A NOTE ON PHRASEOLOGICAL TENDENCIES IN THE CORE VOCABULARY OF ENGLISH

MICHAEL STUBBS

University of Trier

Recent work, particularly in various areas of applied linguistics, has emphasized the pervasive occurrence of phrase-like units of idiomatic language use. Pawley and Syder (1983) claimed that native speakers know "hundreds of thousands" of such multi-word units. Such complex lexical units are now well documented in corpus-based dictionaries, teaching materials, and theoretical work, by Willis (1990), Sinclair (1991), Nattinger and DeCarrico (1992), and others, and the area is well reviewed by Cowie (1994), Weinert (1995) and Howarth (1998). Nevertheless, we still lack an entirely convincing account of the constituency and structure of such units.

This paper provides a preliminary analysis of the collocations of a sample of 1,000 word-forms from the core vocabulary of English. I discuss ways of modelling extended lexical units, including the relations within these units, and the strength of node-collocate attraction. Such an analysis has implications for a theory of language *production* which takes account of the idiomatic nature of much language use, a theory of language *comprehension* which takes account of the contribution of collocational units to textual cohesion, and a theory of language *per se* which takes account of the variable and probabilistic nature of such language units.

1. The data-base

The publication of *Collins COBUILD English Collocations on CD-ROM* (Cobuild 1995a) provides a huge data-base for investigating the collocational behaviour of the core vocabulary of English. The data-base was created in the following way. In a 200-million word (tokens) corpus of contemporary English,

the 10,000 most frequent word-forms (types) were identified. These 10,000 word-forms provide the head-words in the data-base. For each head-word, the 20 most frequent collocates were identified, in a span of 4 word-forms to left and right. And for each of these collocates, 20 concordance lines were selected at random. This comprises a very large amount of data indeed: 10,000 word-forms times 20 collocates times 20 concordance lines, or approximately 4 million concordance lines, which illustrate frequent uses of frequent words.

So, for each word in the head-word list, the user can see: its frequency in the corpus; a list of its 20 most frequent collocates, with the joint head-word/collocate frequency; a selection of concordance lines (the head-word or collocate in an 80-character context); and a slightly larger context with a broad indication of text type (such as British financial journalism or American radio broadcasts).

(To be strictly accurate, there is slightly less data than the 4-million concordance lines. A maximum of 20 collocates for each head-word is given: down to a cut-off point of 15 joint head-word/collocate occurrences. And, since the concordance lines are chosen at random for each collocate, the same line is sometimes chosen twice: so there might be somewhat fewer that 400 (20 times 20) lines for a head-word.)

I have studied a sample of 1,000 of the headwords with their associated collocates and concordance lines: I simply took every tenth word alphabetically.

2. Conventions

I use the term collocation to mean habitual co-occurrence. I will talk of collocates co-occurring with a node within a span of word-forms to left and right. All the data below are from a span of 4:4. I will show lemmas in upper case, and corresponding word-forms in lower case: for example, UNDERGO is realised by the word-forms *undergo*, *undergoes*, *undergoing*, *undergone*, *underwent*. Frequent collocates of a node are shown in diamond brackets. And where relevant, the absolute frequency of the node in the corpus, and the frequency of individual collocates or sets of collocates as a percentage of the node, are shown, for example:

node 1,999 <top collocate 10%, other collocates ...> 20%

Stubbs (1995a) and Barnbrook (1996) describe statistical techniques for measuring lexical attraction between nodes and collocates.

3. Lemmas and word-forms

A major finding of corpus linguistics is that the size of lexical units may be either smaller or larger than has often been assumed in traditional lexical description. In dictionaries (even those based on corpora), most head-words are

lemmas. For example, the head-word SEEK, will treat together the word-forms *seek*, *seeks*, *seeking* and *sought*. Corpus data show however that these different word-forms occur in different text types and have quite different sets of collocates. The word-form *seeks* occurs almost exclusively in lonely hearts ads, such as

(1) female 31, single, *seeks* well educated gentleman

and its most frequent collocates include

(2) seeks <female, male, attractive, caring, professional>

The word-form *seeking* can occur in such texts, but its most frequent collocates show quite different uses:

(3) seeking <asylum, help, advice, support>

So, it is clear that the unit of use and of meaning may be smaller than the lemma (strictly speaking, one of the members of the class lemma). Equally clearly, the unit of language use often comprises a string of lemmas and/or word-forms in a complex lexical unit, and this is now my main topic.

