Plural common noun	cats, lights
PN1	Indefinite pronoun	everyone, somebody
PNP	Personal pronoun	you, they
PNX	Reflexive pronoun	myself, himself
VVZ	The -s form of lexical verbs	sends, lights

POS tagging a corpus is often of great help to an analyst as it can simplify work in many instances. If, for example, a query (or corpus search) of the word *light* or *lights* is done on an untagged corpus, it will retrieve all instances of either of these words in the corpus. But the analyst might be interested in only instances of *light* as an adjective (not as a noun) or *lights* as a verb (not as a noun). In such a

situation, the analyst has to painfully inspect – especially if the query is done on a very large corpus – all hits (the query output) and delete all the unwanted noun uses. In a POS tagged corpus, however, the query can right away be restricted: the query for *light* as a general adjective will be *light_AJ0*, and the query for *lights* as a verb will be *lights_VVZ*.

Analysing a Corpus

Corpus analysis software

Once compiled and stored electronically, a corpus can be subjected to all kinds of linguistic analysis. Corpus analysis is facilitated by corpus analysis software tools which have been improving by way of sophistication since the 1980s. Three of the well-known corpus tools in use now are *WordSmith* (Scott, 2013), *AntConc* (Anthony, 2005) and *ConcGram* (Greaves, 2009). These can be downloaded onto a researcher's computer either free of charge (e.g., AntConc) or at a subscribed fee (e.g., WordSmith). These tools are built in a way as to allow a researcher to upload his or her own corpus for analysis. A more sophisticated tool, *Sketch Engine* (Kilgarriff, Rychly, Smrz & Tugwell, 2004), is a web-based tool, and it has a good number of already-existing corpora of different languages on it, and a researcher can still upload his or her own corpus onto the tool. Sketch Engine allows a free trial version where a researcher is allowed access to a small set of corpora on the tool for analysis. However, to access the tool in its entirety and use all or any of the corpora on it, as well as upload your own corpus, subscription at a fee is required. The AntConc tool, which is freely accessible, is a popular tool for beginners.

Every corpus analysis tool comes as a package that has separate independent functions, each of which can be exploited for analysis. Thus, even though corpus linguistics is often said to be a methodology, in the strictest sense, it involves a variety of independent (analysis) methods, including frequency lists, keyword lists, concordance analysis, cluster/n-gram analysis, and list of collocates and collocational analysis. All these analysis methods are part of, for example, the AntConc or WordSmith package, and any corpus linguistic investigation may not make use of every method. When I looked at epistemic devices in RAs (Ngula,

2015), I used only two methods, frequency lists and concordance analyses, in the WordSmith package. I will, in the next section, briefly discuss the methods for frequency, keyword and concordance analysis.

Frequency lists, keyword lists, concordances

Generating raw frequency lists, keyword lists, and exploring concordance lines are perhaps the most fundamental kinds of analysis for any researcher using corpus methods. These methods often lead a researcher to examine more complex and exciting linguistic items and patterns in a corpus. The word list facility is mainly used to generate a list of all the words in a corpus, ranking them in order of frequency. Frequency is a very important concept in linguistic analysis, as researchers sometimes want to know the most frequent words in a corpus. A simple word list carried out on a corpus invariably shows, for example, that the most frequently occurring words are functional or grammatical words. For example, a simple word list run on my Economics RA sub corpus by native authors, using WordSmith's Wordlist tool, retrieved the top ten words in Table 2.

