4

Words in Phrases 2: A Case Study of the Phraseology of English

In the previous chapter, I discussed the concepts, data and methods which can be used to study extended lexical units. In this chapter, I discuss the collocational behaviour of a sample of frequent English words, using data from the large data-base (Cobuild 1995b) described in chapter 3. I will present findings from a 1,000-word sample of the 10,000 head-words. (Starting from a random word in the first ten, I took every tenth word in the alphabetic list.) I will illustrate the kinds of phraseological constraints that words are subject to, show that the collocational attraction between words is much stronger than often realized, and discuss ways of representing extended lexical units more formally.

4.1 Frequency of Phraseological Units

One phenomenon, by its sheer frequency, shows the strength of phraseological tendencies across the most frequent words in the language. Suppose we take all 47 word-forms which begin with *f* in the sample. In 41 cases, the following easily recognizable combinations account for the collocation of node and top collocate. In some cases, I have added function words (which are omitted from the collocations lists); in the case of *fools* <*suffer*, *gladly*>, I have taken the top two collocates.

- despite the *fact* that; *faded* away; *fair* enough; short-*falls*; football *fans*; *farmer*'s wife; anti-*fascist*; mother and *father*; old *favourite*; to *feather* one's nest; *fellow* members; wire *fence*; a *few* years; *fiercely* competitive; *fighter* aircraft; semi-*final*; *finding* a way; *finish* off; said *firmly*; keep *fit*; natural *flair*; *flavour* of the month; ground *floor*; *flown* back; *focused* attention; suffer *fools* gladly; *forcing* down; rain *forest*; *former* minister; heavily *fortified*; backwards and *forwards*; *founding*

fathers; old and *frail*; *free* trade; more *frequently*; close *friendships*; *fruit* and vegetables; *fuelled* speculation; more *fun*; government *funding*; the sound and the *fury*

In the remaining six cases, collocates further down the lists occur in recognizable phrases, such as:

- natural *fabrics*; animal *feed*; *filing* cabinet; space *flight*; closely *followed* by; beg (for) *forgiveness*

With many words, many more of the top 20 collocates are due to recognizable phrases. Here are examples from *fact* and *fair*:

- despite the *fact* that; as a matter of *fact*; a *fact* of life; *fact* finding; the *fact* remains that
- *fair* enough; *fair* share; a *fair* amount of; a *fair* trial; *fair* play; a *fair* chance; *fair* game

I can think of no reason why a sample of words beginning with *f* might be untypical of the whole 1,000-word sample. We therefore have initial evidence that all of the most frequent lexical words in the vocabulary have a strong tendency to occur in well-attested phraseological units.

4.2 Strength of Attraction: Word-forms, Lemmas and Lexical Sets

To estimate the extent and strength of collocational attraction across the sample of words, we can also calculate how strongly a node attracts a single word-form, its top collocate. In the following examples, the top collocate co-occurs with the node in 20 per cent and over of cases. Around 4 per cent of nodes fall into this category.

- brightly 1,467 <coloured 26 %>; calorie 846 <low 29 %>; classical 5,471 <music 22 %>; pepper 4,389 <salt 37 %>; profile 5,584 <high 28 %>; shuttle 3,453 <space 33 %>; tricks 2,202 <dirty 25 %>

In the following examples, the top collocate co-occurs with the node in between 10 and 20 per cent of cases. Around 20 per cent of nodes fall into this category.

- angrily 1,388 <reacted 18 %>; announcement 9,180 <made 10 %>; bitterly 1,782 <disappointed 11 %>; cheering 1,226 <crowd 13 %>; communique 1,060 <issued 15 %>; doses 1,687 <large 13 %>

In the following examples, the top collocate co-occurs with the node in between 5 and 10 per cent of cases. Around 40 per cent of nodes fall into this category.

- advisory 2,593 <group 7 %>; afternoon 16,204 <late 7 %>; alarming 1,711 <rate 8 %>; amid 4,649 <reports 5 %>; applause 2,207 <round 6 %>; autumn 9,307 <last 9 %>

For almost all other node-words (i.e. something over 30 per cent), the top collocate co-occurs with the node in at least one in fifty cases.

It is interesting to look also at the extremes: the relatively small number of cases where the top collocate accounts for under 2 per cent of occurrences of the node, or for 26 per cent and over. Examples include:

- continental 4,085 <breakfast 1.9 %>; explained 10,966 <never 1.2 %>; favourite 12,223 <old 1.5 %>; followed 23,270 <other 1.6 %>; issue 76,632 <rights 1.5 %>; nick 9,775 <time 1.6 %>; sun 36,118 <down 1.2 %>; victor 2,510 <emerged 0.7 %>
- backdrop 1,214 <against 33 %>; bodily 1,303 <harm 38 %>; cleansing 2,072 <ethnic 45 %>; coronary 1,228 <disease 43 %>; curriculum 4,116 <national 35 %>; efficiently 1,414 <more 29 %>; enforcement 2,990 <law 42 %>; esteem 2,021 <self 76 %>; harassment 2,731 <sexual 45 %>; hesitation 937 <without 30 %>; illusions 865 <no 35 %>; liberties 1,212 <civil 67 %>; warring 1,586 <factions 49 %>; whatsoever 1,950 <no 60 %>