4. Extended lexical units

There are different relations between lemmas and word-forms in recurring extended lexical units. Sinclair (1996) discusses several aspects of these relations and much of my paper develops points which Sinclair makes.

The classic definition of collocation as the company that words keep is from Firth (1957: 11, 14): "quite simply the mere word accompaniment", "actual words in habitual company". In this strict sense, collocation refers to the relation between individual word-forms, which is illustrated in cases such as

(4) applause <loud, thunderous, rapturous, spontaneous, polite, warm, enthusiastic>

In this strict sense, the relation of collocation makes no reference to grammar, though if we make the description of this example a little more delicate, we could add that an ADJ-NOUN (*thunderous applause*) combination is more frequent than a DET-NOUN-BE-ADJ (*the applause was thunderous*) construction.

The collocates listed above for *applause* are all adjectives, though the most frequent collocates also include nouns and verbs: as in *round of applause*, *applause and cheers*, *greeted with applause*. Sometimes a word-form most frequently co-occurs with words from a particular grammatical category, such as quantifiers:

(5) cases <some, many, most, more, both, several>

This collocation is due to frequent phrases such as *in some cases*, *in many cases*, etc. Lower down the list come other collocates which show other meanings (e.g., *court cases*).

A third type of relation is between a word-form and a lexical set, in the sense of closely semantically related words from a well-defined semantic field. For example, the word-form *large* occurs most frequently (in over 20 per cent of cases) with words for "quantities and sizes":

(6) large <number(s), scale, part, amounts, quantities, area(s)>

A fourth type of example shows a relation which expresses the speaker's attitude to what is being talked about. For example, if a speaker talks of something being *provided*, the implication is that the speaker approves of it: "good things" are provided.

(7) PROVIDE <information, service(s), support, help, money, protection, food, care>

In another work (Stubbs 1995a), I discuss such "semantic prosodies" (Sinclair 1991; Louw 1993) in more detail, with reference to the lemma CAUSE, which occurs overwhelmingly with unpleasant words:

(8) CAUSE <problem(s), damage, death(s), disease, concern, cancer, pain>

We are dealing, therefore, not with lists of specific collocations, but with more abstract units whose constituents stand in four types of relation to each other: lexical (purely word-word relations), grammatical (relations between words and grammatical categories), semantic (relations between words and lexical sets), and pragmatic (relations between words and speaker attitude). A more formalized model, which makes these relations more explicit, would show that lexis is by no means idiosyncratic, but can be brought fully within the kinds of relations more traditionally associated with the study of syntax.

5. Corpus representativensess and corpus methods

In any such corpus-based study, it is important to pay attention to the representativeness of the data. The 200-million word corpus comprised a wide selection of texts and text-types: British and American English, spoken and written, fiction and non-fiction, and so on. However, some 65 per cent of the running text for the data-base was taken from the mass media, both printed newspapers and broadcast language. And this bias certainly shows up in the frequency of some words in the head-word list and in frequent phrases, such as *ethnic cleansing*, *issued a communique*, *crippled the economy*, *composite trading*.

It might well be that such words would not be appropriately listed as part of the core vocabulary. And the data-base could not be taken as entirely representative of average language use. However, first, in one sense there is no such thing as average language use: any given instance of language use occurs in a specific text-type and a specific social setting. Second, usage in the mass-media is very influential. And third – and crucial for my argument here – whilst the frequency of some head-words and specific collocations is biassed towards mass-media topics (politics, war, economics, and so on), it is unlikely that the general collocational phenomena which I discuss (such as strength of collocational attraction) are affected by such topical bias.

