名古屋学院大学論集　言語・文化篇　第19巻　第2号（2008年3月）

Keywords in Academic Prose: Collocation Patterns

Phillip R. Morrow

1.0 Introduction

Teaching at a Japanese university, I have often had occasion to read graduate students' theses as an advisor or thesis committee member. Although it is not part of the duties of the advisor or committee member to edit or proofread theses, I have sometimes edited or proofread sections of theses for students when they have been unable to find any other native speaker to do so. Most of the graduate students who write theses in English are highly proficient in English, but have had limited experience with academic writing. When I read drafts of their theses, I make suggestions about reorganizing paragraphs or sections and I correct the grammatical errors, which are generally few. But most of my emendations consist of rewordings of sentences or parts of sentences that are not grammatically incorrect, but which sound unnatural, at least to me. Explaining why these rewordings are preferable is difficult since they are often involve word usages that are not described by grammar rules or discussed in reference books about English usage. Making sentences "sound natural" is largely a matter of using words in the collocation patterns and grammatical patterns that are typically associated with them. There are words that frequently collocate and there are other words that don't, though there often isn't any very clear reason why they couldn't. Gaining a sense of which words are commonly used together — i.e. collocation patterns — is part of acquiring proficiency in a foreign language, but it is difficult for most learners because it is not explicitly taught in most courses nor dealt with in most textbooks.

Language corpora provide a valuable resource for information about collocations. By searching language corpora, especially specialized language corpora, we can get a more accurate picture of the way words are used in any given type of texts. We can easily find the most common collocates for any lexical item in a given register. This information can then be used to develop instructional materials to aid learners in learning the collocation patterns associated with frequently used terms in their fields.

The purpose of this paper is to identify the keywords used in academic prose and to examine the collocation patterns associated with a few words that are frequently used in academic writing. The lists of keywords provided in this study can be of use in courses for writing academic English and the investigations of collocation patterns provide a model for further investigation of these research keywords or others.

2.0 Collocation information in textbooks and reference materials

Many grammar textbooks (e.g. Eastwood 1999, Murphy 2000, Stern 2000, Swan & Walter 1997) deal with collocation patterns insofar as they are related to grammatical points, for example which verbs are followed by —*ing* forms and which are followed by infinitives. There are now a number of vocabulary textbooks (e.g. Redman 1999, McCarthy & O'Dell 2001) for advanced learners of English which are based on corpus research of collocation patterns. These textbooks generally present vocabulary items together with the lexical items with which they are most frequently used. One recent textbook (McCarthy & O'Dell 2005) entitled, *English collocations in use: How words work together for fluent and natural English*, deals specifically with teaching collocation patterns for expressions in a number of common areas, e.g. weather, travel, sport, music. Two units of this textbook deal with common collocations in academic writing. Another excellent resource is the Longman Language Activator (2002). This is a dictionary designed for students at upper intermediate and advanced levels. It is organized differently from other dictionaries. Rather than presenting definitions in alphabetical order, entries are organized around 866 keywords with similar meanings. However, these textbooks deal mostly with general vocabulary items of the sort that one uses for everyday life. There are few materials aimed at the needs of non-native English users who use English for academic writing.

3.0 Data and method

Several registers of the British National Corpus were used for this study. A two-million word "Sampler" from the British National Corpus is available on CD along with various software tools for using it. The two million words on the Sampler include one million words of spoken text and one million words of written text. The corpora on the Sampler were used for some of the initial investigations of collocation patterns in general texts. In addition, some registers from the 100 million word corpus of the British National Corpus (BNC) were used. (The BNC can be accessed freely at: http://corpus.byu.edu/bnc/.) The texts in the BNC (including those on the Sampler) have been tagged for part of speech and are organized according to register. It is possible to search for an individual lexical item and to see its overall frequency in the corpus or in a register, and to compare the frequency of an item in, for example, spoken and written registers. Many words are used as more than one part of speech, e.g. as a noun and as a verb, and in the texts they are tagged accordingly, which makes it possible to search for instances when they are used as one part of speech rather than another. Tagging also makes it possible to check, for example, which adjectives precede a given noun in a corpus.

