13
What are concordances and how are they used?

Christopher Tribble

1. What is a concordance?

It's likely that anyone who has even a passing interest in the use of corpora in language studies and language teaching will have come across the term *concordance*. It's equally likely that many of these people will have seen a printout for a concordance such as the one in Figure 13.1.

In this example the search word (*cat*) is presented at the centre of a fixed context of words or characters – a format commonly known by the acronym KWIC (Key Word In Context), and there are many commercial, free or on-line products available to assist those who wish to look at language in use in this way (Barlow 2002; Anthony 2007; Scott 2008; and see Lee, this volume). However, it is important to remember that a KWIC concordance is only one way of looking at corpus data, and that the definition of a concordance offered by Sinclair (1991) is one that we should continue to bear in mind:

> A *concordance* is a collection of the occurrences of a word-form, each in its own textual environment. In its simplest form it is an index. Each word-form is indexed and a reference is given to the place of occurrence in a text.
>
> (Sinclair 1991: 32)

2. Concordances before the computer age

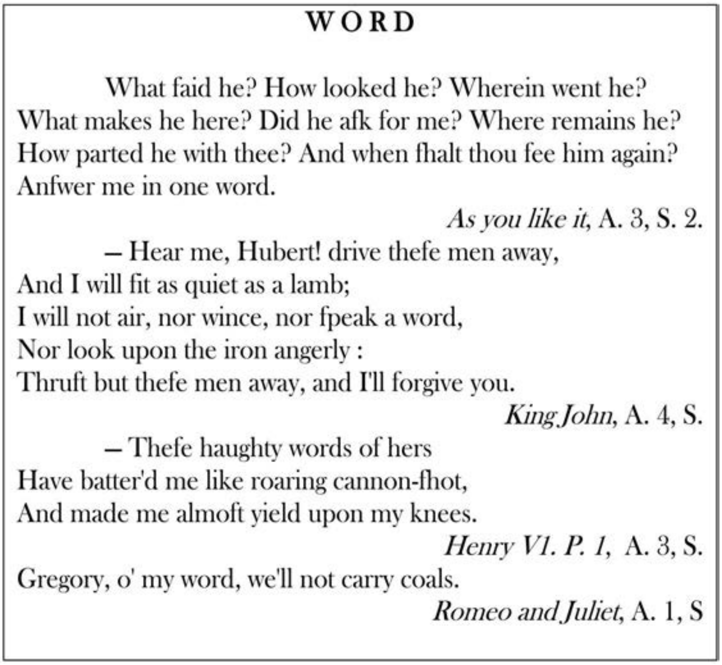
Sinclair's definition is important because it reminds us that, originally, a concordance was a manually prepared list of the word-forms found in a text or set of texts along with references to their precise locations (by book, verse, line, etc.). We stress *word-form* here because with most corpora and corpus tools it would require two searches to find the singular and plural form of e.g. *cat* and *cats*. More of this later.

The first recorded concordance in the Western tradition was based on the work of Cardinal Hugo of St Caro (also referred to as St Cher), who, 'with the help of hundreds of Dominican monks at St James convent in Paris, compiled a word index of the

Vulgate in the year 1230' (Bromiley 1997: 757). Given the huge human effort involved in such a project, it is not surprising that in the pre-computer age, these were only developed for a few culturally valued texts (e.g. Cruden's 1737 *A Complete Concordance to the Holy Scriptures*, Strong's 1890 *Exhaustive Concordance of the Bible,* or Becket's 1787 *A Concordance to Shakespeare*). These books provided scholars with two kinds of resource. The first was an exhaustive account of where words were used in a closed set of texts (in Strong 1890, the 8,674 Hebrew root words in the Old Testament and the 5,624 Greek root words in the New Testament). The second, (e.g. Becket 1787) was

```
1     ht, I should think!' (Dinah was the cat .) 'I hope they' ll remember her
2      And yet I wish I could show you our cat Dinah: I think you' d take a fa
3     to talk about her pet: 'Dinah' s our cat . And she' s such a capital one
4     n here, and I' m sure she' s the best cat in the world! Oh, my dear Dina
5      sneeze, were the cook, and a large cat which was sitting on the heart
6       s for her to speak first, 'why your cat grins like that?' 'It' s a Ches
7     grins like that?' 'It' s a Cheshire cat ,' said the Duchess, 'and that'
8    tle startled by seeing the Cheshire Cat sitting on a bough of a tree a
9      ough of a tree a few yards off. The Cat only grinned when it saw Alice
10    where you want to get to,' said the Cat . 'I don' t much care where - ' s
```

Figure 13.1 KWIC concordance
Source: Lewis Carroll *Alice in Wonderland* http://www.gutenberg.org/files/11/11.txt, accessed 15 August 2008.



WORD

What ſaid he? How looked he? Wherein went he?
What makes he here? Did he aſk for me? Where remains he?
How parted he with thee? And when ſhalt thou ſee him again?
Anſwer me in one word.
As you like it, A. 3, S. 2.

– Hear me, Hubert! drive theſe men away,
And I will ſit as quiet as a lamb;
I will not air, nor wince, nor ſpeak a word,
Nor look upon the iron angerly :
Thruſt but theſe men away, and I'll forgive you.
King John, A. 4, S.

– Theſe haughty words of hers
Have batter'd me like roaring cannon-ſhot,
And made me almoſt yield upon my knees.
Henry VI. P. 1, A. 3, S.

Gregory, o' my word, we'll not carry coals.
Romeo and Juliet, A. 1, S

Figure 13.2 WORD in Becket's concordance.

designed as a source of insight and illumination for a wider readership, the author claiming that:

> the Editor has endeavoured to exhibit the most striking sentiments of the 'great poet of nature', cleared of all impurities, of all 'eye-offending' dross. He has broken and disjointed several of the speeches, but this must not be urged against as a fault: – The nature of the work demanded it; and as the reader is referred to the act and scene of every play, in which the more beautiful of such speeches are to be found, and as there are likewise innumerable compilations in which they are given entire, there is consequently the less occasion for apology.
>
> (Becket 1787: vi)

In Becket's concordance, an example of the word is given, along with its linguistic context and location in the Shakespeare canon (Play, Act, Scene) as in the example for *WORD* given in Figure 13.2.

In computer assisted linguistic analysis, concordances continue to be, at heart, indexes of instances, but they can be generated for a range of new purposes and across a range of ever-expanding texts and text types. A printed concordance of a Greek root word in Strong (1890) would have assisted scholars concerned with biblical exegesis by giving a comprehensive account of how often and where this word was used across the King James Authorised Version of the New Testament. Indeed, this tradition continued until the 1980s with printed concordances of Henry James (e.g. Bender 1987), Joseph Conrad (e.g. Bender 1979) and T. S. Eliot (Dawson 1995).

