The role of corpus and collocation tools in practical lexicography

Elizabeth Walter & Andrew Harley
English Language Teaching & Dictionaries
CambridgeUniversity Press
ewalter@cambridge.org; aharley@cambridge.org

Abstract

This paper looks at the corpus tools developed at Cambridge University Press, particularly those concerned with identifying and analysing collocations. It discusses the pros and cons of different tools, looking particularly at how each may best be used in a commercial environment where lexicographers must meet challenging deadlines whilst producing thorough, authoritative, and accessible text, and argues that while such tools provide invaluable empirical evidence, native speaker intuition still has an important role to play when time for research is limited.

1 Introduction

The dictionary department at Cambridge University Press is located within the English Language Teaching department, and specializes in dictionaries for non-native speakers of English, both monolingual and bilingual. Collocation is a crucial element in learning English, and the identification and analysis of collocations is therefore of great importance in the compilation of dictionary texts. We are fortunate to have a team of systems developers working alongside a team of lexicographers in the same office, which ensures both that there is a deep and ongoing co-operation and exchange of ideas, and that each team has a good understanding of the needs and capabilities of the other. This has led to the development of fast, practical, and flexible corpus tools which are in constant use for compiling dictionaries.

It is undeniable that corpus use has revolutionized lexicography, and the availability of enormous amounts of corpus data and sophisticated analysis tools enable us to gather reliable empirical evidence about words and their environment. However, for lexicographers working for a commercial publisher, who have to work to strict deadlines, the sheer scope of the data available can pose practical problems. At Cambridge, lexicographers are typically expected to compile entries for around 6 senses per hour. By some commercial standards this is leisurely; to an academic who may spend many days teasing out the properties of a single word, it may seem positively barbaric.

This paper will look at the tools available to Cambridge lexicographers, and discuss how different tools may be chosen for different types of words and to elicit different types of information. It will look at the advantages and pitfalls of these tools, and argue that native speaker intuition still has an important role to play when time for research is limited.

