

33. Searching and concordancing

1. Introduction
2. Searching for words, phrases and other patterns
3. Concordances
4. Displaying annotations
5. Using annotation to search
6. Expanding the co-text
7. Sorting
8. Searching in concordances
9. Thinning
10. Categorising
11. Hiding the node word
12. Showing collocates
13. Using metadata
14. Wordlists
15. Keywords
16. Searching for larger units
17. Searching and concordancing beyond the monolingual text corpus
18. Tools and corpora
19. Literature

1. Introduction

This article will deal with the basic techniques of linguistic analysis which involve searching for and finding words and phrases in a corpus, and displaying the results in useful ways. The most common way to display the results of a search in a corpus is in the form of a concordance. In corpus linguistics, a simple concordance is a list of the occurrences of a word, presented one per line along with its immediate context, as in Figure 33.1.

arket by setting up more cost effective production facilities based on
ognition that markets are an effective way of generating wealth and
a national state to exercise effective control of its own affairs has
s from undertaking the most effective means of monitoring the Sibe
as notes. He always felt that effective musical criticism began by bu
it smallpox vaccine remains effective even when stored at relatively
aig Robertson, they have an effective and, at times, elegant midfiel
r his criticisms of the lack of effective policing of what he defends a
of 1986. The extraordinarily effective popular figure of the masked
ers Situationist film to be an effective oppositional practice. One of
ncept of the spectacle is an effective term which now has a wide cu
les. This work made an effective bridge to the equally spare an
unching an inquiry into how effective competition had been in impr
of people already believe an effective education system is the key t
e and get away with it. To be effective this kind of refereeing has to

Fig. 33.1: Concordance of 'effect' in a corpus

Concordancing has a long history. One of the first uses of the computer for processing texts was started in 1946 by Father Roberto Busa with the help of IBM to do searches in and generate concordances from the works of Saint Thomas Aquinas. Busa produced the Index Thomisticum, which is available online (Bernot/Alarcón 2005). This work can be seen both as a precursor for work in modern corpus linguistics, but also as the continuation of a long tradition of non-computational work in generating concordances from important texts. In the older tradition, a concordance is an alphabetical list of the principal words used in a book, or body of work, with their immediate text surrounding them.

For many linguists, searching and concordancing is what they mean by doing corpus linguistics. The availability of an electronic corpus allows the linguist to use a computer to search quickly and efficiently through large amounts of language data for examples of words and other linguistic items. When the results of these searches are displayed as a concordance, as in Figure 33.1, the linguist can view the data in a convenient format and start to analyse it in various ways. For other types of analysis and research, searching and concordancing may only be the start of a linguistic investigation. They are ways to verify, identify or classify examples in a corpus, in order to start to develop a hypothesis or a research methodology. Searching and concordancing are important elements in the basic toolkit of techniques which the linguist uses. They are essential for checking results derived by automatic procedures and to examine examples in a text in more detail. These techniques will be examined in more depth in this article.

How do linguists read and analyse concordances? Tognini-Bonelli (2001) explores the theoretical basis for reading concordances. She draws attention to the differences between, on the one hand, a linguist reading a text in the usual linear fashion from beginning to end, and on the other hand, a linguist reading the lines of a concordance from a corpus. When reading a concordance, the linguist is looking for patterns of similarity or contrast in the words surrounding the search term. In structuralist terms, when the linguist reads a text, they are reading *parole*, or the way meaning is created in this particular text, and when analysing concordances from a corpus, they can also gain insights into *langue*, or the way that the language system works. In functionalist terms, reading texts allows the reader to concentrate on the poetic, emotive, rhetorical, referential and phatic functions, while concordancing a corpus can foreground the metalingual function of a text (Jakobson 1960).

When the linguist reads concordance lines, the focus of attention is usually on repeated patterns in the vertical direction, or paradigmatic plane. It is also necessary to be able to read each one horizontally, from left to right, to interpret the meaning of the particular example. Reading a set of concordance lines vertically, from top to bottom of the screen, and sorting them in various ways, allows the linguist to see lexical, grammatical and textual paradigms. Simply searching through a corpus and looking at examples one by one is to treat the corpus like a text; it is through concordancing that the patterns of usage and the paradigms are revealed.

Each of the sections below in this article will examine one of the various functions which are available for searching a corpus and for generating and analysing concordances. This article will focus on searching and concordancing in a monolingual text corpus, and the examples given are from English. Some different functions may apply to other types of corpora, and to work on other languages. Some of these differences are referred to briefly in section 17 below. This article is not concerned with search logics, indexing strategies or other techniques from information retrieval – see e.g. Baeza-Yates/Ribeiro-Neto (1999) or Manning/Schütze (1999).

2. Searching for words, phrases and other patterns

2.1. Description

Corpus linguists will typically wish to find certain linguistic items, or sets of linguistic items, in a corpus. The item may be a word, a phrase or some other more complex entity.

The computer can search a corpus quickly for words and phrases. At the simplest level, a search may display the one occurrence at a time. More usefully, all the occurrences may be found and displayed for the user.

The process of searching for words in the corpus underpins all of the functions which are described in this article below. They are all either more elaborate ways of searching, or more elaborate ways of displaying the information extracted by a search routine. This section will deal only with the most simple methods of searching in a corpus.

2.2. Example

Figure 33.2 displays the results of opening a corpus file in a word-processing program and searching for the word 'effect'. Note that only one 'effect' is highlighted on screen. References to all corpora and tools mentioned in this article are given in the literature section.

2.3. Analysis with this function

It is possible to search for and find real examples of linguistic items in a corpus. Authentic examples are useful evidence for dictionaries, grammars, textbooks or lecture notes. With the use of a corpus, linguists can find examples which were really used, rather than invented examples. And rather than simply using real examples found when reading a text, using a corpus makes it possible to quickly search across a large collection of texts and find examples which are more typical, and reflect habitual and frequent usage.

Searching for occurrences of a word can also be used to test for its existence in a particular corpus. In this way, a hypothesis that a given feature does not occur can be disproved. However, the non-existence of a linguistic feature in a particular variety of language cannot be proved by its non-existence in a corpus, as the corpus is only a sample and the feature might occur in the much greater population of texts not present in the corpus. Indeed, Chomsky (2002 [1957]) claimed that a corpus (meaning any collection of utterances) can only ever represent a trivial number of the infinite number of possible sentences in a language. Corpus linguists argue, however, that frequencies are interesting and important (for a discussion of these issues see article 3).

While simply searching for and reading examples can be useful, it is preferable to obtain a concordance because then, if results are found, the types of usage which are present can be examined together and compared.

A further use of searching, rather than concordancing, is to search the vast amounts of electronic text available on the web, where the texts cannot all be loaded into a

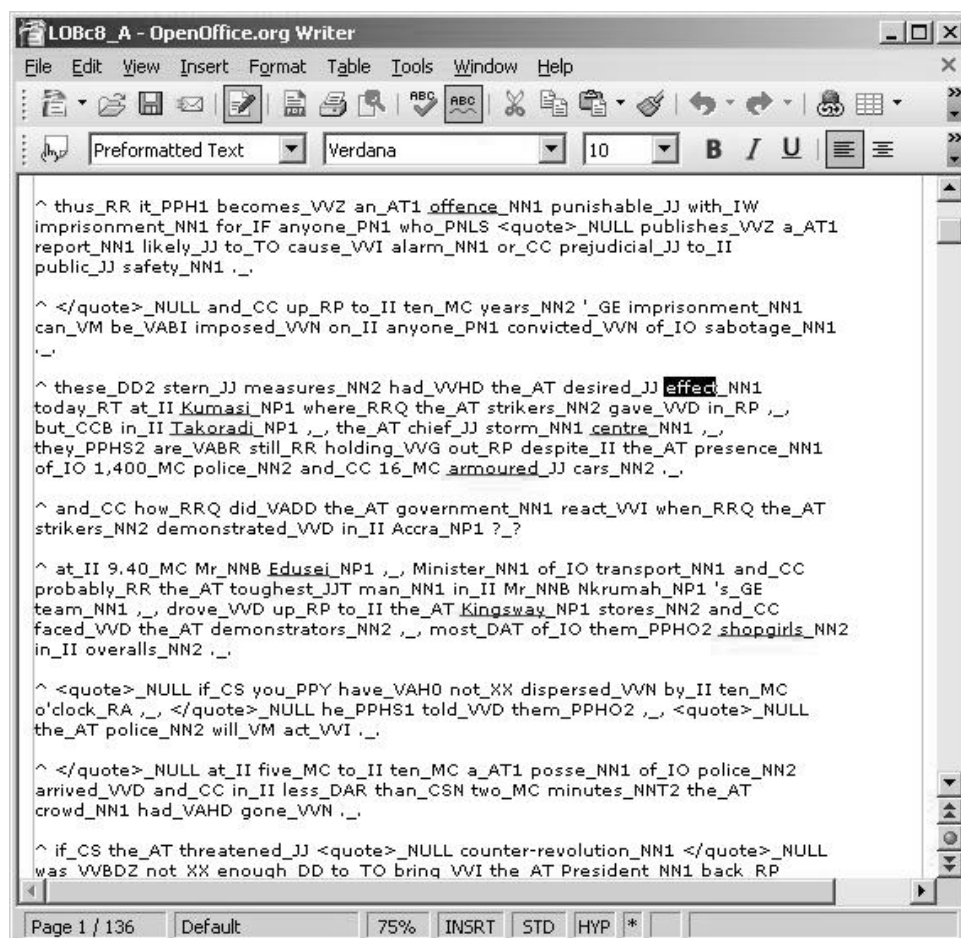


Fig. 33.2: Searching for 'effect' in the Lancaster-Oslo/Bergen (LOB) corpus with a word-processing program

concordance program on the user's computer. Using search engines, the user can search for words or phrases and find many examples. This can be particularly useful for using online text collections, examining new usages in languages which are not yet represented in available corpora, and for investigating emerging modes and styles of electronic communication. Some websites do also offer concordancing online, and systems of returning the results of online searches in the form of concordances are also under development (Renouf/Kehoe/Banerjee 2007). See also article 18.