Table 2. Top ten words in Economics RA corpus

N	Word	Freq.	%
1	THE	9, 829	6.63
2	OF	5, 590	3.77
3	IN	3, 857	2.60
4	AND	3, 668	2.47
5	TO	3, 306	2.23
6	A	2, 904	1.96
7	IS	2, 612	1.76
8	THAT	2, 012	1.36
9	FOR	1, 941	1.31
10	ARE	1, 201	0.81

While these functional words may not contribute much to telling what the Economics RA texts are about, their high frequency reveals their grammatical and

textual role: they make the grammar of texts. It is not surprising that *the* is invariably the most frequent word in any corpus. Calculating keywords is a related form of frequency analysis, but unlike a simple word list, a keyword list tells what the texts in a corpus are essentially about. Baker (2010, p. 104) defines a keyword as "a word which occurs statistically more frequently in one file or corpus, when compared against another comparable or reference corpus". The results of a keyword list for the Economics RA sub corpus, in Table 3, show a markedly different list of words, compared to the simple word list in Table 2.

Table 3. Top ten keywords in Economics RA corpus

Rank	Freq.	Keyness	Keyword
1	713	2354.264	Price
2	472	1483.621	Model
3	289	964.984	Firms
4	383	932.154	Level
5	276	851.206	Growth
6	266	793.626	Effects
7	305	789.433	Changes
8	202	784.506	Inflation
9	275	765.243	Firm
10	261	737.095	Data

As Table 3 makes clear, the keywords give a sense of what the text is about. We can safely relate these words to Economics. The reason the grammatical words in Table 2 disappear in the keyword list is because they are not unusually frequent in the target corpus (the Economics RAs) when compared with the reference corpus used (the LOB). In other words, they are just as frequent in the target corpus as they are in the reference corpus. Keyword rankings are based more on the keyness values rather than on frequency values: the more the keyness value, the more key the word. As Table 3 shows, *price* is the most salient keyword with the highest keyness value. Exploring concordance lines is another fundamental corpus analysis perspective, and while concordances of specific words and phrases can be

generated, simple word lists and keyword lists can be further analysed in terms of concordance outputs. According to Baker (2010, p. 106), "[a] concordance is simply a list of a word or phrase, with a few words of context either side of it, so that we can see at a glance how the word tends to be used"

A concordance thus helps us to study words in context; hence it is also referred to as key word in context (KWIC). Corpus analysis tools allow for concordance lines to be sorted variously so that meanings and patterns associated with words can be more effectively arrived at. When I explored epistemic modal verbs in RAs written by Ghanaian and international scholars (see Ngula, 2015), one noticeable finding in the Law articles of the international writers is that the modal *may* very often co-occurred with *well* to mark 'epistemic probability', a slightly higher degree of epistemicity than when only *may* is used. The Ghanaian Law writers, however, did not use *may well* at all, although they used *may* alone to express 'epistemic possibility'. It was after various sorting and a close inspection of concordance lines that this finding became apparent. Figure 1 is a sample concordance of the *may well* pattern in the Law RAs of the international writers.

N	Concordance	File
1	adhere rigidly to a formalist approach may well threaten to undermine one of	LAW NA14.txt
2	. From a nation-state perspective, this may well seem a defect, although	LAW NA17.txt
3	again of the eggshell skull rule). This may well represent an explicit higher	LAW NA16.txt
4	directly as moral issues. Then we may well not think that courts are the	LAW NA20.txt
5	distinction that economic transactions may well "involve a fundamental public	LAW NA02.txt
6	bonds of varying intensity. These links may well include ethnic, religious,	LAW NA17.txt
7	from one another. Indeed, they may well have contributed to the	LAW NA07.txt
8	litigation under the Hague Convention may well have been avoided had the	LAW NA07.txt
9	contractual claims. ICSID jurisdiction may well extend to purely contractual	LAW NA02.txt
10	allowing legitimacy-based claims may well entail defeat, at least in the	LAW NA14.txt
11	objectives, a legitimacy-based system may well ensure greater overall	LAW NA14.txt
12	return proceedings in the first instance may well be discouraged from	LAW NA07.txt
13	. Under such circumstances, pluralism may well be better served by dividing	LAW NA17.txt
14	arise in moral life. Such presentation may well be artificial compared to	LAW NA20.txt
15	that this is not always the case. There may very well be a growing consensus	LAW NA11.txt

Figure 1: Screenshot of the *may + (very) well +V* pattern in Int. Law RAs

From concordance lines, one can effectively study how a word is used, uncover patterns associated with a word or phrase, determine discourses (representations

or meanings) around a word, and so on. The kinds of insight derived from a reading of concordance lines may be easily missed in a manual analysis.