Even some of the nodes with only a low probability (under 1 in 50) of occurring with a given collocate also form well-known phrases (e.g. ***continental breakfast***, ***an old favourite***, ***in the nick of time***). The nodes with a high probability (over 1 in 4) of occurring with one single collocate are themselves comparatively infrequent, thus decreasing their likelihood of co-occurrence with a wide range of collocates. In addition, some phrases here are certainly due to topics in the British and international press in the 1980s and 1990s (e.g. *ethnic cleansing*, *National Curriculum*).

These figures show the extent to which words are co-selected in phraseological units. However, in many cases, this crude calculation will underestimate the strength of attraction of a node, since the figures show only the relation between the node and a single word-form. If the collocates list is

lemmatized, then strength of attraction is immediately seen to be greater for some words, as in:

- cheering 1,226 <crowd 13 %, crowds 6 %> 19 %
- frail 944 <old 9 %, elderly 6 %> 15 %
- resemblance 1,085 <bears 18 %, bear 11 %, bore 11 %, bearing 4 %> 44 %

Nevertheless, lemmatization perhaps makes less of a difference than might be thought. I looked at all 56 words in the sample beginning with *g* and *h*. Only in 12 cases is the strength of attraction to the top lemma greater than the attraction to the top word-form. Examples are:

- golf 12,026 <course 12 %, courses> 15 %
- graphic <design 5 %, designer> 10 %
- grim 2,755 <faced 5 %, face, faces> 9 %
- himself 55,418 <found 3 %, finds, find> 6 %
- homework 1,310 <done 12 %, doing> 19 %
- honorary 1,233 <degree 9 %, degrees> 14 %

A reason for these modest increases is that the relative frequency of forms of a lemma is often very different, with the result that different forms often do not appear among the top 20 collocates.

However, what makes a much larger difference – though calculations are correspondingly more subjective – is the strength of attraction between a node and lexical sets of words which are semantically closely related to each other. Illustrative figures are:

- breakaway 1,379 <republic(s) 35 %, group, faction, party> 45 %
- cheering 1,226 <crowd(s) 19 %, people, supporters, fans, audience> 30 %
- deadlock 1,236 <BREAK 41 %, END, resolve> 50 %
- doses 1,687 <large 13 %, high, small, low, higher, lower, massive, heavy, larger> 48 %
- gathering 4,464 <information 5 %, intelligence, data, evidence> 11 %
- heated 2,470 <debate 10 %, argument(s), exchange(s), discussion> 16 %
- humanitarian 3,933 <aid 23 %, relief, assistance, help> 39 %
- obey 1,097 <orders 10 %, order, law(s), rules, command(s), instructions> 38 %
- warring 1,586 <factions 49 %, parties, sides> 73 %

In many cases it is not difficult to find a single syntactic-semantic descriptor for these lexical sets of related collocates, such as:

- deadlock <VERB meaning "end"> 50 %, often at N − 2
- doses <ADJ denoting "size"> 48 per cent, usually at N − 1

Such figures are also likely to be at least a small underestimate, since other collocates below the top 20 will also fall into these lexical sets. Occurrences of individual words may be low, but together they may provide many more semantically related words.

In summary so far: words across the whole of the everyday vocabulary of English have frequent, typical, central uses. Words are not chosen freely, but co-selected with other words in a span of a few words to left and right. After these characteristic uses, there is a long tail of word-forms which occur rarely, though these also often realize a frequent semantic pattern. The semantic patterns are typically simple and common, although the lexical realizations may be very diverse. That is, the units which this method identifies are not fixed phrases, but abstract semantic schemas, which have frequent and less frequent lexical exponents.

4.3 Lexical Profiles: Comprehensive Coverage of Data

So far I have picked out examples of node–collocate pairs which illustrate particular relations. However, a method which looks only at one or two words in the collocates list is hardly adequate, since it does not meet the important criterion of comprehensive coverage of data. For the head-words we have the following data-sets: the top 20 collocates, and 20 random concordance lines for each of the 20 collocates. We must at least account for all of these occurrences.

We would then have a profile of the characteristic uses of the node word: a lexical frame and its typical variants. The purpose of profiles (Crystal 1991) is to summarize and present information in a coherent and systematic manner, so as to facilitate comparisons and the discovery of significant patterns: a numerical dimension helps here. In principle, profiles should be comprehensive: in the present case, down to a frequency cut-off point, thereby automatically giving due weight to the most frequent cases. We are always dealing with repeated events: often hundreds of joint occurrences of node and collocate, but (given the organization of the data-base) always more than fifteen. To do this kind of analysis for each of 10,000 nodes would be a major enterprise. Every word is idiosyncratic, in the sense that its collocates are different from those of every other word. However, some initial simple examples provide a clue how to proceed more systematically.

