Corpus linguistics for indexing

Gavin Brookes and Tony McEnery
Lancaster University

Abstract

This methodological paper demonstrates how methods from corpus linguistics – a collection of computer-assisted approaches to the analysis of large volumes of text – can be used in the creation of indexes. We begin this article by introducing corpus linguistics, including its main principles and advantages, before demonstrating how corpus methods can be used by indexers using a case study in which we create an index for an academic journal article using the established corpus techniques of frequency, keywords, collocation and concordance. This case study shows how, when combined with human input and intuition, corpus linguistics methods have the power to provide indexers with new perspectives on the texts they are working on, all the while increasing the systematicity, replicability and objectivity of the indexing process itself.

Keywords
Corpus linguistics, corpora, computational methods, computer-assisted linguistics, corpus-assisted indexing

1. Introduction

In this article, we introduce corpus linguistics as a set of methods that can be used by human indexers to assist them in the compilation of document indexes. The words 'index' and 'indexing' have accrued a wide range of definitions and uses over time (Wellisch, 1988; Day, 2014; Fetters, 2014). For the purposes of this article we follow Booth (2013: 2-3), who defines an 'index' as 'an organized (usually alphabetical) sequence of entries, each of which can lead a user to the desired information within a document, or to the required document within a collection', and 'indexing' as 'the process of creating [,] compiling, or writing the index'. In an editorial published in the inaugural issue of this journal, Harold Smith (1958: 2) observed how '[i]ndexing, like book classification, bibliography, documentation and abstracting, is a method - all too often haphazard and unsystematic - of making known information which would otherwise remain hidden and buried'. Although the methods of indexing have, of course, come a long way since the 1950s, with technological and computational innovations allowing the work of indexing to be carried out ever more swiftly and accurately (cf. Booth, 2013: 375-406), our aim in this article is to introduce a set of methods, known as corpus linguistics, which have the potential to increase the systematicity, replicability and objectivity of this process even further.

This article is arranged into four sections. Following this introduction, the next section provides a more detailed background to corpora and corpus linguistics, including outlining its

main strengths, limitations and briefly reviewing existing areas of application. Following this, Section 3 provide a worked demonstration of how corpus linguistics can be used in indexing. Using the established techniques of frequency, keywords, collocation and concordance, this case study will show how these staple corpus techniques can be used by human indexers to abstract candidate index items and learn about their meanings and the ways they are used in the text or texts being indexed (in this case, an academic journal article). In the fourth and final section, we conclude by reflecting on the case study and considering the opportunities and limitations of corpus linguistics for indexing.

2. Corpora and corpus linguistics

Introduced briefly in the previous section, the term *corpus linguistics* refers to a collection of computer-assisted methods for analysing large amounts of naturally-occurring, machine-readable text (McEnery and Wilson, 2001). Such a collection of texts is known as a corpus (plural *corpora*) – the Latin word for *body* (as in, a 'body' of texts). Despite several established points of departure, corpus linguistics comprises a wide range of approaches to the exploration of language. As McEnery and Hardie (2012: 1) put it, '[c]orpus linguistics is not a monolithic, consensually agreed set of methods and procedures for the exploration of language … Differences exist within corpus linguistics which separate out and subcategorise varying approaches to the use of corpus data'. For the purposes of this article, rather than provide an exhaustive introduction to corpus linguistics, we will restrict our focus to those features and techniques which we deem to be most relevant to its potential application to indexing.[1]

Usually developed for the purposes of research in linguistics, one of the main appeals of using large collections of data – it is not unusual for corpora to run into millions and even billions of words – is that they allow researchers to base their analyses on more substantial and representative bodies of textual evidence. The concept of representativeness is important, for corpora are not simply randomly-compiled collections of texts. Rather they are carefully designed in such a way that they represent a particular language or variety at-scale. Corpora are considered to be representative of the varieties they contain if findings based on their contents can be generalized to those varieties (Leech, 1991). The variety that a corpus is intended to represent will therefore dictate its design regarding its size, content and balance in terms of how much each text or genre contributes to the corpus as a whole (McEnery and Hardie, 2012: 8-11). To demonstrate this, it is useful to briefly consider the distinction between general and specialised corpora. General corpora are designed to represent entire languages or varieties (usually at a particular point in time). Because they aim to represent language on a broad scale, general corpora tend to be very large. An example of such a corpus is the 100 million-word British National Corpus (BNC), which represents written and

[1] For more comprehensive introductions to corpus linguistics, we direct readers to McEnery and Wilson (2001), McEnery, Xiao and Tono (2006) and McEnery and Hardie (2012).

spoken British English during the 1980s and 1990s (Aston and Burnard, 1998).[2] Large general corpora like the BNC are often designed according to sampling frames which help to ensure that they represent the various genres or registers that make up the target language or variety.

Specialised corpora, on the other hand, are designed to represent language use in more specific contexts. Specialised corpora tend to be considerably smaller than general corpora, comprising fewer texts representing a single or at least more restricted range of textual genres and registers. For example, Wright and Brookes (2019) built a 1.8 million-word corpus of UK newspaper articles about immigrants who can't speak English to examine how this group was linguistically represented by the press. However, even specialised corpora can still be very large. In an earlier study of press discourse around immigration, Baker et al. (2013) constructed and analysed a specialised corpus containing 140 million words of UK newspaper articles on the topic of Muslims and Islam. Generally speaking, the larger and more representative the corpus is, the more confidence the researcher can have that their findings translate to the wider population or variety under study. However, corpus techniques do not have to be applied to large volumes of text but can also be effective for studying smaller texts and collections of texts, too.

