University of Huddersfield

Corpus Linguistics: An Introduction

1. Introduction.

Corpus Linguistics is a hugely popular area of linguistics which, since its beginnings in the late 1950s, has revolutionised our understanding of language and how it works. Our aim in this handout is to provide an introduction to some of the basic ideas and methods of Corpus Linguistics.

2. What is Corpus Linguistics?

Within the Corpus Linguistics community, there are two main views on how Corpus Linguistics should be defined. Some linguists see Corpus Linguistics simply as a **methodology** for studying large quantities of language data using computer software. Hardie and McEnery (2010) refer to corpus linguists who take such a view as 'methodologists'. Other linguists, however, see Corpus Linguistics as a **sub-discipline of linguistics** in its own right, concerned with explaining the relationship between meaning and structure in language. Hardie and McEnery (2010) call such people 'Neo-Firthians', because their work builds on principles established by J. R. Firth, the first British Professor of Linguistics. In this handout, we adopt the methodologist position, partly because this is the view of Corpus Linguistics that we take ourselves, and partly because we think it is easier to understand the differences between the two views (which are often set in opposition to each other) if you first familiarise yourself with the methods that corpus linguists use.

3. What is a corpus?

A corpus is a collection of texts that have been selected to be representative of a particular language or language variety. Corpora (the plural of 'corpus') are stored electronically to facilitate analysis using computer software. Corpora are now commonly used within linguistics as sources of data for language analysis. Many professionally produced corpora are commercially available though it is also relatively easy to create your own. Here are some examples of particularly well-known corpora:

The BNC (British National Corpus) A 100 million word corpus of written and spoken British English from the early 1990s, produced by the universities of Lancaster and Oxford.

The Brown Corpus A one million word corpus of early 1960s written American English, produced by Nelson Francis and Henry Kučera at Brown University in the USA and composed of 2000-word samples of text from a variety of genres. This was the first electronic corpus to be produced.

The LOB Corpus A one million word corpus of early 1960s written British English produced by the universities of Lancaster, Oslo and Bergen (hence, LOB). It was designed to match the Brown Corpus, so as to facilitate comparative analysis of British and American English.

The Survey of English Usage An ongoing project at University College London to collect and analyse large quantitative of naturally occurring language, which has resulted in numerous corpora. However, the Survey also gives its name to one the earliest corpora to be built. Randolph Quirk's 'Survey Corpus' is a one million word corpus of written and spoken British English produced between 1955 and 1985. Quirk was the first Director of the Survey and the original Survey Corpus pre-dated modern computers and was stored on file cards. Later on, an electronic version of the spoken data in the corpus was produced in collaboration with Lund University in Sweden, and is known as the **London-Lund Corpus**.

The Bank of English A monitor corpus (i.e. a corpus that is constantly being added to), produced initially under the direction of John Sinclair at the University of Birmingham, and currently containing approximately 650 million words. The Bank of English was used to create the Collins COBUILD series of dictionaries and grammars (see section 13).

4. Why use a corpus?

The advantages of using corpus data (as opposed to, say, the invented examples traditionally favoured by some linguists) include the following:

- Corpora allow linguists to access quantitative information about language, which can often be used to support qualitative analysis.
- Insights into language gained from corpus analysis are often generalisable in a way that insights gained from the qualitative analysis of small samples of data are not.
- Using corpus data forces us to acknowledge how language is really used (which is often different from how we think it is used).

5. Sampling and representativeness: the key to good corpora.

The key to producing a good corpus is to ensure that it is indeed representative of whatever language variety you are trying to represent. For example, if you wanted to build a corpus of Yorkshire English, you would need to decide what time frame you are going to sample, specify whether you are collecting speech, writing or both, decide on the source of your language data, and so on. If you decided you wanted your corpus to be a reflection of Yorkshire English generally, then a corpus which included only speech from males born in Yorkshire between 1950 and 1960 would not be representative of this (though if you sampled it properly, it might well be a good representation of Yorkshire English from that particular time period and those particular speakers). We will come back to the issue of sampling and representativeness later on in section 11, when we discuss the practicalities of building your own corpora.

6. Adding value via annotation and mark-up.

Sometimes, corpus linguists find it useful to tag their corpus data. Tagging can take the form of both annotation and mark-up. Annotation refers to the addition of metalinguistic information to corpus texts. For example, codes (generally known as 'tags') can be inserted into the corpus to indicate the part-of-speech of each word. For instance, **the_AT** would indicate that the word *the* is an article (marked by the tag AT) while **popularity_NN1** indicates that *popularity* is a singular common noun. Mark-up is similar in principle to annotation but involves tagging textual and contextual elements rather than metalinguistic features. Mark-up can be used to indicate attributes such as the identity of the speaker of a particular stretch of language (in a corpus of speech, say), the genre of the constituent texts of a corpus, or graphological features of texts such as paragraph breaks, bold and italic font, font size and much more.

7. A brief history of Corpus Linguistics.

Although the term Corpus Linguistics was not commonly used until the late 1960s, some linguists had been engaged in what might be seen as early forms of this work since the beginning of the twentieth century. However, following the publication of Noam Chomsky's *Syntactic Structures* in 1957, the Chomskyan approach to language, with its focus on non-empiricist methods, quickly became the dominant paradigm within linguistics, particularly in the USA. A consequence of this was that the study of naturally-occurring language (which is, of course, what corpus linguists study) was to some extent sidelined.

However, despite Chomsky's massive influence on linguistics, some linguists continued to pursue the idea of studying language by describing naturally-occurring data (Sociolinguistics, for example, is a sub-discipline of linguistics that is particularly interested in language performance rather than language competence). This was partly in response to the fact that Chomsky's ideas about linguistic structure, which were illustrated by invented examples, often turned out not to adequately describe instances of real language in use.

Surveying English.

The first significant project in what we might term modern Corpus Linguistics was the Survey of English Usage, instigated by Randolph Quirk at University College London in 1959. The Survey led to the creation of a corpus of one million words of written and spoken British English, made up of 200 text samples of 5000 words each. The corpus data was all on paper and indexed on file cards. Each text sample was manually annotated for prosodic and paralinguistic features, and file cards for each sample recorded grammatical structures. Searching the corpus thus meant a trip to the Survey to physically search through the many filing cabinets in which the corpus data was stored.