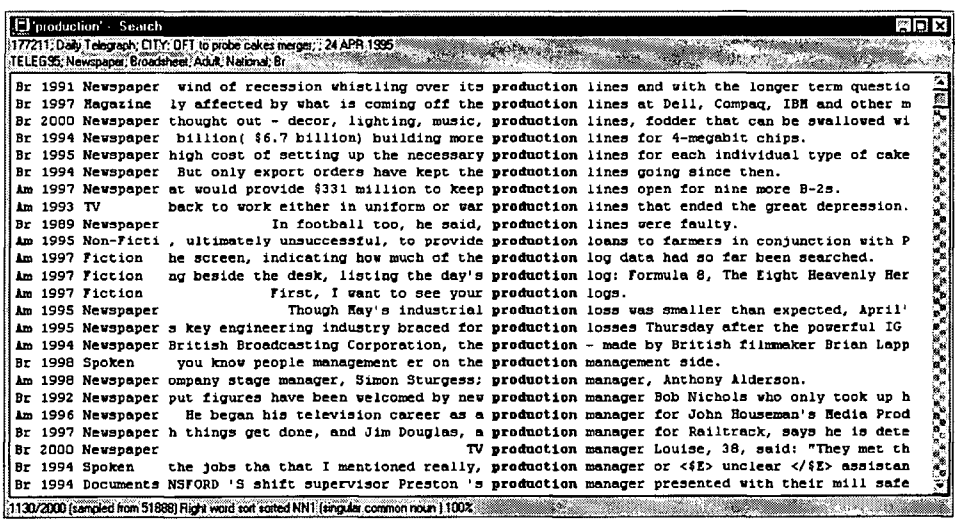
2 Collocation tools

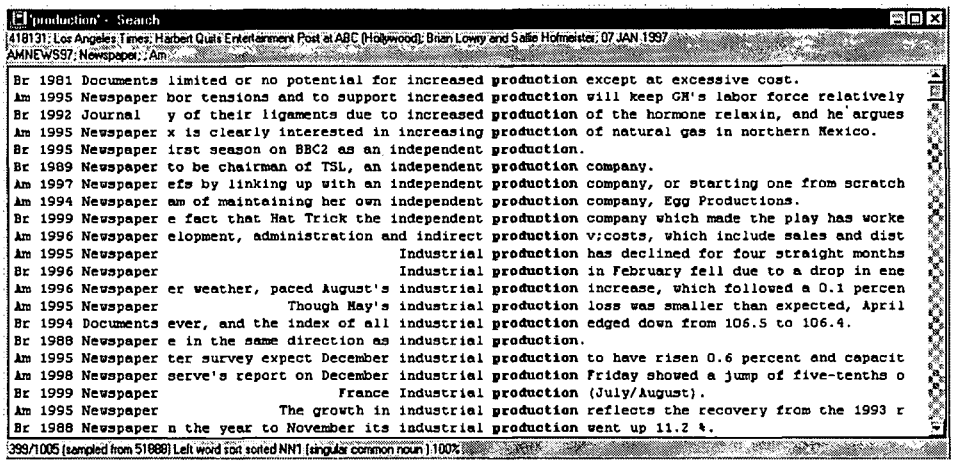
In this section, the three main ways of searching for collocations (concordances, collocation searches, and statistics) are discussed. This is not intended to be a comprehensive description of the Cambridge corpus tools, but a few general points relevant to the discussion of collocational analysis apply to all searches:

1) It is possible to search for all inflections of any word if desired. This is a major advantage over systems which allow searches on one form only, as collocational information may vary significantly according by inflection.
2) Because the corpus is grammar-tagged, it is possible to select a part of speech for any word, thus, for instance, separating the search for duck the noun and duck the verb.
3) Many other refinements of searches are possible, for instance omitting proper nouns, making a search case sensitive, excluding occurrences of less that a specified number, excluding 'empty words' such as common function words.
4) At present, we have no reliable method of automatic sense discrimination. This is a major drawback, particularly when analysing a minor sense of a common word, though analysis of collocation is itself a useful aid to sense discrimination.

2.1 Concordances

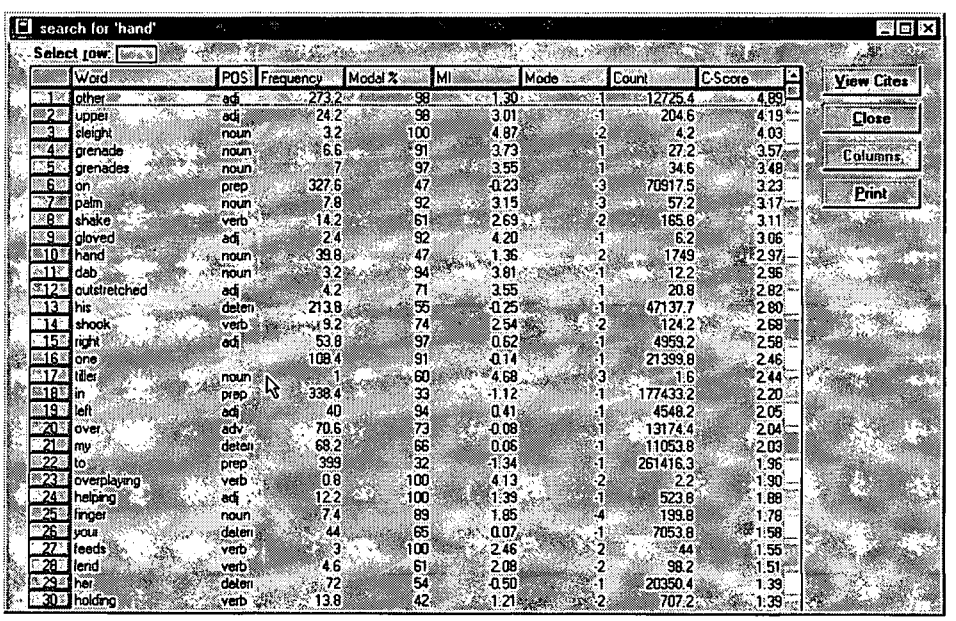
The simplest kind of collocation search is a concordance search which can be scanned by the lexicographer to search for recurring patterns. To search for collocations, words can be sorted alphabetically to the left or right of the Key Word, thus highlighting collocations by presenting them together.





