DESCRIBING ENGLISH LANGUAGE

SERIES EDITORS

JOHN SINCLAIR · RONALD CARTER

Corpus, Concordance, Collocation

John Sinclair

Oxford University Press 1991

8
Collocation

Introduction

This chapter concludes the description of word co-occurrence as we currently conceive it. The next stage is to write a dictionary of collocations, and the project is in hand (Sinclair *et al.* forthcoming). The argument brings together a number of themes that have been developing throughout the book, in particular, the notions of dependent and independent meaning, and the relation of texts to grammar.

Two models of interpretation

It is contended here that in order to explain the way in which meaning arises from language text, we have to advance two different principles of interpretation. One is not enough. No single principle has been advanced which accounts for the evidence in a satisfactory way. The two principles are.

The open-choice principle

This is a way of seeing language text as the result of a very large number of complex choices. At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness.

This is probably the normal way of seeing and describing language. It is often called a 'slot-and-filler' model, evisaging texts as a series of slots which have to be filled from a lexicon which satisifies local restraints. At each slot, virtually any word can occur. Since language is believed to operate simultaneously on several levels, there is a very complex pattern of choices in progress at any moment, but the underlying principle is simple enough.

Any segmental approach to description which deals with progressive choices is of this type. Any tree structure shows it clearly: the nodes on

the tree are the choice points. Virtually all grammars are constructed on the open-choice principle.

The idiom principle

It is clear that words do not occur at random in a text, and that the open-choice principle does not provide for substantial enough restraints on consecutive choices. We would not produce normal text simply by operating the open-choice principle.

To some extent, the nature of the world around us is reflected in the organization of language and contributes to the unrandomness. Things which occur physically together have a stronger chance of being mentioned together; also concepts in the same philosophical area, and the results of exercising a number of organizing features such as contrasts or series. But even allowing for these, there are many ways of saying things, many choices within language that have little or nothing to do with the world outside.

There are sets of linguistic choices which come under the heading of register, and which can be seen as large-scale conditioning choices. Once a register choice is made, and these are normally social choices, then all the slot-by-slot choices are massively reduced in scope or even, in some cases, pre-empted.

Allowing for register as well, there is still far too much opportunity for choice in the model, and the principle of idiom is put forward to account for the restraints that are not captured by the open-choice model.

The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. To some extent, this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation. However it arises, it has been relegated to an inferior position in most current linguistics, because it does not fit the open-choice model.

At its simplest, the principle of idiom can be seen in the apparently simultaneous choice of two words, for example, *of course*. This phrase operates effectively as a single word, and the word space, which is structurally bogus, may disappear in time, as we see in *maybe*, *anyway*, and *another*.

Where there is no variation in the phrase, we are dealing with a fairly trivial mismatch between the writing system and the grammar. The *of*

in *of course* is not the preposition *of* that is found in grammar books. The preposition *of* is normally found after the noun head of a nominal group, or in a quantifier like *a pint of* ... . In an open-choice model, *of* can be followed by any nominal group (see Chapter 6 for details). Similarly, *course* is not the countable noun that dictionaries mention; its meaning is not a property of the word, but of the phrase. If it were a countable noun in the singular it would have to be preceded by a determiner to be grammatical, so it clearly is not.

It would be reasonable to add phrases like *of course* to the list of compounds, like *cupboard*, whose elements have lost their semantic identity, and make allowance for the intrusive word space. The same treatment could be given to hundreds of similar phrases—any occasion where one decision leads to more than one word in text. Idioms, proverbs, clichés, technical terms, jargon expressions, phrasal verbs, and the like could all be covered by a fairly simple statement.

