Contextual Dependency and Lexical Sets

PATRICK HANKS
Oxford English Dictionaries

Preliminary findings from corpus analysis suggest that the semantics of each verb in the language are determined by the totality of its complementation patterns. Accurate description of those patterns requires a level of analytic delicacy which was not possible until the advent of large bodies of data, along with techniques for distinguishing significant patterns from mere noise. Such analysis is in its infancy, but it is already clear that, in order to analyse the semantics of verbs empirically, it is necessary to identify typical subjects, objects, and adverbials and to group individual lexical items into sets within those clause roles. The nature of lexical sets is discussed and an attempt is made to indicate the range of semantic and syntactic phenomena likely to be encountered in lexical analysis of this kind.

KEYWORDS: complementation pattern, verbs, semantics

Verb Complementation

Figure 1 is an example of a type of exercise which is now widespread where corpus data is available and used for teaching purposes. It proves popular and instructive with students who have reasonably sophisticated intuitions about English. Exercises of this sort are comparatively easy to construct nowadays, given the availability of concordancing programs such as MICROCONCORD.

The exercise is to identify the one word (there is only one) which fits all 30 of the concordance lines, a selection from the 300-odd citations for this word in the Oxford Hector Corpus (a pilot study of 18 million words, compiled at Oxford University Press in 1990-92, for what is now the British

International Journal of Corpus Linguistics 1:1 (1996), 75–98. DOI 10.1075/ijcl.1.1.06han
ISSN 1384–6655 / E-ISSN 1569–9811

1.	back to the ferrymaster . "Tell me ," he	..... , "did you take anyone else across the Forth tha
2.	lip . "Drink some of this tea ," Maggie	..... , handing it to Laura . Laura took several sips
3.	en't you going to say something ? " Laura	..... . He took a deep breath , and managed to smiled
4.	school with their children , Princess Anne	..... yesterday . She spoke out after a new survey di
5.	e Archbishop of Cape Town , Desmond Tutu ,	..... Israelis to forgive the Nazis for killing six mi
6.	ckles white . Minnie poured her some tea ,	..... her to remove her coat if she felt the room was
7.	lable and have refused to say if they have	..... him to resign . However , Tory leader Councillo
8.	ected the claims of the offer document and	..... shareholders to hold on to their shares which ,
9.	n a stolen car crashed into a tree , today	..... young people to turn their back on joyriding .
10.	. All teachers in England and Wales were	..... by the National Union of Teachers to have their
11.	lture and lived within it . They were not	..... by their teachers to achieve educational success
12.	e held good for young refugees . They were	..... to learn English , to refrain from speaking or
13.	y might know the identity of the driver is	..... to contact Colchester CID on (0206) 762212 .
14.	e step further . Sisters and brothers are	..... to help smash sexist and classist legislation at
15.	ieved to be the first time a developer has	..... the Government to schedule an archaeological sit
16.	ite right . Our campaign has consistently	..... the Government to undertake a full and proper
17.	dent Bush , for the first time , last week	..... the Iraqis to take matters into their own hands
18.	use arrests of main opposition figures and	..... the military government to restore democratic ri
19.	here 's a sucker born every minute , Klein	..... the suckers of the world to unite in protest .
20.	nineteenth century other Indian Christians	..... that Christianity in India should assume an Indi
21.	artment to look at physical education . It	..... that all children should be taught to swim . The
22.	iropody service which is now available and	..... that pensioners should have their own pocket mon
23.	inds up . " Last week association leaders	..... acceptance of the deal , which had been rejected
24.	rty chairman Chris Patten , who repeatedly	..... caution , was not to know that Treasury forecast
25.	hat he had done , James took the reins and	..... his horse out of Corporation Lane with all speed
26.	oes not wish to go past , and yet is being	..... on by his rider . His confusion will be indicate
27.	ping furiously down the high road . Sharpe	..... the black stallion on , but the footing under th
28.	-driver what to expect . Evans , 30 , then	..... his Sierra through Grizedale faster than McRae ,
29.	July , jumping aboard at the last moment ,	..... on by a friend who thought she might tell the po
30.	dent which saw the flag-waving Mongolian ,	..... on by a gaggle of grannies , invading the ring t

Figure 1. Spot the missing word

Above are thirty short extracts from a cross-section of different English texts, with one word — the same word in each case — blanked out. What is the missing word?