2.4. Technical requirements

Searching in corpora is a task which can be done by a variety of means including the use of scripts, programs, web browsers, word processors, or specialised corpus analysis applications, such as concordancers.

It is possible for users to write their own scripts or programs to search a corpus, if they have the expertise and software tools available to them. In order to obtain a listing of only and all the required examples, in a form that is easy to use, dealing appropriately with markup, along with a suitable amount of co-text, and in a form in which they can easily be read, re-sorted and further analysed, intensive programming and refining of the search program is required. Concordance programs aim to do all of these jobs for the user and present the results of searching a corpus in a useful way. However, a concordance software package might not be able to deal with all types of corpus, and might not be able to carry out all the functions which a user requires, so some programming might be necessary.

A simpler way for linguists to start to search for an example of a string in a corpus is to use a text viewing or editing application, such as a web browser or a word processor. This can be done by loading the corpus text into the application, and then carrying out a simple search (often using the 'Find' command in these programs), where the user types in a word or phrase, and the program displays the section of the text where the next example of the word or phrase occurs. Many applications will allow more complex patterns to be entered as the search term, including wild cards, character ranges, and optional elements. Some applications, such as the IMS Corpus Workbench, will allow regular expressions, which are a powerful way of using a variety of special characters to obtain matches with a set of patterns (using regular expressions one may be able to search, for example, for inflected forms of a word in a non-lemmatized corpus, or for all compounds containing a certain string).

The result of such a search will take the user to the next place in the text where the search term occurs. However, a corpus may be stored in numerous text files and maybe directories, and many applications will not search across more than one file. Opening a corpus in a text editor or word processing program may also be risky, because the user can easily edit the text, perhaps inadvertently, and the program may attempt to correct spelling, silently insert formatting tags, and perhaps alter the file in other ways. The user should also be aware that loading a corpus into a web browser may lead to the browser trying to interpret tags as HTML or as XML; if there are other types of markup in the text it may cause unpredictable and unwanted results in the way the text is displayed. The basic problem is that these programs are designed for reading or editing text documents, but not for searching text corpora.

Concordance programs aim to deal with these potential difficulties and to present the results of searching a corpus in a useful way.

3. Concordances

3.1. Description

A concordance is a listing of each occurrence of a word (or pattern) in a text or corpus, presented with the words surrounding it. A simple concordance of "Key Word In Context" (KWIC) is what is usually referred to when people talk about concordances in corpus linguistics, and an example is shown in Figure 33.1. Concordances are essentially a method of data visualisation. The search term and its co-text are arranged so that the textual environment can be assessed and patterns surrounding the search term can be

identified visually. Barlow (2004) defines concordances (and wordlists) as transformations of a text, giving the analyst the opportunity to view different perspectives on a text.

Often a concordance of a particular search term in a corpus will produce too many results for a linguist to read and analyse. In this case a reduced number of examples can be selected. One or two screenfuls is often useful for providing the analyst with at least a preliminary view of the relevant patterns, although the number necessary to examine is heavily dependent on the structure of the corpus, the total number of examples and the type of investigation which is being carried out. If such a sample is chosen, it is important to select them either randomly from the total, or to select every n th example. Otherwise, the software is likely to provide by default the first 40 examples, which may all come from one file, and thus there would be a high risk of a highly biased sample, reflecting only the language usage of one text or variety. Selecting every n th examples is one method of *thinning* a concordance (see section 9).

3.2. Example

Figure 33.1 above shows an example of a concordance.

3.3. Analysis with this function

The primary motivation for the use of concordance data in modern corpus linguistics is the belief that interesting insights into the structure and usage of a language can be obtained by looking at words in real texts and seeing what patterns of lexis and grammar surround them.

The use of concordances is essentially a manual task for human analysts, unlike the use of many computer algorithms to automatically extract information about the occurrence and co-occurrence of words in texts. Automatic extraction of wordlists, collocate lists, etc. can lead the analyst to deal only with words abstracted from the texts where they occur, and taken away from the place where meaning is created. Reading concordances involves looking at words in their context of occurrence in texts, and allows the analyst to see how meaning emerges in the particular case.

Furthermore, reading concordances allows the user to examine what occurs in the corpus, to see how meaning is created in texts, how words co-occur and are combined in meaningful patterns, without any fixed preconceptions about what those units are. It can be a method of approaching the corpus in a theory-neutral way. This is part of what Tognini-Bonelli (2001) calls corpus-driven linguistics.

However, interesting results do not spring out as soon as the corpus is loaded into the software. To generate a concordance, the user must select what to search for, and this means approaching the corpus with some pre-conceptions about what words (or other features) will be interesting to look at. One way of avoiding this bias is to make use of a function which some programs have to provide a complete concordance of a text or a corpus, which means making a concordance of every single word. This can be useful for a text, and was the traditional way of making concordances for the study of literary or religious works before the era of the computer. However, a complete concordance of a corpus will usually produce more data than human analysts can cope with.

Even major lexicographic projects are likely to be selective with what words to search for and how many examples to look at in a large corpus. There are other functions, described below, such as making lists of words, collocates and keywords, which can be used as starting points which allow the corpus to suggest things to look for and investigate.

While use of a concordance program may be necessary for corpus-driven work, use of concordances does not necessarily imply that research is corpus-driven. It is perfectly possible to use a concordance program simply to look for data to support a hypothesis which has been arrived at by some means other than analysing the corpus, and most research done using a corpus is probably of this type.

Another important type of work which concordances make possible is data-driven learning. For the language learner, use of a corpus can be a substitute for intuitions which the native speaker acquires through exposure to the mother tongue (e.g. Lamy/Klarskov Mortenson 1999–2007). See article 7 for more on this topic.

There are many other areas where the qualitative analysis of concordances is essential for identifying and analysing patterns in language. There are some increasingly influential theories which rely on examining collocations and concordances, namely semantic prosody (Louw 1993/2004), semantic preference (Sinclair 2004) and lexical priming (Hoey 2005). See article 45 for more on these topics.

3.4. Technical requirements

Corpus analysis tools will either search through the corpus as a set of text files, or corpora may be pre-indexed, allowing for faster retrieval and more powerful queries. Some tools require corpora to be in particular formats (e.g. plain text, XML, or some other format). Care should be taken to ensure that the particular forms of character and text encoding, file format and markup are being interpreted correctly by the program. This will be more straightforward if the corpus itself is constructed in a fairly standard way and the corpus design and encoding are well documented.

Concordances are usually generated by a program for on-screen display, but it may be essential to save them so that they can be examined again. While a concordance can often easily be generated again by submitting the same query to the same corpus, this may not be possible in some cases. If some complex series of processing steps has been taken, such as sorting, categorising, or thinning the lines (see below), then it may be difficult to reproduce the results. Some of this processing may have to be done by manual selection or annotation, and then this work certainly needs to be saved. There are other reasons why concordances may need to be saved: access to the tools or corpus may be temporary; the corpus may be under development and may change; the tools may be updated and change their functionality in subtle ways. Furthermore, it may be necessary to make the concordance available outside of the program which generates the concordance, so that it can be processed with other tools, or used in teaching, on a website or in a publication. It would therefore be necessary to save the concordance in some portable format, such as HTML. A user should consider whether these functions are available or necessary when selecting (or developing) concordancing software.

The following sections deal with further refinements and enhancements to concordances.

4. Displaying annotations

4.1. Description

A corpus may include various tags, which may encode descriptions of the texts constituting the corpus, elements of the text structure (e. g. paragraphs), or linguistic annotations (e. g. wordclass tags). Concordance software sometimes has the option to hide or display markup.

One possibility is for the concordance software to colour the different parts of speech, so that nouns are red, and verbs blue, for example. This is likely to be easier to read than viewing the concordances with the tags displayed inline with the text. The analyst can see the wordclass categorisation without interruption to the stream of words.

The analyst may wish to be shown tags associated not with the individual words or lines in the concordance, but rather the information associated with the whole text from which the particular concordance line is derived. For example, the name of the file, or title of the text may be useful. Another option would be to indicate, again perhaps by the use of colour contrasts, the category of text from which the example comes. For example, in BNC-Baby, a corpus which is a subset of the British National Corpus, containing 4 million words of written and spoken English, and which is designed to compare and contrast newspapers, fiction, academic writing and spoken conversation, examples from each of these four sub-corpora could be coloured differently.

4.2. Example

Figure 33.3 shows a concordance from the BNC-Baby corpus. This concordance has been produced by Xaira, an XML-aware program, which can recognise and use the metadata and annotations in the corpus in sophisticated ways because it is able to process the XML tags. XML is an international open standard for marking up documents (see article 22). Since the annotations in BNC-Baby are XML tags, XML-aware programs can selectively display or hide the tags.

4.3. Analysis with this function

Viewing the markup associated with concordance lines may sometimes be useful in order to help interpret some of the concordance lines, or to make more patterns visible. It is essential for checking the results of searches using annotations (see section 5 below). When unexpected results are obtained from searching for a particular wordclass tag, for example, it may be necessary to read the tags to find out whether they have been incorrectly assigned, or at least to try to understand the ways in which the wordclass tags have been assigned. There is little consensus in linguistics about how wordclasses should be categorised, and therefore there is a lot of variation in the ways in which different analysts or different programs will assign tags.