Building the Brown corpora

Shortly after the Survey of English Usage had been initiated, Nelson Francis and Henry Kučera of Brown University in the USA began work on what would eventually become known as the Brown Corpus. This was a one million word collection of written American English, compiled from texts published in 1961. Unlike the Survey Corpus, the Brown Corpus was electronic. Kučera and Nelson published an analysis of the corpus in 1967, titled *Computational Analysis of Present-Day American English*. One of their findings was that the frequency of words in the corpus was inversely proportional to their rank position. For example, the most frequent word in the corpus is *the*, which accounts for approximately 6% of all the words in the corpus. The next most frequent word is *of*, which accounts for approximately 3% of all the words in the corpus – i.e. half of the percentage frequency of *the*. Kučera and Nelson's findings confirmed what is known as Zipf's Law (Zipf 1935), which states that the most frequent word in a given corpus will occur approximately twice as often as the second most frequent, which will occur approximately twice as often as the third most frequent, and so on.

In 1970, Geoffrey Leech, together with a group of colleagues at Lancaster University, embarked on the creation of a corpus of written British English, to match the structure of the Brown Corpus. By 1976, the corpus was still incomplete, due in no small part to the difficulties of obtaining copyright permission to use the corpus's constituent texts. Eventually, through the combined efforts of three universities – Lancaster, Oslo and Bergen – the corpus was finished in 1978 and became known as the LOB corpus, after the initial letters of the universities involved in its creation. Because LOB paralleled the structure of the Brown corpus, it was ideal for comparing British and American English. Later on, the Brown family of corpora was extended by the addition of FROWN (Freiberg-Brown; written American English from 1991), the charmingly titled FLOB (Freiberg-Lancaster-Oslo-Bergen; written British English from 1991) and, most recently, BE06 (written British English from the early years of the 21st century).

Making sense of meaning

While corpus linguists such as Quirk, Jan Svartvik (who had also worked on the Survey), Kučera and Nelson, Leech and Stig Johansson (another member of the LOB team), were interested primarily in grammar, John Sinclair, the first Professor of Modern English Language at the University of Birmingham, was primarily interested in what corpora could reveal about meaning. His earliest investigations in this area had taken place prior to his appointment at Birmingham, in a project known as *English Lexical Studies* undertaken at the University of Edinburgh in 1963 with his colleagues Susan Jones and Robert Daley. Sinclair analysed a small corpus of spoken and written English with the aim of investigating the relationship between words and meaning. This was the beginning of Sinclair's conviction that words are not the core unit of meaning in language. If this sounds counter-intuitive, it is because most people's understanding of how language works has largely been shaped by traditional accounts of grammar. In *English Lexical Studies* (recently republished by Krishnamurthy 2004) Sinclair began to explore the idea that meaning is best seen as a property of words in combination. For example, the negative meanings associated with the word *commit* are not inherent in the word itself but arise from the fact that the word *commit* is often found in descriptions of negative events (e.g. 'commit crime'). In this work, Sinclair was building on J. R. Firth's concept of collocation (the notion that some words have a tendency to co-occur with other words more often than would be expected by chance alone). Sinclair's work on collocation and what corpora can reveal about meaning were instrumental in the COBUILD (Collins Birmingham University International Language Database) project which he initiated at Birmingham in 1980. COBUILD resulted in the creation of the Bank of English corpus which was the data source that underpinned a series of dictionaries and grammars edited by Sinclair and his team, including the *Collins COBUILD English Language Dictionary* (1987), the *Collins COBUILD English Grammar* (1990) and the *Collins COBUILD English Dictionary* (1995).

Advances in annotation

In 1991, a consortium including research teams from Oxford University, Lancaster University and the British Library began work on the British National Corpus (BNC). This corpus consists of 100,000,000 words of written and spoken British English from the later part of the 20th century and was completed in 1994. Part of the process of creating the BNC involved annotating the data for part-of-speech. This process, commonly known as tagging, involves assigning a code (or 'tag') to every word in the corpus to indicate what part-of-speech it is (e.g. possessive pronoun, coordinating conjunction, wh-determiner, etc.). Part-of-speech tagging is done automatically using specialist software and has been a staple of methodologist corpus linguistics since the days of the Brown corpus, when early tagging software achieved a success rate of around 70% on the Brown data. Tagging of the BNC was done using a part-of-speech tagging system called CLAWS (Constituent Likelihood Automatic Word-tagging System), which has a success rate of around 97%.

Tagging is not restricted to part-of-speech information. Indeed, any aspect of language can be tagged in a corpus, though whether this can be done automatically depends on whether the focus of the annotation is a formal feature of language. For example, it is relatively straightforward to tag for part-of-speech, since the grammatical rules of a language can be used to determine the likelihood of a word belonging to a particular part-of-speech category. Tagging for, say, speech presentation, on the other hand, is much more difficult to do automatically, since categories such as direct speech (*He said 'I love corpus linguistics'*) and indirect speech (*He said that he loved corpus linguistics*) are functional rather than formal categories; consequently, annotation of this type needs to be done manually. The value of annotation, however, is that it allows us to quantify other features of language (for example, prosody, clause type, etc.) which can give us insights into how language works as a system.

Recent advances in tagging have involved the development of a system for annotating corpus data for semantic information. Paul Rayson's *WMatrix* software package automatically assigns a semantic tag (as well as a part-of-speech tag) to every word in whatever corpus is uploaded. This allows the user to search the corpus not just for words but also for particular semantic fields.

Looking ahead

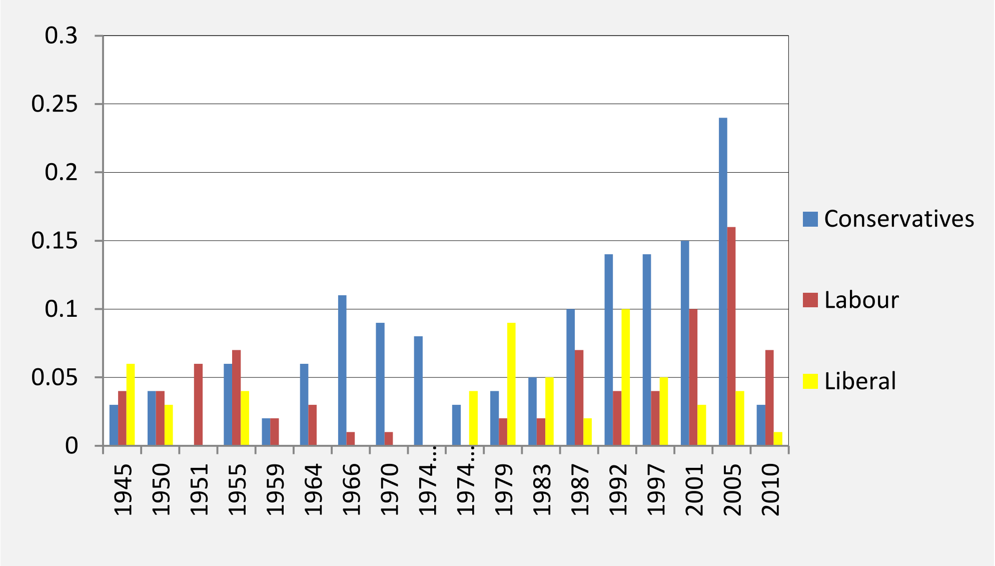
Advances in computing power are having a significant effect on the development of tools and technologies for Corpus Linguistics. At the same time, corpus techniques are finding their way into other disciplines and the broader field of Digital Humanities is developing fast. Recent advances include the development of multimodal corpora such as that built as part of the University of Nottingham's *Headtalk* project (Knight et al. 2008). This focused on the construction of a corpus of video streams of an academic supervision session, marked-up to indicate aspects of body language. This allowed researchers to investigate the relationship between language and gesture. And a project at Lancaster University is focusing on integrating Corpus Linguistics and Geographical Information Systems. This involves extracting place-names from a corpus, searching for their semantic collocates (see 'What can you do with a corpus?' below for an explanation of this term), and creating maps to allows users to visualise how concepts such as war and money are distributed geographically (Gregory and Hardie 2011). Interdisciplinarity is key to all of these recent developments yet language and linguistics remains at the heart of corpus-based research.