Corpora can comprise texts from one or many communicative modes (i.e. speech, writing, computer-mediated communication, gestures, etc.) and genres (e.g. spoken conversation, books, e-books, e-mails, etc.). Whatever modes and genres a corpus represents, it is essential that the texts it includes are electronic, i.e. in a machine-readable format. This allows the corpus of texts to be both stored and analysed using a computer. Due to their large size, it is usually not practical for entire corpora to be analysed by hand. Specialist corpus software packages, such as *WordSmith Tools* (Scott 2016), *AntConc* (Anthony, 2019) and *#LancsBox* (Brezina et al., 2015) provide human users with the facility to carry out a range of analytical procedures with levels of speed, accuracy and replicability that would not be possible without computer assistance. These procedures, some of which will be demonstrated in this paper, allow human users to search for every occurrence of a word or combination of words, generate frequency information about phenomena of interest (e.g. words, chains of words, grammatical types), carry out statistical tests on those frequencies (to measure the significance or strength of relationships between phenomena) and present the texts in the corpus in ways that render them more amenable to human inspection.

As well as convenience, such software packages also bring the added benefits of rendering visible patterns that might run counter to human intuition or which feature sparingly in one or two texts but become significant when considered as part of a larger collection of texts (McEnery et al., 2006). Corpus methods can also help to produce accounts of text(s) – for indexing and other purposes – that rely less on human institution and are guided by more objective criteria like frequency and statistical salience. This increased objectivity shouldn't be overlooked, since added neutrality can be advantageous to indexers, as Booth (2013) points out:

[2] An updated version of the BNC, which represents British English between 2012 and 2016, is currently under construction at Lancaster University. The spoken component, which comprises 11 million words of spoken British English, can be accessed here: http://corpora.lancs.ac.uk/bnc2014/ (see also: Love et al., 2017). Information about the written component can be found here: http://cass.lancs.ac.uk/bnc2014/.

> Every indexer comes to a document with a mental bundle of attitudes, beliefs, prejudices, received ideas, 'facts', general knowledge and 'conventional wisdom'. Much of this bundle is helpful in aiding understanding, interpretation and representation of the document content. Sometimes, with documents that are polemic in style, or that deal critically or controversially with a subject, indexers may have to cope with material that contrasts with their personal views. […] Although the index is a work in its own right, created by the indexer, and exhibiting the general and specialist knowledge and technical expertise of the indexer, it must not reveal the indexer's personal beliefs, attitudes or judgements.
>
> (Booth, 2013: 36)

Despite the increased objectivity offered by corpus linguistics, human intuition and input nevertheless have an important role to play when using corpus methods. As the forthcoming case study will show, human users of corpus linguistic programs are required to make several important decisions: from building or choosing a corpus. to selecting which analytical techniques to use and deciding on their parameters, to, ultimately, interpreting the significance of the computational output. In indexing terms, this final step also includes deciding on whether or not items should be included in an index and, if so, how.

The impact of corpus linguistics on the study of language is easy to understate. It is no exaggeration to claim, as Leech (2000: 677) does, that the availability of large corpora – and the tools to analyse them – has 'revolutionised' the ways in which language is currently studied and even conceptualised. Corpora and corpus linguistic techniques have been employed by researchers across a wide range of disciplines and sub-disciplines within linguistics, and increasingly across areas outside linguistics (see O'Keeffe and McCarthy (2010), McEnery and Hardie (2012) and Biber and Reppen (2015) for overviews of current applications of corpus linguistics). This impact has been felt outside of academic study, too. For example, many lexicographers now rely on large general corpora to identify frequent and new words to include in dictionaries, as well as using the attested examples of word use that corpora provide when developing usage-based definitions (Kilgarriff et al., 2008; Hanks, 2012). Meanwhile, in the domain of language learning, corpora offer vast repositories of authentic language use which learners can study and on which teachers can base their materials (Boulton, 2017). Relatedly, corpora are also used increasingly in the teaching and learning of languages for specific purposes, for example in the teaching of English for medical professionals (Crawford and Brown, 2010). Corpus linguistics thus constitutes a diverse collection of methods that can, in theory, be applied to any area of work or study where language and text are central concerns. Given its versatility, it is somewhat surprising, then, that the potential contribution of corpus linguistics to indexing (a discipline concerned with extracting meaning and content from texts and collections of texts) has, to our knowledge, yet to be explored. We aim to address this gap in this paper.

3. Corpus linguistics for document indexing: A case study

In order to demonstrate the potential contribution of corpus linguistics to indexing, we will now present a case study in which we utilise the established corpus techniques of frequency, keywords, collocation and concordance to identify and learn more about candidate index items in an academic journal article. Although this case study focuses on the context of one type of text in particular – a journal article – it should be noted that the techniques introduced can be applied, in principle, to any type of text where language is the primary mode of communication. As we hope to demonstrate, these techniques can *assist* human indexers in terms of generating candidate index items and in learning about those items' meanings in the target text(s). They will not automatically identify index items. Neither can they contribute to the actual creation or formatting of indexes.