4.3.1 *Example 1: lexical profile for* resemblance

Here is a case where almost all of the top collocates fit into a simple lexical schema:

- resemblance 1,085 <(bears, bear, bore, bearing) 45 %, little, no, striking, between, passing, uncanny, any, more, strong, family, remarkable, physical>

Almost all of these collocates are due to the occurrence of phrases such as:

- BEAR no *or* little resemblance to . . .
- BEAR a passing *or* physical resemblance to . . .
- BEAR a strong *or* striking *or* uncanny resemblance to . . .

These are not the only possibilities. Although these are the typical, central cases, BEAR co-occurs with ***resemblance*** in only 45 per cent of cases: it is also possible to say, for example, *HAVE **a resemblance to***.

4.3.2 *Example 2: lexical profile for* reckless

Here is a case where all the most frequent collocates of a node fit easily into just two schemas. The node ***reckless*** has only five collocates with more than 15 occurrences each:

- reckless 1,045 <driving 19 %, death, causing, admitted, disregard 2 %>

It occurs in almost one case in five in the phrase ***reckless driving***, and often in longer phrases such as

- admitted reckless driving; admitted causing death by reckless driving

In one case in fifty, it occurs in the phrase ***reckless disregard***, and hence in longer phrases such as

- displayed a reckless disregard for safety; with reckless disregard of the consequences

Again, this obviously does not mean that all occurrences of ***reckless*** are in these combinations: only around 20 per cent are. It means that these are

collocations which frequently recur, and that other nouns at N + 1 (as in *reckless expansion* or *reckless outpouring*) are, individually, infrequent.

4.3.3 *Example 3: lexical profile for* backdrop

Here is another simple case, in which only eight collocates co-occur more than 15 times with the node:

- backdrop 1,214 <against 33 %, set, provide, perfect, place, provides, form, provided>

Typical phrases are

- PROVIDE the perfect backdrop for
- TAKE place against a backdrop of
- set against a majestic backdrop of

The most frequent adjective is *perfect* (3 %). However, as well as backdrops which are:

- attractive, beautiful, dramatic, effective, epic, flattering, stunning

there are also *dismal* and *gloomy* backdrops of *disunity*, *turbulence*, *uncertainty* and *violence*.

4.3.4 *Example 4: lexical profile for* doses

In this example, I have grouped the top 20 collocates of the node into syntactic-semantic classes.

- doses 1,687
  <large, high, small, low, higher, lower, massive, heavy, larger> 48 %
  <daily> 4 %
  <radiation, vitamin, drugs> 13 %
  <given, taken, used, received, taking> 14 %
  <very, even> 6 %

The most frequent verb-forms are past participles. The most frequent grammatical words are: *of*, *in*, *are*, *can*. Typical phrases are:

- received massive doses of radiation
- given in very small daily doses

- taken in large repeated doses

The blend of collocation, colligation and semantic preference in the basic pattern can be stated informally as follows. There is typically a verb meaning "give" or "receive", often followed by a size adjective, followed by *doses of*, followed by a medical noun.

This example illustrates two points. First, it shows that even words which appear to have an independent denotation, and which are hardly ambiguous even as decontextualized individual words, may nevertheless have a strong tendency to occur within predictable lexico-syntactic frames. Second, it is a piece of evidence about the range of meanings which are typically encoded. That is, when people talk about *doses* of something, then these are the meanings which frequently get expressed. As G. Francis (1993: 155) puts it:

> [A]s we build up and refine the semantic sets associated with a structure, we move closer to a position where we can compile a grammar of the typical meanings that human communication encodes, and recognise the untypical and therefore foregrounded meanings whenever we come across them.

Corpus analysis shows what are frequent or typical uses. There are, of course, other non-medical uses, often in ironic phrases such as ***large doses of sarcasm*** or ***can take politicians only in small doses.***

4.4 A Model of Extended Lexical Units

These examples are still presented informally. So, how might we more formally define units which are highly conventional in their semantic patterns, but also highly variable in their potential lexical realizations – which have one or more clear central tendencies, but different ranges of variation? The relations defined in chapter 3 give us the basis for a model of extended lexical units. In order to define a linguistic unit, we have to specify its possible constituents, and the possible relations between them. The constituents define the semantic content of the unit. The relations define its structure. (This section develops proposals in Sinclair 1996, 1998.) We have the following model.