National Corpus). The British National Corpus itself, consisting of 100 million words of text, contains 4590 citations for the word in question.

What is the missing word? A typical seminar discussion goes something like this:

LINES	COMMENTS
1-4	What are the obvious candidates? — 'said', 'asked', and perhaps 'suggested'.
5	It can't be 'said' or 'suggested'. Neither 'said' nor 'suggested' take PERSON as a direct object.
6-8	Seems to confirm 'asked'. Other possibilities? 'encouraged'?? 'proposed'??
9	It can't be 'proposed', because it doesn't fit line 5. Can't be 'encouraged' because it doesn't fit lines 1-4. 'asked' looks a bit weak here, too.
10-13	'told'? But 'told' doesn't fit lines 1-3. 'asked' is still the front runner.
14, 17, 19	'asked' would be very weak in such dramatic contexts.
23, 24	Rules out 'asked'. You can ask a question, but you can't 'ask acceptance' or 'ask caution' in modern English. (You can, of course, 'ask *for* caution'.) Is it 'requested'? But 'requested' is even weaker than 'asked' for lines 14, 17, and 19.
25-26	By this point someone will have suggested 'urged'.
28	Note the Ford Sierra as a metaphorical steed.
29-30	Note the use with 'on' — almost a phrasal verb. The missing word must be 'urged'.

What can we learn from such an exercise? Most importantly, it seems that the set of normal complementations for **urge** is unique. If the word *urged* is deleted from any one text (e.g., for a Cloze test), two or three plausible candidates — perhaps more — may fit the slot that it leaves. But cumulatively the complementations add up to a unique set of patterns, ruling out all other candidates. The same seems to be true of most if not all verbs in English. The implications of this phenomenon have not been fully explored. One such is that the semantics of a verb are determined by the totality of its complementation patterns. So, for example, the systemic choice that was made by the utterer of *urged* in lines 1 and 2 (in preference to, e.g., "said" or "asked" or

"told her" or "proposed") was in part motivated by his or her subconscious knowledge that **urge** is associated on the one hand with riders, horses, and forward movement (as in lines 26 and 27), but also with petitioners, governments, and positive actions (as in lines 15 and 16).

Norms of Usage

Teasing out the totality of the complementation patterns in this way is consistent with the Firthian programme of "knowing a word by the company it keeps" (Firth 1957). It has only become possible to attempt this with any confidence recently, with the advent of large corpora. Corpus evidence does not enable us to determine all possible usages. No amount of corpus evidence can prove that some linguistic phenomenon *cannot* exist. Corpus evidence does, however, enable us to determine what patterns of usage are normal, central, and typical and even (with due caution) to say which patterns are the most frequent. Corpus analysts are, therefore, concerned with the regular and the normal, and not at all with the boundaries of linguistic possibility.

While broad subcategorizations such as "transitive" and "intransitive" can be a helpful first step towards a lexicosyntactic analysis, more delicate subcategories are necessary for the proper understanding of a lexical item. Some of these subcategories may be quite idiosyncratic. Note, for example, that at lines 24 and 25 we ruled out *asked* on the grounds that the expressions *ask acceptance* and *ask caution* are not conventional in modern English. In the British National Corpus (taking all forms of the lemma together: *urge, urged, urges, urging*) there are 25 matches for *urge caution* and six for *urge acceptance,* whereas there are none at all for *ask caution* or *ask acceptance*. It is not, of course, the transitivity of **ask** that is in question. Transitive uses such as "ask a question" and "ask a person a question" are quite normal. The problem lies in the selection of the abstract nouns *acceptance* and *caution* as the direct object of *asked.* But what precisely is the problem? A scan of the corpus shows that postulating a selection restriction ruling out abstract nouns as the direct object of **ask** would be too strong. *Permission* is an abstract noun, and it is normal enough to talk about *asking permission.* Perhaps, therefore, what is needed is a simple list of the nouns which typically or normally occur as direct object of **ask**, by some definition of "typical" and "normal", based on evidence of actual usage. A list of this sort might well provoke more

precise generalized explanations than has been possible hitherto. Be that as it may, it would certainly provide a more satisfactory base for predictive purposes in natural-language processing and language teaching than attempts to define lexical classes a priori by the use of words such as "abstract" (or even [ABSTRACT] in capital letters and square brackets).