3.1. Creating a corpus and selecting a tool

The first step in any application of corpus methods involves the selection or construction of a corpus. For indexers, this will involve converting the document or documents they are working on into a corpus or series of corpora. In practical terms, this involves storing a copy of each target document in a plain text (.txt) format so that it can be processed by corpus linguistic software. For texts already available in digital format, such as document proofs or ebooks, this is a fairly straightforward task. However, for indexing texts in other formats, such as hard copies or written texts or audio files, their contents will have to be transcribed in a computer-readable format, or subject to reliable optical character recognition (OCR), and then stored in a plain text file before they can be processed by the corpus software. This step results in texts' non-linguistic parts, such as images, fonts and sounds, being lost. Including these elements in an index is therefore best approached manually or using other techniques, perhaps alongside the use of corpus methods for linguistic items.[3]

For this case study, we will create an index for a single academic journal article, selected at random from the learned writing section (containing academic prose in various disciplines) of BE06, a general corpus representing written, published British English in the mid-2000s (Baker, 2009). The article in question is titled 'Social Housing in Jersey: An Analysis', authored by Chris Steel and published in 2007 in the journal, *Accountancy Business and the Public Interest*.[4] Once the corpus has been constructed, it is then uploaded to the corpus analysis software package of choice. There is a plethora of such packages available, while new programs continue to be developed and existing ones are updated all the time. Some packages are free to use but others require users to purchase a license for a relatively modest fee. In addition to the aforementioned *WordSmith Tools*, *AntConc* and *#LancsBox*, popular programs include, among others, *SketchEngine* (Kilgarriff et al., 2014), *CQPWeb* (Hardie, 2012) and *Wmatrix* (Rayson, 2008). For this case study, we have elected to

[3] It is not impossible to account for non-linguistic, including visual, elements in texts using corpus methods. For example, corpora can be annotated or 'tagged' for additional information about the texts or the language in texts. 'Tags' could therefore be created to indicate the presence and content of images as well as other features of text format and structure. Annotation is, however, a more advanced aspect of corpus compilation and analysis. We would therefore recommend that newcomers to the field adopt alternative, manual approaches to accounting for images in indexes, at least at first. Readers interested in corpus annotation should consult Garside et al. (1997).

[4] BE06 does not contain full versions of texts but balanced samples of approx. 2,000 words. We therefore downloaded the full version of this text which was freely available online.

use *#LancsBox*, as this tool is free and relatively easy to use for people with limited knowledge or experience of corpus linguistics.[5] While there is no standard approach or set of procedures in corpus linguistic methodology, our case study will introduce and demonstrate four well-established techniques in corpus linguistics – frequency, keywords, collocation and concordance. Note that although we are accessing these techniques through *#LancsBox*, all are available in the majority of packages, including those listed above.

3.2. Frequency

The frequency technique provides a list of all the words in the corpus, along with their frequency of occurrence. Frequency information can provide a good starting point for the creation of an index, as it gives a rapid overview of the thematic content of the corpus. This information is typically presented in a vertical list, in descending order of frequency (i.e. with the most frequent items at the top). Frequency information can be generated to account for individual words or recurrent sequences of two or more words. The most frequent words in written and spoken language tend to be grammatical words (e.g. *the*, *is*, *a*) (Leech et al., 2009) which, unlike lexical or 'content' words (nouns, verbs, adjectives and lexical adverbs), do not reveal too much in terms of texts' thematic content (Baker, 2006: 54). Therefore, it is advisable that, for the purposes of indexing, users scan through the frequency list and remove grammatical words, leaving only content words which can offer good candidates for index items. Table 1 shows the ten most frequent content words in the journal article.

Table 1. Top ten content words in the article, ranked by frequency[6]

Rank	Word	Frequency
1	*housing*	160
=	*social*	160
3	*jersey*	103
4	*states*	35
5	*policy*	28
6	*steel*	26
7	*public*	24
8	*uk*	23
9	*business*	22
=	*income*	22

By removing grammatical words, the resultant frequency list provides a number of promising candidates for index items, for example *housing*, *jersey*, *policy*, *uk*, *business* and *income.* Meanwhile, *steel* relates to the author's name, which could be useful if we wanted to include

[5] Note that the guidance provided in this article is generic and applies to how indexing could be carried out with the assistance of any corpus software package rather than #LancsBox specifically. Readers looking for specific guidance on how to use *#LancsBox* can consult the user guide and a series of instructional videos. The user guide, instructional videos and the tool itself can all be downloaded here: http://corpora.lancs.ac.uk/lancsbox/.

[6] Note that the frequency function also tells us how many texts a particular word occurs across within our corpus. However, since this list is based on a single text, we didn't feel it necessary to include this column here.

proper names in our index or create an index of names for our document. Other items, however, are less clear, including the noun, *states*, and the conceptually vague adjectives, *social* and *public*. Determining how these vaguer words are used, and whether or not any of the items in Table 1 should in fact be included in an index, requires us to go beyond this solitary list of words. As well as looking at the frequencies of individual words, we can also generate frequency information for sequences of two or more words, which can be useful for finding multi-word expressions that are frequent in our document and so could be included in the index. In *#LancsBox*, this tool is referred to as N-grams (where N denotes the number of words in the sequence – i.e. 2-grams for sequences of two words, 3-grams for three-word sequences, and so on). We removed expressions that consisted just of grammatical words (e.g. 'of the') or of an article preceding a noun (e.g. 'the public'), as these types of construction didn't advance our view of the themes granted by the words in Table 1 and were ultimately less revealing in terms of identifying potential index items. Table 2 therefore shows the top ten content word 2-grams in the article.

