CORPUS LINGUISTIC METHODS

1. CONCORDANCES

A **concordance** is a listing of all occurrences of a particular linguistic item in a corpus (or a representative sample), together with their contexts of occurrence. The context may be presented in different ways; a typical one is the so-called KWIC (Key Words In Context) format, shown in Table 1: the search word appears in the middle of each line, surrounded by a fixed amount of context (often 7-8 words or 40 characters).

Table 1. *KWIC concordance of the verb lemma* CAUSE *(selection from BROWN)*

…governor's race forward a few months,	**causing** the campaigning to get started earlier…
…not do both"? Military power does not	**cause** war; war is the result of mistrust an…
…the high priests of the cult would have	**cause** to tremble for their personal safety, …
…diation is not like a flu virus which	**causes** temporary discomfort and then dies. The…
…rbon-14 from the fusion process would	**cause** four million embryonic, neonatal or c…
…a strong one, from the outside, might	**cause** it to snap. ## The planners in Taiwan…
…The amazing thing is that this too is	**caused** by the dearth of teachers. Teaching i…
…r the disruptions which it inevitably	**causes**". In my own case, I submit that such re…
…ned. Nervousness at the start must have	**caused** the blemishes of her first scene, or…
…es sparks on occasion and their light	**causes** all else to be forgotten. There is a…
…the business of starting and stopping	**caused** occasional raggedness, as with the firs…
…the discovery that many vegetable fats	**cause** blood cholesterol levels to drop radi…
…ls to drop radically, while animal fats	**cause** them to rise. Here Keys and others, s…
…view of reality in general, which now	**cause** us much difficulty, could be responded…
…t which gives life to a community and	**causes** it to cohere. It is the spirit which is…
…] sin, by interposing death, and thus	**causing** sin to cease, putting an end to it by t…
…ual four? Obviously, something suddenly	**caused** them to start thinking in terms of fi…
…ther than pretexts for them, that are	**causing** the trouble, and do everything possible…
…<effects,> of which the specific action	**causes** directly the one and indirectly the o…
…before or while putting it forth and	**causing** these consequences. He does not expec…

The KWIC format is useful for identifying the types of grammatical structures and set phrases associated with the search word. KWIC concordances can usually be sorted by the word appearing directly to the left or the right of the search word.

EXERCISE: Identify the patterns with which the verb *cause* occurs. Do you notice any semantic classes of words appearing in particular slots in these patterns?

2. COLLOCATE LISTS

A **collocate** list is a list of the words occurring at a particular position relative to the search word, for example ranging from the third word to the left to the third word to the right, together with their frequencies of occurrence at this position. There are various ways of representing collocate lists, a typical one is shown in Table 2. Note that each column must be read in isolation from top to bottom, i.e. it does not make sense to read across lines. Collocate lists are useful summaries of vast amounts of data for rough semantic analyses of words.

EXERCISE: Does the collocate list confirm your ideas about semantic classes?

Table 2. *Collocate list of the verb lemma* CAUSE *(from BROWN)*

L3	L2	L1	R1	R2	R3
24 the	8 the	15 and	36 the	27 to	29 to
7 my	7 hand	8 to	23 by	12 the	16 of
6 to	5 which	7 would	16 a	7 seal	7 in
6 that	4 it	7 which	8 him	6 a	5 and
5 of	4 of	7 had	6 them	5 great	4 that
5 in	4 and	6 may	5 to	4 of	4 the
4 a	4 or	5 have	5 it	3 much	2 than
3 and	3 their	4 has	5 us	2 system	2 rise
2 it	3 This	4 is	3 all	2 entire	2 build
2 with	2 conditions	3 be	2 so	2 lot	2 deal
2 this	2 signals	3 or	2 more	2 his	2 trouble
2 many	2 was	3 that	2 widespread	2 be	2 concern
2 other	2 this	3 might	2 such	2 damage	2 shear
	2 no	3 can	2 any	2 particular	
	2 process	2 it	2 her	2 more	
	2 must	2 but	2 an	2 aerator	
	2 in	2 will	2 increased	2 trouble	
	2 has	2 are			
		2 could			
		2 being			
		2 fats			
		2 been			
		2 hymen			
		2 was			
		2 thus			

3. FREQUENCY

A frequency list is a list of all words in a corpus together with their frequency. Frequency lists can usually be sorted by frequency or in alphabetical order. Frequency lists are useful for investigating global properties of texts, or even of language in general. Table 3 shows the 75 most frequent words in the BROWN corpus.

Table 3a. *Frequency list (based on BROWN)*

69377	the	6688	on	3541	which	2235	who	1749	can
36146	of	6325	be	3270	were	2219	will	1725	only
28708	and	5351	at	3259	one	2204	more	1689	other
25926	to	5272	by	3221	you	2178	no	1609	some
23321	a	5135	i	2997	her	2162	if	1604	new
21222	in	5114	had	2947	all	2071	out	1587	could
10517	that	5110	this	2847	she	1967	so	1575	time
9993	is	4553	not	2706	there	1951	said	1565	these
9755	was	4354	from	2705	would	1875	up	1401	two
9491	he	4353	are	2653	their	1858	what	1390	may
9435	for	4350	but	2630	we	1845	its	1359	then
8655	it	4194	or	2567	him	1806	about	1346	first
7244	with	3909	have	2464	been	1784	into	1336	any
7211	as	3711	an	2421	has	1780	than	1318	do
6955	his	3592	they	2316	when	1751	them	1296	such

In comparison, Table 3b shows the 75 most frequent words in a (much smaller) specialized corpus dealing with a single topic.

Table 3b. *Frequency list (based on a specialized corpus)*

283	the	20	it	13	could	10	only	7	is
83	of	18	cargo	13	he	9	aircraft	7	made
74	to	17	not	13	passenger	9	have	7	some
68	a	17	smoke	13	they	9	out	7	system
61	was	16	all	12	captain	9	which	7	tank
51	and	16	plane	12	crashed	8	fuel	7	through
48	in	16	with	12	i	8	minutes	6	3
45	fire	15	emergency	12	into	8	no	6	air
35	that	15	landing	11	airlines	8	ntsb	6	before
27	an	14	at	11	as	8	while	6	but
27	on	14	be	10	by	7	been	6	caused
23	after	14	were	10	cabin	7	burned	6	determined
23	from	13	airport	10	compartment	7	co2	6	engine
20	flight	13	board	10	crew	7	hazardous	6	evacuated
20	had	13	cockpit	10	for	7	his	6	faa

EXERCISE: (i) What do you notice about the frequency list based on the BROWN corpus (what kind of word make up the most frequent words, what do you notice about their frequencies). (ii) How does this contrast with the specialized frequency list (and can you guess what the topic of the specialized corpus was?

4. DISTRIBUTION

A particular kind of frequency list is the **distribution** frequency list, which gives raw frequencies as well as the distribution of a word across subcorpora or files. Table 4 shows the first ten words from the official BROWN frequency list.

Table 4. *BROWN frequency list*

69971-15-500 THE
36411-15-500 OF
28852-15-500 AND
26149-15-500 TO
23237-15-500 A
21341-15-500 IN
10595-15-500 THAT
10099-15-485 IS
9816-15-466 WAS
9543-15-428 HE

The first figure gives the raw frequency, the second figure gives the number of genres in which a word occurs, the third figure gives the number of texts in which a word occurs.

EXERCISE: Why might this type of information be useful?