8. What can you do with a corpus?

Most software packages for corpus linguistic analysis offer the functions described below, which are the main analytical tools used in corpus linguistics:

Frequency analysis

Frequency analysis offers an insight into how often particular words are used in a data set. This simple statistical measure can be indicative of the overriding concerns expressed in a text. Frequency analysis can also be used to investigate lexical change across time or differences between texts. For example, the graph below shows the changing frequency of the word *choice* in a corpus of party political manifestos between 1945 and 2010:



From this we can see that *choice* rose in popularity from the post-war years to its peak in 2005, just prior to the credit crunch and following recession. The fall in frequency in 2010 may indicate that *choice* is a luxury that politicians do not feel able to offer in times of austerity. In the case of this example, the frequency with which particular concepts are referred to by the main political parties is revealing of their respective linguistic practices. This information might also have a practical value for organisations and pressure groups aiming to position themselves in relation to the mainstream. It may also be useful to the parties themselves since it offers an objective view of what are likely to be perceived as their main concerns.

Frequency analysis can also be used to look at the relative focus on a particular word or words in different texts or types of text. For example, in a sample of political news report from British newspapers, the relatively high frequency of the words *crime*, *criminal*, *criminals* and *criminality* suggests a strong focus on the reporting of crime in the tabloids. Lower frequencies of all these word in the broadsheets suggest that there is less of a focus on this particular topic in these types of newspapers. A high frequency of the words *yobs* and *thugs* in the tabloids also indicates how these types of newspapers stereotype people who commit criminal acts. The frequencies of these words tend to be much lower in the broadsheets.

Keyword analysis

Frequency analysis is useful for investigating the occurrence of a particular word that we already think might be interpretatively revealing. However, if we have no prior expectations about which words are likely to be informative, then we can use keyword analysis to discover the statistically significant words in a corpus. A list of keywords is a list of all those words in a corpus that occur more (or less) frequently than they do in a comparator corpus (generally called a reference corpus).

For example, in a corpus of texts relating to the national day of action on November 30th 2011 collected from various unions (including press releases and information issued to members), the keywords are *pensions*, *action*, *members*, *justice*, *ballot*, *government*, *consultation*, *support*, *campaign*, *negotiations* and *guidance*. However, in a collection of news stories in the British press covering the day of action, the keywords are *strike*, *strikes*, *walkout*, *chaos*, *anger*, *disrupted*, *protestors*, *taxpayers*.

This particular example shows the tendency for reporting of industrial action to focus on the consequences of strikes rather than their underlying causes. In the case of this example, such information could be used to inform communication strategies for press officers and union representatives.

Concordance analysis

The best way to carry out closer analysis of a keyword is to investigate its meaning in context. To do this you can use corpus linguistic software to generate concordance lines. A concordance is a list of all the sentences in which our target word occurs. Concordances are helpful because they can allow us to see patterns of language use. For example, below is an extract from a concordance of the word *slim* (excluding metaphorical usages), taken from a corpus of British English:

to shake, honey blonde hair cascading over	**slim**	shoulders. The girl just has to laugh. She's talk
hen I have a bath,' she laughed. Anthea - tall,	**slim**	and breathtakingly pretty - is nearly ten years
wigs, as he calls her, would want to drag this	**slim**	six-footer off the street. He had the same
s about six foot tall and very attractive. She is	**slim**	with blonde hair and looks like a catwalk mod
anda, of Chadwell Heath, Essex, had hoped to	**slim**	before the first wedding. Then she discovered
t again, Joy walked over to the table with two	**slim**	girls seated there. They were arguing about

the left read' These seats are for those of very	**slim**	build only'. Joy was now standing, reading the
d even though their chance of happiness was	**slim**	and Wickham was disliked, the marriage still
with nine centuries. Six feet (1.8 metres) tall,	**slim**	and athletic, his right-handed batting was less
to his two companions.' Aye,' replied another,	**slim**	and small as a child but with a face centuries
hree young children. She was dark-haired and	**slim**	, 32 years of age and pleasant. Her husband,
I could see no special contact between them.	**Slim**	in her dungarees, with her long, curly, chest
ecute impressively unimpressive water-tricks:	**slim**	, brown sprites. We return to the village by a
and fat. The other was a Sikh, very small and	**slim**	. They looked like a comic turn. There was a
give up totally, but settled for three or four	**slim**	cigars a day instead of ten to fifteen cigarette
built-in disappointments, ounce by ounce the	**slim**	frame turned to flab, and in the end Baxter

It is apparent from reading the corpus lines that *slim* is an adjective used more often to describe women than men. The concordance lines thus indicate an element of the meaning of *slim* that we may or may not have been aware of. If we suspected already that *slim* tended to be used in this way, then the corpus has provided empirical support for this view. If we were not aware of this prior to looking in the corpus, then the concordance lines have provided a new perspective on meaning. Corpus data is particularly useful when it comes to finding out about meaning. For this reason, corpus linguistic methods are now common in dictionary production and are increasingly being applied in the teaching of foreign languages.