Table 2. Top ten content word 2-grams in the article, ranked by frequency

Rank	2-gram	Frequency
1	*social housing*	98
2	*chris steel*	21
3	*accountancy business*	20
=	*public interest*	20
5	*housing property*	17
=	*property plan*	17
7	*first time*	16
8	*social policy*	15
9	*housing stock*	13
10	*housing department*	11

The 2-grams in Table 2 provide some more context in which some of the vaguer items in Table 1 are used. For example, the 2-gram *social housing* indicates that just under two-thirds of the occurrences of these constituent terms (*social* and *housing*) occur together in this expression, reflecting the title of the article but also a potential theme in the text. Rather than indexing them separately, it might therefore be wise to create a single entry for *social housing* but also look at how these words are used in the cases where they do not occur in this context. Other 2-grams in Table 2 are revealing in this regard. The phrases *housing property*, *housing stock* and *housing department* respectively indicate the second, third and fourth most frequent contexts in which *housing* occurs, and all appear to offer potential index items. The same can be said for *social policy*, which indicates the most common use of *social* outside of *social housing*.

The 2-gram *chris steel* reflects the authors name, confirming our interpretation of the high-frequency item *steel* in Table 1, meanwhile the phrases *accountancy business* and *public interest* reflect the title of the journal in which this article was published, so are less likely to be of use for an index. In this way, the more contextualised view afforded by the n-gram

function can help to filter out ‘false friends’ which might at first appear suitable for an index but, on closer inspection, aren’t.

The 2-gram *first time* provides a particularly good example of the potential for this kind of output to reveal smaller segments of longer sequences. In this case, the phrase *first time* is consistently followed by either *buyer* or *buyers*, forming the 3-grams *first time buyer* and *first time buyers*. Taking these constructions together, the phrase *first time buyer(s)* could therefore provide a suitable index item. A similar case is provided by the phrases *housing property* and *property plan*, which consistently occur together to form the 3-gram *housing property plan*, as hinted by their identical frequencies. Again, this could provide a suitable index item. To reach these conclusions, we have gone beyond the word pairings revealed by the 2-grams, widening our search to 3-grams. And we can continue in this vein, looking at frequent chains of 3 or 4 words. However, the longer this chain of words becomes, the less likely we are to find so-called ‘lexical units’ (Zgusta, 1967) that are likely to be suitable as index entries (Biber et al. (2004) recommend looking at n-grams of up to four words in length).

In Tables 1 and 2, we have looked at just ten items. However, because the frequency technique gives the frequencies of every single word or n-gram, in a full analysis we can continue working our way down each list, removing grammatical items and exploring the suitability of content words for inclusion in the index. However, aside from being a rather laborious endeavour, corpus linguists are generally reticent to *prima facie* disregard any word or pattern in the corpus on the basis of it containing grammatical items, particularly if that word or n-gram occurs with a high frequency. A more statistically robust way of identifying words which reveal the content of the texts in our corpus is to use the keywords procedure.

3.3. Keywords

Keywords are words that occur with either a significantly higher frequency (positive keywords) or significantly lower frequency (negative keywords) in our target text or corpus compared against another corpus, known as the reference corpus. Words are deemed to be keywords by the computer based on statistical comparisons of the word frequency lists for each corpus. The choice of reference corpus is important here, as it shapes the keywords generated. When selecting a reference corpus, we usually want one that is similar in size to, or larger than, our target corpus. Ideally, the reference corpus should also represent a norm or ‘benchmark’ for the type of language contained in the document(s) for which we are creating an index. The benefit of this is that it helps to ensure that the resultant keywords flag up what is lexically distinctive about the text(s) in our corpus compared to other texts of a similar type.

For this case study, we compared our academic journal article against a purpose-built reference corpus comprising the remaining 79 ‘learned writing’ texts in BE06. The keywords generated by the computer in this instance will therefore represent those words that are distinctive of the journal article we are indexing compared against a random sample of other journal articles from different disciplines, written in the same language during the same period of time. This helps to safeguard against our keywords simply reflecting words that are indicative of the register of academic writing or the genre of journal articles. It is possible to

create such a bespoke reference corpus in much the same way that we prepared the target document for corpus analysis (Section 3.1), except this time you will want to include texts that represent the wider genre or register to which the document(s) being indexed belongs. Alternatively, there are a number of publicly-available word lists for existing corpora which could serve as suitable reference corpora, including the aforementioned BNC and its written and spoken components. If using more general reference corpora such as these, users will likely have to filter out keywords that indicate the distinctiveness of the genre or register of the target document(s), rather than its thematic content.

Once we have selected our reference corpus, we also have to choose a statistic to measure the 'keyness' of each word in our text(s) and decide whether or not we want to impose a minimum frequency threshold for candidate keywords. We generated keywords using log-likelihood (Dunning, 1993), a statistic which indicates how confident we can be that a keyword is indeed 'key' and characteristic of the text(s) in our target corpus. Corpus analysis programs offer a range of statistics for ranking keywords, all of which measure keyness in slightly different ways (see Gabrielatos (2018) for an overview). Log-likelihood is advantageous for indexing because it tends to produce high-frequency keywords which are likely to indicate the most characteristic themes – or 'aboutness' (Scott and Tribble, 2006: 59-60) – of the text(s) in our corpus. We also imposed a minimum frequency threshold of 5, meaning that a word had to occur at least 5 times for it to be considered as a possible keyword by the computer. Both the statistic used to rank keywords and the minimum frequency threshold can be adjusted by the user. Table 3 shows the top ten keywords, ranked by log-likelihood.