2.2 Collocation searches

Collocation searches perform statistical analysis on the cites containing the search string to identify which other words co-occur most frequently and most significantly with the search word. When a search is completed, a collocation table appears:



	Word	POS	Frequency	Modal %	MI	Mode	Count	C-Score
1	other	adj	273.2	98	1.30	-1	12725.4	4.89
2	upper	adj	24.2	98	3.01	-1	204.6	4.19
3	sleight	noun	3.2	100	4.87	-2	4.2	4.03
4	grenade	noun	6.6	91	3.73	-1	27.2	3.57
5	grenades	noun	7	97	3.55	-1	34.6	3.48
6	on	prep	327.6	47	0.23	-3	70917.5	3.23
7	palm	noun	7.8	92	3.15	-3	57.2	3.17
8	shake	verb	14.2	61	2.69	-2	165.8	3.11
9	gloved	adj	2.4	92	4.20	-1	6.2	3.06
10	hand	noun	39.8	47	1.36	-2	1749	2.97
11	dab	noun	3.2	94	3.81	-1	12.2	2.96
12	outstretched	adj	4.2	71	3.55	-1	20.8	2.82
13	his	deter	213.8	55	0.25	-1	47137.7	2.80
14	shook	verb	9.2	74	2.54	-2	124.2	2.69
15	right	adj	53.8	97	0.62	-1	4959.2	2.58
16	one	pron	108.4	91	0.14	-1	21399.8	2.46
17	tiller	noun	1	60	4.68	-3	1.6	2.44
18	in	prep	338.4	33	1.12	-1	177433.2	2.20
19	left	adj	40	94	0.41	-1	4548.2	2.05
20	over	adv	70.6	73	0.08	-1	13174.4	2.04
21	my	deter	68.2	66	0.06	-1	11053.8	2.03
22	to	prep	399	32	1.34	-1	261416.3	1.96
23	overplaying	verb	0.8	100	4.13	-2	2.2	1.90
24	helping	adj	12.2	100	1.39	-1	523.8	1.88
25	finger	noun	7.4	89	1.85	-4	199.8	1.78
26	your	deter	44	65	0.07	-1	7053.8	1.58
27	feeds	verb	3	100	2.46	-2	44	1.55
28	lend	verb	4.6	61	2.08	-2	98.2	1.51
29	her	deter	72	54	0.50	-1	20350.4	1.39
30	holding	verb	13.8	42	1.21	-2	707.2	1.39

Statistics given in the table are:

Frequency: the number of occurrences, per 10 million words, of the collocate and search string together.

Modal %: The frequency, expressed as a percentage, with which the collocate appears in the 'mode' position.

MI: Mutual information: an indicator of the strength of the collocation, obtained by calculating how often the word co-occurs with its collocate as compared to how the two words would co-occur if the words in the corpus were arranged randomly.

Mode: The position that the collocate most frequently takes in relation to the search string.

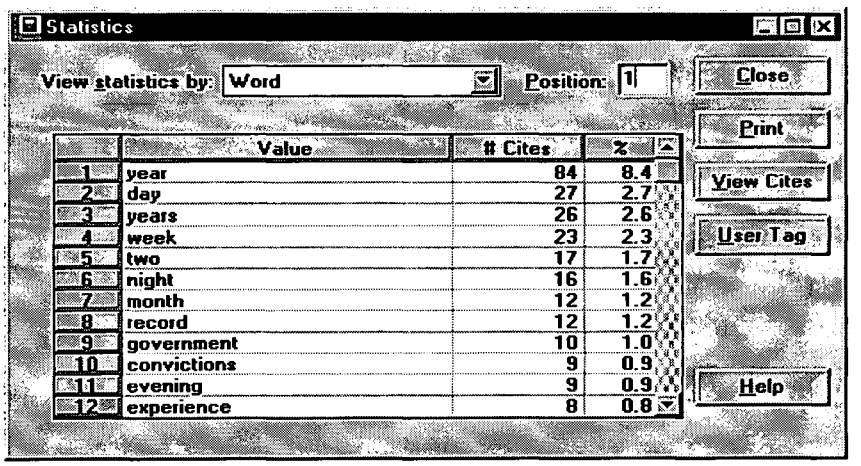
Count: The frequency, per 10 million, of the collocating word in the corpus, regardless of collocations.