However, the principle of idiom is far more pervasive and elusive than we have allowed so far. It has been noted by many writers on language, but its importance has been largely neglected. Some features of the idiom principle follow:

a. Many phrases have an indeterminate extent. As an example, consider *set eyes on*. This seems to attract a pronoun subject, and either *never* or a temporal conjunction like *the moment*, *the first time*, and the word *has* as an auxiliary to *set*. How much of this is integral to the phrase, and how much is in the nature of collocational attraction?

b. Many phrases allow internal lexical variation. For example, there seems to be little to choose between *in some cases* and *in some instances*; or between *set x on fire* and *set fire to x*.

c. Many phrases allow internal lexical syntactic variation. Consider the phrase *it's not in his nature to* ... . The word *it* is part of the phrase, and so is the verb *is*—though this verb can vary to *was* and perhaps can include modals. *Not* can be replaced by any 'broad' negative, including *hardly*, *scarcely*, etc. *In* is fixed, but *his* can be replaced by any possessive pronoun and perhaps by some names with *'s*. *Nature* is fixed.

d. Many phrases allow some some variation in word order. Continuing the last example, we can postulate *to recriminate is not in his nature*, or *it is not in the nature of an academic to* ... .

e. Many uses of words and phrases attract other words in strong collocation; for example, *hard work, hard luck, hard facts, hard evidence.*

f. Many uses of words and phrases show a tendency to co-occur with certain grammatical choices. For example, it was pointed out in Chapter 5 that the phrasal verb *set about*, in its meaning of something like 'inaugurate', is closely associated with a following verb in the *-ing* form, for example, *set about leaving ...* . What is more, the second verb is usually transitive, for example, *set about testing it.* Very often, *set* will be found in co-occurrence patterns.

g. Many uses of words and phrases show a tendency to occur in a certain semantic environment. For example, the verb *happen* is associated with unpleasant things—accidents and the like.

The overwhelming nature of this evidence leads us to elevate the principle of idiom from being a rather minor feature, compared with grammar, to being at least as important as grammar in the explanation of how meaning arises in text. Support comes unexpectedly from a different quarter.

Evidence from long texts

In the current lexical analysis of long texts, a number of problems have arisen, not all of which were anticipated:

1 The 'meanings' of very frequent, so-called grammatical words are a headache in any lexicography, but the problem they typify fits in with some of the newer difficulties.

2 Some 'meanings' of frequent words seem to have very little meaning at all, for example, *take*, in *take a look at this*; *make* in *make up your mind.*

3 The commonest meanings of the commonest words are not the meanings supplied by introspection; for example, the meaning of *back* as 'the posterior part of the human body, extending from the neck to the pelvis' (*Collins English Dictionary* (*CED*) 2nd edition 1986 sense 1) is not a very common meaning. Not until sense 47, the second adverbial sense, do we come to 'in, to or towards the original starting point, place or condition', which is closer to the commonest usage in our evidence.

I think most speakers of English would agree with the *CED*'s ordering of senses, whatever the evidence from frequency. What is disquieting is the apparent lack of good reason for the enormous discrepancy

between the sense to which our intuitions give priority, and the most frequent one.

4 The commonest meanings of many less common words are not those supplied by introspection. Sense 1 offered in the *CED* for *pursue* is 'to follow (a fugitive etc.) in order to capture or overtake', yet by far the commonest meaning is sense 5, 'to apply oneself to (one's studies, hobbies, interests etc.)'.

From this evidence, we can put forward some tentative generalizations:

1 There is a broad general tendency for frequent words, or frequent senses of words, to have less of a clear and independent meaning than less frequent words or senses. These meanings of frequent words are difficult to identify and explain; and, with the very frequent words, we are reduced to talking about uses rather than meanings. The tendency can be seen as a progressive delexicalization, or reduction of the distinctive contribution made by that word to the meaning.

2 This dependency of meaning correlates with the operation of the idiom principle to make fewer and larger choices. The evidence of collocation supports the point. If the words collocate significantly, then to the extent of that significance, their presence is the result of a single choice.

3 The 'core' meaning of a word—the one that first comes to mind for most people—will not normally be a delexical one. A likely hypothesis is that the 'core' meaning is the most frequent independent sense. This hypothesis would have to be extensively tested, but if it proved to hold good then it would help to explain the discrepancy referred to above between the most frequent sense and what intuition suggests is the most important or central one.

4 Most normal text is made up of the occurrence of frequent words, and the frequent senses of less frequent words. Hence, normal text is largely delexicalized, and appears to be formed by exercise of the idiom principle, with occasional switching to the open-choice principle.

5 Just as it is misleading and unrevealing to subject *of course* to grammatical analysis, it is also unhelpful to attempt to analyse grammatically any portion of text which appears to be constructed on the idiom principle.