Table 3. Top ten keywords in the article, ranked by log-likelihood

Rank	Keyword	Frequency	Log-likelihood score
1	*housing*	160	988.03
2	*jersey*	103	703.33
3	*social*	160	573.47
4	*states*	35	185.97
5	*steel*	26	162.90
6	*accountancy*	20	136.27
7	*chris*	21	125.20
8	*property*	22	124.37
9	*stock*	19	121.58
10	*policy*	28	99.42

Although frequency is an important factor in identifying keywords, high frequency alone is not sufficient for a word to be judged as 'key'. Most important here is a word's frequency in the text(s) we are indexing relative to its frequency in the reference corpus. So, even though the word *social* is more frequent than *jersey* in our corpus, the latter was assigned a higher keyness value by the computer because its frequency relative to the rest of our data was higher than its relative frequency in the reference corpus. Because grammatical words like *and*, *the* and *of* have a comparable relative frequency across the corpora being compared, they

have been automatically filtered out of this keyword list. The keyword list therefore required less manual intervention on the part of the indexer, as the keywords flagged up by the computer are already revealing in terms of the content, or 'aboutness' of our text.[7]

We have seen the majority of these keywords in the word frequency list in Table 1. This includes the aforementioned *jersey* and *social*, but also *housing*, *states*, *steel* and *policy*. So, the emergence of these words as keywords indicates that they are not only frequent but also statistically salient in our text, confirming their suitability as potential index items. Meanwhile, that *accountancy*, *chris*, *property* and *stock* – all words we encountered as part of the 2-grams in Table 2 – were also keywords suggests that these terms are not just frequent company of high-frequency words like *social* and *housing* but were actually characteristic of our text compared to other journal articles. These items are therefore also worthy of inclusion in our index and are, by dint of their keyness, likely to warrant closer inspection. In a full analysis, we could expand this list, accounting for more and more keywords, all the while introducing more and more keywords that weren't indicated by the frequency measures.

As well as providing a rapid and replicable overview of the characteristic themes in our text(s), the keywords technique can also be useful for identifying words that are characteristic of a particular issue or edition of a serial publication. For example, rather than generate keywords for our article by comparing it against a random sample of other journal articles, we could construct a reference corpus of other articles published in the same journal. The result of this comparison would be keywords that were distinctive in our article when compared against other texts to do with the same topic; in this case, accountancy. The resultant keywords would therefore indicate words and themes that were characteristic of the particular article being indexed. In a similar vein, we could generate keywords for an entire issue by comparing all the articles in it against other issues from the same journal. Likewise, we could take a chapter within a book or edited volume and compare it against the rest of the chapters in that text, with the resulting keywords showing us what was linguistically and thematically characteristic of that particular chapter or section. In each case, it would be up to the user to create their own bespoke reference corpus that was comparable to the text(s) they were indexing.

Whichever way we choose to use keywords, one issue with the frequency and keywords techniques is that they both present candidate index items in isolated, relatively decontextualized lists. In other words, the word frequency and keyword lists reveal nothing about how their constituent words are actually used, while the 2-gram list provides more context but even this is limited. This is problematic in terms of determining these words' meanings and for deciding whether or not they should actually be included in the index. For these reasons, indexers typically have to go beyond lists of solitary words and expressions, as Booth (2013: 49) argues, '[i]ndexing is usually much more concerned with the meanings of words in combination, and with their relationships to other words, than with individual words as graphic or spoken items'. To gain such insight, we need to adopt a more contextualised

[7] Grammatical words are not precluded from being keywords; if a grammatical word has a relatively high frequency and so occurs more than what might be expected (based on the comparison with the reference corpus), then it can be key. Given this unexpectedly high frequency, such grammatical keywords would be worthy of consideration as index items.

view of our words of interest and inspect their use in-situ. We can begin to do this using the collocation technique.

3.4. Collocation

Collocation is a linguistic device whereby words, in associating strongly with one another, become bearers of meaning by virtue of co-occurrence. Collocation is typically judged to exist using a word association measure that tells us how often two or more words occur alongside one another, and whether this association is notable as a sizeable effect in our data (i.e. the words have a measurably strong preference to occur together as opposed to being randomly associated). Following Firth's (1957: 6) dictum that 'you shall know a word by the company it keeps', corpus linguists have long sought to learn about words' meanings and patterns of use by examining those words with which they tend to co-occur, or 'collocate'. Analysing a word's collocates can therefore provide insight into the textual context surrounding that word in the text(s) we are indexing, which can be useful not just for learning about its meanings but also whether or not it has a tendency to occur in frequent or fixed pairings that might not have been flagged up in searches of n-grams. To demonstrate this, we will look at the collocates of the joint-most frequent content word, and one of our top keywords, *social*.

