10
What are the basics of analysing a corpus?

Jane Evison

1. Analysing a corpus: the basics

Manipulating corpus data: frequency and concordancing

In themselves, corpora can tell us nothing but because they are collections of electronic texts, they are susceptible to computerised analysis using corpus access software. As Hunston (2002: 3) puts it, 'a corpus does not contain new information about language, but the software offers us a new perspective on the familiar'. In order to gain this new perspective, the first analytical steps generally involve two related processes: the production of frequency lists (either in rank order, or sorted alphabetically) and the generation of concordances (examples of particular items in context; see Tribble, this volume). There is an increasing range of software available to carry out such processes, from established commercial software such as *WordSmith Tools* (Scott 1999), *Monoconc Pro* (2000) and *Word Sketch Engine* (Kilgarriff *et al*. 2004) to freeware downloadable from the internet. These two corpus-handling techniques – generating frequency lists and concordances – are built on the very basic foundation that electronic collections of texts can be searched very rapidly. This means that automatic frequency list generation can quickly produce a complete list of all the items in a corpus, ranging from the most ubiquitous ones, the frequency of which may run into millions in the largest corpora, to those more unusual items which occur just once in a particular corpus. Concordance analysis, also a basic technique, begins with a specific item that the researcher has decided to search for. This search brings onto the screen all the examples of the searched-for item, in context. These two basic operations represent two core corpus-handling techniques, but of course, simply counting items or displaying their occurrences does not actually tell us anything in itself; it is the associated analysis, which may be both quantitative and qualitative, which provides the insights.

Considering the issues

First of all, there is the question of how much data is needed (see Nelson, this volume). In the early years of corpus linguistics, there was certainly a drive towards the analysis of larger and larger corpora, with the unwritten assumption that 'biggest is best' (Kennedy 1998). This can be attributed in part to the excitement engendered by the possibility of collecting a million words of data (a figure that was for some time put forward as being the minimum size for a corpus), and to the specific needs of the dictionary-compilers who were important early users of corpus data (see Walter, this volume). These lexicographers needed large data sets so that they could extract sufficient examples of infrequent words to allow the production of reliable descriptions of their use. However, subsequent analysis of corpora by applied linguists with different interests, such as the investigation of high-frequency grammatical patterns or discourse features, has shown that having very large corpora can mean that too much data is generated if one is searching for very frequent items or interested in carrying out detailed analysis. Solutions to the problem of generating too much data include random sampling and the construction of smaller sub-corpora. This latter approach has proved successful for the analysis of high-frequency grammatical features such as pronouns and verb forms. For example, Biber (1990) demonstrates that just 1,000 words of data are able to produce results that are reliable, and Tribble (1997) argues convincingly that if a register is very specialised, a smaller corpus will be adequate to provide insight into the features of that register. Two examples of this kind of small study are Koester's (2006) investigation of workplace discourse, which uses a corpus of just under 34,000 words and O'Keeffe's (2003) study of radio discourse, based on a sample of 55,000 words of phone-in data.

If one does not wish to study a small quantity of data, it can be useful to randomise searches for particularly frequent items. For example, the corpus software can be asked to display concordances for one occurrence in 50, 100 or 500 (Scott 1999). On the other hand, Sinclair (1999) suggests that, if one is looking for patterns of occurrence, a possible solution is to repeat the process of generating thirty random lines (a number that can be comfortably viewed on the screen at one time) until no new patterns are observed. The results of the analysis of a particular corpus can also be validated against those from other comparable corpora. For example, in their paper on basic spoken vocabulary, McCarthy and Carter (2003) validate the results of the analysis of the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) with a frequency list generated for a five-million-word spoken element of the British National Corpus (the BNC). It is also possible to triangulate corpus findings. For example, in the case of the analysis of spoken language, this can be achieved by observing the kinds of spoken encounters which make up the corpus, or by questioning participants though interview or questionnaire.

The sections which make up the rest of this chapter exemplify the basics of frequency and concordance analysis. Frequency analysis can be done by anyone who has a collection of electronic texts, basic computer skills and the appropriate software. Concordancing is even more accessible, and can be carried out on corpora searchable via the internet as well as on those stored on a personal computer. It is important to point out here that the example analyses that are presented in this chapter are limited to basic techniques involving counting and searching for single items, rather than the kinds of multi-word units referred to in Greaves and Warren (this volume).