It is important also to comment on the limitations of purely automatic methods. A recurrent misunderstanding is that "computer-assisted corpus-based methods" means that everything is left to the machine. But this cannot be done. For example, the top 20 collocates of *somewhat* are as follows:

(9) somewhat <more, different, less, also, though, still, although, seems, similar, since, better, seemed, become, feel, higher, view, seem, later, however, perhaps>

If the analysis was restricted to this list of collocates, then the impression would be that *somewhat* is a word used in the context of comparisons (more, different, etc.) and often when the speaker is making some concession to a different point of view (though, still, etc.). However, a glance at the concordance lines shows the frequency of examples such as

(10) somewhat cruelly; somewhat more awkward and conspicuous; somewhat negatively; somewhat on the defensive.

Such examples reveal a negative semantic prosody: *somewhat* is also frequently used to mitigate a critical reference to someone. But the individual word-forms which realise this prosody are too diverse to appear in the list of the top 20 collocates. I have analysed these uses in more detail elsewhere (Stubbs 1995b).

Corpora and associated data-bases provide systematic information, but this information must be carefully interpreted, and sometimes the human being can spot a pattern which is invisible to the computer. The interpretations do not arise automatically from the data; or if you like, there is no such thing as the pure induction of significant patterns from data.

6. Strength of attraction

The data-base provides the necessary information to calculate the strength of attraction between word-pairs, for example between a given head-word and its

top collocate. In the following cases, given the head-word, there is a very high probability (over 20 per cent) that a specific single top collocate will occur:

(11) breakaway 1,379 <republic 24%>; deadlock 1,236 <break 29%>

Such a strong attraction (over 20 per cent) is found with over one head-word in twenty. In the following cases, given the head-word, there is a slightly lower probability of the occurrence of a single top collocate:

(12) angrily 1,388 <reacted 18%>; cheering 1,226 <crowd 13%>

Such an attraction (between 10 and 20 per cent) is found with over one head-word in five. In the following cases, given the head-word, there is lower probability again of the occurrence of a single top collocate:

(13) alarming 1,711 <rate 8%>; applause 2,207 <round 6%>

Such an attraction (between 5 and 10 per cent) is found in over four cases out of ten of the head-words.

There are very few cases indeed, where a head-word predicts a single top collocate with less than one chance in fifty. And even in such cases, the collocation may nevertheless be a familiar one. One such example is

(14) continental 4,085 <breakfast 1.9%>

but of course *continental breakfast* is a well-known phrase!

This very simple calculation, between pairs of individual word-forms, shows a strong tendency for given words to co-occur with other individual words. However, a calculation which is based on individual word-pairs greatly underestimates the strength of such collocational attraction. If we lemmatize the collocates and also look at lexical sets, then the power of attraction of the head-words is seen to be much stronger again. Thus, compare the figures above on *breakaway*, *deadlock* and *cheering* with the following:

(15) breakaway 1,379 <republic(s) 35%, group, faction, party> 45%
(16) deadlock 1,236 <BREAK 41%, END, resolve> 50%
(17) cheering 1,226 <crowd(s) 20%, people, supporters, fans, audience> 30%

Here, with *deadlock* for example, there is a fifty-fifty chance that it co-occurs with one of only three verbs which are close synonyms of each other. A major finding from such work is that underlying semantic patterns are very clear, although the surface lexical variation may be considerable.

7. Other patterns of attraction

There are other striking patterns of co-occurrence in the data. For example, there is a well above chance probability that antonyms co-occur:

(18) answers 6,166 <question(s) 19%>; births 1,010 <deaths 10%>.