In the literature, such subcategorizational phenomena are often referred to as selectional restrictions. It is preferable, though, to think of them as selectional preferences. A restriction prevents or forbids you from doing something, whereas it is often the case that locutions excluded by a selectional preference are, nevertheless, perfectly grammatical, psychologically acceptable, and communicatively adequate. They are just not conventional. They deviate from an established norm.

But what is this "established norm"? It is not to be found described fully in any published work. Indeed, because of the flexible, variable nature of the lexicon, even attempting a full and accurate description of the norm for any given usage may be impossible, for principled reasons: how can one define a phenomenon whose boundaries are shifting and variable? Distinguishing the norm from the possible takes us into the unfamiliar world of probabilities and three-valued logics. As Zadeh (1965) pointed out, classical logic can deal with questions such as "What is the set of numbers greater than one?" but not with questions such as "What is the set of numbers *much* greater than one?'' Questions about lexical sets are of the latter kind.

If we are determined to persevere in the attempt to encapsulate the invariant core that lies at the heart of that elusive, variable phenomenon, the conventional use of a word in a natural language, we are committed to behavioural profiles of normal usage. What would such a profile look like?

Figure 2 is a behavioural profile of the verb **urge** arising out of the work of the Hector Project: an attempt to encapsulate its established norms (patterns of usage) on the basis of analysis of a body of evidence of actual usage (a corpus). Note that nothing here is said about meanings. Corpora give only indirect evidence for meanings. Before attempting to attach meanings, definitions, or indeed translations to a lexical item, we first summarize the various different syntactic and collocational patterns in which the word regularly participates.

How was this summary arrived at? The Oxford Hector Corpus was searched for all occurrences of the lemma **urge**. The matches were then painstakingly classified according to the contexts in which they occur. The

Lemma: **urge, urges, urging, urged**

PATTERN 1 (61%):
[PERSON$_i$] urges [PERSON$_j$] to-INF [DO]
PATTERN 1.1 (*passive*):
[PERSON$_j$] is urged to-INF [DO]

PATTERN 2 (15.4%):
2.1:
[PERSON$_i$ *or* SPEECHACT] urges [ACTION *or* ATTITUDE] ((up)on [PERSON$_j$])
2.2:
[PERSON$_i$ *or* CIRCUMSTANCE] urges [REASON_FOR [ACTION *or* ATTITUDE]] ((up)on [PERSON$_j$])

PATTERN 3 (4.3%):
[PERSON] urges that [CLAUSE]

PATTERN 4 (5.4%):
4.1 [QUOTE], urged [SUBJECT [PERSON]]
4.2 [QUOTE], [SUBJECT [PERSON]] urged
4.3 [QUOTE], [SUBJECT [PERSON$_i$]] urged [OBJECT [PERSON$_j$]]

PATTERN 5 (3.5%):
5.1 [PERSON$_i$] urges [PERSON$_j$] [A-DIR]
5.2 [PERSON] urges [STEED] [A-DIR]

Note 1: [CLAUSE] in pattern 3 often contains 'should' or a subjunctive.

Note 2: [A-DIR] in pattern 5 is generally realized as a prepositional phrase headed by *on, along, into, to, towards*, etc.

Note 3: [PERSON] includes groups, organizations, nations, and other such. The person or organization *urged* to do something is generally in a position of authority to take the action mentioned. The urger is, as it were, a supplicant.

Note 4: [QUOTE] in pattern 4 follows the general pattern of speech-act verbs: in 4.1, the 'urge' clause may also be embedded in the quote; in 4.2 and 4.3 it may also be embedded or it may come first as well as last.

Figure 2. A behavioural profile of "urge" (verb senses only)

most common pattern — a person urging another person to do something — accounts for 61% of the uses, while a person urging a steed or another person onwards accounts for only 3.5%. Such imbalance is typical in the distribution patterns of most polysemous words, although standard dictionaries give no hint of it, giving equal weight to all senses, even the rarest.

A couple of details concerning Figure 2 are worth highlighting. Firstly, the percentages do not sum to 100%. This is because about 10% of the uses are either what I call exploitations — metaphors, figurative uses, etc. — or simply unclassifiable on the available data. All interpretable uses of a word are either norms (unremarkable uses of one of the regular patterns in which it occurs) or exploitations (uses in which one or more aspects of the norm are flouted).