2. Exploring word frequency lists

Displaying frequency data

The first basic corpus technique that we will consider is that of frequency analysis. When we generate a frequency list for a particular corpus, the software searches every item in that corpus in order to establish how many tokens there are in total – at the simplest level a token and a word can be considered to be the same thing – and how many different types constitute this total. The software then outputs the final counts as a frequency list, which can be displayed in rank order of frequency or in alphabetical order (or, in the case of *WordSmith Tools*, in reverse word order, useful, for example, if one is interested in the frequency of suffixes or inflexions). In Table 10.1 we can see the beginning of a rank order frequency list for a small corpus of just over 4,000 words of paired discussion tasks. In addition to the rank order (N) and the raw frequency (actual number of occurrences) of each token, we can see the percentage of tokens in the whole corpus that each frequency count represents.

Alphabetical frequency lists can be generated at the same time as rank order ones, but give a different picture of the same frequency distribution. In order to exemplify this, Table 10.2 shows the tokens in positions 265–74 in the alphabetical list for the same small corpus of discussion tasks.

The alphabetical frequency list extract in Table 10.2 raises two important issues in relation to frequency list generation. First, it shows that the corpus software counts tokens, which may or may not be words in the traditional sense. As well as including vocalisations such as *huh* and *hum*, the first two items on the list, most software will count

Table 10.1 Frequency list (rank order)

N	*Token*	*Freq.*	*%*
1	the	203	4.76
2	I	129	3.02
3	a	116	2.72
4	and	109	2.55
5	it	89	2.09
6	to	86	2.02
7	think	81	1.9
8	of	80	1.87
9	you	78	1.83
10	yeah	76	1.78

Source: Data from Evison (2001).

Table 10.2 Frequency list (alphabetical)

N	*Token*	*Freq.*	*%*
265	huh	14	0.33
266	hum	4	0.09
267	hundred	1	0.02
268	hundreds	2	0.05
269	I	129	3.02
270	I'd	5	0.12
271	I'm	6	0.14
272	idea	10	0.23
273	if	15	0.35
274	imagine	1	0.02

Source: Data from Evison (2001).

the contractions *I'll* and *I'd* as single items even though they are traditionally seen as contractions of two separate words. Although corpus linguists tend to consider contractions such as *I'll* and *I'd* as single words, as major corpus-based grammars attest (Biber *et al.* 1999; Carter and McCarthy 2006), more traditional grammars may consider them as two. Second, there is the issue of lemmatisation. The alphabetical frequency list in Table 10.2 shows that when it was generated, the program listed all the different tokens in the corpus under investigation separately, rather than grouping together related forms such as *hundred* and *hundreds,* which instead have separate entries. This is generally the way that basic frequency list generation works. However, some frequency software (e.g. *WordSmith Tools* version 5) can generate frequency lists for lemmas (e.g. one count for the lemma *smile*, which would include *smile(s), smiled* and *smiling*). Nevertheless, the software used for this process is not always sophisticated enough to pick up other similarities such as comparative and superlative forms of adjectives, or to avoid conflating the counts for unrelated words (e.g. the verb *sit* and the unrelated noun *site*). A useful discussion of the limitations of lemmatisation can be found in O'Keeffe *et al.* (2007: 32–3).

Exploiting frequency data

Having established what frequency lists look like, we can now turn our attention to what frequency lists can tell us. Frequency lists can be useful documents for lexicographers (see Walter, this volume) and language syllabus and materials designers (see McCarten, this volume). Their importance is underlined by the range of frequency information that is available. In the case of the BNC, for example, a selection of lists (Leech *et al.* 2001) is supported by updated information on the BNC website. The Compleat Lexical Tutor (available online) utilises the Academic Wordlist (see Coxhead, this volume) and the much older General Service List (West 1953) as the basis of its lexical profiling programs. Such programs rely on the establishment of frequency bands, something that Sinclair (1991) highlights as particularly useful. Research into spoken language such as McCarthy (1998, 1999) and McCarthy and Carter (2003) exploits frequency bands, using the sudden drop-off in frequency which occurs after about 1,800 words in a rank frequency list generated from the CANCODE corpus to argue that a basic spoken vocabulary of English must include these 1,800 items. Further discussion of frequency bands can be found in O'Keeffe *et al.* (2007: 31ff). Finally, frequency lists form the basis of more complex statistical measures such as Mutual Information (MI) scores, *t*-scores and *z*-scores, which express the strength of collocations; in other words the likelihood of items – such as *dark* and *hair* – co-occurring. See chapters by Moon and Greaves and Warren, this volume, for further discussion, as well as McEnery *et al.* (2006).

Comparing frequency lists

It can be useful to compare the rank order of items in two or more corpora by looking at them side by side, as in Table 10.3, which shows the top ten most frequent items in 50,000 words of conversation extracted from the BNC, and the top ten items from a corpus of 54,000 words of podcast talk (TESOL Talk from Nottingham, or TTFN). The TTFN corpus is made up of informal broadcast conversations between two university lecturers and occasional guests about topics relating to the subjects that they are teaching on an MA programme for English Language Teachers.