Useful as these concordances have been, a computer can now be used to create an equivalent to the Strong concordance in the blink of an eye, and it can also support a much wider range of analytic purposes and does not suffer from the limitations of its paper counterpart. In the following sections we will be reviewing the different ways in which computerised concordances can be generated, and how they can be used to present and analyse language data.

3. Computer generated concordances: approaches, tools and resources

As we have seen, before the days of digitised texts and modern computers, concordances were made by dedicated individuals or teams, working often over long periods of time. Team members would read the text, identify words that mattered to the analysis and painstakingly build up tables which allowed one to record where each instance of that word was found. Whether lines, verses, chapters, books, scenes, acts or other units were used would depend on the text type being worked with. At the simplest level, a paper-based concordance for *rabbit* (underlined in Figure 13.3, in an extract from Lewis Carroll's *Alice in Wonderland*) could take two main forms.

Assuming a text such as the example above, where chapter, page and line numbering are known, a concordance in the Strong 1890 tradition might give us an entry for *rabbit* such as in Table 13.1.

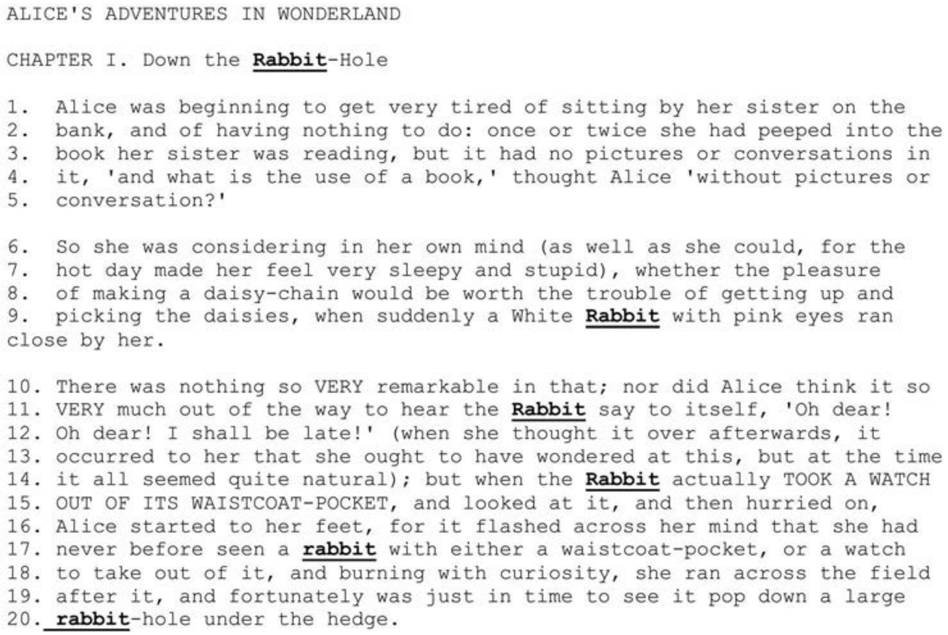
Each time the word *rabbit* appeared in a page, the tally would be increased and a cumulative log would be kept of the chapter, page and line reference so that a cumulative total and list of locations could be compiled at the end of the research process.

A concordance in the Becket 1787 tradition might look more like:

> … when suddenly a White Rabbit with pink eyes ran close by her. (Chapter 1, p.1, l.9)
>
> Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it … (Chapter 1, p.1, l.18)

… the editor having decided that these two uses of *rabbit* were more interesting than other instances in the text.

This paper and pencil approach to concordance-making has not entirely disappeared, having its parallels in projects like the *Oxford English Dictionary*'s 'Reading Programme'. This is still the basis for many of the millions of quotations on which examples in the dictionary are based (see 'The Reading Programme' on the *OED* website), but it has been largely superseded by computer concordances of electronically readable texts.



```
ALICE'S ADVENTURES IN WONDERLAND

CHAPTER I. Down the Rabbit-Hole

1.  Alice was beginning to get very tired of sitting by her sister on the
2.  bank, and of having nothing to do: once or twice she had peeped into the
3.  book her sister was reading, but it had no pictures or conversations in
4.  it, 'and what is the use of a book,' thought Alice 'without pictures or
5.  conversation?'

6.  So she was considering in her own mind (as well as she could, for the
7.  hot day made her feel very sleepy and stupid), whether the pleasure
8.  of making a daisy-chain would be worth the trouble of getting up and
9.  picking the daisies, when suddenly a White Rabbit with pink eyes ran
close by her.

10. There was nothing so VERY remarkable in that; nor did Alice think it so
11. VERY much out of the way to hear the Rabbit say to itself, 'Oh dear!
12. Oh dear! I shall be late!' (when she thought it over afterwards, it
13. occurred to her that she ought to have wondered at this, but at the time
14. it all seemed quite natural); but when the Rabbit actually TOOK A WATCH
15. OUT OF ITS WAISTCOAT-POCKET, and looked at it, and then hurried on,
16. Alice started to her feet, for it flashed across her mind that she had
17. never before seen a rabbit with either a waistcoat-pocket, or a watch
18. to take out of it, and burning with curiosity, she ran across the field
19. after it, and fortunately was just in time to see it pop down a large
20. rabbit-hole under the hedge.
```

Figure 13.3 Sentence concordance — *Alice in Wonderland*
Source: Project Gutenberg Edition (http://www.gutenberg.org/files/11/11.txt accessed 13 August 2008).

Table 13.1 Basic index

count	*word*	*chapter*	*page*	*line*
1	rabbit	1	1	9
2				12
3				15
4				18

Table 13.2 One-word context concordance

the	cat	sat
the	mat	
sat	on	the
cat	sat	on
	the	cat
on	the	mat

Simple concordances

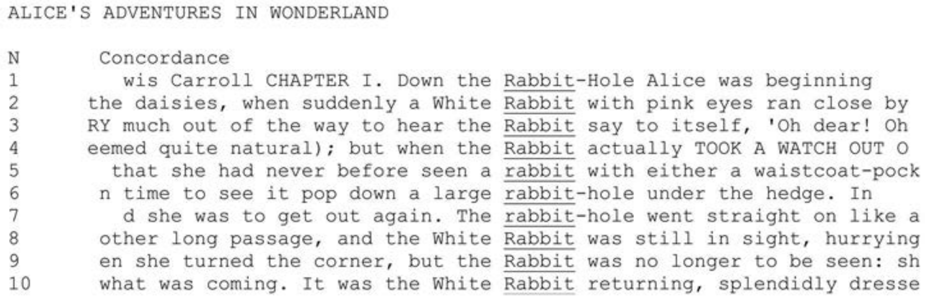
In contemporary computer-assisted analyses of texts, we expect to be able to access all the information we found in a paper concordance (i.e. frequency of occurrences and exact location in the text), along with a great deal more information. Using a modern concordancer such as *WordSmith Tools version 5* (Scott 2008), it is possible to look at a word-form in a number of ways, each of which has its value for the researcher.