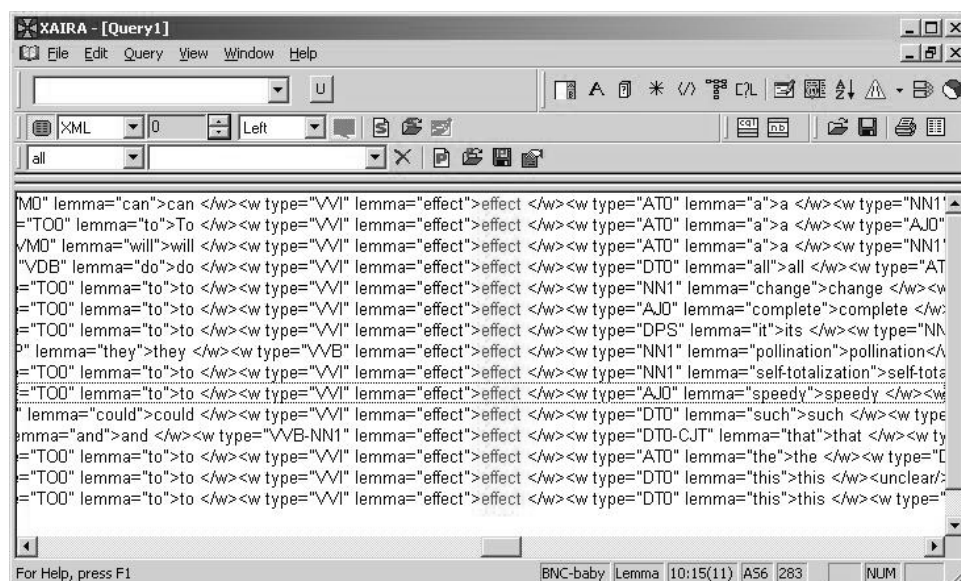


Fig. 33.3: Displaying XML annotations in the BNC-Baby

4.4. Technical requirements

Applying this function usually requires that the corpus text has been annotated. The possibility of implementing this function, and of it being useful to the analyst, depends on the manner in which the markup has been encoded and documented. It is possible in principle that markup could be applied by the software on the fly, but relying on automatic tagging is likely to involve problems of accuracy and consistency as well as additional computational processing.

Displaying information about the source text can help the user to understand the lines of a concordance. Many programs will show the file or text name alongside each concordance line. Others will allow the user to select a line and then view the metadata associated with the text, such as the title, author, date of publication, etc.

The possibilities for selecting elements of the markup and using them to display words or lines differently will become increasingly possible as standards for the encoding of corpora in XML are developed and stylesheet functionality is incorporated into XML corpus analysis tools, to enable the user to adjust and control the display of the output.

5. Using annotation to search

5.1. Description

Corpora often contain various types of tagging. These tags exist in the files in addition to the words which make up the texts, and can include tags which encode descriptions

of the corpus and its constituent texts (descriptive metadata), tags which encode information about the text structure, formatting and appearance (structural markup), and tags which encode various levels of linguistic categorisation or analysis of the text (linguistic annotation). It can be useful for the purposes of linguistic analysis to search for examples of words or other units which have been categorised by the use of these tags. In particular, the analyst may wish to exploit the linguistic annotation in a text, such as wordclass tagging (see article 24), or lemmatisation. If a concordancer offers the necessary functionality to interpret the tagging in a sensible way, then it should be possible to search for all examples of a particular word when it is tagged with a particular wordclass categorisation, for example 'effect' as a verb (see Figure 33.4).

Methods for using the descriptive metadata and structural markup are described in the section 'Using metadata' below.

5.2. Example

Figure 33.4 shows a concordance of the word 'effect' where it has been tagged as a verb in the BNC-Baby corpus. The fact that 'cause and effect' is included here is due to a tagging error.

months that it's that it's really started to effect this but I know what it is that's because .
 actually interview if I do effect all the Well you're all g
 d with sledgehammers and crowbars 'to effect speedy entry'. Compensation of £8,500 fr
 enditure of the 1960s and 70s. To effect a new social discipline, a new relationshi
 its own alterity and duplicity in order to effect its deconstruction. In this context, we ma
 ims that certain forms of literature could effect such a critical, reflective detachment. Thi
 e a totality. Although Sartre's inability to effect self-totalization is often presented as a fa
 o the interests of the exploiter. This will effect a displacement or dissolution of self-resp
 ulla] as before. The student is invited to effect this reciprocation by means of the fourth
 is a baseline against which attempts to effect change can be measured.
 ore ample maintenance and authority to effect the same, We do command the said Chri
 s the chain motion becomes too slow to effect complete untangling of the polymer coils.
 time. It is this delay between cause and effect that is fundamental to the observed viscc
 by ny back up to the canopy, where they effect pollination, any slight wind drifting them t
 :s in the pattern of home ownership, can effect a change in partisan support. There

Fig. 33.4: 'effect' used as a verb

5.3. Analysis with this function

Exploiting the annotation to specify search terms can help to make more refined, and more grammatically targeted searches. For example, grammar books may say that it is not permissible to say 'less books' or 'less examples', and that it should be 'fewer books' and 'fewer examples'. It is possible to test this prescriptive rule by looking in a corpus at the evidence of what native speakers really say and write. Searching for 'less' immediately followed by a plural noun in the corpus yielded no results, while there were 40 examples of 'fewer' immediately followed by a plural noun.

This does not give conclusive proof that ‘less’ does not occur before plural nouns, nor that the prescriptive rule is correct. A slightly more sophisticated search pattern (allowing adjectives to occur between ‘less’ and the plural noun) yielded the following example from this corpus: “even if on the lower rungs with less promotion chances than white men”. It is also worth noting that wordclass tags were assigned in this corpus automatically, and it is a possibility that the tagging program would not have been likely to assign a noun tag to a word following ‘less’, because it might have been programmed with the “rule” which does not permit this sequence.

Searching using the annotation can help to reveal grammatical patterns in the corpus, and it can also be used to find the grammatical patterns which tend to occur with certain words. This tendency for certain grammatical patterns to associate with certain words is known as *colligation* (Firth 1968, see also article 43).

It can also be argued that there is a danger of analysts focussing on the annotations rather than words of the text. Indeed there is a danger of circularity in this methodology, if users simply retrieve the information which has been inserted in the form of annotations by other linguists, or even by themselves, without retrieving any other useful information about the text.

5.4. Technical requirements

Searching with annotation depends on the presence of markup tags in the corpus. The usefulness of the function depends on the quality of the markup, and on the accessibility and quality of the documentation of the markup which is available to the analyst. If the user does not know the tags, or understand the ways in which they have been applied, then it is very difficult to use them and it is easy to misinterpret results. In order to exploit the tagging with software, the software needs to know how to identify and process the tags in the text. (To put this in a more technical way, it is necessary for the corpus analysis application software to be interoperable with the text markup formalism.) For this reason it is useful for the tagging in the corpus to be inserted into the text in a reasonably standardised way, for example as XML tags. XML is a standard way of inserting metadata and markup in a document. If a non-standard form of tagging is used, concordance software is less likely to recognise the markup, and may be unable to differentiate it from the text, or to make use of it in any useful ways. Using non-standard markup can mean that the user is tied to software written specifically to process that markup, which is likely to restrict access to the corpus and reduce its usefulness, especially if the documentation and software specific to the corpus do not survive in the long term.

If the tags are stored separately from the corpus text, in a separate file as “stand-off” markup, then the risk of the tags interfering with the processing of the text is diminished. On the other hand, the computational task of using the tags is made more difficult, and there may be few, if any, standard corpus analysis tools available which can successfully process the tags. However, this type of markup is likely to become more standardised and widely used in the future.

It is not always necessary for a corpus to contain detailed annotation in order to search using wordclass tags and other linguistic categories. An alternative to the use of

annotation to carry out analysis of this type is to use lexical and grammatical information which is separate from the text, but which the searching and concordancing programs can make use of. For example, morphological tables may hold lists of inflected forms, which can be used for searches for verb paradigms, as an alternative to annotating all the lemmas in the corpus. One important disadvantage of this approach is that ambiguous occurrences in the corpus are unresolved, or must be resolved “on the fly” by the software each time that the user wishes to make use of them. An important advantage is that the user can conduct searches using annotation on texts which have not been annotated.

6. Expanding the co-text

6.1. Description

Expanding the co-text is a function which can be applied to concordance lines. A concordance line will often be presented to the user as one line on the screen, with perhaps 4 or 5 words visible on either side of the search term. In order for the analyst to be able to read and understand a particular concordance line, it is often necessary to be able to read beyond this limited amount of co-text. It is therefore useful to be able to expand the amount of co-text which is available. Some concordancers only give access to a few extra words or lines; some restrict the scope of the context to some textual unit such as the sentence or paragraph; others allow the analyst to start from the concordance line and read as far as they wish.

6.2. Example

Figure 33.5 shows the concordance of ‘effect’ in the BNC-Baby corpus (displayed in this case by the MonoConc program), with the expanded co-text for the selected line displayed in the box above the concordance lines.

6.3. Analysis with this function

Being able to read more of the co-text is essential when the analyst wishes to take account of meaning in the text. It may be necessary to read a long way in a text to get enough information to be able to explain the occurrence of a particular linguistic feature in a text.

6.4. Technical requirements

Most concordance programs which make use of a corpus installed locally offer the functionality to expand the co-text. Online concordance programs may restrict the amount

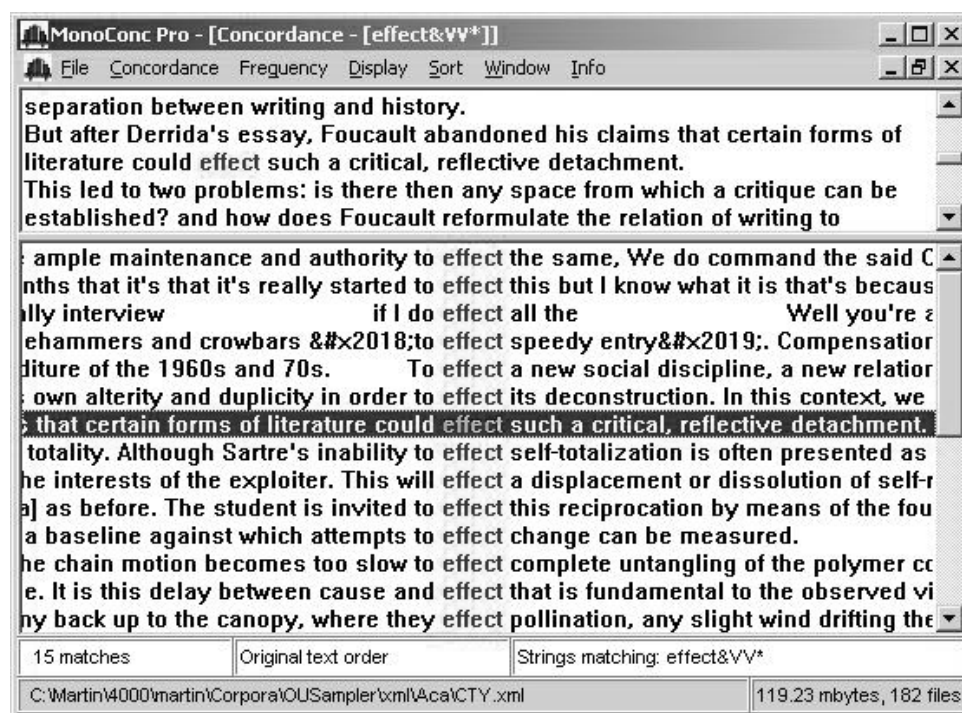


Fig. 33.5: Concordance of 'effect' used as a verb, with expanded context for the selected line

of co-text available in order to prevent the user from downloading an entire text or corpus. This may be necessary because of licensing restrictions on the distribution of the corpus. In this case, it may be necessary to attempt to negotiate access to the full corpus text for more detailed analysis of concordance lines.