Table 10.3 Comparison of rank frequency

N	*BNC*	*TTFN*
1	**I**	the
2	**you**	and
3	it	of
4	the	**I**
5	and	a
6	a	to
7	to	that
8	that	**you**
9	yeah	in
10	oh	it

Source: BNC data extracted from the BNC Baby. TTFN data from Evison (2009).

In Table 10.3, we can see that while eight out of the ten items are common to both lists, the first and second person pronouns *I* and *you* occur higher up the list for intimate conversations and lower down that for the more academic conversations. We can understand the relatively high frequency of the interactive pronouns *I* and *you* in casual conversation in terms of the topicality of the participants themselves, and their orientation towards each other (McCarthy 1998; Biber *et al.* 1999). In the frequency list for the academic podcast conversations, however, we find that *I* and *you* have been 'displaced'; the higher positions of *the* and *of,* suggesting that topics are being referred to using noun phrases (e.g. ***the*** *teaching* ***of*** *speaking;* ***the*** *importance* ***of*** *grammar*). Of course, these hypotheses need to be investigated further by more detailed analysis of examples in context.

Normalisation

In order to compare frequency counts across corpora of different sizes, a process of normalisation is required. This process involves extrapolating raw frequencies from the different-sized corpora which are being compared so that they can be expressed by a common factor such as a thousand or a million words. For example, the pronoun *we* occurs 2,142 times in a sub-corpus of meetings extracted from the BNC Sampler corpus, and 2,666 times in another sub-corpus of the BNC Sampler made up of casual conversation. However, because the two corpora are of such different sizes, these raw frequencies mean very little relative to each other. In order to normalise the figure for the meeting sub-corpus, the raw frequency of 2,142 is divided by 148,624 (the total word count of the meeting sub-corpus) and multiplied by 1,000, giving a figure of fourteen occurrences per thousand words. In order to compare this normalised frequency with that of *we* in the sub-corpus of casual conversation, we take the raw frequency of 2,666, divide it by 483,913 (the total word count in the conversational sub-corpus) and multiply by 1,000, which results in a normalised count of six occurrences per thousand words. We can now see that *we* is more than twice as frequent in the sub-corpus of meetings (fourteen occurrences per thousand) than in that of casual conversation (six occurrences per thousand). Its frequency in the meeting sub-corpus is related to both the demands of immediate group reference and the construction of corporate authority and responsibility, issues which McCarthy and Handford (2004) discuss in relation to the use of *we* in business English specifically (see also Handford, this volume).

3. Exploring key-word lists

Keyness

Key words are not necessarily the most frequent words in a corpus, but they are those words which are identified by statistical comparison of a 'target' corpus with another, larger corpus, which is referred to as the 'reference' or 'benchmark' corpus. This identification involves the automatic comparison of word lists using software such as the *WordSmith Tools Keyword* program. A key-word list includes items that are either significantly frequent (positive key words) or infrequent (negative key words), and is a useful starting point for many corpus linguistic analyses (Scott 1999 and this volume; Hunston 2002; Reppen and Simpson 2002; McEnery *et al.* 2006). Although there are several ways of calculating statistical significance available, a test of 'keyness' is especially useful for the analysis of corpus data because, being based on a log-likelihood (LL) test (Dunning 1993), it is not predicated on the assumption that data have a normal distribution (see McEnery *et al.* 2006: 55–6).

Positive key words

Table 10.4 shows the top ten (positive) key words generated when the frequency list for a 75,000-word sub-corpus of sociology and history essays extracted from the British Academic Written English (BAWE) corpus of student writing is compared with a larger, more general corpus made up of five million words of written English (the reference corpus) extracted from the BNC Sampler. The table details both raw frequencies and percentages; where there is no percentage given for the reference corpus, this is because the value is too small to be of use to this comparison. The zero figures in the *p*-value column simply indicate that the results are significant; it is the keyness values in the preceding column which are useful for comparative purposes.

We can see in Table 10.4 that four of the nouns with the highest keyness values (*class*, *society*, *women*, *power*) and the only adjective (*social*) reflect typical sociological or historical topics of the essays in the corpus. At first sight, however, the reason for the high keyness value of *archer* (a person who fires an arrow) is not apparent. This is a case where the analyst is likely to examine the item in context (through a concordance search for *archer*)

Table 10.4 Positive key words in sociology and history texts

N	*Key word*	*Freq.*	*%*	*RC Freq.*	*RC %*	*Keyness*	*P Value*
1	social	372	0.5	229	0.02	1,269.90	0.000
2	p	394	0.53	294	0.03	1,258.83	0.000
3	class	259	0.35	159	0.01	884.6	0.000
4	society	222	0.3	179	0.02	688.54	0.000
5	women	263	0.35	341	0.03	658.91	0.000
6	power	209	0.28	269	0.03	525.51	0.000
7	archer	87	0.12	1		465.55	0.000
8	of	3,408	4.59	33,798	3.15	410.23	0.000
9	ibid	67	0.09	0		366.84	0.000
10	that	1,110	1.5	8,555	0.8	333.6	0.000

Note: RC = Reference Corpus.
Source: Data extracted from BNC Sampler and BAWE Corpus.