As a way of demonstrating the principle which underlies a concordance, Sinclair (1991: 33) gives an example of a concordance of a complete short text, '*The cat sat on the mat*.' Here, each word is treated as a node word-form (i.e. the string of characters which the computer has been instructed to search for in the corpus). These notes have then been sorted in alphabetic order (rather than being presented in text sequence) and a one-word-form context at either side has been provided (Table 13.2).

Although such full text concordances continue to be a possibility, it would be unusual to use this approach for larger texts or text collections. A more common format is one in which a single *node* or search term (this can be a word-form or phrase) is looked for in all its contexts across a text. Given in Figure 13.4 is a KWIC concordance for the word-form *rabbit* in *Alice in Wonderland* by Lewis Carroll. In the current display, the first ten occurrences of the word-form are presented in text sequence at the centre of a context of seventy characters (spaces and punctuation are counted as characters).

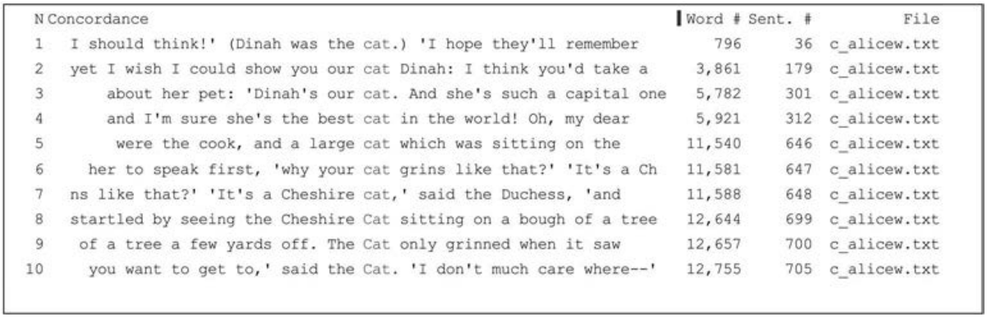
In the next example the same concordance is shown, but this time with the additional information of the position at which the word-string occurs in the text (word number / sentence number / file name) (Figure 13.5).

It should be remembered that the KWIC format is not the only way of displaying concordance data, and that it is not always the best. The concordance extract in Figure 13.6 gives an alternative view of the data in which complete sentences are shown and the node word is underlined.



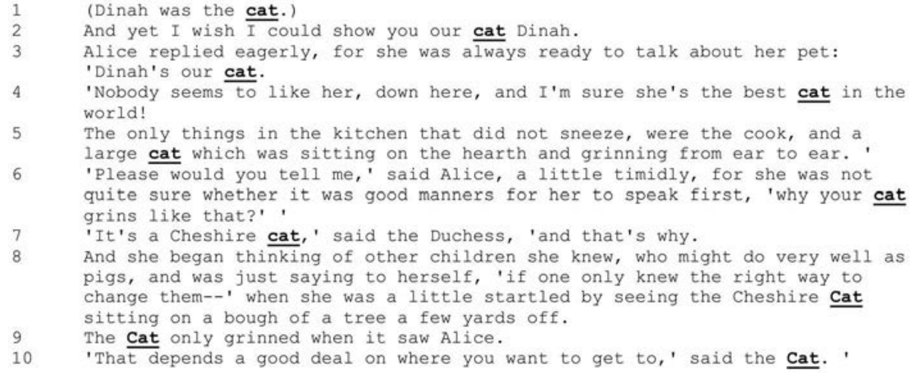
```
ALICE'S ADVENTURES IN WONDERLAND

N       Concordance
1         wis Carroll CHAPTER I. Down the Rabbit-Hole Alice was beginning
2       the daisies, when suddenly a White Rabbit with pink eyes ran close by
3       RY much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh
4       eemed quite natural); but when the Rabbit actually TOOK A WATCH OUT O
5         that she had never before seen a rabbit with either a waistcoat-pock
6        n time to see it pop down a large rabbit-hole under the hedge. In
7         d she was to get out again. The rabbit-hole went straight on like a
8        other long passage, and the White Rabbit was still in sight, hurrying
9        en she turned the corner, but the Rabbit was no longer to be seen: sh
10       what was coming. It was the White Rabbit returning, splendidly dresse
```

Figure 13.4 KWIC concordance for *rabbit*.



```
N Concordance                                                    Word # Sent. #        File
 1  I should think!' (Dinah was the cat.) 'I hope they'll remember       796      36  c_alicew.txt
 2  yet I wish I could show you our cat Dinah: I think you'd take a    3,861     179  c_alicew.txt
 3      about her pet: 'Dinah's our cat. And she's such a capital one  5,782     301  c_alicew.txt
 4      and I'm sure she's the best cat in the world! Oh, my dear      5,921     312  c_alicew.txt
 5       were the cook, and a large cat which was sitting on the      11,540     646  c_alicew.txt
 6    her to speak first, 'why your cat grins like that?' 'It's a Ch  11,581     647  c_alicew.txt
 7  ns like that?' 'It's a Cheshire cat,' said the Duchess, 'and      11,588     648  c_alicew.txt
 8  startled by seeing the Cheshire Cat sitting on a bough of a tree  12,644     699  c_alicew.txt
 9   of a tree a few yards off. The Cat only grinned when it saw      12,657     700  c_alicew.txt
10    you want to get to,' said the Cat. 'I don't much care where--'  12,755     705  c_alicew.txt
```

Figure 13.5 KWIC concordance + provenance data.



```
1     (Dinah was the cat.)
2     And yet I wish I could show you our cat Dinah.
3     Alice replied eagerly, for she was always ready to talk about her pet:
      'Dinah's our cat.
4     'Nobody seems to like her, down here, and I'm sure she's the best cat in the
      world!
5     The only things in the kitchen that did not sneeze, were the cook, and a
      large cat which was sitting on the hearth and grinning from ear to ear. '
6     'Please would you tell me,' said Alice, a little timidly, for she was not
      quite sure whether it was good manners for her to speak first, 'why your cat
      grins like that?' '
7     'It's a Cheshire cat,' said the Duchess, 'and that's why.
8     And she began thinking of other children she knew, who might do very well as
      pigs, and was just saying to herself, 'if one only knew the right way to
      change them--' when she was a little startled by seeing the Cheshire Cat
      sitting on a bough of a tree a few yards off.
9     The Cat only grinned when it saw Alice.
10    'That depends a good deal on where you want to get to,' said the Cat. '
```

Figure 13.6 Sentence concordance.