7. Sorting

7.1. Description

Sorting is a function which can be applied to concordance lines from a corpus. A concordance program will usually by default present lines in the order in which they occur in the corpus (often called the 'original order'). Sorting the lines on the basis of various criteria may be necessary to reveal patterns in the words surrounding the search term. The criteria according to which lines can be sorted include:

- alphabetically by node (which is only relevant if there is some variation in the node word, such as when wildcards have been used, or when there have been multiple search terms, or search by wordclass or lemma, etc.);
- alphabetically by co-text: words in certain positions around the node word, e. g. one word to the right or left of the node;

- annotations (e. g. grouped together by wordclass tag, or user-defined annotations);
- metadata categories (e. g. text type, sex of speaker);

A further set of options for each of these searches may be made in ascending or descending order (e. g. a-z or z-a in English).

When searching by co-text, it is often useful to try sorting by words in various positions around the search term, for example 1 to 5 places to the left or right. Where the search term is part of a long fixed phrase, it may be necessary to look further away than this, in order to find variation and patterns in the co-text.

It is also possible to sort on the basis of more than one criterion, so that lines which are grouped together according to the first sort criterion are then further ordered, as in Figure 33.6.

bility of charging for more services. The effect on the demography of the inner cities could be. How does it change the meaning or effect on the text, as far as you as a reader ... this difference is not likely to have any effect on the way in which the individuals will be. The additions to this B-tree have had no effect on the index entries on the left-hand side ed by some invisible gale which had no effect on the branches of the little trees and even government and that has inevitably had an effect on the level of the charge. 'This is first turn, the Scientific Revolution had some effect on the visual arts. Military engineers, who of personal reference has a noticeable effect on the structure of contributions in conversation were talking about had had a very bad effect on the Quigleys. Mrs Quigley was hyperactive, but is something which also has an effect on the way people behave towards others. s on the prince's function, it also had an effect on the way he fulfilled it. In 1140 the monastic entanglements begin to have a significant effect on the relaxation times. The undiluted result, road safety campaigns have little effect on them because they are seen as being variations were found to be variable in their effect on timing. IBM files performed better than it is introduced, will have a significant effect on transitions. The movement of individual competitive encounter I had an immediate effect on United's performance. Suddenly 80pc of its visitors, thus having a major effect on visitor income and support for the estate is in some of those early films had their effect on Wil he promoted himself to a star on the design of the course, conservation and effect on wildlife. The course boasts a large this week as part of a series showing the effect on wildlife of the lowering water table. number three so they don't know.' The effect on women, in a Moslem culture which permits or female family members can have the effect on women sectioned, of producing less leisure. I felt flattered I don't usually have an effect on women like that I and thought I loved I ce in April 1988 have had a devastating effect on young people. At the stroke of a pen that at his words were starting to have some effect. One or two of the older members were not

Fig. 33.6: Concordance sorted by first right then second right

7.2. Example

Figure 33.6 shows a screenful of concordance lines for 'effect' from BNC-Baby, sorted by first word to the right, then second word to the right.

7.3. Analysis with this function

Sorting is often necessary to reveal the patterns of words surrounding the search term. These patterns can more easily be seen by the analyst when repeated occurrences of

relevant features are grouped together on the screen. Sorting can help to bring these examples together. In this way, simply sorting on various positions and viewing the lines can reveal patterns which were previously invisible.

7.4. Technical requirements

A concordance program will usually provide the functionality to sort concordance lines. There may be variation in the number of criteria allowed, which criteria may be applied, and the range of co-text over which they may be applied.

8. Searching in concordances

Much of the work of searching and concordancing is about finding out which words tend to occur in the vicinity of other words, and searching in the co-text surrounding the search term in concordance lines may be useful if the analyst is looking for examples of a particular word or pattern. This is a way of searching for the co-occurrence of words and phrases.

The linguist can start to find the words which occur with the search term by sorting the concordance lines and by computing collocations. When some potentially interesting words are suggested by the list of collocates, or by the examination of the concordance lines, the linguist will want to search for the lines in which the word occurs.

Not all concordance programs provide the functionality to search within the concordance lines for a particular word or pattern, or to find co-occurrences of words at all. It may be necessary to use sorting to find the word, although this can be difficult if it is appearing in many different positions around the search term in the KWIC concordance. Words that occur above a certain frequency will appear in a list of collocates, and some programs will allow the user to switch from the collocates list to show a concordance of all the lines in which the collocates appear with the search term.

9. Thinning

9.1. Description

Thinning concordance lines is a function which reduces the number of lines in a concordance, by selecting a subset of the lines based on some criterion. This may be done in order to reduce the number, if there are too many to analyse, or because the analyst is only interested in a particular subset.

Ways of thinning concordance lines include reducing the set to every n th occurrence, to n per text, or to the first n examples (where n is any positive integer). A set of concordance lines may also be thinned on the basis of user annotations (see section 10 below).

Searching in results to produce a reduced number of concordance lines (see section 8) can be one way of thinning the concordance lines. Some programs allow the user to search for a string in the concordance lines, and then thin the set of concordances to only those which contain the search string.

9.2. Example

tanglements begin to have a significant effect on the relaxation times. The undiluted sy even more doses. Although its effect on the circulation of wild polioviruses ha their properties would have a beneficial effect on the overall scheme, members heard. as rabbits or sheep, has a devastating effect on the fine-leaved bouncy turf rich in spe ist, such groups must have had a major effect on the structure of the forest. The v ish whether artemether has a beneficial effect on the objective and unambiguous prima ernment and that has inevitably had an effect on the level of the charge. 'This is fi og-meat and biscuits had had a ruinous effect on the housekeeping. Happily Herbert ha / were talking about had had a very bad effect on the Quigleys. Mrs Quigley was hyper oleoresins of the dipterocarps have an effect on the bacteria of the fore-stomach of col n but progressive and compensatory in effect. On the circumference of that circle are n ility of charging for more services. The effect on the demography of the inner cities coi ce in April 1988 have had a devastating effect on young people. At the stroke of a pen tl ur to her to worry about the devastating effect Paula was having on Edward. Behin and for public health activities. Thus in effect reference centres are indistinguishable f a matrix between 'knowledge of a cause/effect relationship between participation progra nds, detecting a marked distance decay effect. Research p rease in blood volume in the lungs I an effect shown by transthoracic impedance techn time. It is this delay between cause and effect that is fundamental to the observed visci ; so great variety") give an overall effect that the conclusion is a promotional, or u e per se , there is some authority to the effect that trespass to goods requires proof of : Tc interval confirming a largely additive effect; the dose response curves for salbutamc ial solution are further examples of this effect. The fundamentals of light scatterin x up together than the cross-cousins. In effect, the parallel cousins are as familiar as s hat if a placebo is to have a therapeutic effect, the patient must believe that it will. Nevi

Fig. 33.7: The concordance from Figure 33.6 (sorted on right co-text), thinned to display only every 5th occurrence

9.3. Analysis with this function

Thinning lines is often part of the heuristic process of focussing the analysis on a particular area of usage in the corpus. A corpus-driven enquiry will typically start with a search for a particular form, followed by analysis of its meaning and contexts, and then searching for a longer phrase.

Thinning concordance lines is used chiefly for providing an appropriate number of examples for a human analyst to be able to view. This may be done for use in the classroom, so as not to intimidate the student with too many examples. Manually selecting lines is also possible, and may be useful for illustrative or pedagogic purposes, but there is a danger of making a biased selection, and it is important that the person reading the concordance knows that the lines have been manually selected.

If it is intended to generalize from the analysis of the sample, then it is necessary to be aware of the way in which the corpus is structured, and to decide whether the sample is likely to be representative of all the examples. In a similar fashion, if the intention is to generalise about the language on the basis of a corpus, the linguist must also always bear in mind the way in which the texts in the corpus itself have been sampled from the overall population of texts. Analysing only a limited number of the concordance lines may be necessary from a practical point of view, but the analyst must bear in mind that the analysis is based on a sample of a sample.

It is also possible, at least in principle, to apply automatic procedures to thin concordance lines by selecting one or two examples which exemplify typical patterns of usage. This is an attempt to automate the work of finding typical patterns of usage in concordance lines, and may be useful for pedagogical, or for lexicographic applications. Concordance output thinned in this way may be able to show something of the variety of different usages, but will not show patterns of repeated usage in and around the search term. Such a concordance must be read differently. The analyst should not look for repeated occurrences as evidence of typicality, because lines displaying some similarity will have been deleted, and a single typical example allowed to stand for them.

9.4. Technical requirements

A concordance program may be able to thin results, or the same result may be possible by re-running the query with a different search term, or with more filters, for example by searching for a phrase, or by limiting the results to every n th occurrence, as described in section 3.