Table 10.5 Negative keywords in sociology and history essays

N	*Key word*	*Freq.*	*%*	*RC Freq.*	*RC %*	*Keyness*	*P Value*
615	she	20	0.03	2,485	0.23	–209.44	0.000
616	I	99	0.13	6,218	0.58	–356.58	0.000
617	you	17	0.02	4,044	0.38	–415.28	0.000

Note: RC = Reference Corpus.
Source: Data extracted from BNC Sampler and BAWE Corpus.

in order to find some kind of explanation for its relatively high frequency. In this example, such an examination shows that the key item is in fact *Archer,* a very commonly cited reference in a number of the essays. The statistical significance of the two abbreviations in the essay corpus (*p* and *ibid*) is also related to referencing: the convention of writing *p* for page number, and *ibid* to indicate reference to a previously cited work. The remaining two items in Table 10.4 are the grammatical items *of* and *that.* Both have strong associations with academic writing: the former because it is a constituent in post-modified noun phrases (e.g. *the end* ***of*** *the Cold War*) and the latter because of its multi-functionality – not only does *that* function as a subordinator, it also follows reporting verbs, often as part of *it* patterns such as *it is reported* ***that*** (see Biber 2006; O'Keeffe *et al.* 2007).

Negative key words

We can also identify negative key words, or those items which occur significantly less often in the target corpus than in the reference one. Table 10.5 shows the three most significantly infrequent words in the same corpus of sociology and history essays for which we displayed the positive key words.

Here we can see that the three most significantly underrepresented words are pronouns (*you*, *I*, *she*) and that their unusually low frequency reflects the impersonal style of academic writing compared with the more personal focus of the reference corpus, which contains a broader range of written genres including many which tend to be more individually oriented, such as works of fiction, letters and journalistic prose. This difference in orientation of the writers in the essay corpus – less focus on individual women and more on women as a group – is nicely exemplified by the contrasting negative value for the third person singular female subject pronoun *she* (-209.44) and the positive keyness value (+658.91) for the more general female plural noun *women*.

4. Exploring concordance lines

Online concordancing

Also known as KWIC (key word in context) analysis, concordance analysis is probably the first basic corpus analytic technique that many people interested in corpus analysis undertake. This is because of the increasing numbers of websites which offer internet users the chance to search their corpora for specific words or phrases. The COBUILD Concordance and Collocations Sampler and the Corpus-based Concordances which make up part of The Compleat Lexical Tutor (available free online) are two examples of online concordancing programs particularly popular with language teachers and learners.

These websites, like many others, allow users to carry out concordance searches of the corpora which they hold, although they do not let them download the corpus files themselves. Corpora such as the BNC, in addition to being purchased, can be freely searched online. In general there are restrictions on how many results are displayed when doing online concordancing, and only random samples may be available. However, multiple searches for the same item can be carried out in order to generate different random samples which can then be accumulated, a procedure which can help validate the results. Some online corpora, such as the Michigan Corpus of Academic Spoken English (MICASE), do not limit the number of hits when concordancing, although if a very frequent item is being searched for, the software does give the option of limiting the numbers of results that are displayed.

Concordancing is a valuable analytical technique because it allows a large number of examples of an item to be brought together in one place, in their original context. It is useful both for hypothesis testing and for hypothesis generation. In the case of the latter, a hypothesis can be generated based on patterns observed in just a small number of lines, and subsequently tested out through further searches. However, as Hunston argues, when using concordancing to test a hypothesis, it is important to consider items which do not appear to support the hypothesis being tested, and, if necessary, review one's hypothesis rather than rejecting the forms themselves (2002: 52ff) (see also Hunston, this volume).

Searching and sorting

A concordance program allows any item (a single word, a wild-card item or a string of words) to be searched for within a corpus, and the results of that search displayed on the screen (see Tribble, this volume). These results are known as concordances or concordance lines. All the occurrences of the target item (or node word) are displayed, vertically centred, on the screen along with a preset number of characters either side (however, see Tribble, this volume, for discussion of a range of different concordance types). For example, if we search – with the wildcard asterisk – for the target item *shop** in a corpus of discussion tasks, all words beginning with these letters will be displayed as in the list below, which contains *shop, shops* and *shopping.*

```
 1    t know about that, erm, the shopping mall. I' m not so sure about the
 2    Bournemouth has got enough shopping centres I suppose … The people won' t go
 3   ' t it really? Cos they like shopping more than boys. Yeah. I suppose so …
 4    uppose really … and time to shop, and money to shop. How' s it gonna
 5     . I' m not so sure about the shopping mall myself … I can' t imagine it on
 6   n' t there? There' s loads of shops isn' t there? Hundreds of things. There' s
 7    three options we have are a shopping centre, a park or entertainment
 8   ey don' t have really enough shop, er big shopping malls in Bournemouth.
 9     ppose really … and time to shop, and money to shop. How' s it gonna
10  k their cider. Erm, OK … This shopping mall. shopping mall. It will attract,
```