There are advantages and disadvantages to this kind of display as we shall see (see Sripicharn, this volume), but in the early stages of introducing concordancing to learners and other students of language it is sometimes the case that a sentence view of concordance data can be more useful than the KWIC display as it presents fewer reading challenges to newcomers to corpus analysis.

Word-forms and lemmas

The first concordance examples we looked at in this chapter were for single word-forms. While such searches can be very revealing, there are times when a researcher needs to move beyond individual word-forms. Thus, in many studies, it will be important to investigate a *lemma* rather than a simple word-form. Sinclair defines *lemma* as follows:

> A lemma is what we normally mean by a ‘word’. Many words in English have several actual word-forms – so that, for example, the verb *to give* has the forms *give, gives, given, gave, giving*, and *to give*. In other languages, the range of forms can be ten or more, and even hundreds. So ‘the word *give*’ can mean either (i) the four letters **g**, **i**, **v**, **e**, or (ii) the six forms listed above.

> In linguistics and lexicography we have to keep these meanings separate; otherwise it would not be possible to understand a sentence like '*Give* occurs 50 times in this text'. For this reason, the composite set of word-forms is called the lemma.
>
> (Sinclair 1991: 173)

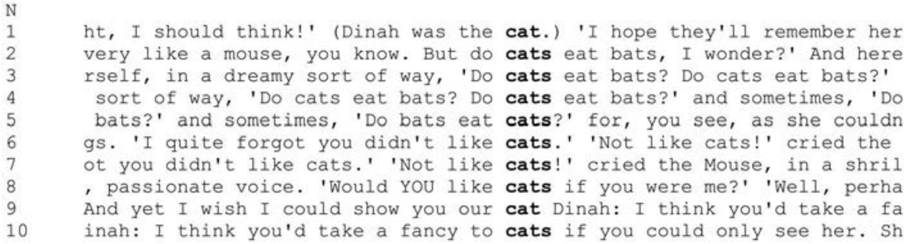
In other contexts, researchers may need to search for particular phrases (fixed combinations of word-forms) which have importance for them or even for non-contiguous patterns where the node is separated from a list of required context words by a set span of word-forms or characters. Modern concordancing software offers resources which make both of these tasks relatively straightforward. With an unmarked-up corpus (i.e. one that does not contain part-of-speech tags, lemma information or other codes) two strategies can be used. The first requires the user to enter a list of the word-forms which constitute the lemma of a stem: e.g. cat/cats or smile/smiles/smiling/smiled. Figure 13.7 shows what a concordance for cat/cats in *Alice in Wonderland* will look like.

An alternative approach to typing all the word-forms you wish to find is to use the *wild-card* facility (*regular expression* in UNIX environments) which different concordancing programs offer. A *wild-card* is a symbol which can be used to stand for one or many alpha-numeric characters. In a *Windows* operating system environment wild-cards might include examples such as shown in Table 13.3.

There is no point in giving an exhaustive list of such symbols here as wild-card symbols can and do change from one program or operating system to another. Finding out what wild-card searches are available to you is one of the first things you need to do when working with a new concordancing program. Once you have learned what wild-cards are available in the software you are using, the next thing to remember is that wild-card searches will usually produce a mix of wanted and unwanted results, especially in an unmarked-up corpus. Thus the search string *cat** will give you *cat* and *cats*, but it also produces *catch*. Similarly, although a search using **ing* will generate concordances for all the present participles, it will also find word-forms such as *sing* or *nothing* which will have to be edited out.

Phrases

Concordancing software does not restrict you to searching for individual word-forms. It is also possible to look for closed and open phrase patterns, using a mix of full word-forms and wild-cards to create search algorithms that most closely meet your needs. Thus, in a corpus of business correspondence (for example, look up Mike Nelson's Business English Lexis Site) a search for *thank you for* will produce results like those in Figure 13.8, while a search for *do not * to* will produce Figure 13.9.

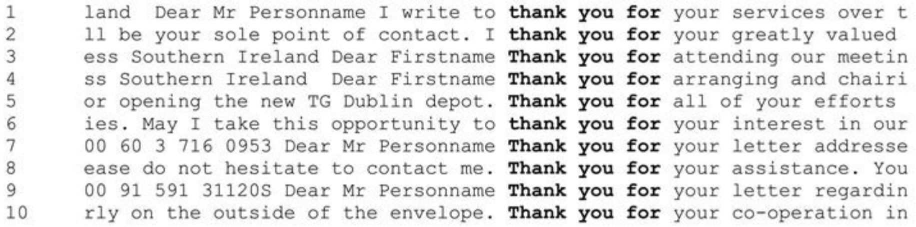


```
N
1     ht, I should think!' (Dinah was the cat.) 'I hope they'll remember her
2     very like a mouse, you know. But do cats eat bats, I wonder?' And here
3     rself, in a dreamy sort of way, 'Do cats eat bats? Do cats eat bats?'
4      sort of way, 'Do cats eat bats? Do cats eat bats?' and sometimes, 'Do
5      bats?' and sometimes, 'Do bats eat cats?' for, you see, as she couldn
6     gs. 'I quite forgot you didn't like cats.' 'Not like cats!' cried the
7     ot you didn't like cats.' 'Not like cats!' cried the Mouse, in a shril
8     , passionate voice. 'Would YOU like cats if you were me?' 'Well, perha
9     And yet I wish I could show you our cat Dinah: I think you'd take a fa
10    inah: I think you'd take a fancy to cats if you could only see her. Sh
```

Figure 13.7 Concordance sample for *cat/cats* in *Alice in Wonderland*.

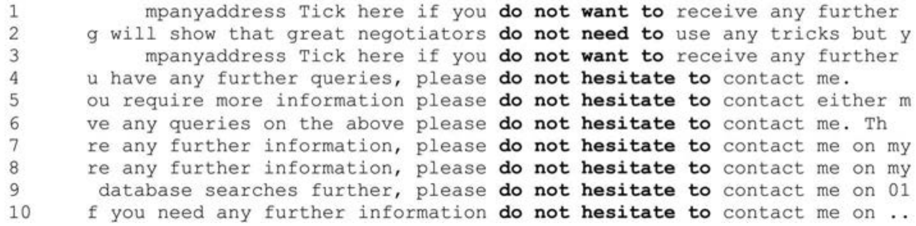
Table 13.3 *Windows* wild-cards

Wild-card	*Search*	*Result*
★ any character at the end of a word (including punctuation)	cat★	I’ m afraid, but you might **catch** a bat, and that’ s a mouse, you know. But do **cats** eat bats, I wonder? a dreamy sort of way, ‘Do **cats** eat bats? Do cats e ay, ‘Do cats eat bats? Do **cats** eat bats?’ and some d sometimes, ‘Do bats eat **cats?’** for, you see, as
★ any character at the beginning of a word (including punctuation)	★ing	the Rabbit-Hole Alice was **beginning** to get very tired ning to get very tired of **sitting** by her sister on the ister on the bank, and of **having** nothing to do: once o n the bank, and of having **nothing** to do: once or twice o the book her sister was **reading**, but it had no pictu
★ any whole word	have ★ to	poky little house, and **have next to** no toys to e! I do wonder what CAN **have happened to** me! When an—but then—always to **have lessons to** learn! Oh eshire Cat: now I shall **have somebody to** talk to. what you had been would **have appeared to** them to
? any single character	Engl?	. The further off from **England** the nearer is to and yet it was certainly **English.** ‘I don’ t quite c remedies—’ ‘Speak **English**!’ said the Eaglet oon submitted to by the **English**, who wanted leade ps it doesn’ t understand **English**,’ thought Alice;