For this study I first investigated what the most frequently occurring nouns, verbs and adjectives were in the written academic register of the BNC. I then examined the collocation patterns associated with some words that are frequently used in academic writing. From this examination some characteristics of the selected keywords became apparent, and this information is of use to those learning and to those involved in teaching academic writing in English.

4.0 Keywords in academic prose

Words are used with different frequencies in different registers. The structure of the BNC makes it possible to search individual registers to ascertain the frequency of a word's usage in that register. For this study I made a search of the nouns, verbs and adjectives that occur most frequently in the register of academic writing. The most frequent 100 tokens of each of the three word types are shown in Table 1:

Table 1. The most frequent nouns, verbs and adjectives in academic writing

	Nouns	Verbs	Adjectives
1.	time	see	other
2.	case	need	social
3.	people	make	new
4.	way	seem	different
5.	patients	know	important
6.	work	take	local
7.	law	provide	political
8.	system	become	possible
9.	children	say	general
10.	government	think	particular
11.	number	suggest	public
12.	years	include	large
13.	court	appear	likely
14.	state	give	small
15.	part	tend	high
16.	information	find	certain
17.	act	consider	similar
18.	women	let	national
19.	use	come	good
20.	group	show	common
21.	evidence	want	economic
22.	form	use	early
23.	cases	believe	clear
24.	fact	occur	major
25.	study	remain	necessary
26.	power	mean	further
27.	section	go	British
28.	education	require	human
29.	language	feel	available

30.	effect	exist	great
31.	development	get	legal
32.	life	contain	significant
33.	world	indicate	difficult
34.	data	represent	higher
35.	party	apply	main
36.	point	put	able
37.	society	arise	central
38.	child	form	present
39.	class	set	individual
40.	control	live	greater
41.	groups	wish	concerned
42.	school	agree	specific
43.	problems	suppose	various
44.	process	allow	normal
45.	areas	look	true
46.	problem	produce	single
47.	period	involve	young
48.	person	argue	special
49.	question	follow	old
50.	order	vary	real
51.	analysis	develop	personal
52.	care	reflect	long
53.	words	try	appropriate
54.	chapter	refer	recent
55.	theory	constitute	private
56.	view	continue	low
57.	authority	work	natural
58.	level	fail	following
59.	position	begin	simple
60.	practice	differ	relevant
61.	health	assume	English
62.	others	relate	whole
63.	policy	note	sexual
64.	family	seek	full
65.	sense	operate	modern
66.	terms	describe	free
67.	men	write	basic
68.	figure	receive	considerable
69.	community	accept	lower
70.	nature	ask	effective
71.	research	depend	physical
72.	action	carry	primary
73.	age	leave	total
74.	knowledge	like	complex
75.	area	understanding	direct
76.	studies	offer	only
77.	services	cf.	positive
78.	place	read	international
79.	role	learn	medical
80.	interest	turn	previous

81.	approach	add	useful
82.	change	serve	strong
83.	century	bring	current
84.	changes	express	traditional
85.	result	call	poor
86.	population	create	late
87.	history	share	American
88.	conditions	choose	cultural
89.	results	lie	serious
90.	structure	hold	civil
91.	end	compare	later
92.	year	define	foreign
93.	treatment	reduce	educational
94.	service	discuss	independent
95.	property	contribute	professional
96.	experience	keep	industrial
97.	value	open	original
98.	range	play	open
99.	basis	determine	short
100.	word	tell	black

A few observations about these lists are in order. First, these lists and the frequency figures on which they are based are derived from the texts contained in the corpus. While standard corpora are designed to be representative, their lexical content and characteristics reflect the lexical characteristics of the texts that comprise them. The lists here are based on frequency figures for lexical items in the written academic register of the BNC. The written academic register includes texts from several academic fields: humanities and the arts, medicine, natural science, politics, law, education, social science, technology and engineering. While this would seem to be a balanced sampling of academic texts, it is certainly true that a different sampling of texts or the inclusion of texts from other academic areas would yield different results. However, the BNC's academic register is relatively large: it includes more than 15 million words of text. Thus we may reasonably assume that a different sampling of texts, or inclusion of different types of texts would not alter the results significantly. However we may note that some items on the lists are associated with the geographical area in which the texts were produced, for example, three of the 100 most frequent adjectives are: *British, English* and *American*.