(Data from Evison 2001)

Concordancing is particularly useful because the lines displayed can be sorted. Although the node item is relatively easy to see in randomly generated concordance lines, if we sort them, regularities of occurrence can be identified more easily. For example, below the same concordance lines for *shop** have been sorted alphabetically first by the centre item and then by the first and second words to the right (usually expressed as centre, R1, R2 in the options offered by the software). Now they have been sorted, we can see any regularities more clearly.

```
 1    ppose really … and time to shop, and money to shop. How' s it gonna
 2   ey don' t have really enough shop, er big shopping malls in Bournemouth.
 3   uppose really … and time to shop, and money to shop. How' s it gonna
 4   three options we have are a shopping centre, a park or entertainment
 5   Bournemouth has got enough shopping centres I suppose … The people won' t go
 6   t know about that, erm, the shopping mall. I' m not so sure about the
 7    . I' m not so sure about the shopping mall myself … I can' t imagine it on
 8  k their cider. Erm, OK … This shopping mall. shopping mall. It will attract,
 9    ' t it really? Cos they like shopping more than boys. Yeah. I suppose so …
10   n' t there? There' s loads of shops isn' t there? Hundreds of things. There' s
```

These very simple concordance lines demonstrate the versatility of concordance programs, and show the potential that they have to provide insight into the *typicality* of item use. In particular, concordance analysis can provide evidence of the most frequent meanings, or the most frequent collocates (co-occurring items) such as *shopping centre* or *shopping mall* (see Biber *et al.* 1998; Scott 1999; Tognini Bonelli 2001; Hunston 2002; Reppen and Simpson 2002; McEnery *et al.* 2006). A more detailed discussion of concordancing can also be found in Tribble (this volume), and exemplification of its role in data-driven learning in chapters by Gilquin and Granger, and Sripicharn, this volume.

5. Exploring discourse

Corpora and discourse

While it is true that large-scale analysis of part-of-speech tagged corpora, such as that carried out by Biber and his colleagues (see Biber, this volume), can tell us a great deal about the differences and similarities between different types of discourses (or registers, as Biber prefers to call them; see Biber, this volume), it is also true that the use of basic corpus analysis of untagged corpora (using freeware or relatively inexpensive proprietary software) can yield useful insights into discourse-level features of language (see O'Keeffe *et al.* 2007 for a recent survey of such analysis). For this reason, the final section of this chapter focuses on what basic corpus techniques can tell us about discourse, although more details about the kinds of methodology and data that are required to analyse discourse can be found in Thornbury, this volume. Here we will exemplify how basic corpus techniques can be used to explore the discourse functions of some common items in spoken language, focusing primarily (but not exclusively) on concordancing because this technique does not necessarily require the user to compile or purchase their own corpus; as we observed earlier, internet users can search a range of corpora online (see Lee, this volume).

Exploiting basic corpus techniques

Investigating the discourse functions of particular forms is complicated by the fact that these items often have clause-level, as well as discourse-level, functions. The word *now* is a good example. Although most dictionary entries for *now* highlight its temporal meaning of 'at the present time' (i.e. *now* as a temporal adverb), many users of English will be very well aware of its use as a focusing device (e.g. ***Now****, what did we do in class yesterday – can anyone remember?*). Researchers such as Swales and Malczewski (2001) have in fact investigated this use of *now* in the MICASE corpus of academic talk, and a quick look at a frequency list generated for that corpus shows us that *now* is in 68th position overall, occurring over 4,000 times and being used at least once in every academic event in the corpus. If we go on to generate concordance lines for *now* in MICASE and sort them three places to the left (L1, L2, L3), we are able to see when *now* is functioning as a temporal adverb and when it has broader discourse-level functions:

```
 1  t, accumulation. <PAUSE DUR = ″:05″/> now let′s, we′ve now covered the t
 2   t we know today. <PAUSE DUR = ″:05″/> now you will see something that wi
 3   it′s an agonist. <PAUSE DUR = ″:04″/> now, let′s see there were a couple
 4     s. and look how many we′ve already, now i don′t know if you wanna read
 5     e strategy we′re kind of engaged in now is a strategy of making sure p
 6    and they′re writing what they know. now do we fault them for writing w
 7    e, just give them some stars. okay? now. let′s, let it sit for a while
 8   relevance, in revolutionary Paris. now i don′t know whether Schikaned
 9      yeah </U> and </U> <U WHO = ″S1″> now you see this here, this is a c
10   of, um, anatomy. </U> <U WHO = ″S1″> now do you believe that you′re gon
11   hat′s the (skull) </U> <U WHO = ″S1″> now do you notice what happened wi
12    WHO = ″S3″> yeah </U> <U WHO = ″S1″> now, you′re going to be meeting wi
13     ″> right. </U> i the state here we now have a bike helmet law that sa
14    s happened in Minnesota, is that we now have a huge population of Cana
15      you know it seems to, seems to you now that you′re in this classroom
```