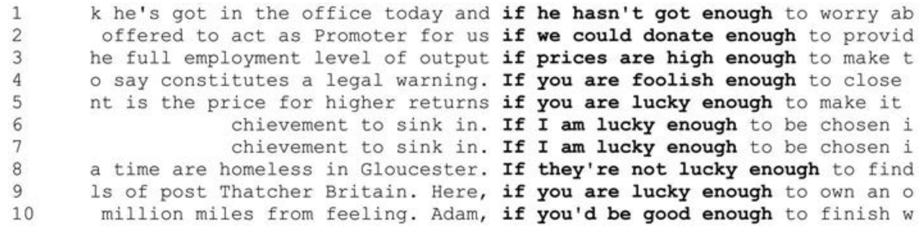
```
1     land  Dear Mr Personname I write to thank you for your services over t
2     ll be your sole point of contact. I thank you for your greatly valued
3     ess Southern Ireland Dear Firstname Thank you for attending our meetin
4     ss Southern Ireland  Dear Firstname Thank you for arranging and chairi
5     or opening the new TG Dublin depot. Thank you for all of your efforts
6     ies. May I take this opportunity to thank you for your interest in our
7     00 60 3 716 0953 Dear Mr Personname Thank you for your letter addresse
8     ease do not hesitate to contact me. Thank you for your assistance. You
9     00 91 591 31120S Dear Mr Personname Thank you for your letter regardin
10    rly on the outside of the envelope. Thank you for your co-operation in
```

Figure 13.8 Phrase search.



```
1         mpanyaddress Tick here if you do not want to receive any further
2     g will show that great negotiators do not need to use any tricks but y
3         mpanyaddress Tick here if you do not want to receive any further
4     u have any further queries, please do not hesitate to contact me.
5     ou require more information please do not hesitate to contact either m
6     ve any queries on the above please do not hesitate to contact me. Th
7     re any further information, please do not hesitate to contact me on my
8     re any further information, please do not hesitate to contact me on my
9      database searches further, please do not hesitate to contact me on 01
10    f you need any further information do not hesitate to contact me on ..
```

Figure 13.9 Wild-card phrase search.



```
1     k he's got in the office today and if he hasn't got enough to worry ab
2      offered to act as Promoter for us if we could donate enough to provid
3     he full employment level of output if prices are high enough to make t
4     o say constitutes a legal warning. If you are foolish enough to close
5     nt is the price for higher returns if you are lucky enough to make it
6                 chievement to sink in. If I am lucky enough to be chosen i
7                 chievement to sink in. If I am lucky enough to be chosen i
8     a time are homeless in Gloucester. If they're not lucky enough to find
9     ls of post Thatcher Britain. Here, if you are lucky enough to own an o
10     million miles from feeling. Adam, if you'd be good enough to finish w
```

Figure 13.10 Wild-card multiword phrase search.

In a larger corpus, it begins to be possible to search for quite extended patterns and obtain a surprisingly large number of results. Thus a search in the BNC for *if * * * enough to* produces 156 results. (NB in the example in Figure 13.10 the contracted word-form *hasn't* counts as a one word. This is a feature common to most concordancing programs.)

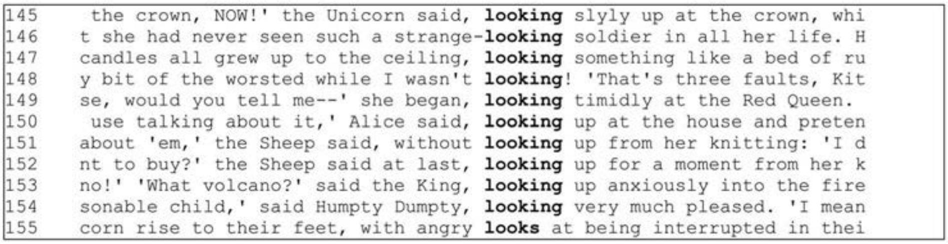
And this brings us to our next section. How do you manage the data once you start to get more than you can conveniently see on your screen?

4. Working with corpus data

Sorting

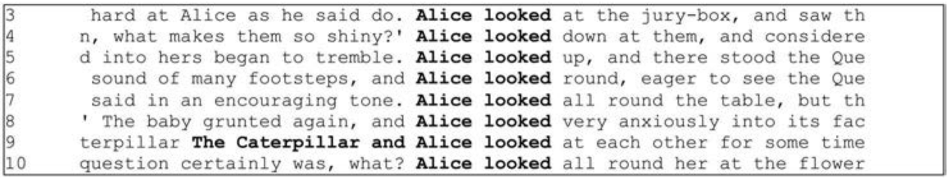
The first way of dealing with a surfeit of results is to take advantage of the fact that your concordance is electronic. You can sort the output so that like is grouped with like. There are at least three ways of re-sorting concordance data: by the node word itself, by the left context of the node and by the right context. Of course, sorting can be done in ascending or descending order. If further information is available, the data can also be sorted by text, by tag or by any other available category. The three examples for the search string *look** below demonstrate the potential for this approach. A search for *look** in *Through the Looking Glass* produces 155 results. This is not a huge amount of data, but it is still more than you can take in without some further processing. As Figure 13.11 shows, sorting the node word first reveals which form of *look** occurs most frequently (*looking*, with seventy-one occurrences) and which form is least frequent (in this instance *looks*, which only occurs once).

Sorting by the left context shows us the typical subjects for *look** as verb, as in Figure 13.12, while sorting by the right context can show typical collocating adverbs, prepositions, complements, etc. (Figure 13.13).