Collocation and semantic prosody analysis

Investigating patterns of usage in concordance lines involves looking not just at the ways in which the target word is used but also at its typical collocates. Collocates are words that have a tendency to co-occur with a given word. Collocation can be analysed systematically using statistical tests. For example, in the British National Corpus that comprises 100,000,000 words of British English, the word *unemployed* has the following collocates: *long-term*, *homeless*, *unskilled*, *unemployed*, *disabled*, *redundant* and *unsuccessful*. That is, whenever the word *unemployed* is used, it's statistically likely that these other words will be used somewhere in the surrounding context. One consequent problem with using the word *unemployed* is that it is strongly associated with negative concepts. Corpus linguists call this 'negative semantic prosody'. (Some words, by contrast, might have a positive semantic prosody). Collocation analysis can be revealing of why there might be a stigma attached to particular words. It can also be used to determine alternative stylistic choices with more positive associations.

9. Key discoveries from Corpus Linguistics.

Research in corpus linguistics has led to a number of insights into the nature of language that would have been difficult to determine without corpus linguistic and computational methods. From the methodologist perspective, Corpus Linguistics has made it possible to empirically test hypotheses from many sub-disciplines of language study. Here are just some of the many discoveries that Corpus Linguistics has led to.

Semantics

John Sinclair's corpus work has led to a revised understanding of how language is structured and how meaning operates. Sinclair's insights into collocation and phraseology led him to develop the Idiom Principle, which suggests that language consists for the most part of semi-fixed expressions. According to Sinclair, speakers only resort to an open choice of words when there is no semi-fixed expression available, and meaning is a property of words in combination rather than in isolation.

Find out more: Sinclair (1991) and (2004)

Syntax

Insights from corpora have been used to support many developments in syntactic theory. One of the major innovations has been in the development of grammars of spoken language. Biber et al.'s (1999) grammar of spoken English, for instance, discards the notion of the sentence as an appropriate unit for describing speech and replaces this with the C-unit, a clause-like unit of around 5 to 6 words. C-units can be analysed grammatically but cannot be connected to anything else to form a longer syntactic unit. For example, here is a string of spoken language divided into C-units: 'Well, | you know | what I did, | I looked in the trunk | I was going to Vicky's' (Leech 2000: 699). Carter and McCarthy's work on spoken grammar, on the other hand, sees discourse as the driver of grammar, rather than the other way round (as is the traditional view). Whatever theoretical position you adopt, corpus methods have dispensed once and for all with the notion that spoken language lacks order and is chaotic.

Find out more: Biber et al. (1999), Carter and McCarthy (1998) and (2006), Leech (2000)

Historical Linguistics

Corpus methods have been adopted enthusiastically by historical linguists and most research in this area now relies on corpora at least to some extent. The Brown family of corpora, for example, have enabled linguists to determine changes in British and American English across time. Leech (2004), for instance, has examined such linguistic elements as verb phrases and modal auxiliaries in British and American English. Insights from this research have led to the identification of a number of social trends in English, including tendencies towards colloquialisation (i.e. language becoming more informal), democratisation (i.e. speakers and writers avoiding inequality in interaction) and Americanisation (i.e.

the process of adopting usages from American English into other forms of English). Corpus Linguistics has also enabled historical linguists to investigate internal developments in language such as grammaticalisation (the process by which lexical items develop to serve grammatical functions; e.g. the change of *would* from being a main verb to an auxiliary verb).

Find out more: Leech (2004) and Lindquist and Mair (2004)

Stylistics

In Stylistics, corpus methods have provided support for foregrounding theory by making it possible to determine the extent to which a word, phrase or other element of linguistic structure stands out against a perceived norm. For example, the key word function in corpus linguistic software packages makes it possible to calculate which of the words in a given corpus are statistically over- or under-used in comparison to their distribution in a reference corpus. There is not always a correlation between statistical significance and interpretative significance, but the technique does at least allow stylisticians to define a norm against which to measure foregrounding.

Find out more: McIntyre (2012) and McIntyre and Walker (2010)

10. Corpus Linguistics in action

Dictionary writing

Samuel Johnson was one of the first lexicographers, and in the mid-eighteenth century, he famously prepared a dictionary of English. However, the entries were largely impressionistic and subjective, clearly showing Johnson's own prejudices about language and society. Some examples of entries from Samuel Johnson's *Dictionary of the English Language (1755)* can be found below:

> **Distiller:** One who makes and sells pernicious and inflammatory spirits.
>
> **Dull:** Not exhilaterating (sic); not delightful; as, *to make dictionaries is* dull *work.*
>
> **Oats:** A grain, which in England is generally given to horses, but in Scotland appears to support the people.
>
> **Patron:** One who countenances, supports or protects. Commonly a wretch who supports with insolence, and is paid with flattery.

Not all dictionaries were so flagrantly opinionated as Johnson's, but many dictionaries were *prescriptive* rather than *descriptive* because they told us how language **should** be used rather than how it **is** used. Such dictionaries were often based on the language intuitions of the dictionary writers. However, in the

1980s, advances in computer technology changed the way in which dictionaries are written, when Collins, in collaboration with linguists at the University of Birmingham (led by John Sinclair), created a new breed of dictionaries.

In a ground-breaking project, Collins and the University of Birmingham jointly developed an electronic corpus of around 2.5 billion words of contemporary British English, called **Co**llins **B**irmingham **U**niversity **I**nternational **L**anguage **D**atabase, or COBUILD. The Bank of English™ forms part of COBUILD, and contains around 650 million words. It comprises written materials that were chosen to give a balanced and accurate reflection of English usage, and included samples from newspapers, magazines, books and websites, and spoken material transcribed from radio, TV and everyday conversations.

The COBUILD corpus and the Bank of English™ were used to create a new generation of English dictionaries, with the first COBUILD dictionary being published in 1987. The entries COBUILD dictionaries are not based on the intuitions of the dictionary writers, but on evidence from the corpus. The team at Birmingham investigated how each word entry in the dictionary was actually used using examples drawn from the real-life data held in COBUILD and the Bank of English™, thus allowing them to see the different meanings a word had. They were also able to investigate which senses of a word were used most frequently, whether words regularly co-occurred with other words, and they were able to provide authentic examples of words in use, rather than made-up examples.

The COBUILD project revolutionised dictionary writing, particularly dictionaries written for learners of English as a second language. In such dictionaries, information from the corpus informs word selections, on the basis that the words that are used most frequently in a language are the words that will be most useful to a language learner.

11. Building a corpus

Introduction

As we have already noted, a corpus is a collection of texts that have been selected to be representative of a particular language or language variety. McEnery and Wilson (2001) expand on this definition, saying that a corpus is:

> [...] a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration.
>
> (McEnery and Wilson 2001:32).

Additionally, Sinclair (2005) states along similar lines that a corpus is:

> [...] a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

One important notion made explicit in these definitions is that of **representativeness**, or that a corpus should attempt to represent a particular language or language variety under investigation. As we said earlier, this is key to producing a good corpus.