(Data from the MICASE corpus)

By sorting to the left in this way, we are able to see easily that in only four lines is *now* clause-medial, and acting as a temporal adverb (lines 5, 13–15). In the remaining eleven lines it is clause-initial and has discourse-level functions, signalling a change of focus or topic (see Carter and McCarthy 2006: 112). In three of these cases, we can also see that *now* occurs immediately after a pause (denoted by < PAUSE DUR = " … "/ >), and in four after a new speaker tag (denoted by < U WHO = "S1" >). If we now right sort these eleven lines (R1, R2, R3), we can see more clearly how *now* functions at a discourse level.

```
 1    and they′re writing what they know. now do we fault them for writing w
 2   of, um, anatomy. </U> <U WHO = ″S1″> now do you believe that you′re gon
 3   hat′s the (skull) </U> <U WHO = ″S1″> now do you notice what happened wi
 4     s. and look how many we′ve already, now i don′t know if you wanna read
 5   relevance, in revolutionary Paris. now i don′t know whether Schikaned
 6    e, just give them some stars. okay? now. let′s, let it sit for a while
 7   it′s an agonist. <PAUSE DUR = ″:04″/> now, let′s see there were a couple
 8  t, accumulation. <PAUSE DUR = ″:05″/> now let′s, we′ve now covered the t
 9      yeah </U> and </U> <U WHO = ″S1″> now you see this here, this is a c
10    WHO = ″S3″> yeah </U> <U WHO = ″S1″> now, you′re going to be meeting wi
11   t we know today. <PAUSE DUR = ″:05″/> now you will see something that wi
```

(Data from the MICASE corpus)

With the concordance lines sorted in this way, we can see four patterns associated with *now*. The signalling of question forms shows up more clearly because the occurrences of the auxiliary verb *do* appear underneath each other (lines 1–3). Lines 4 and 5 exemplify the indirect structure *I don't know if/whether …* , while in lines 6–8 we can observe *now* introducing *let*-imperative structures, and in lines 9–11 the use of *now* + *you* which allows the speaker to address the listeners specifically in order to focus their attention. Interestingly, we can see how, in these concordance lines, *now* precedes language such as the personal pronouns *I* and *you,* questions, and *let*-constructions which has been shown to be part of an interactive teaching style (e.g. Morell 2004).

This discussion of the functions of *now* in pedagogic discourse illustrates how we can use concordancing to look more deeply behind initial quantitative results gained from frequency analysis, or begin the research process by searching for examples of a specific word that is of special interest, perhaps because of a hunch or because of previous research (or a mixture of the two). When used in this way, these basic techniques can be understood as part of a corpus-based approach which 'does not go to the extreme of rejecting intuition while attaching importance to empirical data' (McEnery *et al.* 2006: 7).

Combining corpus techniques with other approaches: the case of turn-taking

In addition to using quantitative data analysis, many researchers also explore their data through qualitative analysis (e.g. McCarthy 1998; Carter 2004; O'Keeffe 2006; O'Keeffe *et al.* 2007) and small domain-specific corpora such as those discussed by Koester (this volume) are particularly susceptible to qualitative readings. In fact, there are growing numbers of researchers who suggest that combining automatic corpus analytic techniques with more fine-grained qualitative investigation, such as Conversation Analysis (CA), is a robust methodology for dealing with the intricacies of spoken language in particular (e.g. Tao 2003; O'Keeffe 2006; Walsh and O'Keeffe 2007). In this respect, Walsh and O'Keeffe (ibid.: 123) argue that there can be a 'synergy of CA and corpus approaches', which they suggest can 'offer a greater understanding of … interactional processes'. Such processes include turn-taking, a feature of discourse very strongly associated with CA, but one which is increasingly being investigated using corpus techniques in conjunction with fine-grained analysis. Turn-openings are one position in the discourse which have been exploited in this way (e.g. Tao and McCarthy 2001; McCarthy 2002, 2003; Farr 2003; Tao 2003; Evison 2008).