The last point contains an implication that a description must indicate how users know which way to interpret each portion of an utterance. The boundaries between stretches constructed on different principles will not normally be clear-cut, and not all stretches carry as much evidence as *of course* does to suggest that it is not constructed by the normal rules of grammar.

It should be recognized that the two models of language that are in use are incompatible with each other. There is no shading of one into another; the switch from one model to the other will be sharp. The models are diametrically opposed.

The last two points taken together suggest one reason why language text is often indeterminate in its interpretation and hence very flexible in use. If the 'switch points' between two modes of intepretation are not always explicitly signalled, and the two modes offer sharply contrasting ways of interpreting the data, then it is quite likely that an utterance will not be interpreted in exactly the same way in which it was constructed. Also, two listeners, or two readers, will not interpret in precisely the same way.

For normal texts, we can put forward the proposal that the first mode to be applied is the idiom principle, since most of the text will be interpretable by this principle. Whenever there is good reason, the interpretive process switches to the open-choice principle, and quickly back again. Lexical choices which are unexpected in their environment will presumably occasion a switch; choices which, if grammatically interpreted, would be unusual are an affirmation of the operation of the idiom principle.

Some texts may be composed in a tradition which makes greater than normal use of the open-choice principle; legal statements, for example. Some poems may contrast the two principles of interpretation. But these are specialized genres that require additional practice in understanding.

It thus appears that a model of language which divides grammar and lexis, and which uses the grammar to provide a string of lexical choice points, is a secondary model. It cannot be relinquished, because a text still has many switch points where the open-choice model will come into play. It has an abstract relevance, in the sense that much of the text shows a potential for being analysed as the result of open choices, but the other principle, the idiom principle, dominates. The open-choice analysis could be imagined as an analytical process which goes on in principle all the time, but whose results are only intermittently called for.

This view of how the two principles are deployed in interpretation can be used to make predictions about the way people behave, and the accuracy of the predictions can be used as a measure of the accuracy of the model. Areas of relevant study include: the transitional probabilities of words; the prevalent notion of *chunking* (see Chapter 9); the occurrence of hesitations, etc., and the placement of boundaries; and the behaviour of subjects trying to guess the next word in a mystery text.

Collocation

The above is the framework within which I would like to consider the role of collocation. Collocation, as has been mentioned, illustrates the idiom principle. On some occasions, words appear to be chosen in pairs or groups and these are not necessarily adjacent.

One aspect of collocation has been of enduring interest. When two words of different frequencies collocate significantly, the collocation has a different value in the description of each of the two words. If word *a* is twice as frequent as word *b*, then each time they occur together is twice as important for *b* than it is for *a*.This is because that particular event accounts for twice the proportion of the occurrence of *b* than of *a*.

So when all the occurrences of *a* with *b* are counted up and evaluated, one figure is recorded in the profile of *a*, and another figure double the size, is recorded in the profile of *b*.

By entering the same set of events twice, once as the collocation of *a* with *b* and again as the collocation of *b* with *a*, one incurs the strictures of Benson, Brainerd, and Greaves (1985) who say 'there are two problems here: double counting of nodes and double counting of collocates. The parts now add up to considerably more than the whole, which makes computation under any statistical model inaccurate'. In practice, the possibility of double entry allows us to highlight two different aspects of collocation.

I would like to consider separately the two types of collocation instanced above, using the term *node* for the word that is being studied, and the term *collocate* for any word that occurs in the specified environment of a node. Each successive word in a text is thus both node and collocate, though never at the same time.

When *a* is node and *b* is collocate, I shall call this *downward collocation*—collocation of *a* with a less frequent word (*b*). When *b* is

node and *a* is collocate, I shall call this *upward collocation*. The whole of a given word list may be treated in this way.

There appears to be a systematic difference between upward and downward collocation. Upward collocation, of course, is the weaker pattern in statistical terms, and the words tend to be elements of grammatical frames, or superordinates. Downward collocation by contrast gives us a semantic analysis of a word.

Collocation of *back*

Let us illustrate collocational patterns, in a provisional way, with the word *back*. I shall make no attempt to differentiate separate senses, but will put the collocates into *ad hoc* groups.