Representativeness

Representativeness is important because, even though corpora can be very large (in theory, there is no limit to the size of a corpus), they can never hold every single example of a particular language variety (except, perhaps, in the case of a dead language). This is because, practically, it would be impossible to collect every utterance spoken or written from one particular language variety. For example, if we were investigating Yorkshire spoken English, it would be impossible to collect every instance of English ever spoken in Yorkshire. Even if we said that our corpus was of Yorkshire English spoken in one school in Yorkshire on May 1st 2013, it would be almost impossible to collect every word of every utterance. This is not least because collecting speech is particularly difficult since you first have to record it, which often has a number of practical issues; and then the recordings need to be transcribed, which can take a long time. Therefore, a corpus, no matter how big, will only ever be a sample of a particular language variety. So, the idea, in theory at least, is to collect enough samples of a language variety to **represent** the *entirety* of that language variety. By 'represent' we mean that, **proportionally**, there should be the same quantity of any linguistic item we care to examine (for example, a particular word or grammatical part-of-speech) in the corpus as there is in the language variety as a whole.

An issue with representativeness and the whole notion of corpus studies is that it actually seems unlikely that a whole language variety (such as British English) could be represented by a small sample of texts. It is inevitable that lots of examples of particular language usages (certain words or phrases or grammatical structures) will be missing from the sample. This matter did not go unnoticed by Chomsky (1962) in a famous paper criticising corpus linguistics; while a corpus might be representative of some common language features, less common language features might be completely absent from the corpus, so not represented at all. Corpus linguists such as Douglas Biber have worked extensively on this problem in terms of statistical probability, using many different samples of texts from many different language varieties (see, for example, Biber 1993). However, it remains a complex area of research with, as yet, a number of unresolved problems. An associated issue is that the only way to be absolutely sure whether a particular linguistic item occurs proportionally the same number of times in a corpus as it does in the entire population the language variety is to count the number of occurrences in each. We have already established, though, that this is almost impossible, and definitely impractical.

So, we can never know for sure whether any corpus we build is representative of the language variety it contains or not. Therefore, we can only strive to be as representative as is practically possible. Both McEnery and Wilson's and Sinclair's definitions (above) recognise this when they state, respectively, that a corpus should be 'maximally representative', 'as far as possible'. This does not mean that we should abandon corpus research altogether because, as we have already explained, it can provide valuable insights into many aspects of language. Knowledge of the language variety under investigation coupled with careful corpus design can help us to build corpora that, while not being complete pictures of a language variety it contains, are, nevertheless, as maximally representative as possible and able to reveal general tendencies within that language variety.

Sampling: working out what to put in a corpus.

As should be clear by now, creating a corpus that represents a language variety is no simple matter. Since it is practically impossible to investigate entire language varieties, corpus building will inevitably involve gathering together enough samples of that language variety (i.e. texts) to adequately represent it. **Sampling**, then, is crucially important in corpus building.

The first step in sampling is to deciding on a **sampling frame**. This means working out what your corpus needs to consist of in order to maximally represent the language variety under investigation. A corpus, no matter how large, will always consist of individual texts, some of which will be large (e.g. a political speech; a newspaper editorial, a whole book), while others might just be a few words (e.g. the message on the back of a postcard, an email, a poem). Narrowing down which texts you need to collect in order to represent the language variety you are investigating could start with some simple questions, such as:

- What language or language variety is being studied? (e.g. English, Australian English, Yorkshire English.)
- What is the location of the texts you need to collect? (e.g. UK, Australia, Yorkshire.)
- What is the production/publishing date of the texts you want? (This might be one day, one year, or a span of years)

These questions are quite broad, but nevertheless need careful consideration. Further questions you might ask to help you specify the sorts of texts you want to collect could include:

- What is the mode of the texts? (spoken or written, or both)

- What sorts of texts (text-types) need to be included? (e.g. newspaper articles, short stories, letters, text messages)
- What domain will the texts come from? (e.g. academic, legal, business)

The answers to such questions become the sampling parameters, which help to identify the sorts of texts that you will eventually collect to populate your corpus. The number of parameters needed depends largely on the nature of your corpus and the language variety that is being represented. Therefore, the questions above do not form an exhaustive list; there might be many other questions that need to be asked (e.g. what is the nationality, or the gender, or the age of the producer of the text?). Typically, though, the questions (and hence the sampling parameters) will start off very broad and become gradually more focused.

The sampling parameters specify the components of the corpus and their relationship to each other. To illustrate this, consider a corpus of British English. The first sampling parameter of such a corpus is, rather obviously, British English. That is to say, the person building the corpus will only sample from British English texts, so non-British English texts can immediately be removed from consideration. Of course, British English has been around for quite a long time, so we could add a time parameter and specify, say, the year 2012. The overarching component of the proposed corpus, then, would be British English, which is then immediately narrowed down with a year of production/publication parameter of 2012. The next sampling parameter might be to do with mode. For a corpus to be representative of British English it would need to contain examples of both spoken and written English. Therefore, the next two sampling parameters would be ‘spoken English’ and ‘written English’, and these would form components that sit beneath the overarching components of ‘British English’ and ‘2012’. Having decided that the corpus will include samples of spoken English, it might be useful to distinguish between, say, everyday, naturally occurring conversations, and conversations influenced by the context they occur in (e.g. doctor-patient interviews, lectures, business meetings). Thus, the sampling parameters ‘everyday conversation’ and ‘context influenced conversation’ could be added, and two further components are therefore added. These two components might need to be split further. So, for example, the ‘everyday’ component could be split by gender, while the ‘context influenced’ component could be split into, say, business and educational domains. Similar sorts of structural components could be added to the written section. For example, a distinction could be made between books and periodicals, and books could be spilt further into fiction and non-fiction, while periodicals could be split into newspapers and magazines. This process could continue with further parameters, and therefore further components, being added.

The components form a hierarchy, with one component at the head of the hierarchy (‘British English’ in this example) under which all the other components are arranged in a series of levels. The bottom level will have the most components, and these sit closest to the texts that will eventually be collected. Thus, the resulting componential structure of the proposed British English corpus would be something like that shown in Figure 1. Sampling parameters therefore specify the components of a corpus and its eventual structure.