RELATION	*constituent*
(1) COLLOCATION	collocate: individual word-form or lemma
(2) COLLIGATION	grammatical category

(3)	SEMANTIC PREFERENCE	lexical set: class of semantically related word-forms or lemmas
(4)	DISCOURSE PROSODY	descriptor of speaker attitude and discourse function

As Sinclair (1998) points out, relations (1) to (4) are increasingly abstract. Collocation refers to individual word-forms, which are directly observable in texts. Colligation refers to classes of words (such as past participles or quantifiers), which are not directly observable: they are abstractions based on generalizations about the behaviour of the word in the class. The classes are often small, and always closed (for example, there is a small, finite number of quantifiers in English). Semantic preferences refer to a class of words which share some semantic feature (such as words to do with "medicine" or "change"). Such a class is also abstract, and will have frequent and typical members, but will be open-ended. Discourse prosodies are even more open-ended and typically have great lexical variability.

These four relations are all probabilistic and non-directional. Two further relations specify the probabilities and the positions of occurrence. And finally, we must also say how widely our description applies.

(5) STRENGTH OF ATTRACTION. This is defined in percentage terms: given the occurrence of a node, what is the probability of occurence of a collocate, grammatical category, lexical set or discourse prosody?

(6) POSITION AND POSITIONAL MOBILITY. Relations (1) to (5) are non-directional: two constituents simply co-occur. However, it may be that one sequence always occurs (e.g. ***spick and span***, but not ****span and spick***), or that relative position is variable.

(7) DISTRIBUTION IN TEXT-TYPES. We must specify whether the lexical unit occurs widely in general English, or whether it is restricted to broad varieties, such as journalism or technical and scientific English, or to specialized text-types, with a narrow speech-act function, such as recipes or weather forecasts.

This model brings lexis fully within the traditional concerns of linguistic theory. Much twentieth-century linguistics has assumed that lexis is not amenable to systematic treatment, because the vocabulary is merely 'a list of basic irregularities' in a language (Bloomfield 1933: 274). For much of Chomskyan linguistics, it is syntax which is concerned with general rules, whereas lexis is largely dismissed as being concerned with isolated and idiosyncratic facts. However, relations (1) to (4) correspond to

the classic distinctions between syntactics, semantics and pragmatics, which were drawn by Morris in the 1930s (Morris 1938). Syntactics (or syntax) deals with how linguistic signs relate to one another (here collocation and colligation), semantics deals with how linguistic signs relate to the external world (here lexical sets and the phenomena they denote), and pragmatics deals with how linguistic signs relate to their users (here expression of speaker attitude).

The examples of collocation above also show that there is much in the behaviour of words which is automatic and not open to conscious reflection. This means that introspection about lexical meaning is often unreliable or at least incomplete. Also in terms of its automaticity, lexis is seen to be in line with many aspects of phonology and syntax (Channell 2000). In the model, lexis has acquired a primary role, and syntax a reduced role, in determining aspects of positional mobility (for example, in active versus passive variants of a unit), and in linking phraseological units to each other in running text. This revised division of labour between lexis and syntax will require much working out in detail.

A more stringent procedure – not entirely formalized, but at least a check on rank subjectivism – can be defined as follows (see Sinclair 1991: 54ff, 84ff, 105ff; 1996; De Beaugrande 1996: 515ff; and Clear 1996, for related suggestions). (1) Group the 20 collocates into semantic subsets, using criteria which are as explicit as possible. (2) Calculate what percentage these semantic subsets comprise of the whole collocates list. (3) Check the positional variability of the constitutents. The data-base averages information across a span of 4 : 4. However, positional information can easily be retrieved from the concordance lines, by using positional frequency tables: see table 4.1. (4) Check whether independent corpus data (not restricted to the top 20 collocates) reveal further uses. That is, check the recall of information (see chapter 3.6.3).

4.4.1 *Example 5: lexical profile for* UNDERGO

The following analysis follows this procedure. The collocational data are as follows:

- undergo 1,205 <surgery 108, tests 67, treatment 62, change 53, training 43, test 41, medical 40, before 37, changes 35, operation 34, women 31, forced 26, further 25, testing 25, major 24, examination 23, extensive 21, heart 20, required 19, transformation 17>

There is a simple pattern and discourse prosody: people involuntarily ***undergo*** serious and unpleasant events, such as medical procedures.

The 20 collocates can be arranged into sub-lists. Some words, mainly nouns, are medical (*surgery*, *treatment*, *medical*, *operation*, *heart*); some have to do with training and testing (*training*, *examination*, *test*, *tests*, *testing*); some concern change (*change*, *changes*, *transformations*); some adjectives concern the seriousness or extent of the events (*further*, *major*, *extensive*); some verbs concern their involuntary nature (*forced*, *required*). The two remaining words do not obviously fit into these sub-lists, but must also be accounted for (*before*, *women*).