Another point to be noted is that in searching a corpus one makes certain choices about search parameters and the results reflect these parameters. For instance, on the list of nouns, singular and plural forms are listed separately, so *result* and *results* are counted as separate items. For the verb list, I used verb base forms and thus obtained quite different results than I would have had I used "all verb forms" for the search parameters. With those parameters, the first 15 items — in terms of frequency — would have been forms of *be* or of modal verbs.

The value of lists such as those in Table 1 is that they provide a starting point for the description of the lexical characteristics of a particular register, in this case, the written academic register. Having once identified the keywords in a particular register, one can move on to looking at how they are used, in what collocation patterns. In Section 5.0 the collocations associated with some selected research keywords are examined.

5.0 Collocations of research keywords

5.1 *Results* and *findings*

The word, *results*, frequently occurs in academic writing and *findings* is also used though less frequently. (It is not among the 100 most frequent nouns in the written academic register.) There are 5,248 tokens (occurrences) of *results* in the BNC register of written academic texts which consists of 15,429,582 words. Thus, there is an average of 340.13 tokens per one million words in the written academic text register of the BNC. In the full BNC corpus of 100 million words in 70 different registers, there are 15,321 tokens of *results*, for an average frequency of 153.21 tokens per million words. *Findings* is much less frequent: There are 1,650 tokens in the written academic register (average frequency: 106.94 tokens per million words), and 3,276 tokens in the full corpus (average frequency: 32.76 tokens per million words). The greater frequency of *results* may be due in part to the fact that it has more meanings or senses than *findings*. Collins CoBuild English Dictionary (1995) lists six meanings for *result* (two of them for its use as a verb), but only two for *finding*.

Analysis of distribution patterns showed that both *results* and *findings* were more frequent in written registers than in spoken registers. 32 of the 38 registers in which *results* had an average frequency of more than 100 tokens per million words were registers of written texts, and 12 of the 18 registers in which *results* had an average frequency of less than 50 tokens per million words were registers of spoken texts. Similarly, 11 of the 12 registers in which *findings* had an average frequency of more than 50 tokens per million words were written registers, and 17 of the 29 registers in which *findings* had an average frequency of less than 10 tokens per million words were spoken registers. However, it should be noted that of the 70 registers in the BNC, 46 are written and 24 are spoken.

In some registers, the frequency of *results* and *findings* is particularly high. In the written academic register of medicine there were 1,370.5 tokens of *results* per million words and 465.5 tokens of *findings* per million words. The highest frequency for *results* in the spoken registers was 728.2 tokens per million for the spoken lectures on commerce register. For *findings* the highest frequency was 56.3 tokens per million for the spoken register of lectures in social sciences. From data about the distribution of the two lexical items according to register it is clear that both are much more strongly associated with written registers than with spoken registers.

By checking the occurrence of adjectives that occurred within five words to the left or five words to the right of *results* and *findings* a comparison of the adjectives that frequently collocate with these two lexical items could be made. In general there was considerable overlap in the adjectives that occurred with *results* and *findings*. Both items tended to occur with adjectives that were used in order to compare (e.g. *similar, consistent, other, recent, conflicting*) and adjectives used to evaluate (e.g. *main, important, normal, clear*). There were however some differences: Some adjectives which collocated with *results* did not seem to collocate with *findings*, e.g. *visible, valuable, enormous, fatal*. This does not mean that they cannot or never collocate with *findings*, but the fact that there were no such collocations either in the academic register or in the 100 million word BNC corpus suggests that these items do not collocate as a general rule.

As for verbs which collocate with the two lexical items, there was much overlap, as might be expected. Verb forms such as *obtained, suggest, indicate, showed* and *presented* collocated with both *results* and *findings*. But one noteworthy difference involved the verb forms, *produce, produces* and *produced*. These were among the most frequent verb forms to occur before *results*, but they did not occur before *findings*, either in the written academic register or in the full BNC corpus.