C-score: A calculation of the strength of the collocation. It is calculated by multiplying the MI score by the frequency (with a few other refinements), thus giving more frequent collocations a higher score.

The table is sorted by C-score as the default, but it can be sorted on any of the other columns if desired.

2.3 Statistics

The statistics feature enables analysis of cites by several criteria. The one relevant to collocation is statistics by 'word'. The lexicographer can select any word position (e.g. 1 after the key word), and instantly receives a list of words in that position, ordered by frequency.



From this table, it is possible to see the citations for any line by clicking on the 'View Cites' button.

3 Using collocation tools in practical lexicography

3.1 Using concordances

Concordances are very fast, and useful when information is needed quickly, while collocation searches take longer to run.

The main advantage of concordances is that it is easy to scan down the screen and spot recurring patterns, as opposed to the single-word tables generated by the concordance and statistics searches, in which the reason for an item's position is often not obvious, and requires further investigation. It is therefore often worth scanning a left and right sorted collocation search as an initial aid to understanding collocation tables.

One problem faced by a lexicographer is that the amount of data available can be daunting, and it is difficult to know how to limit research in a practical way, in order to meet schedules. Our searches are limited by default to display 1,000 hits for a search string. These thousand citations are drawn randomly from the full available number, and are usually sufficient, though it is possible to select a higher number if needed, for instance when looking for a less frequent sense of a commonly occurring word.

The biggest drawback of concordances is that they do not offer a statistical analysis of frequency of collocation, so the lexicographer is given only an impressionistic sense of numbers, although in some cases, patterns are so prevalent that they will be clear even using this method. When the left/right sorting is complete, it is possible to search quickly for collocating words by keying the first letter of the word. For example, for a check on whether 'forthright comments' are more frequent than 'forthright manner', a right sort on 'forthright' followed by 'c' to find 'comments' and 'm' to find 'manner' will quickly show that 'manner' occurs more often.

It is often the case, particularly when editing rather than compiling entries, that a quick answer to a specific question like this one is what is required, rather than a full analysis of the whole word.

3.2 Using collocation searches

The collocation search gives the most detailed and statistically reliable analysis of collocation. One of the most important advantages is that it picks up collocation irrespective of the position of the collocate in relation to the keyword, thus giving the overall picture of each collocate.

In the following sentences:

The committee awarded her the prize.
The prize was awarded to her by the committee.
The committee, after long deliberations, awarded her the prize.

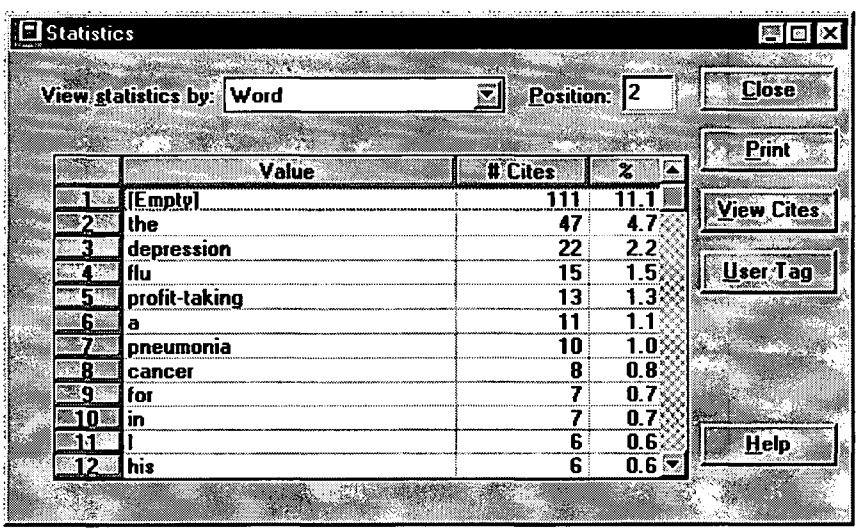
The word 'award' collocates with 'prize'. The collocation search will recognise this, whereas a left/right sort of a concordance would not group them together. In addition, the collocation search will calculate which position of 'prize' in relation to 'award' is most frequent, so the most typical example can be chosen.