Conclusion

In this chapter, corpus linguistics has been discussed as a research methodology for studies in language, addressing issues of its theory, methods and procedures, and practice. There is no doubt that corpus linguistics has brought a massive boost to the study of language. Its theoretical credentials are assured, its results are accurate, insightful and objective, and its applications are now attested in nearly every sub field of linguistics including lexis, grammar, discourse, pragmatics, sociolinguistics, stylistics, register linguistics, and many more. Even theoretical linguists, who before would have nothing to do with corpora, now see interesting ways the approach can enrich their work (McEnery & Hardie, 2012). Considering the current trends of the approach, it is most likely that the future of corpus linguistics will see a greater sophistication of corpus analysis tools, the building of much larger corpora, and an expansion of applications, especially to many other languages. It seems an even more promising future awaits this versatile approach to language studies.

References

Anthony, L. (2005). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In Proceedings of IWLeL 2004: An interactive workshop on language learning (pp. 7 – 13). Tokyo, Waseda University.

Baker, P. (2006). 'The question is, how cruel is it?' Keywords, foxhunting and the House of Commons. *AHRC ICT Methods Network, Centre for Computing in the Humanities*.

Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti (ed.), *Research methods in linguistics* (pp. 93–113). London: Continuum.

Baker, P. (2014). *Using corpora to analyze gender*. London: Bloomsbury.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge MA: MIT Press.

Garside, R., Leech, G. & McEnery, A. (eds.) (1997). *Corpus annotation*. London: Longman.

Garside, R., Leech, G. & Sampson, G. (eds.) (1987). *The computational analysis of English: A corpus-based approach*. London: Longman.

Greaves, C. (2009). *ConcGram: A phraseological search engine.* Amsterdam: John Benjamins.

Handford, M. (2010). *The language of business meetings*. Cambridge: Cambridge University Press.

Handford, M., & Matous, P. (2011). Lexicogrammar in the international construction industry: A corpus-based case study of Japanese-Hong-Kongese on-site interactions in English. *English for Specific Purposes*. 30, 87–100.

Hardie, A., & McEnery, T. (2010). On two traditions in corpus linguistics, and what they have in common. *International Journal of Corpus Linguistics*. 15(3), 384–394.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (eds.). *Sociolinguistics* (pp. 26 –293). London: Penguin.

Kilgarriff, A., Rychly, P. Smrz, P., & Tugwell, D. (2004). *The sketch engine*. Proceedings of Euralex. Lorient, France, July: 105–116.

Koester, A. (2010). Building small specialised corpora. In A. O'keeffe & M. McCarthy (eds.). *The Routledge handbook of corpus linguistics* (pp. 66– 79). London: Routledge.

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech & A. McEnery (eds.). *Corpus annotation* (pp. 1–18). London: Longman.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh: Edinburgh University Press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Methods, theory and practice.* Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd ed.). Edinburgh: Edinburgh University Press.

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

Meyer, C. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Ngula, R. S. (2015). *Epistemic modality in social science research articles written by Ghanaian authors: A corpus-based study of disciplinary and native vs. non-native variations.* Unpublished *Doctoral* dissertation, Lancaster University, Lancaster, UK.

Reppen, R. (2010). Building a corpus: What are the key considerations? In A. O'Keeffe & M. McCarthy (eds.). *The Routledge handbook of corpus linguistics* (pp. 31–37). London: Routledge.

Scott, M. (2013). *WordSmith tools* (version 6.0). Oxford: Oxford University Press.

Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature*. 14(1), 5–24.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.