No standard of statistical significance is claimed at present, because many typical collocations are of such low frequency compared with the overall length of a text. Because of the low frequency of the vast majority of words, almost any repeated collocation is a most unlikely event, but because the set of texts is so large, unlikely events of this kind may still be the result of chance factors.

However, no speaker of English would doubt the importance of these patterns. One recognizes them immediately, because they are features of the organization of texts; often subliminal, they cannot be reliably retrieved by introspection.

In distinguishing upward and downward collocation I have made a buffer area of (plus or minus) 15 per cent of the frequency of the node word. For example, let us take a word occurring 1,000 times; when it is examined as a node, collocates are grouped into:

a. upward collocates—those whose own occurrence is over 115 per cent of the node frequency (that is, 1,150);
b. neutral collocates—between 85 per cent and 115 per cent of the node frequency (in this instance, 850 and 1,150), this is the buffer area;
c. downward collocates—less than 80 per cent (in this instance, 850).

Neutral collocates are added on an *ad hoc* basis to upward or downward groups, and are given round brackets. Since this has to be a summary account of a very large set of data, I have removed some items which seem to be of little general significance. These include personal names, contracted forms like *I'll*, and word-forms whose

co-occurrence with *back* is infrequent and carries no conviction of any general significance. Of the last category, the form *anger* only occurs in the title of the play *Look Back in Anger*.

The nouns and verbs listed below as collocating with *back* are representative only. Given the uncertainty at the limits of statistical significance, it could be more misleading to include doubtful contenders. Thus, while *get, go*, and *bring* are unlikely to be challenged, *beach, box*, and *bus* are much less convincing when the actual instances are examined.

The qualification for an instance being scrutinized is co-occurrence within four words of *back*, on either side, this being the cut-off point established some years ago (Jones and Sinclair 1974). No account is taken of syntax, punctuation, change of speaker, or anything other than the word-forms themselves.

No doubt the studies which succeed this one will sharpen up the picture considerably. For example, the evidence of *back* suggests that few intuitively interesting collocations cross a punctuation mark. But it would be unwise to generalize from the pattern of one word, particularly such an unusual one as *back*. Now that tagged and parsed texts are becoming available, the co-patterning of lexical and grammatical choices is open to research. But it is still important to draw attention to the strength of patterning which emerges from the rawest of unprocessed data.

In pushing forward into new kinds of observation of language, the computer is simultaneously pulling us back to some very basic facts that are often ignored in linguistics. The set of four choices, a,b,c,k, from the alphabet, arranged in the sequence b,a,c,k with nothing in between them, that is, *back*, is an important linguistic event in its own right, long before it is ascribed a word-class or a meaning. It is difficult for users of English to notice this, but it is the computer's starting point.

Analysis of the collocational pattern of *back*

Upward collocates: back

Prepositions/adverbs/conjunctions: *at, (down), from, into, now, on, then, to, up, when*

Pronouns: *her, him, me, she, them, we*

Possessive pronouns: *her, his, my, (your)*

Verb: *get, (go), got*

The meaning of *back* as 'return' attracts expressions of time and place; *after* and *where* are also prominent. The presence of four subject pronouns may have a more general explanation than anything to do with *back*, but the absence of *you* and *I* from the list may be worth pursuing. Possessive pronouns suggest the anatomical sense of *back* and would explain why *they* and *their* do not figure prominently. The two verbs *get* and *go* are superordinates of a large number of verbs of motion, many of which will be found in the downward collocates.

I have selected a few examples of these words to show the way in which the basic syntax of *back* is established. The sets of examples follow the four categories mentioned above:

It really was like being back *at* school
He drive back *down* to the terrace
When our parents came back *from* Paris
I followed him back *into* the wood
A hefty slap *on* the back

He turned back to the bookshelf
When can I have *him* back home, doctor?
She went back to her typing
It would be nice to have *them* back
We went back to the bungalow

She has gone back to *her* parents
He want back into *his* office
I ran back to *my* cabin
Go back to *your* dormitory at once