Because new turns in corpus data are identified by speaker tags such as < U WHO = "S1" > (the tags used in the MICASE corpus and exemplified earlier), it is possible to establish the regularised ways in which turns open by generating concordance lines for all the new speaker tags in a corpus using a wildcard search. In the case of MICASE, we simply need to insert a wildcard (indicated by an asterisk) so the search item becomes < U WHO = "S*" >. By sorting to the right, we are able to see turn-initial regularities, such as the occurrence of *yeah no* as shown below.

```
1    this, it's just to show you know, <U WHO = ″S4″> yeah no no i know i'm
2    o-math i didn't mean to there but <U WHO = ″S4″> yeah no no no no no w
3  ay oh i though you said on Tuesday <U WHO = ″S1″> yeah, no. okay and yea
4    okay, okay okay? does that help? <U WHO = ″S3″> yeah no panicking it's
5   it in the economics part? um yeah <U WHO = ″S4″> yeah? no problem, i was
```

```
 6         hi, can i pay a fine here? <U WHO = "S52"> yeah. no problem. the
 7   's what i was kinda looking into. <U WHO = "S1"> yeah, no that i mean t
 8  then we have one over Z-bar right? <U WHO = "S1"> yeah no. then we have
 9     maybe it's a new term they have. <U WHO = "S3"> yeah no they do call it
10     thanks. yeah. sorry i forgot to <U WHO = "S1"> yeah no uh'uh. yeah u
```

(Data from the MICASE corpus)

Tao (2003) uses this method to investigate American English and Evison (2008) to study British English. Their results are strikingly similar; in both cases the twenty most frequent forms account for 60 per cent of all turn-openings. There are some differences between the two varieties: *yeah*, *mhm* and *right* are more frequent in American English compared with British, and *yes*, *no* and *mm* are more frequent in British English than American (Evison 2008) – these comparative findings confirming those of Tottie's earlier corpus-based study of British and American English (Tottie 1991).

Once concordance analysis has been able to establish patterns of turn-initial items, further detailed analysis can of course be carried out on specific examples that have first been identified. For example, the talk immediately preceding the turn-initial *yeah no* (line 9 in the list above) is shown in Extract 1. Here we can see that three students are working hard to co-create an understanding of their subject, and that they signal strong other-orientation; in fact, each of the three turns opens with a responsive item. It is the importance of acknowledging prior talk that results in the sequence *yeah no*; first S3 acknowledges S2's contribution (*yeah*), before signalling suggesting that it may in fact be incorrect.

Extract 1 [Data from MICASE Transcript ID: SGR175SU123]

S1: okay. so where does it say Q cycle?
S2: oh, in our book it was like later in the_ was it, when they talk about, complex three, how ca-i don't know ma-maybe it's a new term they have.
S3: **yeah no** they do call it the Q, something or other.

By examining the sequence *yeah no* in context, we get a clear sense that the turn-initial position is 'the locus of choice where speakers frequently select items which contribute to the non-transactional stratum of the talk, and where [a] set of "small" items does its work of supporting, converging, bridging and facilitating transitions' (McCarthy 2003: 38). This example shows us how detailed analysis of corpus transcripts can be used in conjunction with basic corpus techniques such as frequency analysis and concordancing to give us greater insight into discourse.

Further reading

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. (This book contains useful chapters on corpus data interpretation with plenty of examples.)

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge. (This book provides a useful overview of corpus analytical techniques as well as a set of case studies which exemplify how analysis works in practice.)

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press. (This book summarises recent corpus-based studies which are particularly important to corpus-informed pedagogy.)

Sinclair, J. McH. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press. (This classic volume by the late John Sinclair exemplifies how basic corpus analytical techniques can reveal much about language and is a vibrant account of the emergence of corpus linguistics.)

References

Biber, D. (1990) 'Methodological issues regarding corpus-based analyses of linguistic variation', *Literary and Linguistic Computing* 5: 257–69.

——(2006) *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

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

Carter, R. A. (2004) *Language and Creativity: The Art of Common Talk*. London: Routledge.

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

Dunning, T. (1993) 'Accurate Methods for the Statistics of Surprise and Coincidence', *Computational Linguistics* 19(1): 61–74.

Evison, J. (2001) 'The Language in First Certificate Discussion Tasks: Are the Exponents of Agreement and Disagreement Presented in Exam Preparation Materials the Same as Those Used by Post-First Certificate Level Non-native Speakers and Native Speakers during Discussion?', unpublished MA dissertation, University of Portsmouth.

——(2008) 'Turn-openers in Academic Talk: An Exploration of Discourse Responsibility', unpublished PhD thesis, University of Nottingham.