Automatically thinning a concordance to produce typical examples, as discussed above, requires software to implement complex algorithms to interpret the patterns in the co-text and to select typical examples.

10. Categorising

10.1. Description

This is a function which can be applied to concordance lines from a corpus. It is sometimes useful for the analyst to be able to manually categorise the concordance lines, for example to classify different senses of a word which the analyst is able to assign by reading the concordance. Categorising concordance lines can also be used as a way of manually thinning the concordance.

10.2. Example

In the example in Figure 33.8, the analyst has assigned letter codes ('i', 'j', 'r', 'o' and 'n') to each of the 21 concordance lines for 'fast' (every 15th example sampled from BNC-Baby).

easily cultured.

The tubercle bacillus has long been recognised to exist in various guises and seems able to exist interchangeably with and without its cell wall.

When a concerted and invasive effort has been made to find acid fast rods in sarcoid tissue they seem to be present, and acid fast bacteria without cell walls and tuberculostearic acid have also been isolated from lesions of patients with sarcoidosis.

r	Skoda then? No good? if I go too fast!	you know, if they put cou
j	ld make that. Ow	Think how fast it's gonna be on that. Although game. th
j	tions were reduced to the occasional fast break and the low-percentage shot. Peter S	
r	e, for example, was growing twice as fast as the United States' zone, and now empl	
i	europod could avoid becoming stuck fast in the soft, muddy bottom of a lake. If the ..	
j	t's rate of descent was half again as fast as the rest, taking him past the others, and	
i	In no time at all Miss Beard was fast asleep. She lay on her back, her usually s	
r	start to a dull market going nowhere fast. By late afternoon the FT-SE 100 had r	
j	ne batsman but the real need is for a fast bowler.' Man in the middle RICHIE Ri	
r	.. way I and he couldn't get rid of me fast enough. I felt then as if my whole life I	
r	n has caught its radiance. It is rising fast I swear I can see its motion I above ...	
j	ir engagement. Angus reckoned that fast business expansion was absorbing all her	
o	be they seem to be present, and acid fast bacteria without cell walls and tuberculoste	
j	s disgust with the restless owners of fast cars, a temperate man's contempt for drink	
r	He knew he would have to work fast. There were already police whistles soundi	
n	s of red satin. As I said, he broke his fast and left within the hour.' Corbett rose :	
r	being asked. But we've got to move fast.' 'This haste,' said Paul, 'it's ...	
j	usiness tax. We are also encouraging fast payment by large companies.' He sai	
r	in hand to take advantage of another fast growing market. He was optimistic ab	

Fig. 33.8: Categorisation codes assigned to concordance lines (using the MonoConc program)

The categorisation has been done as follows: 'r' indicates 'fast' is an adverb, meaning 'quickly'; 'j' indicates that 'fast' is an adjective, meaning 'quick'; 'n' indicates that 'fast' is a noun, meaning 'to go without food', and 'i' indicates that 'fast' is part of an idiomatic expression, partially or fully de-lexicalised. One line has also been tagged 'o', for 'other', and it is often useful to have such a category for problematic examples. Examining more concordances would probably yield more evidence, making it possible to categorise this and other difficult examples, and would involve increasing the number of categories.

10.3. Analysis with this function

Categorising lines manually is necessary where it cannot be done by specifying formal criteria when searching or thinning, either because the functionality is not available, or the necessary level of annotation is not present, or, most likely, because the desired categorisation requires human intervention and analysis. This type of categorisation may therefore be seen as a type of research where the concordance is a tool to help manual, qualitative linguistic analysis.

10.4. Technical requirements

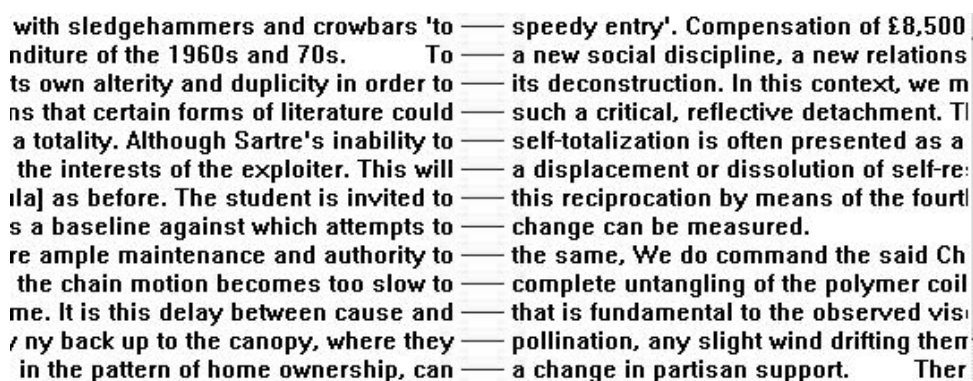
This type of manual annotation of concordance lines is often done on concordance printouts with a pen. Software which allows the annotation to be done on the electronic concordance data makes it possible to sort on the basis of the annotations, and to thin the concordance to leave only those lines with or without a certain manual categorisation.

11. Hiding the node word

11.1. Description

A simple but powerful pedagogic exercise can be created by hiding the search term (or node word) in a KWIC concordance. A human subject can then be shown the concordance lines with the node word invisible, and they must try to guess what the word is. An alternative, or additional, task is to ask the student to identify its wordclass.

11.2. Example



with sledgehammers and crowbars 'to	speedy entry'. Compensation of £8,500
nditure of the 1960s and 70s. To	a new social discipline, a new relations
ts own alterity and duplicity in order to	its deconstruction. In this context, we m
ns that certain forms of literature could	such a critical, reflective detachment. Tl
a totality. Although Sartre's inability to	self-totalization is often presented as a
the interests of the exploiter. This will	a displacement or dissolution of self-re:
ila] as before. The student is invited to	this reciprocation by means of the fourtl
s a baseline against which attempts to	change can be measured.
re ample maintenance and authority to	the same, We do command the said Ch
the chain motion becomes too slow to	complete untangling of the polymer coil
me. It is this delay between cause and	that is fundamental to the observed vis
/ny back up to the canopy, where they	pollination, any slight wind drifting ther
in the pattern of home ownership, can	a change in partisan support. Ther

Fig. 33.9: Concordance with node word concealed

11.3. Analysis with this function

The use of this technique is usually pedagogically motivated. It can be used as a language awareness training exercise for native or non-native speakers.

11.4. Technical requirements

It is useful if the software can do this and print or save the lines with the node word hidden. Otherwise the user can save the concordance and then edit it in another applica-

tion (such as a text editor or word processor), or even simply print it out and black out the node word with ink.

A related technique that could be useful for teaching purposes would be to conceal the significant collocates where they occur in the concordance lines, and some concordancers can do this automatically.

12. Showing collocates

Collocates are words which tend to occur frequently in the vicinity of the search term. Some concordance software applications can compute the significant collocates of the search term in the corpus, and highlight these words in a particular way in the concordance view, for example by colouring them. In Figure 33.10, the collocates are shown in bold and italic type to differentiate them from other words in the co-text.

market by setting up more **cost effective** production facilities based on L
 cognition that markets are **an effective way of** generating wealth and
 of a national state **to exercise effective control of** its own affairs has le
 sts from undertaking **the most effective means of** monitoring the Siberi
 : as notes. He always felt that **effective** musical criticism began by beir
 that smallpox vaccine remains **effective** even when stored at relatively
 Craig Robertson, they have **an effective and**, at times, elegant midfiel
 for his criticisms of the lack **of effective** policing **of** what he defends as
 y of 1986. **The** extraordinarily **effective** popular figure of the masked Si
 ders Situationist film to **be an effective** oppositional practice. One of th
 concept of the spectacle **is an effective** term which now has a wide cur
 ides. This work made **an effective** bridge **to** the equally spare and
 launching an inquiry into how **effective** competition had been in improv
 t of people already believe **an effective** education system is the key to

Fig. 33.10: Collocates of the node word

This can help to identify patterns of co-occurrence in the concordance lines, particularly where there are too many examples to see in one screenful, or where the position of the collocate is variable. This is a useful function, because while the linguist may be able see repeated co-occurrences of words and structures, it is not possible to assess the statistical significance of these features simply by looking at them (see article 36).

This method of computing and displaying the collocates does risk obscuring the process of calculation from the user. The linguist should remember that there are various ways to calculate collocates, and choices need to be made regarding, among other things, the collocation window, the basis for establishing what is the expected frequency of co-occurrence, the metric for assessing significance and the thresholds for frequencies and significance. Showing collocates in the concordance window should be seen as only a quick or preliminary indication of potential collocates, which are likely to require more focussed investigation and verification.

Concordance programs will also typically be able to generate lists of significant collocates, sometimes lists of positional collocates, showing which words tend to co-occur in particular positions to the left and right of the node word. Such lists can be invaluable for suggesting further searches to produce concordances and examine patterns of usage. Investigating collocation is a very important part of the corpus linguistics basic toolkit, and is covered in article 58.

13. Using metadata

13.1. Description

With certain corpora and analysis tools, it is possible to restrict the scope of searches to texts (or elements of texts) with particular characteristics. For example, some corpora contain texts which originate in both written and spoken modes. If a corpus is marked-up in such a way that the component texts are clearly marked-up as written and spoken, then searches can be restricted to one section or the other, and frequencies compared.

13.2. Example

Figure 33.11 shows a pie chart from the Xaira program which displays graphically the distribution of the word 'effect' in the BNC-Baby corpus. BNC-Baby is divided into four subsections, each of approximately 1 million words, representing spoken conversation, fictional prose, academic writing and newspapers.

13.3. Analysis with this function

Using metadata to search in particular texts or elements of texts can be particularly useful for research which aims to exploit differences in register, genre, mode and text type among the texts in a corpus.

It should, however, be borne in mind that the design criteria of a corpus may have aimed to sample certain categorisations in a balanced and representative way, while others may simply be indicated and were not carefully sampled and balanced. Sinclair (2005) argues that only elements which are designed to be balanced and representative should be contrasted in research using a corpus.