The data-base gives 400 (20 × 20) randomly selected concordance lines, but these lines may, by chance, be selected twice, and in this case there are 343 different lines. In descending frequency: 181 involve people undergoing medical procedures, including medical tests, such as:

- major heart surgery; conventional medical treatment; mandatory drug tests

Some 72 involve people and things undergoing *changes*, *transformations* and *metamorphoses*. Some are explicitly unpleasant: *agonies of readjustment*, *malignant transformation*. Many others are by implication unpleasant, since they are

- considerable, dramatic, drastic, extensive, fundamental, major, profound, radical, significant

Some 46 involve non-medical testing, again often by implication unpleasant, since it may be *compulsory* or *rigorous*, or may involve *tough scrutiny* or *police checks*. Twenty-two involve people undergoing *training*: often military, and often *extensive*, *intensive* or *lengthy*, and again, therefore, not necessarily pleasant. Eleven additional cases are explicitly unpleasant: people or things undergo, for example, *cutbacks*, *humiliation*, *imprisonment*, *trauma*. The remaining examples are technical: see below: *bifurcations*, etc.

These exponents of the discourse prosody "unpleasant" almost always occur to the right of *undergo*. Exponents of a related prosody, "involuntary", occur mainly to the left. The lexical realizations *forced* and *required* occur in the top 20 collocates, and *must* is one of the most frequent collocates amongst the stop-words.

I have now said something about all 20 top collocates, except *before* and *women*. *Before* occurs amongst the top 10 collocates, in 3 per cent of occurrences, and provides a hint of the characteristic discourse in which *undergo* occurs. In many cases, a sequence of actions, which happen before or after surgery, tests or training, is being reported: around half the con-

cordance lines contain references to the time when events happen, to sequences of events, or to events being planned:

- must wait 24 hours before they can undergo the procedure; undergo his fourth operation inside a year; undergo several systems checks; is planning to undergo; due to undergo; scheduled to undergo

The reason why *women* is so frequent a collocate (3%) is less obvious. It seems partly due to the frequency of mention of events such as *abortions*, *fertility treatment* and *hysterectomies*. It may also be partly because a sex-neutral collocate such as *patients* can refer to men or women, but when women are meant, they are explicitly mentioned.

The 343 concordance lines are not a random selection of all occurrences of *undergo*, since they all contain one of the top 20 collocates. I therefore checked an independent 2.3-million-word corpus. The word-form *undergo* is not very frequent ($n = 14$), and in this case there are no obvious differences in use across different forms of the lemma ($n = 42$), which I therefore looked at as a whole. The percentages are different, but the patterns are confirmed, and one pattern becomes clearer. In this smaller corpus, objects of the verb were from the semantic fields of "change" (16) or "medicine" (8), or were "unpleasant" (9):

- ordeals; a crisis; a savage sentence for a crime; a traumatic experience; bizarre eighteenth-century initiation rites

UNDERGO also occurs in technical English with no necessarily unpleasant connotations. Almost all other cases (8) were scientific and technical, as marked by collocates such as *bifurcations*, *diapause*, *nucleon*. The sole remaining case is *the spring-cleaning which it had undergone*: a humorous reference to a landing strip, which then *shone like black glass*.

Further corpus data reveal further specific lexical items, and show how the simple patterns can be realized by a great variety of lexis (Sinclair 1996: 95). For example, in this case, the "unpleasant" prosody is implied by the text following *pessimism*:

- why Voltaire's ideas *underwent* this *change* is not clear – possibly his new *pessimism* was a result of the great earthquake of 1755

Similarly, great lexical variety is possible in expressing the "involuntary" prosody. As well as explicit lexical items (*forced to*; *required to*; *have to*; *must*; *will have to*), the prosody may be only implied as in

- police said he would ***undergo*** psychiatric ***examination***

Further corpus data would be certain to reveal further lexical variation, but unlikely to reveal other major semantic preferences. This is a prediction about how the word is used, and is open to empirical testing. In summary, the main semantic patterns are simple. (1) In general English, people are forced to undergo unpleasant experiences, especially medical procedures, or tests and (often arduous) training. (2) People and things undergo (usually radical and often unpleasant) changes. (3) In scientific and technical English, the word is usually neutral.

The central uses of the word, with its typical collocates, can easily be stated: see figure 4.1. The "involuntary" and "unpleasant" prosodies are usually encoded to the left and right respectively. They express the discourse function of the extended lexical unit: why is this being mentioned now? And, despite the variation, there are preferred lexical selections, down to the choice of individual words (Sinclair 1996: 88–9).

Characteristic examples from the concordance lines are:

- he was forced to ***undergo*** an emergency operation
- his character appeared to ***undergo*** a major transformation
- each operative had to ***undergo*** the most rigorous test
- will ***undergo*** extensive skills and fitness training
- forced to become refugees, to ***undergo*** further migration and further suffering

Concordance 4.1 shows a larger random selection of 50 concordance lines (from amongst those lines which contain one of the top 20 collocates).

passive or modal + *undergo* +	adjective +	abstract noun
forced to	typical	typical
required to	adjectives	lexical fields
must		
etc.	*further*	"medical procedure"
	extensive	"testing"
	major	"training"
	severe	"change"
	etc.	"a trauma"
		etc.