5.2 Talking about causation: *influence, affect* and *cause*

A typical goal of research is to identify the factors that are related to an outcome, in other words, causation. Factors may be more or less directly related to an outcome. The words, *influence* and *affect* are used in cases where other factors are present or where the effect of one particular factor is unclear, while *cause* is used in cases where there is a direct link between one factor and a particular outcome. All three words are relatively frequent in academic prose: Forms of *cause* (*cause, causes, caused, causing*) occurred with an average frequency of 477.28 tokens per million words, forms of *influence* (*influence, influences, influenced, influencing*) had a frequency of 344.28 tokens per million words, and forms of *affect* (*affect, affects, affected, affecting*) occurred with a frequency of 248.62 tokens per million words.

Influence and *affect* are often listed as synonyms and can be used interchangeably in many cases. There are many nouns that collocate with both verbs frequently, e.g. *behavior, people, way, health, development, performance, rate, life, outcome*, *function*. But there are also nouns that collocate primarily with *influence* or with *affect*, but tend not to collocate with both. *Policy* has a high frequency of collocation with *influence*, but rarely collocates with *affect*: There are 59 tokens of *policy* which occur after *influence*, but it does not collocate with *affect* in the academic register. In the 70 registers of the 100 million word corpus there are only 3 instances in which the collocation *affect* + *policy* occurs. There are several other nouns which frequently collocate with *influence* in the academic register, but do not collocate with *affect* in the academic register, and collocate with *affect* only rarely in the 100 million word BNC. These nouns include, e.g. *government, development, perceptions, groups, work, world, events, practice, health, public*, *life*. Then there are a few more which collocate with *influence* in the academic register, but never collocate with *affect* in any of the 70 registers of the BNC. These include, e.g. *process, factors, ideas, age, law, research, system*. Of course, the fact that they do not collocate with *affect* in the BNC does not mean that they could not; collocation patterns are best regarded as tendencies rather than exception-less rules.

There are a smaller number of nouns that collocate with *affect* but not with *influence*, e.g. *ability, validity, property, quality*. And there are a few that do not collocate with *affect* in the academic register, but do collocate, infrequently, in other registers, e.g. *results, management.* There are several adjectives that frequently precede *influence* when it is used as a noun. Adjective collocates from the academic register include, e.g. *undue, political, important, considerable, major, significant, strong, able, other.*

As for *cause*, in examining concordance lines one immediately notices one characteristic of its collocation pattern: the tendency to occur with expressions that have a negative meaning. In the following eight concordance lines (selected semi-randomly) from the academic register this tendency is amply illustrated:

> But he then goes on to remark that a further **cause** of inadequate response is the reader's unfamiliari
> case it's only a symptom, certainly not the root **cause** of his disease. (In Italian Fascism it showed up
> the hook —; for him the anti-Jewishness is symptom, not **cause**: There it is. It stops you. You feel him
> the London of the second decade of this century made common **cause** and thereafter —; despite ever
> he conceived of himself. However, his allegiance to the **cause** and principles of international modernis
> fella that if there's something to be done that might **cause** him paperwork, he'll do anything to avoid it
> The compassion and sympathy which the victims of these offences naturally **cause** in all policemen,
> of these programmes among the section police in Easton is both **cause** and effect of the evaluations th

Not all of the tokens of *cause* occur in negative contexts, but the tendency is evident. This tendency is referred to as negative prosody, and it is more evident when one looks at the nouns that occur after *cause*. Most, though not all, of the nouns that occurred after *cause* in the academic register have a negative sense as can easily be seen in Table 2 which shows nouns that occurred within 5 words after *cause.*

Table 2. Noun collocates of *cause*

Word	# of times nearby
Action	171
Death	133
Effect	122
Problems	78
Damage	70
Concern	56
Harm	50
Difficulties	39
Injury	38
Anaemia	34
Disease	33
Loss	28
Mortality	25
Difficulty	22
Failure	21
Matter	20
Problem	20
Patients	19
Person	18
Change	18

From the concordance lines and from the list of collocates, it is clear that *cause* is used especially in negative contexts. This may be considered part of its meaning although it is not something that is necessarily intuitively obvious, nor is it always mentioned in textbooks. It is however, an important point that a learner needs to be aware of in order to use this item appropriately. The case of *cause* and the contrast in its usage with *influence* and *affect* offers an example of the value of corpora for learners who are learning to do academic writing in English and for those who teach them.