Careful attention should always be paid to the documentation of the design and implementation of the corpus metadata. For example, Burnard (2000) indicates what the design criteria for the British National Corpus were. He also indicates that, for the spoken, conversational part of the corpus, there is a metadata category which could be used to indicate the sex of the speaker, but which is often only recorded for the main respondent (the person carrying the recording device), and not for the other interlocutors. It is therefore difficult to carry out research to compare the speech of men and women using the BNC, although not impossible (e.g. Rayson/Leech/Hodges 1997).

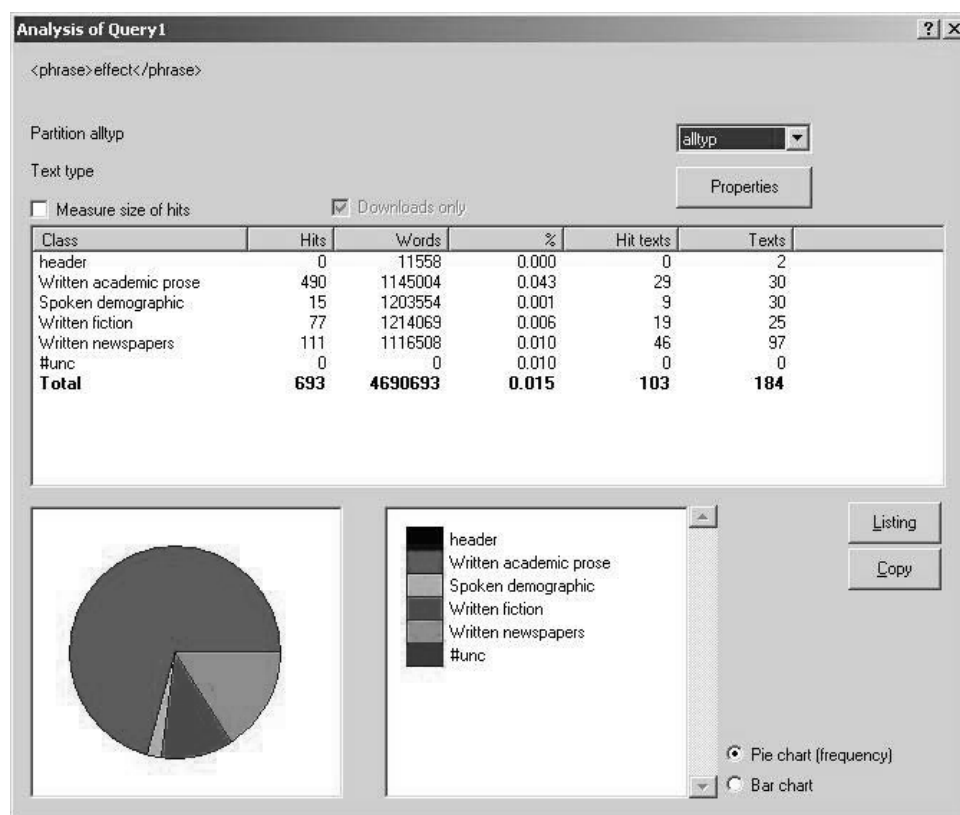


Fig. 33.11: Distribution of 'effect'

13.4. Technical requirements

In common with other functions which exploit the markup in a text, application of this function requires access to software which is capable of interoperating with the markup. Successfully doing this is rarer than successful processing of structural markup such as paragraph and page breaks, and annotations associated with individual words, such as wordclass tags. This is because it is necessary for software to recognise and process annotations with a wider, and more complex scope. There may be descriptive metadata which applies to the whole corpus, to sections of it, to subsections, to individual texts, to groupings of texts in different sections, or to certain passages within various texts. For this reason, software which can work with descriptive markup in a sophisticated way will often be tied to a particular corpus, or corpus markup framework.

If the limitations of the corpus markup or the corpus analysis software mean that it is not possible to restrict searches on the basis of metadata categories, then an alternative approach is to put files representing texts from different text types in different folders. Similarly, file naming conventions can be used, so that in effect, codes representing metadata values are encoded in the file name. With these methods, the operating system's method of storing files is exploited to allow the analysis of certain corpus components.

While certain types of analysis can be done this way, there are various limitations and potential problems. These problems include the following:

- the user has to load and analyse different subcorpora and find a way of comparing results;
- in transferring between computers, it may be difficult to preserve the file hierarchy and the file names, and for this reason this solution is not suitable for long term preservation of the corpus metadata;
- it can be difficult, and sometimes impossible to apply multiple metadata categories to define sets of texts – the user may be restricted to one pre-defined categorisation which is fixed in the arrangement of files into folders;
- some operating systems and some concordance programs restrict allowable file names and suffixes, and the number of files in a folder, so a solution may not work in all computing environments.

14. Wordlists

14.1. Description

A wordlist (sometimes written ‘word list’) is a list of all the different words in a text or corpus, usually accompanied by the number of times each word occurs.

While it is often possible to simply press a button in a software application to make a wordlist, this ease of use may disguise the fact that making a wordlist depends on having a working definition of what a word is, and of what counts as an occurrence of the same word, and that this will vary from language to language.

It is also possible to make lemma lists, where inflected forms of words are counted together as examples of the occurrence of a head word, or lemma. Wordclass tags, or word and tag pairs, may also be counted and listed.

14.2. Example

See next page.

14.3. Analysis with this function

Examining the wordlist from a large general reference corpus, as in Figure 33.12 above, can be useful for finding out about the words which occur most frequently in a language. Examining the wordlist from a text or a specialised corpus can be a starting point for examining the lexis of a particular text or text type. A wordlist from a text or corpus may be compared to the wordlist from a large reference corpus. It may be interesting, for example, to see whether the most frequent words differ from the norm. The lower than usual occurrence of ‘of’, for example, may be a reflection of a lower amount of

Rank	Frequency	Lemma	Wordclass
1	6187267	the	determiner
2	4239632	be	verb
3	3093444	of	preposition
4	2687863	and	conjunction
5	2186369	a	determiner
6	1924315	in	preposition
7	1620850	to	infinitive-marker
8	1375636	have	verb
9	1090186	it	pronoun
10	1039323	to	preposition
11	887877	for	preposition
12	884599	i	pronoun
13	760399	that	Conjunction
14	695498	you	pronoun
15	681255	he	pronoun
16	680739	on	preposition
17	675027	with	preposition
18	559596	do	verb
19	534162	at	preposition
20	517171	by	preposition

Fig. 33.12: 20 most frequent lemmas in the BNC

post-modification of noun phrases, which may be a sign of a style which is simpler, or less formal in some way.

Wordlists derived from different components within a corpus may also be compared. Rayson/Leech/Hodges (1997) were able to compare wordlists derived from various sets of speakers in the spoken component of the BNC, and to conduct a quantitative analysis of the utterances of speakers categorized by such factors as sex, age, social group and geographical region.

For a systematic way to find words that occur more or less frequently in a text compared to a reference corpus, a method for determining 'keywords' can be used (see section 15 below).

The most frequent words in a language tend to be grammatical or 'closed class' words. When the user wants to look beyond the most frequent grammatical words in order to see which are the 'content' words which are used most often, a 'stop list' may be used. This is a list of words which the program omits from its searches. Such lists typically contain grammatical, or closed class, words, or simply the most common words in a corpus. If the user is interested in the grammatical words, then it may be necessary to make sure that a stop list is not being used by the program by default. Stop lists are sometimes used by software to prevent users from searching for the most common words, as this would be too big a task for the software, and would produce too many results to analyse manually.

Looking at frequently occurring words may also tell you something about the themes or topics in the texts in a corpus (see more in the section on *Keywords* below).

Obtaining lists of the most frequent words in a corpus can also be used in pedagogic research. Knowing which words occur most frequently can help to indicate which words a learner is most likely to encounter. The top of a wordlist can perhaps be seen as the core vocabulary of a language. The selection of words for learner dictionaries, grammars and other teaching materials is often informed by wordlists obtained from corpora.

14.4. Technical requirements

A utility to make wordlists is usually part of a concordancer or a corpus analysis software package. Although it may not be visible to the user, compiling a wordlist depends on recognising words, or ‘tokenising’ the text. Being able to count words depends on being able to recognise what a word is, and so some variation in word frequency counts will result from the differing tokenisation algorithms which are employed by different programs.

It is not necessary to use a corpus analysis software package to obtain wordlists from a text or corpus. A simple program can be written in a variety of scripting or programming languages to produce a wordlist. Typically, what such a program will do is send the contents of a file (or a collection of files) through a series of processing steps to (i) remove punctuation characters and other non-word elements, (ii) identify word boundaries, (iii) sort the words into alphabetical order, (iv) identify and count identical words which are adjacent in the sorted list, and (v) sort the wordlist into descending order of frequency.

It is possible to obtain the result by various means, and to make different decisions about what to include or exclude, and how to identify word boundaries, etc. Software usually includes functions to make a wordlist, with varying levels of control of these options.

15. Keywords

15.1. Description

The keywords of a text, in the sense intended here, are words which can be shown to occur in the text with a frequency greater than the expected frequency (using some relevant measure), to an extent which is statistically significant.

Confusingly, the term ‘keyword’ is used in more than one way in corpus linguistics. It is also used to mean the search term, or node word, as in Keyword in Context (KWIC). The meaning in this section is an important, or ‘key’, word in a text. This latter usage is inspired by Raymond Williams (1988 [1976], Bennett/Grossberg/Morris 2005), who uses it to signify culturally significant words in the discourse of a society. Mike Scott (2001, Scott/Tribble 2006) introduced the term to corpus linguistics and implemented a method of computing keywords in his WordSmith Tools software.

It is usually most relevant to compute keywords for a text (or a set of related texts) in comparison to a reference corpus, to try to obtain an indication of characteristic lexis in the text (or set of texts).

An interesting development of the notion of keywords is *key key-words*. Scott/Tribble (2006) noted that while words are often computed as *key* in a particular text, they may not be significant across a number of texts of the same type. Those that are *key* across a number of texts in a corpus are called *key key-words*.