Statistics on the probabilities of such co-occurrence (calculated on data of a quite independent kind) are discussed in detail by Justeson and Katz (1991). There is also a tendency for words to co-occur with approximate synonyms and with co-hyponyms, for example:

(19) anarchy 1,057 <chaos 4%>
(20) anxiety 4,961 <depression 5%, fear, stress, tension> 12%
(21) bronze 3,229 <silver, gold> 11%
(22) café 6,969 <bar, restaurant> 7%

These co-occurrence patterns contribute to a previously rarely mentioned mechanism of textual cohesion (Sinclair 1992; Bublitz 1996; Stubbs forthcoming). Related phenomena, called variously feature-sharing or feature-copying, include frequent phrases such as:

(23) physical assault, full circle, scientific experiment, heavy load, completely forgot

or the following examples from the data-base:

(24) addition 12,306 <new 2%>; bonus 3,459 <added 6%, extra 2%>.

8. Idiomaticity versus idioms and fixed phrases

It is important for my general argument to emphasize that I am discussing idiomatic language, but that the examples I have given are neither fixed phrases nor idioms. That is, there are strong expected patterns of co-occurrence, but the expectations are probabilistic, and only very rarely deterministic. And in most cases, the extended lexical units are semantically transparent.

The methods certainly do also identify fixed phrasal units which might be separately listed in dictionaries. The Cobuild dictionary (Cobuild 1995b) lists both *awe-inspiring* and *pit bull terrier* as head-words, and the data-base produced

(25) awe 1,128 <inspiring 15%>
(26) bull 3,829 <pit 8%, terrier(s) 7%>

The method also identifies phrases which occur in discussion on particular topics, such as (*potential*) *first time* (*home*) *buyers*:

(27) buyers 5,788 <first 11%, time 11%, home 6%, potential 5%>

And it finds phrases with technical meanings in particular text-types, such as legal language: *charged with assault causing grievous bodily harm* ("GBH") or *intent to cause actual bodily harm*:

(28) bodily 1,303 <harm 38%, grievous 19%, causing 14%, actual 9%, assault 6%, intent 5%, charged 3%>

A few such complex lexical units are genuine idioms, such as *no strings attached*:

(29) attached 5,054 <no 4%, strings 3%>

But such cases are only the very tip of the iceberg. The majority of cases are neither fixed phrases nor idioms. They show the very widespread tendency of the most frequent words in the language to occur in predictable phraseology.

9. A final example

A final example of a single lemma will show important related features of extended lexical units: that the lexical attraction of a lemma or word-form can stretch to several words to the left and right, can involve several types of relation, and that the boundaries of the lexical unit need not correspond to syntactic boundaries. Amongst frequent collocates of UNDERGO are

(30) surgery, treatment, training, medical, operation, heart; examination, test(s), testing, trials; change(s), transformation; considerable, extensive, further, major, radical, traumatic; difficulties, ordeal, pains; forced, required

The lemma occurs in uses such as:

(31) forced to undergo extensive surgery
(32) had to undergo a stringent medical examination
(33) the traumatic experience undergone by ...
(34) now undergoing rapid change
(35) underwent a radical transformation
(36) has undergone several difficulties

The basic semantic pattern and the strong semantic prosody are clear. People involuntarily undergo unpleasant procedures, often medical. People and things undergo changes, often radical, and not always pleasant. The "involuntary" prosody is usually expressed to the left, either by a verb in the passive (such as *forced*) or by a modal verb. The "unpleasant" prosody is usually expressed to

the right, and usually in words from a small number of lexical fields (medicine, testing, change) and often in adjectives which convey the seriousness and/or length of the procedure (such as *considerable*, *stringent*). However, although the semantic pattern is very simple, and although the most frequent occurrences use a small selection of predictable vocabulary, the lexical realizations are open-ended. Thus in an example such as

(37) a prisoner undergoing a savage sentence for a crime

the same "involuntary" and "unpleasant" semantic prosody is observable, but the vocabulary is not predictable and the lexical items do not occur amongst the most frequent collocates.