——(2009) '"It's Goodbye from Me … and It's TTFN from Me": Creating Podcast(er) Identity in Broadcast Academic Conversations', paper given at IVACS Annual Symposium, University of Edinburgh, January.

Farr, F. (2003) 'Engaged Listenership in Spoken Academic Discourse: The Case of Student–Tutor Meetings', *Journal of English for Academic Purposes* 2(1): 67–85.

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. London: Longman.

Kilgarriff, A., Rychl, P., Smrž, P. and Tugwell, D. (2004) 'The Sketch Engine', in *Proceedings of the Eleventh EURALEX International Congress*. Lorient, France: Universite de Bretagne-Sud, 105–16; available at http://nlp.fi.muni.cz/publications/euralex2004_kilgarriff_pary_smrz_tugwell/ (accessed 6 January 2009).

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

Leech, G., Rayson, P. and Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Harlow: Pearson Education.

McCarthy, M. J. (1998) *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

——(1999) 'What Constitutes a Basic Vocabulary for Spoken Communication?' *Studies in English Language and Literature* 1: 233–49; available at www.cambridge.org/elt/touchstone/images/pdf/What%20constitutes%20a%20basic%20vocabulary.pdf (accessed 20 October 2008).

——(2002) 'Good Listenership Made Plain: British and American Non-minimal Response Tokens in Everyday Conversation', in R. Reppen, S. Fitzmaurice and D. Biber (eds) *Using Corpora to Explore Linguistics Variation*. Amsterdam: John Benjamins, pp. 49–71.

——(2003) 'Talking Back: "Small" Interactional Response Tokens in Everyday Conversation', in J. Coupland (ed.) 'Small Talk', special issue of *Research on Language in Social Interaction* 36(1): 33–6.

McCarthy, M. J. and Carter, R. A. (2003) 'What Constitutes a Basic Spoken Vocabulary?' *Cambridge ESOL Research Notes* 13: 5–7; available at www.cambridgeesol.org (accessed 16 October 2006).

McCarthy, M. J. and Handford, M. (2004) '"Invisible to Us": A Preliminary Corpus-based Study of Spoken Business English', in U. Connor and T. Upton (eds) *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins, pp. 167–201.

Monoconc Pro Concordance Software, Version 2 (2000) Houston, TX: Athelstan.

Morell, T. (2004) 'Interactive Lecture Discourse for University EFL Students', *English for Specific Purposes* 23(3): 325–38.

O'Keeffe, A. (2003) 'Strangers on the Line: A Corpus-based Lexico-grammatical Analysis of Radio Phone-in', unpublished PhD thesis, University of Limerick.

——(2006) *Investigating Media Discourse*. London: Routledge.

O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Reppen, R. and Simpson, R. (2002) 'Corpus Linguistics', in N. Schmitt (ed.) *An Introduction to Applied Linguistics*. London: Arnold, pp. 92–111.

Scott, M. (1999) *WordSmith Tools*. Oxford: Oxford University Press.

Sinclair, J. M. (1999) 'A Way with Common Words', in H. Hasselgård and S. Oksefjell (eds) *Out of Corpora: Studies in Honor of Stig Johansson*. Amsterdam: Rodopi, pp. 157–79.

Swales, J. M. and Malczewski, B. (2001) 'Discourse Management and New Episode flags in MICASE', in R. C. Simpson and J. M. Swales (eds) *Corpus Linguistics in North America: Selections from the 1999 Symposium*. Michigan: University of Michigan, pp. 45–164.

Tao, H. (2003) 'Turn Initiators in Spoken English: A Corpus-based Approach to Interaction and Grammar', in P. Leistyna and C. F. Meyer (eds) *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, pp. 187–207.

Tao, H. and McCarthy, M. (2001) 'Understanding Non-restrictive Which-clauses in Spoken English, Which is Not an Easy Thing', *Language Sciences* 23: 651–77.

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

Tottie, G. (1991) 'Conversational Style in British and American English: The Case of Backchannels', in K. Aijmer and B. Altenberg (eds) *English Corpus Linguistics*. London: Longman, pp. 254–71.

Tribble, C. (1997) 'Improvising Corpora for ELT: Quick and Dirty Ways of Developing Corpora for Language Teaching', in B. Lewandowska-Tomaszczyk and J. Melia (eds) *Proceedings of the First International Conference on Practical Applications in Language Corpora*. Łodz: Łodz University Press, pp. 106–17; available at www.ctribble.co.uk (accessed 16 March 2004).

Walsh, S. and O'Keeffe, A. (2007) 'Applying CA to a Modes Analysis of Higher Education Spoken Academic Discourse', in H. Bowles and P. Seedhouse (eds) *Conversation Analysis and Language for Specific Purposes*. Bern: Peter Lang, pp. 100–39.

West, M. P. (1953) *A General Service List of English Words*. London: Longman.