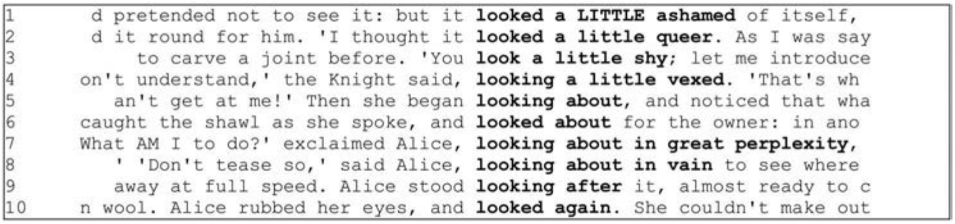
```
145    the crown, NOW!' the Unicorn said, looking slyly up at the crown, whi
146   t she had never seen such a strange-looking soldier in all her life. H
147   candles all grew up to the ceiling, looking something like a bed of ru
148   y bit of the worsted while I wasn't looking! 'That's three faults, Kit
149   se, would you tell me--' she began, looking timidly at the Red Queen.
150    use talking about it,' Alice said, looking up at the house and preten
151   about 'em,' the Sheep said, without looking up from her knitting: 'I d
152   nt to buy?' the Sheep said at last, looking up for a moment from her k
153   no!' 'What volcano?' said the King, looking up anxiously into the fire
154   sonable child,' said Humpty Dumpty, looking very much pleased. 'I mean
155   corn rise to their feet, with angry looks at being interrupted in thei
```

Figure 13.11 *Look** (by the node word).



```
3       hard at Alice as he said do. Alice looked at the jury-box, and saw th
4      n, what makes them so shiny?' Alice looked down at them, and considere
5      d into hers began to tremble. Alice looked up, and there stood the Que
6       sound of many footsteps, and Alice looked round, eager to see the Que
7       said in an encouraging tone. Alice looked all round the table, but th
8      ' The baby grunted again, and Alice looked very anxiously into its fac
9      terpillar The Caterpillar and Alice looked at each other for some time
10     question certainly was, what? Alice looked all round her at the flower
```

Figure 13.12 Left sort.



```
1       d pretended not to see it: but it looked a LITTLE ashamed of itself,
2       d it round for him. 'I thought it looked a little queer. As I was say
3           to carve a joint before. 'You look a little shy; let me introduce
4      on't understand,' the Knight said, looking a little vexed. 'That's wh
5          an't get at me!' Then she began looking about, and noticed that wha
6      caught the shawl as she spoke, and looked about for the owner: in ano
7      What AM I to do?' exclaimed Alice, looking about in great perplexity,
8         ' 'Don't tease so,' said Alice, looking about in vain to see where
9         away at full speed. Alice stood looking after it, almost ready to c
10     n wool. Alice rubbed her eyes, and looked again. She couldn't make out
```

Figure 13.13 Right sort.

Depending on the software you are working with, sorts can also extend beyond the immediate context of the node, with sorts at one, two, three or more words to the right or left of the node being possible.

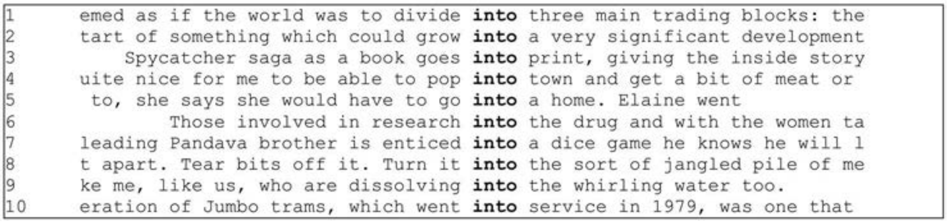
Sampling

If you are still overloaded with data from a concordance search, there are three main ways of reducing the amount of information the program throws at you. One approach is to reduce the amount of data that you are working with. For example, a search for *into* in the BNC will produce 157,925 concordance lines. This is probably too much information for your needs! Biber (1990) has, however, shown that for investigations of word-classes such as prepositions, a corpus of 1,000,000 words can be more than sufficient to make useful linguistic generalisations. So one answer to our problem is to turn to a smaller corpus – in this case the four-million-word BNC Baby Corpus (*The BNC Baby* 2005). This produces a (slightly) more manageable 5,982 results.

If this is still more than you need, but you want to make sure that you are selecting from across a representative sample of the language, an alternative strategy to adopt is to make a randomised selection from the data. If your software permits, you can randomise your selection by a set number. A search with a limit of one in twenty instances across the BNC Baby Corpus produces 300 results, a much easier quantity of data to work with, and one which allows one to begin to select instances that have pedagogic or other kinds of relevance, or which can be used in the first stages of hypothesis development (Figure 13.14).

Restricted searches

If sorting and sampling don't bring the results down to manageable levels – or if you are spending too much time weeding out examples that are not relevant to the research that you are doing – then you need to start to refine the searches that you are making. Even



```
1      emed as if the world was to divide into three main trading blocks: the
2      tart of something which could grow into a very significant development
3          Spycatcher saga as a book goes into print, giving the inside story
4      uite nice for me to be able to pop into town and get a bit of meat or
5       to, she says she would have to go into a home. Elaine went
6              Those involved in research into the drug and with the women ta
7      leading Pandava brother is enticed into a dice game he knows he will l
8      t apart. Tear bits off it. Turn it into the sort of jangled pile of me
9      ke me, like us, who are dissolving into the whirling water too.
10     eration of Jumbo trams, which went into service in 1979, was one that
```

Figure 13.14 Concordance lines of *into* from BNC Baby Corpus.

with an untagged corpus, there are many ways of producing more useful results by making use of the features that most modern concordancing programs provide.

Clearly, with a part-of-speech (POS) tagged corpus such as the British National Corpus, it is possible to specify the word classes that will be included in a search (Aston and Burnard 1998). Thus, rather than looking for, say, *rabbit* (2,571 results), it would be possible to look for it as verb only (thirty-nine results). However, if you have a reasonable knowledge of a structure of the language in question, and a small degree of cunning, it is possible to develop simple algorithms which will greatly improve the searches that you make.

For example, in academic discourse, the ways in which extended noun phrases are post-modified is an important component in the construction of meaning in impersonal, fact-oriented written texts (Biber 2006). Assuming you have access to a corpus of experimental science research journal articles which have not been part-of-speech tagged, how do you go about collecting examples of, say, extended noun phrases in grammatical subject or sentence theme position for research or teaching purposes (see Halliday 1994 for an extended discussion of theme and rheme)?

If you have access to a concordancer which uses asterisk (*) as a whole- or part-word wild-card, it is possible to make the following search algorithm:

> Search for "the * of" in the context of "*." up to four words to the right
>
> [i.e. search for any word preceded by a definite article and followed by the preposition *of* in the context of a preceding full-stop up to four word-forms to the right of the node]

This produces results such as those in Figure 13.15 (shown in sentence concordance format with extended theme in bold and node underlined).

Similarly, should you wish to refine the search so as to find nouns that are both pre- and post-modified, the search algorithm:

> *Search for* 'the * * of' *in the context of* '*.' *up to four words to the left*
>
> [i.e. search for any two word-forms that are preceded by a definite article and followed by the particle *of* in the context of a preceding full-stop up to four word-forms to the left of the node]

… will produce results as in Figure 13.16.