Like generating keywords, deriving collocates requires the human user to make a series of procedural decisions, for example pertaining to span, method of calculation and use of a minimum frequency threshold, all of which will ultimately shape the number and type of collocates identified by the computer. The span refers to the number of words to the left and/or right of the user-determined search word within which we want to look for candidate collocates. Tighter spans will produce a smaller, more manageable number of collocates which occur within closer proximity to the search word. On the other hand, wider spans are likely to produce a higher number of collocates, some of which might not occur in such close proximity to the search word. We searched for collocates of *social* within a window of five words to the left and right of the search word (otherwise expressed as L5>R5). This is a fairly standard span in corpus research, as it provides a 'good balance between identifying words that actually do have a relationship with each other (longer spans can throw up unrelated cases) and [gives] enough words to analyse (shorter spans result in fewer collocates)' (Baker et al., 2013: 36).

We then have to decide how we will rank, score or 'cut-off' the candidate collocates. We can do this by ranking the collocates according to frequency of co-occurrence or by using a statistical measure. Corpus linguistic software packages offer a range of statistics for determining the strength of a collocational pairing (for an overview, see Gablasova et al., 2017). For the purpose of this demonstration, we will focus on the most frequent content word collocates, filtering out grammatical words, as this not only provides an insight into the types of meanings that *social* takes on in our text, but can also flag up potential frequent multi-word expressions featuring this term. Finally, we have to decide whether or not we want to impose a minimum frequency threshold. Lower thresholds produce larger numbers of collocates which can occur alongside the search word sparingly, whereas higher thresholds produce smaller numbers of more selective collocates. Most software packages operate with

default thresholds of between 3 and 5 but this can be adjusted by the user. Because we ranked our collocates by frequency – and so are only looking at the most frequent items – we didn't impose a minimum threshold. Table 4 shows the top ten content word collocates of *social*.

Table 4. Top twenty content word collocates of *social*, ranked by frequency

Rank	Collocate	Collocation frequency (left)	Collocation frequency (right)	Collocation frequency (total)
1	*housing*	19	108	127
2	*jersey*	14	13	27
3	*policy*	3	17	20
4	*social*	9	9	18
=	*property*	0	18	18
6	*stock*	3	14	17
=	*plan*	0	17	17
8	*units*	3	10	13
=	*policies*	4	9	13
=	*allocation*	4	9	13

The collocation measure tells us not only how often two words occur with each other, but also whether or not these collocates tend to occur to the left or right of our search word. For example, the most frequent collocate in Table 4, *housing*, occurs within the five words before and after *social* in our text a total of 127 times. Of these co-occurrences, *housing*, features to the left of *social* 19 times and to the right 108 times. Therefore, *housing* is more likely to follow *social* than precede it. The remaining items in this collocate list could all indicate themes around the word *social* that could serve as candidate index items, so a full analysis would investigate all of them. Like the frequency and keyword measures introduced earlier, we can continue down the list, all the while apprehending a fuller range of meanings of, and word pairings involving, our search word, repeating the process for other potential index items similarly identified using frequency and keywords. We could then use the collocation technique on other words and n-grams of interest, repeating the process to learn more about the phrases in which these words occur and the types of meanings they take on. However, to test and substantiate such hypotheses and to decide on whether or not they should be included in an index, it is useful to take an even more contextually-embedded perspective using a technique known as concordancing.

3.5. Concordance

Concordancing provides a means of viewing the data that allows us to examine every occurrence of a word or chain of words in its original contexts of use throughout the text(s) we are indexing. Building on our analysis of the collocates in the previous section, Table 5 below shows a random sample of concordance lines for the word *social*.

Table 5. Sample concordance lines for *social*

Line number	Context (left)	Search word	Context (right)
1	The direction that	social	housing policy has taken over the last twenty-eight years in
2	in the UK is a residual needs based model, where	social	housing is concentrated on those with the greatest need, but
3	or minimum wage paid employment that tends to severely handicap	social	mobility because of the unbalanced socio-economic mix (Hills, 2007). The
4	unbalanced socio-economic mix (Hills, 2007). The Choice Based Model The	Social	Housing Property Plan, makes no mention of choice based social
5	Social Housing Property Plan, makes no mention of choice based	social	housing allocation. However, we believe that the choice based model
6	that the choice based model provides a fairer approach to	social	housing allocation and that there is a current trend for
7	allocations, and towards community lettings, which aim to widen neighbourhood	social	and/or demographic mix (Cole et al, 2001). Choice based models
8	al, 2001). Choice based models may be used by trained	social	housing allocation staff to overcome some of the inherent problems
9	allocation staff to overcome some of the inherent problems with	social	housing estates, in the form of, Accountancy Business and the
10	deviancy, crime, dysfunctionality, drug & alcohol abuse and problems of	social	and economic deprivation (Murie, 1999). If the above social problems
11	of social and economic deprivation (Murie, 1999). If the above	social	problems are allowed to become manifest social housing becomes highly
12	If the above social problems are allowed to become manifest	social	housing becomes highly stigmatised and marginalised, which then tends to

With the search word running down the centre of the computer screen and a few words of context displayed to the left and right, concordance output can be very useful for spotting patterns that might be less obvious during more linear, left-to-right readings of the text(s) being indexed. Concordance output can be displayed in order of occurrence, in random order, or alphabetically according to the words surrounding the search word (rendering recurrent patterns more observable).