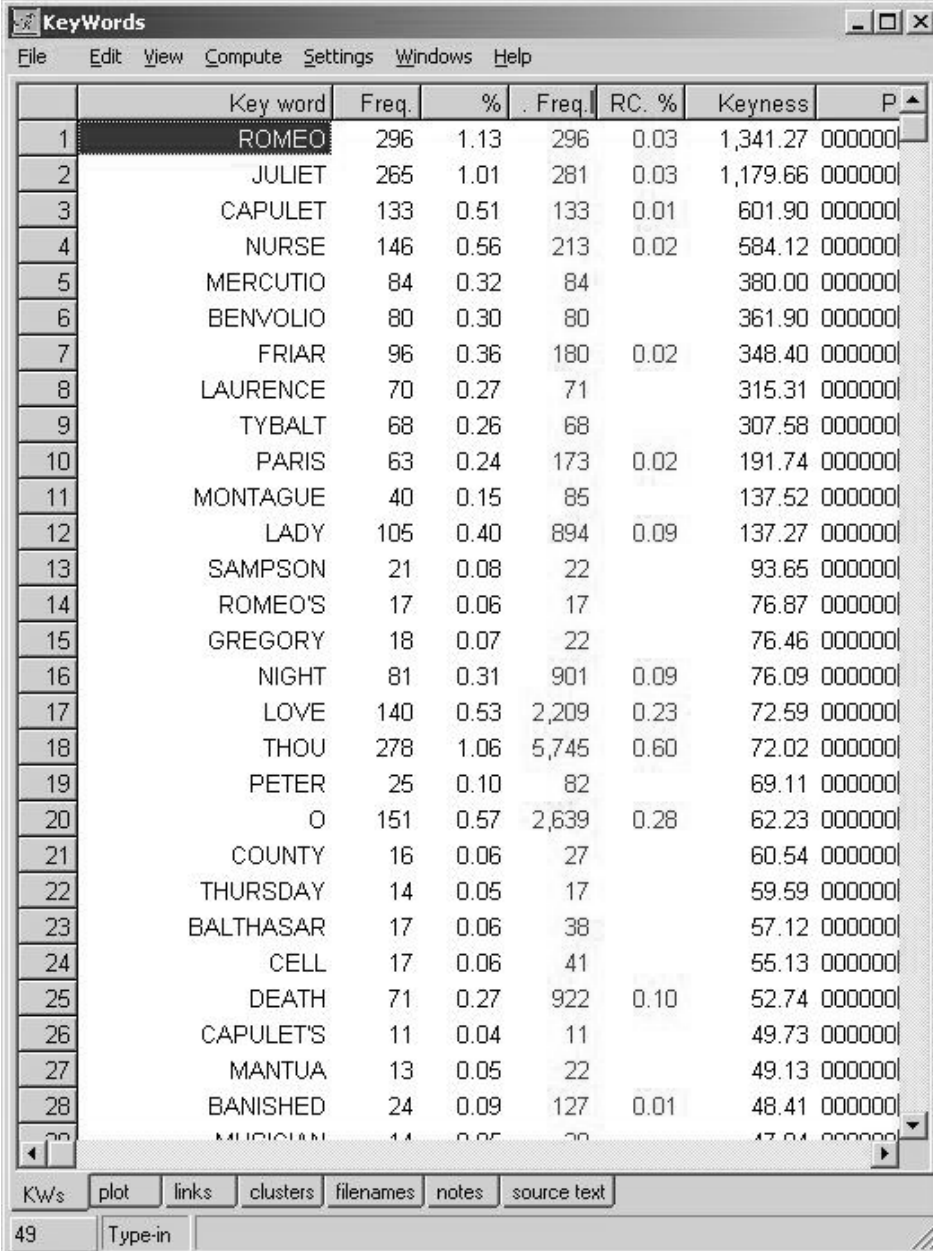
15.2. Example

Figure 33.13 shows a list of keywords generated by WordSmith Tools by comparing *A Connecticut Yankee in King Arthur's Court* by Mark Twain with the written component of the British National Corpus, ranked in descending order of their significance, or *keyness*. (The usefulness of such a comparison is discussed in 15.3 below.)

I, AND, SIR, KING, YE, IT, MY,
 LAUNCELOT, ME, WAS, KNIGHTS, MERLIN,
 KNIGHT, ARMOR, CLARENCE, THING,
 SANDY, HIM, MARHAUS, THAT, UPON,
 TOWARD, MORDRED, GAWAINE, CAMELOT,
 SAGRAMOR, SO, DOWLEY, YES, COULDN'T,
 MILRAYS, THEN, BUT, THEY, HUNDRED,
 PRESENTLY, KING'S, ARTHUR'S, WOULD,
 MAN, HAD, WE, ALL, YONDER, THOU,
 SLAVE, MIRACLE, OUT, ARTHUR, GOOD,
 UNTO, COULD, AH, HATH, MYSELF,
 ERRANTRY, LET, SMOTE, ALONG, WELL,
 MAGICIAN, NOBLE, HIS, GOT, WHEREFORE,
 SWORD, HE, EVERYBODY, THEE, SPEAR,
 YOU, ABBOT, PERADVENTURE, OFFENSE,
 HERMIT, THEM, PROCESSION,
 STRAIGHTWAY, A, YET, MONKS, KAY,
 EVER, GUENEVER

Fig. 33.13: Keywords in *A Connecticut Yankee*

In the example given in Figure 33.14 below, the keywords from Shakespeare's *Romeo and Juliet* are given, as calculated in comparison to the rest of Shakespeare's dramatic works. In this way, words which appear relatively more frequently in this play than in the others should appear at the top of the list. The WordSmith Tools program was used for these examples, and the display includes frequency of the words in the text (*Romeo and Juliet*), and in the corpus (the rest of Shakespeare's drama), as well as the frequencies as a percentage of the total number of words in the text or corpus and a calculation of the 'keyness' of the words.



	Key word	Freq.	%	. Freq.	RC. %	Keyness	P
1	ROMEO	296	1.13	296	0.03	1,341.27	000000
2	JULIET	265	1.01	281	0.03	1,179.66	000000
3	CAPULET	133	0.51	133	0.01	601.90	000000
4	NURSE	146	0.56	213	0.02	584.12	000000
5	MERCUTIO	84	0.32	84		380.00	000000
6	BENVOLIO	80	0.30	80		361.90	000000
7	FRIAR	96	0.36	180	0.02	348.40	000000
8	LAURENCE	70	0.27	71		315.31	000000
9	TYBALT	68	0.26	68		307.58	000000
10	PARIS	63	0.24	173	0.02	191.74	000000
11	MONTAGUE	40	0.15	85		137.52	000000
12	LADY	105	0.40	894	0.09	137.27	000000
13	SAMPSON	21	0.08	22		93.65	000000
14	ROMEO'S	17	0.06	17		76.87	000000
15	GREGORY	18	0.07	22		76.46	000000
16	NIGHT	81	0.31	901	0.09	76.09	000000
17	LOVE	140	0.53	2,209	0.23	72.59	000000
18	THOU	278	1.06	5,745	0.60	72.02	000000
19	PETER	25	0.10	82		69.11	000000
20	O	151	0.57	2,639	0.28	62.23	000000
21	COUNTY	16	0.06	27		60.54	000000
22	THURSDAY	14	0.05	17		59.59	000000
23	BALTHASAR	17	0.06	38		57.12	000000
24	CELL	17	0.06	41		55.13	000000
25	DEATH	71	0.27	922	0.10	52.74	000000
26	CAPULET'S	11	0.04	11		49.73	000000
27	MANTUA	13	0.05	22		49.13	000000
28	BANISHED	24	0.09	127	0.01	48.41	000000
29	MUSICIAN	11	0.05	20		47.04	000000

KW's plot links clusters filenames notes source text

49 Type-in

Fig. 33.14: Keywords from *Romeo and Juliet*

HSK – Corpus Linguistics

MILES, Release 18.02x on Tuesday January 22 18:53:50 BST 2008

gesp. unter: HSKCORSU33 / letzter Rechenvorgang: 11-02-08 15:05:40

15.3. Analysis with this function

Keywords are an attempt to characterise the topic, themes or style of a text or corpus. As such, compared to some other forms of analysis using a corpus, keywords analysis tends to focus on the ways in which texts function, rather than on overall characterisations of a corpus, or focussing on isolated linguistic elements in the corpus. For this reason, it is a technique which is popular in various forms of discourse and stylistic analysis (e. g. Culpeper 2002).

It is possible to obtain a lexical characterisation of a text using keywords analysis. The analyst needs to be careful how to interpret results: words can show up as keywords because they are related to the topic, and this is especially likely with proper nouns. Another potential problem is that the value of a keywords analysis depends on the relevance of the reference corpus. Comparing a US nineteenth-century novel with the BNC should produce some keywords typical of the topic and style of the novel, but will also show words which are more typical of US English, words which are more typical of nineteenth-century English, and words which are more typical of prose fiction than the wide range of texts sampled in the BNC. If the aim of the analysis is to identify typical stylistic features of the author, then the novel should more usefully be compared to prose fiction in English of his US contemporaries. If the aim is to identify features typical of only the one novel, it could be compared to the rest of the author's oeuvre.

A final problem relates to frequency and salience. The words which are perceived by the reader as the most significant in a text are not necessarily only those which occur more frequently than the reader would expect. There are other textual devices which can give a particular importance to a word in a text. The fact that a character in a play does not mention his wife's name can be striking and important, for example. Keywords can suggest ways to start to understand the topics or style of a text, and provide statistical evidence for certain textual phenomena, but cannot provide a list of all the interesting words, reveal all stylistic devices, and cannot explain a text.