Further refinements of this simple algorithm can be devised, making it possible to identify, quantify and describe other kinds of noun post-modification (present- or past-participle, relative clause and other prepositional phrases) and, of course, it is then

1 **To determine the significance of these banding patterns following the MEE analysis of samples of pools of individuals we** compared individual worms with the pools using the rational and allozyme interpretation as detailed by Andrews and Chilton (1999).

2 **The usefulness of the application of MEE to provide answers to parasite systematics** has been reviewed by Andrews and Chilton (1999).

3 **To date, the roles of genetic variation of O. viverrini on this observed variability in infection, transmission and associated disease** are not known.

5 **The specificity of this PCR, in addition to its sensitivity (50 pg)**, demonstrates its usefulness in Leishmania typing.

Figure 13.15 'the * of'—data source Acta Tropica (http://www.sciencedirect.com/science/journal/0001706X accessed 19 August 2008).

1 **The last repeat of the cpb cluste**r was named cpbE for L. infantum and cpbF for L. donovani.

2 **The species profiles of the An. dirus and An. minimus complexes in north-east India** are largely unknown and need investigation for improved understanding.

3 **The trypanosome stock** can play a role (Maudlin and Welburn, 1994) as well as numerous extrinsic factors.

4 **The theoretical digest of the 648 bp Ade2 amplicon** gave 35425638 bp and 610 38 bp fragments for the resistant and sensitive strains, respectively.

5 **The expected sizes of the three PCR products** were 616 bp (Ade1 amplicon), 648 bp (Ade2 amplicon) and 518 bp (Ade3 amplicon) for the first, the second and the third respectively.

Figure 13.16 'the * * of'.

possible to devise other searches to identify other lexico-grammatical or discourse features. Being able to develop such algorithms has long been part of the skill set of the corpus analyst. In earlier days when programs such as *Oxford Concordancing Programme* were the state-of-the-art tools of the trade, you could spend a long time making sure that the syntax of command lines was right – the computer was completely unforgiving if a single character was out of place. Life is a little easier these days, but there is still a need to be able to think through this kind of extended query.

Reading concordances

Once you have a page (or more than one pages) of concordance data in front of you, how do you read them? Sinclair (2003: xvi–xvii) proposes a seven-stage procedure for working with concordance data, and this offers an excellent starting point for carrying out a careful and comprehensive account of the data that you have found through a concordance search. We summarise the procedure below (and strongly recommend that anyone with an interest in this area returns to the much more extensive discussion and activities in the original).

Given a context where a researcher or teacher wished to investigate citation practices in the Academic Writing section of the BNC Baby Corpus, how might you use Sinclair's procedure? The following search algorithm will produce a useful set of concordances.

> Find all instances of '(19★★) NOT in the context of '★.' 5 word-forms to the right [i.e. search for all dates included in round brackets that are NOT at the ends of

> sentences and which are therefore more likely to be associated with sentence integral citation forms]
>
> (See Thompson and Tribble 2001)

Step 1 Initiate

This is a process of looking for patterns to the right or left of the node which have some kind of prominence, and which may be worth focusing on in order to assess their possible salience to the analysis in question (Figure 13.17).

At the *initiate* stage in an analysis of this data you may first notice that there is a major pattern of ***SURNAME*** + *(DATE)*.

Step 2 Interpret

Sinclair comments for this stage:

> Look at the repeated words, and try to form a hypothesis that may link them or most of them. For example, they may be from the same word class, or they may have similar meanings.
>
> (Sinclair 2003: xvi)

In the present context, an initial working hypothesis could be:

> In academic writing, a pattern *SURNAME* + *(DATE)* is used to represent published work to which a reference is being made. Neither first name nor initials are used.

1	r people and relationships. **Maslow**	(1966)	and Hudson (1966 and 1968) s
2	e data. Similarly, **Openshaw et al.**	(1986)	report that the data in the
3	assified by **Flowerdew and Openshaw**	(1987)	and several examples are giv
4	cheme devised by **Ellis and Schmidt**	(1977)	, singularities in maximal, f
5	to recall that **Elementary Matrices**	(1938)	was printed nine times in U.
6	o shapes social reality. As **Davies**	(1981)	has pointed out, speech is t
7	ersities). **Montefiore and Ishiguro**	(1979)	point out: The universities
8	except for those noted by **Fillmore**	(1971)	, notably come and go, which
9	ation estimation because as **Tobler**	(1979)	points out, there is a dange
10	obation Service. Indeed, as **Bochel**	(1976)	has shown, the establishment
11	s originally postulated by **Erikson**	(1965)	and then developed by Marcia
12	exact solution of Khan and **Penrose**	(1971)	described in Chapter 3, ther
13	on has been considered by **Szekeres**	(1972)	, who found that, for collidi
14	his work was generalized by **Sbytov**	(1976)	to plane gravitational waves
15	s-areal interpolation methods. **Lam**	(1983)	suggests that there are two
16	by **Chandrasekhar and Xanthopoulos**	(1986c)	, and in the aligned case by
17	telligent interpolation. **Flowerdew**	(1988)	developed a theoretically so
18	ounterexample proposed by **Stoyanov**	(1979)	has proved to be incorrect a
19	s type, given by Bell and **Szekeres**	(1974)	, the singularity on the hype
20	central question noted by **Levinson**	(1983)	is whether the study of deix

Figure 13.17 Citation concordance.

Base pattern
SURNAME (DATE)

Step 3 Consolidate

In this stage you look further away from the node to assess if there are additional patterns or other variations in the pattern. In this instance you may notice that certain kinds of verb are associated with Pattern A (these are underlined in Figure 13.17). Your conclusion could be:

(a) A small set of verbs is associated with this pattern. These verbs either precede or follow the initial surname + date node, and produce two distinct new patterns:

Pattern A VERB + by + SURNAME + (DATE)
Pattern C SURNAME + (DATE) + VERB

(b) It also appears that there is a difference between those verbs which precede the node and those which follow it, and it may also be the case that these verbs can be classified evaluatively. This could lead to further research questions, e.g. to what extent does the verb choice indicate whether the writer approves or does not approve of the cited source? To what extent does the verb choice indicate the relative authority or certainty of the cited source? These verb forms are listed in Table 13.4.

Step 4 Report

Here Sinclair comments:

> When you have exhausted the patterns you can observe, and have revised your hypothesis so that it is as flexible as it needs to be and as strong as it can be, write it out so that you have an explicit, testable version for the future.
>
> (Sinclair 2003: xvii)

In the present case, a possible hypothesis is as shown in Table 13.5.