Now I must *get* back to work
They *go* back to the same nest

Downward collocates: back

Verbs: *arrive, bring, etc., climbed, come, etc., cut, etc., dates, etc., drew, etc., drove, etc., fall, etc., flew, flung, handed, hold, etc., jerked, lay, etc., leaned, etc., looked, looking, etc., pay, pulled, etc., pushed, etc., put, ran, rocking, rolled, rush, sank, sat, etc., sent, etc., shouted, snapped, stared, stepped, steps, etc., stood, threw, traced, turned, etc., walked, etc., waved.*

Prepositions: *along, behind, onto, past, toward, towards*

Adverbs: *again, forth, further, slowly, straight*

Adjective: *normal*

Nouns: *camp, flat, garden, home, hotel, office, road, streets, village, yard*
bed, chair, couch, door, sofa, wall, window, feet, forehead, hair, hand, head, neck, shoulder, car, seat
mind, sleep,
kitchen, living room, porch, room.

The word-class groupings above are based on frequency with *back*; many words actually occur in more than one word-class. Verbs are given in their most frequent form. Note the preponderance of past tense verbs, reflecting the temporal meaning of *back*.

The prepositions and adverbs suggest some typical phrases with *back*, and the nouns are largely those of direction, physical space, and human anatomy. A few typical examples follow:

Verbs:
You *arrive* back on the Thursday
May *bring* it back into fashion
We *climbed* back up on the stepladder
They had *come* back to England
She never *cut* back on flowers
It possibly *dates* back to the war
The bearer *drew* back in fear
We *drove* back to Cambridge
You can *fall* back on something definite
I *flew* back home in a light aircraft
He *flung* back the drapes joyously
Don't try to *hold* her back
She *lay* back in the darkness
He *leaned* back in his chair
He *looked* back at her, and their eyes met
Pay me back for all you took from me
Pulled back the bedclothes and climbed into bed
I *pushed* back my chair and made to rise
Shall I *put* it back in the box for you
I *rolled* back onto the grass
She *sat* back and crossed her legs
Edward was *sent* back to school
He *shouted* back
The girl *stared* back
They *started* walking back to Fifth Avenue

He *stepped* back and said ...
He then *stood* back for a minute
The woman *threw* her head back
These could be *traced* back to the early sixties
He *turned* back to the bookshelf
She *walked* back to the bus stop
We *waved* back like anything

Prepositions: Hands held *behind* his back
Walked back *toward* the house

Adverbs: Later we came back *again*
Rock us gently back and *forth*
If you look *further* back in my files
The *straight* back to his cabin
He went *slowly* back to his book

Adjective: Things would soon get back to *normal*

Nouns: I crawled back to *camp*
I'll drive you back to your *flat*
Not a bit like his back *garden*
He turned and went back *home*
We had to go back to the *hotel*
You've just got back from the *office*
Set back from the *road*
The back *streets* of Glasgow
All the way back to the *village*
On his *way* back to the apartment
Without even a back *yard*
Go back to *bed*
He leaned back in his *chair*
Stepping outside the back *door*
A man standing by the back *wall*
Tom went back to the *window*
Britain would be back on its *feet*
He brushed back his *hair*
With the back of his *hand*
She put her *head* back against the seat
The hairs on the back of my *neck*
He gestured back over his *shoulder*
They got back into the *car*

There was some beer on the back *seat*
In the back of his *mind*
Then we go back to *sleep* again
You must come back to the *kitchen*
She went back into the *living room*
Beside me here on the back *porch*
He came back into the *room*

Conclusion

All the evidence points to an underlying rigidity of phraseology, despite a rich superficial variation. Hardly any collocates occur more than once in more than two patterns. The phraseology is frequently discriminatory in terms of sense; for example, there are almost as many instances of *flat on her back* as *back to her flat*. Some, like *arrive*, seem characteristic of the spoken language, some, like *hotel*, show the wisdom of allowing a nine-word span for collocation.

Early predictions of lexical structure were suitably cautious; there was no reason to believe that the patterns of lexis should map on to semantic structures. For one thing, lexis was syntagmatic and semantics was paradigmatic; for another, lexis was limited to evidence of physical co-occurrence, whereas semantics was intuitive and associative.

The early results given here are characteristic of present evidence; there is a great deal of overlap with semantics, and very little reason to posit an independent semantics for the purpose of text description.