Collocation searches take more time to run. The most practical method of running the searches is to queue them to run in the background overnight or while other work is going on. This can be done efficiently when working on a known stretch of text, as the lexicographer can set up searches for the next day's work. In a commercial environment there simply will not be time to run collocation searches on every word, and experience comes into play in knowing which words merit a full collocation search. Some concrete nouns, such as 'colander', 'debtor', 'materialism', are unlikely to form significant collocations, while others, such as 'production', will produce worthwhile results. We *boost/increase/ramp/expand* production; we *halt/stop* it; we have things *in* production and *going into* production; the products themselves are strong collocates: *steel/oil/gas/car* production; we have production *capacity/capability/facility/schedule/costs/figures* etc, etc. The lexicographer needs reliable, empirical evidence both to identify the collocates and to evaluate which are the most useful ones, when limited to just a few examples. An experienced lexicographer will have a good feel for which words merit a collocation search.

3.3 Using statistics

Using the statistics is very quick. It is particularly useful for providing answers to specific questions, and the skill of the lexicographer comes in knowing what are the important questions to ask, and also in knowing how to question results which appear in a chart without context, for instance knowing when it is necessary to go beyond the chart and look at the actual cites for clarification of a point.

For instance, a lexicographer working on the word 'bout' is likely to want to know what we most commonly have bouts of. A quick way to find this information is to do a normal concordance search with a 1000 sample, and then ask the stats to find words appearing 2 places to the right of 'bout'.



	Value	# Cites	%
1	[Empty]	111	11.1
2	the	47	4.7
3	depression	22	2.2
4	flu	15	1.5
5	profit-taking	13	1.3
6	a	11	1.1
7	pneumonia	10	1.0
8	cancer	8	0.8
9	for	7	0.7
10	in	7	0.7
11	I	6	0.6
12	his	6	0.6

This screen gives us fast information that in the sense of 'a short period of activity or illness', the most common collocates are 'bout of depression/flu'. However, there are some potential pitfalls in the information, and this is where the intuition of the experienced lexicographer comes in. For instance, the 8th word shown is 'cancer'. While 'bout' often collocates with an illness, someone with good native-speaker intuition will spot that 'cancer' is not a likely collocate in this sense, and on further investigation, by actually viewing those cites (which are just one key press away), will discover that the sentences are for 'bouts *with* cancer', i.e. a metaphorical usage of the fighting sense. The 5th word shown is 'profit-taking'. The lexicographer may be surprised by this, and on viewing the cites will discover that the vast majority of them come from a single source, Reuters Financial News, and may therefore place less importance on them.

As this sort of example shows, where there are pressures of time, it is important for lexicographers to develop the skills to analyse the data efficiently, and this will often mean having the intuitions necessary to ask the relevant questions.

4 Conclusion

The Cambridge collocation tools provide lexicographers with a formidable tool kit. The three major tools, concordance, collocation searches, and statistics provide different functions, and each has advantages and disadvantages, and the lexicographer need to learn how to choose the most appropriate tool for a particular word, while working to strict deadlines. This means that experience and a feel for language is still an important asset, even in an age where so much empirical data is available.