Figure 4.1 The prototypical uses of ***undergo***

The prototypical uses of ***undergo*** can be represented as a lexico-grammatical frame plus:
frequent individual collocates (e.g. ***surgery***)
typical pragmatically specified adjectives (e.g. ***major***)
typical semantically specified lexical fields (e.g. "change")

1. f the Oval Test last summer to undergo a cartilage operation. He was not
2. and international institutions undergo a change – Political observers in
3. ould be aware the system is to undergo a historic transformation. Sometime
4. families of the nation did not undergo a major metamorphosis until the op
5. or the first time will have to undergo a means test and a needs assessmen
6. the applicants, asking them to undergo a medical examination, and prepari
7. court today and is expected to undergo a psychiatric examination. 930430
8. r – Discovery will not have to undergo a special fueling test because it
9. ir work, each operative had to undergo a stringent medical examination ev
10. m-you find romantic are due to undergo a transformation on the 4th, and w
11. h John Fashanu being forced to undergo an Achilles tendon operation. The
12. rawling estate are required to undergo an 'eyescan' before being allowed t
13. ge it's led dozens of women to undergo back-alley abortions in countries
14. arnations that the spirit must undergo before it can achieve release from
15. Mr Forbes subsequently had to undergo brain surgery, and his friends and
16. ill Clinton, style is about to undergo dramatic changes. Out for instance
17. former champion Pat Cash will undergo exploratory surgery on an injured
18. s of alcoholic beverages is to undergo extensive food testing. And only i
19. programme for hostages. He'll undergo extensive medical checks and psych
20. now in Bahrain where they will undergo extensive medical examinations bef
21. e championships tomorrow, will undergo extensive skills and fitness train
22. a hospital and insisted that I undergo extensive tests – There was nothin
23. baden in Germany where he will undergo further medical tests at an Americ
24. inal hysterectomy patients may undergo further surgery at a rate as high
25. Robert Mays allow Kimberly to undergo genetic testing. As fiction, the t
26. r Warren, who was scheduled to undergo his eighth open heart surgery afte
27. nd Howey may even be forced to undergo his fourth operation inside a year
28. h to one half of all women who undergo hysterectomy develop some morbidit
29. management know-how. Employees undergo intensive training on the shop flo
30. gories of children required to undergo language testing. The categories o
31. ster. Yesterday, he was due to undergo major brain surgery. On Friday nig
32. undergone and will continue to undergo major cutbacks. If Japan does not
33. atrick Buchanan is planning to undergo major heart surgery tomorrow – His
34. he first established prison to undergo "market testing" and the first for
35. s suggest that women likely to undergo menopause, at about age 50, ought
36. whether or not a woman should undergo more extensive tests where a diagr
37. t that a pioneer product would undergo more testing, he says. Rissler, wh
38. weapon – Police said he would undergo psychiatric examination before any
39. leaders could not be forced to undergo random drug testing in order to re
40. nly two feet in diameter, will undergo several systems checks before bein
41. use the family butcher shop to undergo significant change in appearance a
42. captain, Villiam Hyravy, is to undergo surgery and will take no part. The
43. wait 24 hours before they can undergo the procedure. Doctors must tell p
44. . So who is actually having to undergo the tests? An oceanographer got te
45. e. But it has not been able to undergo the transformation and economic mo
46. ut her ability to continue and undergo the treatment. It was very clear t
47. attempt last year. As recruits undergo training in a Fortitude Valley fig
48. d deed.' Before she agreed to undergo treatment and completed donor cons
49. ages are found, patients often undergo treatment, including bypass surge
50. e Pendennis Shipyard. She will undergo trials locally before sailing to

Concordance 4.1 Sample concordance lines for ***undergo***

Notes: Lines from the data-base were put in random order, and every 8th line selected. These 50 lines were then ordered alphabetically to the right

Table 4.1 Positional frequency table for ***undergo***, span 3:3

was [18]	forced [26]	to [219]	*	a [85]	medical [22]	and [21]
is [14]	required [21]	will [38]	*	an [26]	surgery [20]	tests [16]
be [13]	have [15]	and [9]	*	further [25]	testing [16]	examination [14]
and [11]	is [11]	must [9]	*	the [21]	treatment [15]	surgery [13]
and [11]	is [11]	must [8]	*	the [21]	treatment [12]	surgery [13]
that [8]	they [11]	he'll [7]	*	major [20]	change [9]	operation [12]
been [7]	about [10]	should [7]	*	surgery [12]	changes [9]	transformation [11]
were [7]	and [9]	who [7]	*	treatment [9]	for [9]	before [9]
where [7]	patients [7]	women [7]	*	medical [7]	heart [9]	test [9]
children [6]	that [7]	often [6]	*	heart [6]	and [8]	medical [8]
he [6]	he [6]		*	his [5]	major [8]	for [7]
in [6]	will [6]		*	testing [5]	operation [8]	in [7]
the [6]	women [6]				examination [6]	on [7]
women [6]	due [5]				extensive [6]	training [6]
will [6]	ordered [5]				transformation [6]	to [6]
for [5]					radical [5]	testing [6]
last [5]					test [5]	the [6]
not [5]					training [5]	a [5]
of [5]					the [5]	as [5]
						by [5]
						changes [5]

Notes: The node ***undergo*** is indicated with an asterisk. Only collocates occurring 5 times and more are shown.