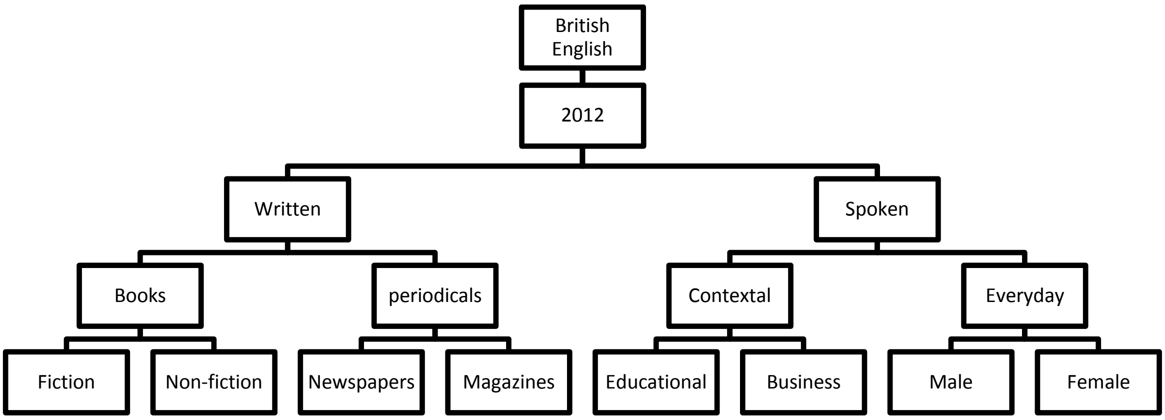


Figure 1. Some of the possible components of a British English corpus.

The structure in Figure 1 shows that the number of components from which texts would be collected is eight: four spoken and four written. However, this example is just illustrative and not exhaustive; many more components are possible and probably necessary. For example, fiction could be split into 'adult' and 'children', and then these two new components could be further split into, for example, 'crime', 'romance', 'science fiction/fantasy', and 'comedy'. Additionally, someone interested in regional variation in spoken British English might want to add further sampling parameters relating to location. These would form further components and further levels in the hierarchy.

Ultimately, building corpora is about collecting texts. Thinking carefully about the components of a corpus helps to decide what sorts of texts constitute a representation of the particular language variety under investigation. This, in turn, helps with the overall probable representativeness of the corpus. The levels of components in a corpus hierarchical structure represent a narrowing down of what will be in the corpus, with each level descended giving finer distinctions about the textual content of the corpus, and each component signifying a smaller and smaller well from which texts will be drawn.

These sorts of decisions had to be made by the teams who built the BNC and LOB corpora when they considered the best ways to represent British English and British written English, respectively. Even though their eventual strategies were slightly different, they nevertheless broke their corpus down into several different components from which they then drew samples.

Sampling: numbers and sizes of texts

Once the components of a corpus have been decided upon, the next consideration is sample size, which concerns both the number of texts that are needed to represent each component, and the size of those texts. With regard to the latter, this will probably involve deciding whether to use whole texts or extracts, or both. This, then, is the 'how much?' question and, as you might have guessed, there is no simple answer to this question. The language variety under investigation will have a bearing on the question, and so too will practical issues such as the amount of time available for text collection. For example, in a corpus that attempts to represent the language of text-messages, it is likely that the key to representativeness is collecting samples of texts from a large number of different respondents. Consequently, the components would probably need to take into account gender, age and different social contexts (so, not just 17 year old college students, for example). The number of texts collected for each person might be quite small, if one assumes that a person is consistent in their texting style (which might not be the case, of course), but the number of respondents would ideally be quite large. The latter is likely to be governed by the investigator's ability to persuade many different people from different walks of life to give them examples of text-messages and the associated ethical issues with this.

Clearly, the more texts that are in a corpus, the more representative the corpus is likely to be, provided the texts are distributed across the components specified in the hierarchy, and not clumped together in just one component. Having said that, the distribution of texts across the components of a corpus does not necessarily have to be equal. For instance, the relative popularity of some texts or text-types might influence the number of texts collected for each component. For example, a corpus of text messages might contain more texts written by, say, 15 to 20 year olds if there is evidence to suggest that this age group send more texts than any other age group. Similarly, in a corpus that contains 21st century British written fiction, you might argue for the inclusion of British soap opera scripts, since these fictional texts are very popular (if viewing figures are anything to go by). Given they are so popular (and provided that the texts can be obtained, and any issues regarding copyright resolved), you might also argue that there should be a greater proportion of these texts in the corpus than other fictional texts that have more literary value, but are less popular. Of course, there are issues concerning what counts as popularity (viewing figures, sales figures, website hits), and what counts as literary value (literary critical consensus?), and these would have to be resolved at the start of the project.

An important consideration, then, when answering the 'how much?' question is whether the amounts of texts in each component of a corpus should attempt to reflect actual proportions in the real-world, and whether this will make the corpus more representative. These sorts of considerations should be included at the corpus design stage, and made explicit in any documentation accompanying the corpus. (See the Leech 2011 reading for further discussion of this).

Another consideration with regard to the 'how much?' question is whether to include whole texts or extracts from texts. Texts can vary massively in size from a few characters (for example, a text message) to thousands, hundreds of thousands, or even millions of words. Sinclair (2005) suggests that, regardless of any differences in size, whole texts should be used in a corpus. This is because any part of a text cannot represent the whole since position within a text affects language choices. That is, language choices available at, say, the start of a text (e.g. *Dear Madam*; *In this chapter we will ...*; *Fire fighters battle mile-wide moorland inferno*; *Can I speak to the homeowner?...*) are different from the choices available at, say, the end of a text (e.g. *Yours sincerely*; *In summary, this chapter ...*; *The cause of the fire has still not been established*; *Goodbye and thank you for your time ...*). Therefore, a corpus that contains just the beginnings of large texts will not adequately represent that text-type; at best it will only represent the start of texts within that text-type. One issue with collecting whole texts, though, is that if the corpus you are building is small, and the variation in text size within the corpus is large, then it is likely that the larger texts, or even one very large text, will skew any results obtained from the corpus. The solution to this issue is either to make the corpus bigger so that the effects of very large texts are watered down, or to select a smaller extracts from large texts. The former might not be possible due to external limits such as time, while the latter needs careful management due to the phenomena of language choices being affected by text position described above. This might mean randomly selecting multiple extracts from a large text from near the beginning, the middle and the end.

A further possible issue with some text-types is deciding what exactly counts as a text. Sinclair (2005) mentions communicative events, but this notion becomes less definite when dealing with, say, interactions on internet bulletin boards or Twitter feeds. This is another matter that would depend on the corpus being built, and would need to be decided at the start of the project. Where there are issues concerning where a particular text begins and ends, then it is important to be consistent (i.e. make a decision and stick to it).

When we build a corpus we are aiming to collect enough samples of a language variety to represent the *entirety* of that language variety. Having an appreciation of what the entire population consists of will help to decide upon answers to the 'how much?' and 'how many?' questions. Giving careful thought to a componential structure (described above) will help with this.