15.4. Technical requirements

Keywords are calculated by comparing word frequency lists. One interesting aspect of the technique is that it is not necessary to have access to the full texts or corpora used, only to the word frequency list. This has the advantage that researchers who do not have access to the corpus for whatever reason may still be able to access the wordlist and thus calculate keywords. It also means that if the researcher is employing a very large reference corpus, only the wordlist needs to be stored on the computer. So this technique can provide a means of using a large reference corpus where restrictions arise due to size, cost or legal issues. The researchers can conduct their analysis with access to only the wordlists, and keywords can be identified and assessed for their significance purely on the basis of the comparison of lists of out-of-context words. However convenient this may be, it can also be a hindrance to thorough analysis of the actual usage of words in texts. Once a word has appeared in a list of keywords, it is likely to be useful, usually necessary, to look at the concordance lines from the corpus to understand more about whether it is part of a larger unit of words, whether it is occurring only in particular texts, and whether it is playing a particular role in the discourse.

16. Searching for larger units

16.1. Description

Up until now this article has focussed on searching for words which are entered as the search term. Researchers are also interested in using computational methods to find out which sequences of more than one word occur frequently in a corpus. This section deals with using the computer to generate lists of the most frequently occurring sequences of words in a corpus.

16.2. Example

Figure 33.15 contains a list of the 20 most frequently occurring sequences of four words in the BNC, as computed by the online resource Phrases in English (<http://pie.usna.edu/explore.html>). Note that # represents a number, and that certain sequences count as two words (e.g. “don’t”) and others as one (e.g. “per cent”), following the ways in which words are defined in the BNC.

```
I don't know
the end of the
at the end of
at the same time
I don't think
for the first time
on the other hand
between # and #
the rest of the
as a result of
in the case of
one of the most
# per cent of the
the Secretary of State
by the end of
from # to #
is one of the
don't want to
to be able to
I don't want
```

Fig. 33.15: 4-grams from the BNC

16.3. Analysis with this function

Various types of multi-word expression may be discovered with these methods. Different sorts of recurring sequences of words are referred to as multi-word expressions, multi-word units, pre-fabricated units (also known as “pre-fabs”), *n*-grams, idioms, phrases and fixed and semi-fixed expressions. Generating lists of such sequences can provide insights into the recurrent phraseology of texts.

The analyst will often find it useful to search for patterns where one of the words is a wildcard, or specified by a part of speech, or a lemma, rather than a literal word form. For example ‘*preposition* the *noun* of the’ proves to be a very common pattern in English. Examining concordances of these variable multi-word expressions will help to show the variety of lexical forms which is produced within and around them, and give some insights into how they are used.

16.4. Technical requirements

Calculating frequencies of multi-word expressions is a computationally complex task. Tools can be optimised for searching for multi-word expressions. All sequences of words of a given length (e.g. all two-, three- and four-word sequences) can be stored in an index, or database. This can give a large improvement to access times and allows computation of the most frequent and significant sequences. It is also possible to make use of lemmatisation and wordclass tagging to find sequences where the forms of some of the words in an expression may vary.

Such work requires considerable pre-processing of the corpus, and then the use of tools designed to access the indexed data.

There is also a complication with identifying the scope and granularity of multi-word expressions. Many patterns appear to be built up of smaller units, with some variable and some fixed elements. For example, some of the entries in the list above overlap, so that for example “I don’t want” and “don’t want to” appear as separate entries, but many if not most of these will be part of the longer sequence “I don’t want to”.

17. Searching and concordancing beyond the monolingual text corpus

This article has concentrated on the functions and tools which are commonly used with an English monolingual text corpus. Other types of corpus exist and are widely used. There are corpora of many languages and writing systems (see articles 21 and 20). With some languages it is useful to focus on different forms of analysis from the ones suggested here. There are multilingual corpora, where the focus of research is on comparing languages; similarly there may be corpora which have different versions of a text in one language (e.g. different translations) which the user wishes to compare (see article 16 and 55). And there are corpora which encode and store different modes, in the form of digital audio or video (see articles 11 and 12). Some corpora have streams which need to be aligned, such as audio, a phonetic transcription and an orthographic transcription.

Similar forms of alignment are necessary for various forms of parallel corpus, such as translation corpora (see articles 16 and 55).

Many of the functions described can be used on corpora of other languages; some of the functions reflect basic techniques which have to be implemented in different ways for other types of language; quite different techniques are necessary for other types of corpus.

Parallel corpora require structural markup to indicate the alignment of equivalents units. In the translation corpus, this alignment is typically done at the paragraph or sentence level. Specialist concordance programs exist for displaying both versions to the user, and may have further functionality, such as suggesting potential translation equivalents for the search term.

It is becoming increasingly possible for multimedia corpora to capture in digital form the audio (and sometimes also video) of language events. Current technology for sound or image-based retrieval (e. g. “find me something which sounds like this ...”) is rarely successfully implemented in language analysis tools, and corpora still generally require the use of the text transcription or markup for retrieval. A user may search in the orthographic or phonological transcription for occurrences of a particular word (or other unit) and then listen to the linked audio stream as well for each of the concordance lines. The techniques for analysis of spoken data are likely to be subject to significant advances in coming years as more resources and tools become available (see article 31).

While it is hoped that some of the principles outlined above will continue to be of relevance, the corpora of the future are likely to reflect a multilingual and multimedia environment in which distributed online access to resources and to analysis tools is increasingly the norm.

18. Tools and corpora

The following resources and tools were used in producing the examples for this article, or were examined to help with the taxonomy of functions:

- British National Corpus (BNC) and BNC-Baby <http://www.natcorp.ox.ac.uk/>.
- Bank of English and the ‘lookup’ tool – this resource is not freely available online. Search for Bank of English and check websites at Collins Dictionaries and at the University of Birmingham for further information.
- Concapp <http://www.edict.com.hk/PUB/concapp/>.
- Concordance <http://www.concordancesoftware.co.uk/>.
- IMS Corpus Workbench <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>.
- Lancaster-Oslo/Bergen Corpus (LOB), see <http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM>.
- MonoConc <http://www.athel.com/mono.html>.
- ParaConc <http://www.athel.com/para.html>.
- Phrases in English (PIE) <http://pie.usna.edu/>.
- Variation in English Words and Phrases (VIEW) <http://view.byu.edu/>.
- Sara and Xaira <http://www.natcorp.ox.ac.uk/tools/>.
- WordSmith Tools <http://www.lexically.net/wordsmith/>.

There are many other very useful tools available. The fact that they are not included here should not be seen as a reflection on their potential usefulness.

19. Literature

- Baeza-Yates, R./Ribeiro-Neto, B. (1999), *Modern Information Retrieval*. Harlow: ACM Press, Addison Wesley.
- Barlow, M. (2004), Software for Corpus Access and Analysis. In: Sinclair, J. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 204–221.
- Bennett, T./Grossberg, L./Morris, M. (eds.) (2005), *New Keywords: A Revised Vocabulary of Culture and Society*. Oxford: Blackwell.
- Bernot, E./Alarcón, E. (2005), *Index Thomisticus by Roberto Busa SJ and Associates*. Available from: <http://www.corpusthomicum.org/it/index.age> [Accessed 2006–10–02].
- Burnard, L. (ed.) (2000), *Reference Guide for the British National Corpus (World Edition)*. Available from <http://www.natcorp.ox.ac.uk/docs/userManual/> [Accessed 2007-01-11].
- Chomsky, N. (2002 [1957]), *Syntactic Structures*. Berlin: Mouton de Gruyter.
- Culpeper, J. (2002), Computers, Language and Characterisation: An Analysis of Six Characters in Romeo and Juliet. In: Melander-Marttala, U./Östman, C./Kytö, M. (eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium 15*. Uppsala: Universitetsstryckeriet, 11–30.
- Firth, J. R. (1968), *Selected Papers of J. R. Firth 1952–59*. Edited by F. R. Palmer. London: Longman.
- Hoey, M. (2005), *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Jakobson, R. (1960), Concluding Statement: Linguistics and Poetics. In: Sebeok, T. (ed.), *Style in Language*. Cambridge, MA: MIT Press, 350–377.
- Lamy, M.-N./Klarskov Mortensen, H. J. (1999–2007), Using Concordance Programs in the Modern Foreign Languages Classroom. In: Davies, G. (ed.), *Information and Communications Technology for Language Teachers (ICT4LT)*, Slough, Thames Valley University. Available from: http://www.ict4lt.org/en/en_mod2-4.htm [Accessed 2007-04-25].
- Louw, W. (1993/2004), Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In: Baker, M./Francis, G./Tognini-Bonelli, E. (eds.), *Text and Technology*. Amsterdam: John Benjamins, 157–176. [Reprinted in Sampson, G./McCarthy, D. (eds.) (2004), *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum, 229–241.]
- Manning, C./Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Rayson, P./Leech, G./Hodges, M. (1997), Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. In: *International Journal of Corpus Linguistics* 2(1), 133–152.
- Renouf, A./Kehoe, A./Banerjee, J. (2007), WebCorp: An Integrated System for Web Text Search. In: Hundt, M./Nesselhauf, N./Biewer, C. (eds.), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 47–68.
- Scott, M. (2001), Comparing Corpora and Identifying Key Words, Collocations, and Frequency Distributions through the WordSmith Tools Suite of Computer Programs. In: Ghadessy, M./Henry, A./Roseberry, R. L. (eds.), *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: Benjamins, 47–67.
- Scott, M./Tribble, C. (2006), *Textual Patterns: Keyword and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Sinclair, J. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (2004), *Trust the Text*. London: Routledge.
- Sinclair, J. (2005), The Text and the Corpus: Basic Principles. In: Wynne, M. (ed.), *Developing Linguistic Corpora*. Oxford: Oxbow Books, 1–16. Also online, available from <http://www.ahds.ac.uk/linguistic-corpora/> [Accessed 2006-10-02].
- Tognini-Bonelli, E. (2001), *Corpus Linguistics at Work*. Amsterdam: John Benjamins.
- Williams, R. (1988 [1976]), *Keywords*. London: Fontana.

Martin Wynne, Oxford (UK)

HSK – Corpus Linguistics

MILES, Release 18.02x on Tuesday January 22 18:53:50 BST 2008

gesp. unter: HSKCOR\$U33 / letzter Rechenvorgang: 11-02-08 15:05:40