Table 4.1 shows a positional frequency table, in which words in positions N − 3 to N + 3 are displayed in descending order of frequency, down to a frequency cut-off of 5 joint occurrences.

4.4.2 *Example 6: lexical profile for* chopped

The next example illustrates further principles. The starting data are:

- chopped 3,602 <finely, fresh, parsley, onion, garlic, tbsp, tomatoes, oz, peeled, add, off, onion(s), pepper, salt, chives, herbs, tablespoons, dried, small, tsp>

It is sometimes argued that co-occurrences between words such as *chopped*, *herbs*, *parsley* and *onions* are not real collocations, but words which co-occur simply because they correlate with states of affairs in the world. Smadja (1993: 150) argues this with reference to word-pairs such as *doctor–nurse* and *doctors–hospitals*. Kjellmer (1991: 114) points out that the phrase *glass of water* is more frequent than *cup of water*, merely because water is usually served in a glass. However, given our present limited knowledge about statistical properties of extended lexical units, it seems unwise to make firm distinctions about what is and is not linguistic. Similarly, Benson's (1990: 26) rejection of *pass the salt* as a collocation, on the grounds that one can pass all sorts of things, seems odd, since *pass the salt* is a highly stereotyped phrase.

In any case, although the extended lexical units around *chopped* are not idioms, they are idiomatic. Recipe writers could talk of *?finely cut* or *?finely sliced fresh parsley*, but by and large they do not. The word-form *add* occurs not only in recipes (I might add). However, when it occurs as a sentence- or clause-initial imperative, it is almost always in a recipe (or instructions for a chemical experiment). If *add* and *chopped* co-occur then the probability that this is a recipe must be near 100 per cent.

Chopped is a case where the node-word is the centre of a tight collocational cluster: the top three collocates are each 12 per cent and over, and even the last collocate is 3 per cent. In the list, 19 out of the 20 collocates are due to the use of the word in recipes.

(The exception is *off*. The collocation *chopped off* occurs almost exclusively in connection with chopping off bits of human body parts. This is confirmed by looking at all instances of *chopped* in an independent 2.3-million-word corpus. Out of 20 instances, 15 were from recipes. The other 5 all involved verb plus particle: *chopped off*, *chopped up*, *chopped at*. Four involved violence to humans. The fifth was a critical reference to music being superficially *chopped up*. So this finding seems not to be due to an over-representation of recipes in the Cobuild (1995b) data-base. I checked further by looking at

occurrences from the 100-million-word British National Corpus (see Notes on Corpus Data and Software). Out of 50 random examples of ***chopped***, 39 were from recipes. Of the other 11 examples, four occurred in the phrase ***chopped off***, one in ***chopped up***, and one in ***chopped down***. Four of these were references to violence to humans. The word-forms ***chop*** and ***chopping*** are also frequent in recipes. Other phrases include ***chop down trees***, ***due for the chop***, ***endless chopping and changing***. The form ***chops*** occurs mainly as a noun in other phrases, such as ***lamb chops*** and ***licking their chops***.)

As has often been pointed out in stylistic analyses, the vocabulary of recipes is distinctive, and the collocates of ***peeled*** and ***garnish*** have a large overlap with ***chopped***. The following are collocates of two or all three of these nodes:

- chives, finely, fresh, garlic, herbs, onion, parsley, pepper, slice/d, small, tomatoes

4.5 Summary and Implications

In this chapter, I have used some simple statistics to describe how words co-occur in text. The data-base (Cobuild 1995b) was produced by an entirely automatic procedure: a computer was programmed to extract the 10,000 most frequent word-forms from a large corpus together with their most frequent collocates and a random selection of concordance lines. Corpus linguistics is based on publicly available data and replicable methods: this is what is meant by empirical linguistics. Nevertheless, the output requires considerable interpretation.

A great deal of language in use consists of extended lexico-semantic units. These units are not just individual phrases which can be listed. Typical instances can be listed, but not all instances are equally representative. The units themselves are abstract: they are semantic schemas, which have default values, and typical realizations, but often no necessary or sufficient features. If we are thinking of the behaviour of a language community, then they are norms. If we are thinking of the competence of individual speakers, then they are mental models.