Table 13.4 Verb forms

Preceding	*Following*
classified (by)	report
devised (by)	pointed out
noted (by)	pointed out
postulated (by)	has shown
considered (by)	described
generalised (by)	suggests
proposed (by)	
given (by)	
noted (by)	

Table 13.5 Hypothesis 1

Hypothesis #1

In certain contexts a researcher may wish to incorporate or comment on the opinions, conclusions, etc., of authorities during the development of their own arguments. This can be done through the use of two main patterns:

Pattern A SURNAME (DATE) + VERB
Pattern B VERB + by + SURNAME (DATE)

Verbs associated with pattern A include: *describe / point out / show / report / suggest*

Verbs associated with pattern B include: *classify / consider / devise / generalize / give / note / postulate / propose*

Step 5 Recycle

This stage involves a further rigorous consideration of the extended contexts in which the node is found. This could lead to the discovery that evaluation or other comment on the authority cited may be shown through additional structures (Table 13.6).

This produces two further patterns:

Pattern C ADVERBIAL [*] + SURNAME + (DATE)
Pattern D VERB + by + SURNAME + (DATE) + VERB + TO INFINITIVE + EVALUATIVE ADJECTIVE

Step 6 Result

These observations can be recorded as a focus for further study and will be incorporated into a fuller report which contains a second working hypothesis, as in Table 13.7.

Table 13.6 Extended patterns

2	e data. **Similarly,** Openshaw et al. (1986) report that the data in the
10	obation Service. **Indeed,** as Bochel (1976) has shown, the establishment
18	ounterexample proposed by Stoyanov (1979) **has proved to be incorrect** a

Table 13.7 Hypotheses 1 + 2

Hypothesis 1

In certain contexts a researcher may wish to incorporate or comment on the opinions, conclusions, etc., of authorities during the development of their own arguments. This can be done through the use of two main patterns:

Pattern A SURNAME (DATE) + VERB
Pattern B VERB + by + SURNAME (DATE)

Verbs associated with pattern A include: *describe / point out / show / report / suggest*

Verbs associated with pattern B include: *classify / consider / devise / generalize / give / note / postulate / propose*

Hypothesis 2

Further qualifying information can be added to Patterns A and B in two ways:

Pattern A (q) ADVERBIAL [+ optional additional word form] + SURNAME (DATE) + VERB
Pattern B (q) VERB + by + SURNAME + (DATE) + VERB + TO INFINITIVE + EVALUATIVE ADJECTIVE

Step 7 Repeat

The seventh stage in this process (but not the final stage!) is to repeat the process with more data. This enables the researcher to test, and then extend, refine or revise the hypothesis in order to render it as robust and useful as possible for your particular purposes.

5. Why concordances are not enough

In this chapter we have seen how concordances have developed since their first use several hundreds of years ago, how the results of searches through electronic corpora can be displayed, and how concordances can be read in order to find answers to questions about how language is used – in general and in particular. We have also discussed some of the strategies you can use to reduce the amount of data that can come churning out of the system if you work with very large corpora. Herein lies one of the limitations of the concordance and the reason why it has been necessary to develop new approaches to corpus investigation which make it possible to identify how lexical items collocate and how they are differentially distributed within and across texts and text collections.

We started this chapter by describing a concordance as the result of many years of labour of a small army of monks. Within the last few years, we have moved to a new situation in which, if researchers have access to the simplest of computing resources or to the internet, it is possible for them to produce enough concordance lines to occupy another army of monks. In later chapters in this book, we will see how new tools have been developed to carry out these new tasks.

Further reading

O'Keeffe, A, McCarthy, M. and Carter, R. (2007) *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press. (Although it ranges widely beyond the topic of concordances, this book gives language teachers an extensive insight into the potential of corpus data in language description and language teaching, with a particularly strong emphasis on the use of spoken language data.)

Sinclair, J. M. (2003) *Reading Concordances*. Harlow: Pearson Longman. (This unique book offers a systematic and authoritative account of how to go about the process of reading concordances. It is accessible to those with little or no background in linguistics, but also highly relevant to teachers and students who wish to become more proficient at reading concordances.)

Thurston, J. and Candlin, C. N. (1997) *Exploring Academic English: A Workbook for Student Essay Writing*. Sydney: NCELTR. (A practical demonstration of how a one-million-word corpus of academic texts can be used as the basis for practical teaching materials to support students who wish to develop academic writing skills.)

Tribble, C. and Jones, G. (1997) *Concordances in the Classroom: A Resource Book for Teachers*. Houston, TX: Athelstan. (This is a reprint of Tribble and Jones 1990. It continues to be one of the few resources which gives practical demonstrations of how to turn concordance output into practical teaching materials. Aimed at the teacher with access to small amounts of data, it shows the reader how to design simple concordance searches and format the results for on-screen or paper-based language teaching purposes.)

References

Anthony, L. (2007) *AntConc 3.2.1w*. Waseda: Waseda University.

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

Barlow, M. (2002) *Monoconc Pro version 2*. Houston, TX: Athelstan.

Becket, A. (1787) *A Concordance to Shakespeare; Suited to All the Editions*. London: Robinson.

Bender, T. K. (1979) *A Concordance to Conrad's* The Secret Agent. New York: Garland.

——(1987) *A Concordance to Henry James's* Daisy Miller. New York: Garland.

Biber, D. (1990) 'Methodological Issues Regarding Corpus Based Analyses of Linguistic Variation', *Literary and Linguistic Computing* 5(4): 257–69.

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

The BNC Baby, version 2. (2005). Distributed by Oxford University Computing Services on behalf of the BNC Consortium, available at www.natcorp.ox.ac.uk/

Bromiley, G. W. (1997) *The International Standard Bible Encyclopedia: Vol I: A–D*. Grand Rapids, MI: William B. Eerdmans.

Cruden, A. (1738) *A Complete Concordance to the Holy Scriptures*. London: Midwinter.

Dawson, J. (1995) *Concordance to the Complete Poems and Plays*. London: Faber & Faber.

Halliday, M. A. K. (1968) 'Notes on Transitivity and Theme in English III', *Journal of Linguistics* 4(2): 179–215.

——(1994) *An Introduction to Functional Grammar*, second edition. London: Edward Arnold.

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

Scott, M. (2008) *WordSmith Tools version* 5. Liverpool: Lexical Analysis Software.

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

——(2003) *Reading Concordances*. Harlow: Pearson Longman.

Strong, J. (1890) *Strong's Exhaustive Concordance of the Bible*. Abingdon: Abingdon Press.

Thompson, P. and Tribble, C. (2001) 'Looking at Citations: Using Corpora in English for Academic Purposes', *Language Learning and Technology* 5(3): 91–105.

Thurston, J. and Candlin, C. N. (1997) *Exploring Academic English: A Workbook for Student Essay Writing*. Sydney: NCELTR.

Tribble, C. and Jones, G. (1990) *Concordances in the Classroom*. Harlow: Longman.

——(1997) *Concordances in the Classroom: A Resource Book for Teachers*. Houston, TX: Athelstan.