6.0 Conclusion

This study outlines an approach to the investigation of keywords in academic prose. This is an issue of considerable importance to many learners of English who need to write academic texts. While writing textbooks provide instruction about grammatical matters, the area of phraseology is mostly untouched. This is an unfortunate omission since students who are attempting to do academic writing have usually mastered the main grammatical structures of English but still have some difficulty expressing their ideas in appropriate words. Through extensive reading in one's field one can gain a sense of which words are used for describing research in that field, but this is not the most direct nor efficient way of learning the lexical dimension of academic writing, and some students do not seem to acquire the vocabulary they need through this indirect, inductive method. It is therefore useful to make a study of the vocabulary used in academic writing with a view to identifying the most frequent words and then following this up with a study of the collocation patterns associated with those words. The existence and availability of standardized corpora such as the BNC make this a feasible project, but it has not yet been done on a large scale.

The present study can be seen as a pilot study in this direction. I have developed lists of the most frequent nouns, verbs and adjectives in the academic register of the BNC. These lists provide a starting point for more detailed investigation of the collocation patterns of individual lexical items. I have illustrated how such investigation of collocation patterns might be carried out. Study of items that collocate yield important information for a non-native academic writer. Analysis of *results* and *findings* revealed some notable differences in the frequency with which these items were used and of the words with which they collocated. It was shown, for instance, that *produce*, the verb which most frequently precedes *results*, does not collocate with *findings*. Similarly, the analysis of collocation patterns associated with three terms that are used for expressing causation, *influence, affect* and *cause*, showed that *cause* was used primarily in negative contexts, and in this way was very different from *influence* and *affect*. Again, this kind of information about negative prosody is of value to learners who do not have intuitive knowledge about a lexical item's typical contexts of use.

This study is limited in that the lists of research keywords that were developed were based on a rather general corpus of academic writing texts from many academic areas. One can expect some overlap in keywords across academic areas, especially for what McCarthy (1991) has called *discourse-organizing words*, that is, words such as *issue, problem* or *question*. However, texts from different academic fields will, of course, contain different keywords that are associated with those particular fields. Thus, it would be very useful to develop keyword lists from more specialized corpora of texts, and there is scope for much further research along these lines.

References

Aijmer, K. & B. Altenberg (eds.) (1991) *English corpus linguistics*. London: Longman.

Carter, R., R. Hughes & M. McCarthy (2000) *Exploring grammar in context*. Cambridge: Cambridge Univ. Press.

Eastwood, J. (1999) *Oxford practice grammar*. Oxford: Oxford Univ. Press.

Hunston, S. (2002) *Corpora in applied linguistics*. Cambridge: Cambridge Univ. Press.

McCarthy, M. (1991) *Discourse analysis for language teachers*. Cambridge: Cambridge Univ. Press.

McCarthy, M. & F. O'Dell (2001) *English vocabulary in use*. Cambridge: Cambridge Univ. Press.

McCarthy, M. & F. O'Dell (2005) *English collocations in use: How words work together for fluent and natural English*. Cambridge: Cambridge Univ. Press.

Murphy, R. (2000) *Grammar in use: intermediate*. Cambridge: Cambridge Univ. Press.

Redman, S. (1999) *Vocabulary in use: intermediate*. Cambridge: Cambridge Univ. Press.

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford Univ. Press.

Stern, G. (2000) *The grammar dictionary*. Greenwood, WA: R.I.C. Publications.

Swan, M. & C. Walter (1999) *How English works: a grammar practice book*. Oxford: Oxford Univ. Press.

Dictionaries

Collins CoBuild English dictionary (1995) London: Harper Collins.

Longman language activator (2nd ed.) (2002) Essex, U.K.: Pearson Longman.