Corpus Size

Corpus size (usually measured in words) can be either (i) decided first and the size of the components divided to fit into that fixed maximum size, or (ii) result from decisions made about the number and size of the components. If, in the former case, the corpus size is set at one million words and the number of bottom-layer components is eight (as in Figure 1), then each component will contain 125,000 words.

If, in the latter case, the number of components at the bottom level of the corpus structural hierarchy is eight, and it is decided that, in order to represent each component, one million words of text is needed for each, then the resulting corpus will be eight million words in size.

The approach used should be decided in the design stage of the corpus and will, inevitably, have implications for the sampling. For instance, it may be necessary to use extracts of large texts in a corpus that has a small fixed size relative to some of the text-types it will contain. This is the case for the Brown family of corpora, which are one million words in size and yet attempt to represent America and British written English. One of the textual components of these corpora is fiction, which includes novels. The size of this component is set to approximately 250, 000 words, and it is easy to see how this component could be filled with just three or four whole novels. Therefore, this component of the Brown family of corpora contains text samples of around 2000 words from 126 works of fiction. As we stated earlier, sampling of this nature requires careful management.

Whatever the approach, sufficient words spread over enough different texts need to be collected in order to adequately represent each component of the corpus, and thus the language variety being investigated.

Further reading on representativeness and sampling: Leech (2011: 158-160). In this reading, Geoff Leech answers the question:

How representative can a corpus be?

Well, how long is a piece of string? A useful definition of *representativeness* is provided by Manning and Schütze (1999:119): a sample is representative if what we find for the sample also holds for the general population. If we replace ‘sample’ by ‘corpus’ and ‘population’ by ‘language’, we see the difficulty of determining whether what is found to be true of a corpus can be extrapolated to the language as a whole. We cannot (in general) study the whole language in terms of use, so we are left with the problem of ensuring that the corpus is as representative a sample of the language (or language variety) as possible.

When I first started drafting this answer, I began by describing representativeness as the Achilles’ heel of Corpus Linguistics. Afterwards, I changed ‘Achilles heel’ to ‘Holy Grail’, to sound a rather more positive note – representativeness is something we are optimistically looking for, but may never exactly find. In this respect it is like truth. Very rarely can complete representativeness, like complete truth, be attained. What we can do, though, is work towards greater *representativity*. I prefer to use this term (Leech 2007:140) to denote a scalar concept (one corpus being *more* representative or *less* representative than another), rather than ‘representativeness’, which tends to suggest an all-or-nothing quality.

If a corpus is a sample of language in use, then obviously the language (the ‘population’) of which it is a sample has also to be considered as consisting of language in use – language as performance, rather than competence. A general corpus of English in the year 2000 should be considered a sample of all utterances/texts that were produced in English at that time. But this is a mindbogglingly impractical population to measure a corpus against, and in practice the ‘population’ of which the corpus is a sample is defined not just as being in a particular language, but as being circumscribed by parameters of language variety. For example, the Brown Corpus was designed to be sample of *written American* English published in the year *1961*.

It is probably uncontroversial that the bigger the corpus is and the greater the variety of text types (genres, registers) it contains, the great it representativity. However, this is not sufficient. We have also to consider the *quantity* of particular varieties that must be contained in a corpus of a given size: how a corpus can become ‘balanced’ by including appropriate proportions of different text types. This is the issue of proportionality: the proportion of a text type in a corpus should ideally equate with the proportion of that text type in the population as a whole. The one-million-word Brown Corpus contains c.88,000 words of newspaper reportage and c. 12,000 words of science fiction writing. We may feel satisfied that the news report genre is a bigger, more important category than science fiction, and that this proportion 88,000 to 12,000 is intuitively reasonable. But how would we demonstrate this, and how can we determine the right quantities more precisely? The most likely answer to this would be to find a library with a classification of all the publication in the US in the year 1961, and replicate the proportions found there in the Brown Corpus.

Biber (1993) rejected proportionality as a means to achieving representativeness, because this would, he argued, lead to a highly skewed sample of language. since about 90 per cent of all linguistic usage is ordinary private conversation (an estimate less convincing now than it was), the corpus would consist largely of conversational data, and some important genres, such as the language of statutes, or of inaugural addresses of American presidents, would have scarcely no existence in the corpus. However, my position is that we should rather concentrate on the addressee’s end of the message, rather than the addresser’s, in calculating the proportion of usage. The author of a message is normally an individual, whereas the number of receivers can vary from one individual to many million individuals (in the case of a popular newspaper or TV broadcast). The number of receivers a message has provides a reasonable way of determining its importance. If this position is adopted, a radio new bulletin listened to by a million people deserves to be in a corpus sample a million times more than a conversation between two private individuals. The conclusion is that the design of a general corpus should ideally be built on extensive language reception research (see Cermák 1997 on the Czech National Corpus). This does not, however, solve our problem of representativeness – it is difficult to obtain precise information on the number of readers of a written text, or the number of listeners to a spoken discourse. But the receiver-centred view of representativity gives us a conceptually sound model for defining a ‘balanced corpus’, and where we have no precise figures (which is very often), we can do our best by relying on estimates. The Holy Grail of complete representativeness is still far away, but we can aim to achieve better coverage of genres and better approximations to proportionality. As a sample of written American English, the Brown Corpus may not be entirely representative, but it is better than a million words of the *Wall Street Journal*, for instance.

12. Step-by-step guide to building a corpus.

The following represents a possible set of steps for building a corpus:

1. The first step when building a corpus should be the corpus design. Thinking carefully about sampling and representativeness issues discussed above will help to build a good corpus, as well as save valuable time and effort.
2. Decide on what your corpus is attempting to represent and, therefore, what will be in it.
3. Use your answer to the above to create a hierarchical model of the components of the corpus
4. For each of the components at the bottom level of the hierarchy, list the possible text-types or texts (depending on the detail of your hierarchy) that you expect to find.
5. Consider whether each text-type should have equal standing in the corpus.
6. Think about the size of each component, taking into consideration the number of text-types available, their real-world importance, and the practical issues in gathering them.
7. Decide whether whole texts will be collected, or extracts, or both. If you are collecting extracts, where possible use random sampling. This can be achieved by using a random number generator (there are many available on the internet) and employing those numbers to select, for example, page numbers. Whatever strategy you decide upon, keep in mind that lexical choice can be influenced by textual position.
8. Choose, locate and gather the texts that will populate the corpus. This can be a very time consuming step, depending on the nature of the texts being collected. For instance, gathering spoken texts will require a number of further steps including making audio recordings of the spoken data, and then transcribing them.
9. Make and keep a security copy of the text in its original form. This copy might be electronic, paper or audio/visual, depending on the nature of the text(s) being collected.
10. Make and keep an electronic, plain text format copy of the text. This format is the simplest electronic format and, at the moment, a requirement of most corpus-tools. It is also the most portable, flexible and future-proof format, so a good bet for archive-copies of the corpus. If required, the plain text copies can be (usually very easily) converted into other formats.
11. Add some information about the text at the beginning of the text file. Sinclair (2005) suggests that the easiest way to do this is to simply add an identification code or serial number, which can then be cross-referenced with a database or spread-sheet that holds useful information about the text (i.e. metadata). This might include details such as the author, date of publication, genre, page-numbers sampled, and so on. Another option is to include this sort of information actually in the text file in what is known as a header (see Figure 2.). The header is separated from the data using mark-up. Currently xml-style mark-up is popular for this. An xml-style