All of the most frequent content words in the language are involved in such patterning. This is not a peripheral phenomenon (collocations are not an idiosyncratic feature of just a few words), but a central part of communicative competence. These semantic schemas can be modelled as clusters of lexis (node and collocates), grammar (colligation), semantics (preferences for words from particular lexical fields) and pragmatics (connotations or discourse prosodies). Such a model brings the study of lexis within the

mainstream of linguistic description: the units are combinations of lexis, syntax, semantics and pragmatics. The findings show that there is a level of organization between lexis and syntax, which is only starting to be systematically studied, and which is not reducible to any other level of organization.

The central problem in linguistic description is how to describe a system which is both highly complex and highly variable. Semantic schemas are general and simple patterns which have considerable lexical variation due to local context and choice.

4.6 Background and Further Reading

For references to the large literature on phraseology, see chapter 3.11. For a range of computational methods for identifying recurrent phrasal units in corpora, see Choueka et al. (1983), Yang (1986), Smadja (1993), and Justeson and Katz (1995).

4.7 Topics for Further Study

(1) It is easy to find further examples which support the claim that UNDERGO has a negative discourse prosody. However, such claims must be tested by searching for counter-examples. Can you find any? For example, study the collocation *UNDERGO **training***: does this always co-occur with further collocates which imply an unpleasant experience?

You could also check examples of the collocation ***willingly** UNDERGO*: do they contradict the claim that people "involuntarily" undergo "unpleasant" experiences? This phrase is not frequent and you may have to search a very large text collection to find enough examples to make generalizations about its use. You might use a search engine which can find phrases in documents in the world-wide web. Here are two examples out of around 175 which I found:

- no-one, short of a severely psychotic masochist, would ***willingly undergo*** what she went through
- why did he ***willingly undergo*** forty years of hardship?

Are these uses typical? If so, what discourse prosody is there around the phrase ***willingly** UNDERGO*?

(2) This chapter has largely ignored the variation in collocations across different text-types. Some individual collocations may signal a specific text-

type: the phrase ***finely chopped*** is probably from a recipe; ***warm*** and ***front*** do not signal any text-type on their own, but ***warm front*** is probably from a weather forecast; ***luxury home*** is probably from advertising by a builder or estate agent. In general English, ***time*** might collocate with ***spend*** or ***waste***; but in sports commentaries, it is likely to collocate with ***half*** and ***injury*** (Partington 1998: 17). Find other examples where a particular phrase or collocation reliably identifies a text-type, and other examples where words have different collocates in different text-types.

On the basis of such differences across text-types, Biber et al. (1998: 234) argue that 'characterizations of ***general English*** are usually not characterizations of any variety at all, but rather a middle ground that describes no actual text or register'. Is this criticism of the concept 'general English' justified?

(3) Words which are rough (denotational) synonyms are usually used in quite different ways: possibly in different collocations, with different connotations, in different text-types, and so on. Study the different patterns around these approximately synonymous adjectives:

- escalating, growing, increasing, mounting, rising, soaring, spiralling

For example, does ***rising*** have mainly positive collocates (***rising prosperity***), or mainly negative collocates (***rising costs***)? Does its discourse prosody depend on the longer phrase it occurs in? What nouns typically follow ***a rising tide of***? Which nouns typically follow ***mounting*** or ***soaring***? Partington (1998: 113–14) provides further data and discussion of these roughly synonymous adjectives.

(4) Data from the Cobuild (1995b) data-base show the nouns which typically follow ***amid***, and adjectives which typically precede the nouns:

- amid 4,649 <reports 5 %, fears, speculation, allegations, signs, concern, scenes, controversy, security, claims, rumours> 28 %
- amid <growing, continuing, mounting> 7 %

Some phrases include:

- amid reports of heavy fighting; amid reports of a Cabinet split; amid tight security; amid signs of growing concern; amid scenes of blood-curdling violence; amid scenes of high emotion
- amid breath-taking scenery; amid beautiful countryside; amid romantic ivy-covered walls; amid the frantic last few days in London; amid much fanfare, the ***Manhattan*** tried to sail

Use these data and other corpus data to make a statement of the discourse prosody predicted by ***amid***, and to discuss whether this prosody is different in different text-types.

(5) It has been claimed that 'all the forms of a verb . . . are often very different from one another' (Sinclair 1991: 8), in the sense that they have different collocates and therefore different uses and different meanings. However, different forms of UNDERGO are very similar in their collocates. And while *seeks* is sharply different from other forms of SEEK, *seek*, *seeking* and *sought* are similar in their uses and all share collocates from the semantic field of "help" (see chapter 2.2.1). Sinclair's claim is an empirical one, but I do not know of work which has investigated how often it is actually the case, and even the best-known corpus-based dictionaries (such as CIDE 1995; Cobuild 1995a; LDOCE 1995; OALD 1995) still use mainly lemmas as head-words. Investigate the different forms of some lemmas. For example, do the different forms of ACHIEVE or PURSUE share a significant number of collocates? Or do the different forms occur in significantly different phrases? What would be 'significant' in such cases?