I should note also that collocational patterns differ in different text types. The lemma UNDERGO is relatively formal: in casual conversation people usually *have an operation*. And in scientific and technical English UNDERGO has no "unpleasant" prosody: for example

(38) electrolytes undergo chemical decomposition

10. Implications and work in progress

The present paper is a brief outline of work in progress which will be reported in more detail elsewhere (e.g., Stubbs forthcoming). This work will formalise the model of extended lexical units (with lexical, grammatical, semantic and pragmatic relations between constituents), and will illustrate the possible form of dictionary entries (Stubbs 1995a, 1995b).

Here, in conclusion, I will note some implications of such analysis. Corpus work has questioned two widespread assumptions of much recent linguistics. One is the argument that grammar deals with general laws, whereas lexis deals only with isolated and idiosyncratic facts: this view has been widespread since it was formulated by Bloomfield (1933: 274). The other is the argument that statistical preferences should be ignored in modelling language: this view derives in particular from Chomsky (1957: 16), and since then, neither variability nor probability have been central concepts in linguistics (unless in certain approaches to sociolinguistic variation), and only isolated voices have argued that "relative frequency in text might have any theoretical significance" (Halliday 1991: 30).

The kind of corpus findings discussed above suggests the need for a model which can represent the balance of variation and norm in language use. The probabilistic structure of collocations is a completely general phenomenon, and such a probabilistic model would have implications for several areas.

- *A theory of connotation.* If all words have predictable collocates, then lexical meaning is distributed over several words, and the balance between propositional and connotational meaning shifts towards the latter. Channell (in press) has claimed that many more words have evaluative connotations than previously recognised. In turn, this has implications for the practice and theory of stylistics and of translating.
- *A theory of textual cohesion.* Findings about the co-occurrence of antonyms and synonyms and about semantic feature-sharing are findings about the frequent semantic redundancy of language in use. A theory of lexical collocation is a theory of the extent to which lexis is predictable, and can therefore serve as one mechanism of textual cohesion (Bublitz 1996; Stubbs forthcoming).
- *A theory of competence.* Such findings also begin to solve the puzzle of why some ways of saying things sound idiomatic, natural and native-like, whilst others just don't (Pawley – Syder 1983; Howarth 1998). The concept of mental lexicon has to include the very large number of predictable combinations which native speakers know: not only specific lexical collocations, but also more abstract semantic patterns. This in turn has implications for language teaching (Willis 1990), and for understanding the balance between routine and creativity in language use.
- *A theory of performance.* It turns out that language use is much more highly organized than often suspected. It appears to be governed, not only by the kinds of rules that linguistics has traditionally dealt with, but by tendencies which can be expressed as (often high) levels of probability. Without large corpora and appropriate software, this was much more difficult to observe, although some scholars, such as Bolinger (1976) repeatedly insisted on the importance of such patterns.
- *A theory of lexis-grammar relations.* Such a model gives lexis a much more central place in language use, and certainly the lexicon acquires a greatly expanded syntagmatic dimension (Sinclair 1991: 65, 137; De Beaugrande 1996). This is a very different orientation from much linguistic theory since the 1950s, and it will take a long time before the appropriately revised division of labour between an expanded lexis and a reduced syntax has been worked out.
- *A theory of natural language use.* Concepts such as "current" or "natural" language use have intuitive plausibility, but are not well understood (Sinclair 1991: 174). A closely related concept (taken for granted in much language teaching) is that a language has a "core" vocabulary and grammar. The kind of evidence I have presented could be used to develop such a concept. I have been discussing the central uses of the most frequent words in English: the core collocations of the core vocabulary.

ACKNOWLEDGEMENTS

For help in extracting the data from the Cobuild data-base, I am very grateful to Kirsten Günther. For comments on an earlier draft, I am very grateful to Gabi Keck and John Sinclair, and to seminar audiences in Bonn, Fribourg and Luxembourg.