header uses angle brackets (i.e. < >) to enclose information about the text. Computer-tools used to analyse corpora are designed to ignore anything enclosed in angle-brackets, so the metadata you add will not affect your results.

```
<header>
<publication title = “The Joy of Corpora”/>
<author = “Emma Pyrikal”/>
<publication date = “01/05/2013”/>
<pages sampled = “38-53”/>
</header>
.......
```

Figure 2 An example of a simple header

12. Add further annotation or mark-up required by the investigation (for example, part-of-speech annotation). Keep copies of the corpus before the annotation is applied, and make copies afterwards.
13. It might be necessary to clean the text before using a corpus tool to process it. This might mean removing any spurious characters from the text left over from the conversion to plain text format. It might also mean that certain everyday characters (such as ampersands, and pound signs) might not be recognised by the software, and might need to be converted to some other format (for example SGML-entities such as: &amp, &pound). The manual for the software you are using should tell you what to do.
14. Once you have a working copy of the corpus, MAKE A COPY OF IT! Ideally, you should make copies along the way as you build the corpus. This means that you can always go back if something goes wrong at some stage.
15. Keep notes of what you do and why you are doing it. Record decisions about corpus design, the data you use, where you got the data from, any annotation you apply to it, and so on. These notes will act as an *aide memoire*, and help others see and understand what you did.
16. Do some analysis!
17. If you need to, repeat any of the above steps in light of your analysis.

References

Baker, P. (2009) 'The BE06 Corpus of British English and recent language change'', *International Journal of Corpus Linguistics* 14(3): 312-37.

Biber, D. (1993) 'Representativeness in corpus design', *Literary and Linguistic Computing* 8(4): 243-57.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. London: Longman.

Carter, R. and McCarthy, M. (1995) 'Grammar and the spoken language', *Applied Linguistics* 16(2): 141–158.

Carter, R. and McCarthy, M. (2006) Cambridge Grammar of English: A Comprehensive Guide: Spoken and Written English Grammar and Usage. Cambridge: Cambridge University Press.

Cermák, F. (1997) 'Czech National Corpus: A case in many contexts', *International Journal of Corpus Linguistics* 2:181-97.

Chomsky, N. (1962) 'A Transformational Approach to Syntax', *Proceedings of the Third Texas Conference on Problems of Linguistic Analysis in English on May 9-12, 1958.* edited by Hill, 124-58. Texas, 1962. (Reprinted in *Structure of Language*, edited by Fodor and Katz. New York: Prentice-Hall, 1964; reprinted as "Une Conception Transformationelle de la Syntaxe." *Language* 4 (December 4, 1966): 39-80; Reprinted in *Classics in Linguistics*, edited by Hayden, Alworth and Tate, 337-71. New York: Philosophical Library, 1967.)

Gregory, I. and Hardie, A. (2011) 'Visual GISting: bringing together corpus linguistics and Geographical Information Systems', *Literary and Linguistic Computing* 26(3): 297-314.

Hardie, A. and McEnery, T. (2010) 'On two traditions in corpus linguistics, and what they have in common', *International Journal of Corpus Linguistics* 15(3): 384-94.

Knight, D., Adolphs, S., Tennent, P. and Carter, R. (2008) 'The Nottingham Multi-Modal Corpus: a demonstration', *Proceedings of the 6th Language Resources and Evaluation Conference*, Palais des Congrés Mansour Eddahbi, Marrakech, Morocco, 28-30th May.

Krishnamurthy, R. (ed.) (2004) *English Collocational Studies: The OSTI Report*. London: Continuum [originally published as Sinclair, J., Jones, S. and Daley, R. (1970) *English Lexical Studies*. Report to OSTI on Project C/LP/08].

Kučera, H. and Nelson, W. N. (1967) *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Leech, G. (2000) 'Grammars of spoken English: new outcomes of corpus-oriented research', *Language Learning* 50(4): 675–724.

Leech, G. (2004) 'Recent grammatical change in English: data, description, theory', in Aijmer, K. and Altenberg, B. (eds) *Advances in Corpus Linguistics: Papers from the 23rd International Conference on English Language Research on Computerized Corpora* (*ICAME 23) Göteborg 22-26 May 2002*, pp. 61-81. Amsterdam: Rodopi. [Available from:

http://www.lancs.ac.uk/fass/doc_library/linguistics/leechg/leech_2004.pdf]

Leech, G. (2011) 'Principles and applications of corpus linguistics' in Viana, V., Zyngier, S. and Barnbrook, G. (eds.) *Perspectives on Corpus Linguistics*, pp155-70. Amsterdam: John Benjamins.

Leech, G. (2007) 'New resources or just better old ones?' in Hundt, M., Nesselhauf, N. and Biewer, B. (eds) *Corpus Linguistics and the Web*, pp 133-49. Amsterdam: Rodopi.

Lindquist, H. and Mair, C. (eds) (2004) *Corpus Approaches to Grammaticalization in English*. Amsterdam: John Benjamins.

Manning, C. D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge MA: The MIT Press.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics (2nd edition).* Edinburgh: Edinburgh University Press

McIntyre, D. (2012) 'Prototypical characteristics of blockbuster movie dialogue: a corpus stylistic analysis', *Texas Studies in Literature and Language* 54(3): 402-25.

McIntyre, D. and Walker, B. (2010) 'How can corpora be used to explore the language of poetry and drama?' in McCarthy, M. and O'Keefe, A. (eds) *The Routledge Handbook of Corpus Linguistics*, pp. 516-30. Abingdon: Routledge.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004) *Trust the Text*. London: Routledge.

Sinclair, J. (2005) 'Corpus and Text-Basic Principles' in Wynne, M. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*, pp.1-16. Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2013-05-12].

Zipf, G. K. (1935) *The Psychobiology of Language*. Boston, MA: Houghton-Mifflin.