From the limited sample of concordance lines displayed in Table 5, we can identify a series of themes around the word *social* that gesture towards potential index entries. These include (corresponding concordance line number in brackets): social mobility (3), social and demographic mix (7) and social problems (11). In the 2-gram and collocation analyses presented earlier, we saw that in around three-quarters of its uses, the word *social* featured as part of the expression *social housing*. While this would therefore likely constitute an index entry in its own right, concordance lines containing this expression also indicate a number of potential themes surrounding this concept which could form the basis of sub-entries in an index. This includes social housing policy (1), a needs based model of social housing (2), the Social Housing Property Plan (4), the allocation of social housing (5), (6), (8), social housing estates (9) and the stigmatisation of social housing (12).

Concordancing is therefore a means to adopting a different perspective on the text(s) being indexed, providing the opportunity for the human indexer to carry out a closer reading of words or phrases of interest, bringing the process back to the manual, more traditional approaches to document indexing. If the perspective in Table 5 proved to be too narrow to ascertain the sense in which a word or phrase was used, it is also possible to expand the number of words displayed to the left or right of the search word and even access the original text in its entirety, usually by simply clicking on the search word displayed in the centre of the concordance line of interest. This final step in the corpus procedure outlined in this article is crucial, as it allows the human indexer to identify more granular themes in their document(s), confirm or revise their hypotheses about words' meanings and whether or not words should actually be included in the index (including discovering red herrings or 'false friends') and to group words and phrases into index headings that accurately reflect the content they relate to. In light of this case study, we now conclude this article by reviewing some of the opportunities and challenges, as we see them, of using corpus linguistics in indexing.

4. Opportunities and challenges of corpus linguistics for indexers

This article has introduced corpus linguistics and demonstrated some of the ways in which indexers can utilise corpus methods in the creation of indexes. In the case study in the previous section, we combined the established corpus techniques of frequency, keywords, collocation and concordance in an approach that involved: (i) initially identifying frequent and characteristic words and themes in our text(s) using frequency (including n-grams) and keywords, (ii) using collocation to gain a sense of the meanings that these frequent and characteristic words and phrases accrued throughout our text and then (iii) close reading of concordance lines containing words or combinations of words of interest in order to confirm or revise our hypotheses about their meanings and ultimately decide on whether or not these items should be included in our index. In addition to this approach, existing corpora can also provide a useful resource for indexers. While many modern dictionaries are usage-based, publicly-available reference corpora provide vast repositories of real-life language in which indexers can search for and scan patterns of use surrounding a particular word or phrase of interest to gain a sense of how it acquires meaning in speech and writing. General reference

corpora like the BNC offer reference sources that can be used alongside more traditional sources, like dictionaries and encyclopaedias, to allow indexers to learn about unfamiliar words' meanings.

Within the scope of this article, we have only been able to introduce some of the most established methods in corpus linguistics, demonstrating their use on just a single text. However, corpus linguistics offers a wide and increasing range of techniques beyond those covered here, all of which can contribute to the development of indexes for, in principle, any type of text. Indeed, while our case study has demonstrated the application of corpus methods to whole-document indexing, the techniques can be used, in theory, for the indexing of (section of) any type of document, pertaining to any subject matter, with any target readership. Although we have alluded to alternative possibilities for corpus linguistics for indexing at various points throughout this paper, we would encourage indexers to go beyond what we've covered here and to engage with the other techniques hosted by *#LancBox* and other tools.

Whatever type of text we are indexing, and whatever techniques we use to do it, critical to the corpus approach is the interplay between computational and statistical measures on the one hand, and human user-led readings of texts on the other. As well as providing novel perspectives on the text(s) being indexed, computational measures like frequency, keywords and collocation bring the added benefit that they provide more objective starting points for the development of indexes, for example by pointing the human user in the direction of words and word combinations that are not only frequent but also statistically salient. In allowing indexers to base their indexes on frequent and statistically-salient parts of documents, corpus techniques can therefore increase the objectivity with which indexers create their indexes, helping them to keep the influence of attitudes, beliefs and prejudices in-check, and advance beyond 'received ideas' and 'conventional wisdom' to produce more systematic indexes. This can be particularly advantageous for indexers working on texts related to topics of which they have limited knowledge or with which they otherwise wouldn't engage.

Although corpus methods can therefore help us to go beyond our intuition, these still have an important role to play in the creation of indexes. Computer software will not create an index for us. The frequency and keywords measures can gesture towards candidate index items. However, these need to be read and interpreted by the human indexer who must then decide on their suitability for inclusion in the index, as well as *how* they will be listed. At this point in the process, corpus methods like collocation and concordance can provide novel perspectives on the texts we're working on. This echoes Booth (2013), who argues:

> Indexing is not a mechanical word-spotting process. It involves intellectual activity - understanding and analysis of texts and their messages, selection of significant references to relevant topics, assembly of references, choice of suitable vocabulary for the representation of topics, and presentation in an accessible format. Headings in an index to a document do not consist solely of words appearing in its text, because part of the indexer's role is to supply additional headings that may be more familiar to certain index-users.
>
> (Booth, 2013: 3)

With all this in mind, although corpus techniques can provide more systematic, replicable and objective techniques for indexing, users of corpus methods should nevertheless take care when making claims about objectivity. While it is the case that computer programs do not make errors and are not subject to the types of ideological and cognitive biases that humans are, both the designers and users of these programs are. All users, including indexers, should therefore avoid uncritical overreliance on corpus techniques and be self-reflexive about the influence that their own choices and biases are likely to have had on the indexes they produce.