REFERENCES

Aijmer, K. – B. Altenberg (eds.)
1991 *English corpus linguistics.* London: Longman.
Asher, R. E. (ed.)
1994 *The encyclopedia of language and linguistics.* Oxford: Pergamon.
Baker, M. – G. Francis – E. Tognini-Bonelli (eds.)
1993 *Text and technology.* Amsterdam: Benjamins.
Barnbrook, G.
1996 *Language and computers.* Edinburgh: Edinburgh University Press.
Bloomfield, L.
1933 *Language.* New York: Holt, Rinehart & Winston. (Page references to British edition, London: Allen & Unwin, 1935.)
Bolinger, D.
1976 "Meaning and memory", *Forum Linguisticum* 1, 1: 1-14.
Bublitz, W.
1996 "Semantic prosody and cohesive company: 'Somewhat predictable'", *Leuvense Bijdragen: Tijdschrift voor Germaanse Filologie* 85, 1-2: 1-32.
Channell, J.
in press "The analysis of evaluative lexis", in: S. Hunston – G. Thompson (eds.).
Chomsky, N.
1957 *Syntactic structures.* The Hague: Mouton.
Cobuild
1995a *Collins COBUILD English Collocations on CD-ROM.* London: HarperCollins.
1995b *Collins COBUILD English Dictionary.* London: HarperCollins.
Cook, G. – B. Seidlhofer (eds.)
1995 *Principle and practice in applied linguistics.* Oxford: OUP.
Cowie, A. P.
1994 "Phraseology", in: R. E. Asher (ed.), 3168-3171.
Davies, M. – L. Ravelli (eds.)
1992 *Advances in systemic linguistics.* London: Pinter.
De Beaugrande, R.
1996 "The 'pragmatics' of doing language science: the 'warrant' for large-corpus linguistics", *Journal of Pragmatics* 25: 503-35.
Firth, J. R.
1957 "A synopsis of linguistic theory, 1930-55", in: *Studies in Linguistic Analysis.* Special volume of the Philological Society, Oxford, 1-31.
Halliday, M. A. K.
1991 "Corpus studies and probabilistic grammar", in: K. Aijmer – B. Altenberg (eds.), 30-43.
Howarth, P.
1998 "Phraseology and second language proficiency", *Applied Linguistics* 19, 1: 24-44.

Hunston, S. – G. Thompson (eds.)
in press *Language in evaluation*. Oxford: OUP.
Justeson, J. S. – S. M. Katz
1991 "Redefining antonymy: The textual structure of a semantic relation", in: [no editor.]
Louw, B.
1993 "Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies", in: M. Baker – G. Francis – E. Tognini-Bonelli (eds.), 157-176.
Nattinger, J. – J. DeCarrico
1992 *Lexical phrases and language teaching*. Oxford: OUP.
Pawley, A. – F. H. Syder
1983 "Two puzzles for linguistic theory", in: J. C. Richards – R. W. Schmidt (eds.), 191-226.
Richards, J. C. – R. W. Schmidt (eds.)
1983 *Language and communication*. London: Longman.
Sinclair, J. McH.
1991 *Corpus, concordance, collocation*. Oxford: OUP.
1992 "Trust the text", in: M. Davies – L. Ravelli (eds.), 5-19.
1996 "The search for units of meaning", *Textus* 9: 75-106.
Stubbs, M.
1995a "Collocations and semantic profiles", *Functions of Language* 2, 1: 1-33.
1995b "Corpus evidence for norms of lexical collocation", in: G. Cook – B. Seidlhofer (eds.), 245-256.
forthcoming "Computer-assisted text and corpus analysis: Collocation, cohesion and communicative competence", in: D. Tannen et al. (eds.).
Tannen, D. et al. (eds.)
forthcoming *The handbook of discourse analysis*. Oxford: Blackwell.
Weinert, R.
1995 "The role of formulaic language in second language acquisition: A review", *Applied Linguistics* 16, 2: 180-205.
Willis, D.
1990 *The lexical syllabus*. London: HarperCollins.
[no editor]
1991 *Using corpora. Proceedings of 7th Annual Conference of the UW Centre for the New OED and Text Research*. Oxford: OUP.