Having foregrounded the advantages of corpus linguistics methods for indexers, it is also worth considering some of their possible limitations with respect to this area of application. One limitation of techniques based on frequency and statistics – like frequency, keywords and collocation – is that they work better for higher-frequency items. However, significance is not always reflected in frequency. While the keywords technique can help to overcome this limitation, as it does not depend solely on raw frequency, accounting for significant but infrequent words and phrases will likely require the human user to inspect the lower reaches of the word frequency and keyword lists. Another limitation of corpus linguistics relevant to indexers concerns what corpora presently have the capacity to represent. As we discussed earlier, the rendering of any text or collection of texts into a corpus is a transformative process, the product of which bears important differences to the original(s). Because corpora are stored in a plain text (.txt) format, the texts they contain are reduced to their linguistic elements only, meaning that modes like gesture and image are removed. For indexers, this means that the conversion of a document into a plain text corpus will exclude all non-linguistic elements (e.g. photographs, images, graphs, emoticons). Neither will it discriminate according to such features as typeface, font and colour of the text. Advances in the development of multimodal corpora (e.g. Adolphs and Carter, 2013) mean that collecting and analysing corpora representing modes like gesture and image is easier now than it ever has been. Yet, for now, the vast majority of corpus software packages and studies of corpora remain monomodal, accounting for language only. However, ongoing efforts to develop corpus methods that are more finely attuned to the visual components of texts could offer promising innovations for indexers, who are often required to index the content of the images as well as language that texts contain. As the title of this section suggests, we would regard these limitations as challenges rather than deterrents, and we would strongly encourage indexers to engage with the literature and techniques introduced in this paper and attempt to use corpus linguistics methods in the development of their own document indexes.[8]

[8] Readers wanting to learn more about corpus linguistics methods and applications might be interested in the freely-available Massive Open Online Course (MOOC), *Corpus Linguistics: Method, Analysis, Interpretation* (https://www.futurelearn.com/courses/corpus-linguistics). This course offers a practical introduction to the methodology of corpus linguistics and is targeted at newcomers to the field.

References

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

Anthony, L. (2019) *AntConc (Version 3.5.8)*. Tokyo: Waseda University.

Aston, G. and Burnard, L. (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

Baker, P. (2009) The BE06 Corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14(3), 312–337.

Baker, P., Gabrielatos, C. and McEnery, T. (2013) *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Cortes, V. (2004) If you look at …: lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.

Biber, D. and Reppen, R. (2015) *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.

Booth, P. F. (2013) *Indexing: The Manual of Good Practice*. Munich: De Gruyter.

Boulton, A. (2017) Corpora in language teaching and learning. *Language Teaching*, 50(4), 483–506.

Brezina, V., McEnery, T. and Wattam, S. (2015) Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2), 139–173.

Crawford, P. and Brown, B. (2010) Health communication: Corpus linguistics, data driven learning and education for health professionals. *International English for Specific Purposes Journal*, 2(1), 1–25.

Day, R. E. (2014) Indexing it all: the subject in the age of documentation, information, and data. Massachusetts: MIT Press.

Dunning, T. (1993) Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Fetters, L. K. (2014) *Handbook of Indexing Techniques: A Guide for Beginning Indexers* (Fifth Edition). New Jersey: Information Today Inc.

Firth, J. R. (1957) *Papers in Linguistics 1934–1951*. Oxford: Oxford University Press.

Gablasova, D., Brezina, V. and McEnery, T. (2017) Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, 67(S1), 155–179.

Gabrielatos, C. (2018) 'Keyness Analysis: nature, metrics and techniques'. In: C. Taylor and Marchi, A. (Eds.), *Corpus Approaches to Discourse: A Critical Review*. London and New York: Routledge, pp. 225–258.

Garside, R., Leech, G. and McEnery, A (Eds.) (1997) *Corpus Annotation*. London: Longman.

Hanks, P. (2012) The Corpus Revolution in Lexicography. *International Journal of Lexicography*, 25(4), 398–436.

Hardie, A. (2012) CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. and Rychlý, P. (2008} 'GDEX: Automatically Finding Good Dictionary Examples in a Corpus', in Bernal, E. and DeCesaris, J. (Eds.), *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra, pp. 425–433.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014) The Sketch Engine: ten years on. *Lexicography*, 1, 7-36.

Leech, G. (1991) 'The state of the art in corpus linguistics', in K. Aijmer and B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, pp. 8–29.

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

Leech, G., Hundt, M., Mair, C. and Smith, N. (2009) *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Love, R., Dembry, C., Hardie, A., Brezina, V., and McEnery, T. (2017) The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319–344.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

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

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies: An Advanced Resource Book*. London and New York: Routledge.

O'Keeffe, A. and McCarthy, M. (Eds.) (2010) *The Routledge Handbook of Corpus Linguistics*. London and New York: Routledge.

Rayson, P. (2008) From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519–549.

Scott, M. (2016) *WordSmith Tools version 7*. Stroud: Lexical Analysis Software.

Scott, M. and Tribble, C. (2006). *Textual Patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Wellisch, H. H. (1988) Indexing and abstracting: a current-awareness bibliography. *The Indexer*, 16(2), 107–110.

Wright, D. and Brookes, G. (2019) 'This is England, speak English!': a corpus-assisted critical study of language ideologies in the right-leaning British press. *Critical Discourse Studies*, 16(1), 56–83.

Zgusta, L. (1967) Multiword Lexical Units. *Word*, 23(1-3), 578–587.