An example of an exploitation is the metaphor in the sentence "urging his Sierra through Grizedale Forest". This is, of course, in no way ungrammatical (it is perfectly natural and interpretable). But it would be wrong to classify a Ford Sierra as a member of the lexical set that includes horses. Canonically, for this sense of **urge**, psychological persuasion is implied. A rider who urges his horse over a jump tries to influence the animal's psyche. But a motor car does not have a psyche for the driver to influence. So it is better to regard *Sierra* as an "honorary" or "nonce" member of this set. Classifying it as a full member of the set would have disastrous consequences for the usability of such sets in lexical analysis: as more and more data piled in, they would become hopelessly broad and all-inclusive. For the most useful identification of the regularities on which successful communication is based, it is desirable to take the narrowest possible criteria for set membership.

Extensional and Intensional Criteria

So corpus analysts face an apparent paradox. On the one hand, there is a need for empirical analysis yielding a simple extensional listing of all the words fulfilling a particular clause role in relation to a given target word. It is often argued that it ought to be possible to do this by some simple computational procedure. But on the other hand, such an extensional listing would be full of odd and unusual uses; so some intensional criterion is required to separate the relevant set members from the noise. But what counts as noise, and how is relevance to be determined? How can we tell an odd and unusual use from a normal or typical one?

It seems intuitively clear that, in relation to **urge**, a horse is a very good example of a member of a relevant lexical set — call it STEED — while a Ford Sierra is not. But then what about camels? Intuitively, it seems equally reasonable to assert that camels are indeed steeds, though only one example

was found in the corpora I have examined. Is it worth adjusting the name of our set in order to indicate that it includes camels (and possibly one or two other rare but plausible items)? It would have seemed plausible enough to respond to the preponderance of the evidence by naming the set HORSE. However, this would have meant that camels would have been excluded, while Przewalski's wild horses would (wrongly) have been included. An intensional criterion (something like ridability or drivability as well as animacy) is needed to help us to sort the corpus data into relevant groups. The intensional criterion gives us grounds for including camels as bona-fide members of the set STEED, even though there may be few actual examples, while still excluding Ford Sierras (however numerous) on the grounds that they are not animate. The Ford Sierra in this context is, therefore, only an "honorary" or "ad-hoc" member of the set of STEEDS.

Rather than getting bogged down too early on in attempts to state intensional criteria on inadequate evidence (i.e., before starting an analysis), it may be better to adopt the following methodology:

- Use some statistical procedure, as proposed by Church, Hanks, et al. (1990, 1991, 1994) and Stubbs (1995), to identify statistically significant collocates of the target word, sorted as far as possible by clause roles.
- Sort this first list of collocates into relevant sets for meaning discrimination; devise approximate intensional criteria for set membership.
- Give each set a name (coined ad-hoc as a mnemonic, and bringing with it no theoretical baggage).
- Sort more concordance lines into groups, according to the intensional criteria just mentioned; extend the sets; refine the intensional criteria; refine the lexical-set mnemonics. Repeat indefinitely as new data becomes available.
- Note correlations among different sets in particular clause roles, with a view to specifying the meaning potentials of the target word.
- (Optionally) add a numeric value such as the number of occurences or (better) the number of different texts in which each set or pair or group of sets is found.
- Explain the relation of any ad-hoc set members to bona-fide set members by appealing to criteria of ellipsis, stylistics, rhetoric, metaphor, etc.

In the context of **urge**, the members of the set STEED in the British National Corpus are as follows:

a horse
his horse (X 8)
his large roan horse
his mount (X 5)
his pony (X 3)
the black stallion
various named horses (Chalon, Contralto, Fontana, Nero, Violet).

The Oxford Hector Corpus also contains a text mentioning "two stalling camels"; this text was not included in the main BNC.

Triangulation and Correlation

"Triangulation" (a land-surveying metaphor suggested by Ken Church) is often a good guide to sense discrimination. By looking at the correlation between two or more lexical sets in different clause roles, it is possible to pick out the contextually determined meaning potentials of a third word, *the target word*, and to say what the relevant contextual features are. (The term "meaning potential", as distinct from meaning as event, is discussed more fully in Hanks 1994.)

An example of triangulation is that the persuasion-to-physical-movement sense of **urge** is correlated with a direct object denoting a STEED plus an adverbial of direction: *on, onwards, forward, along, into the shallow water, up the path, down the rutted lane, through the desert*, or *up the slope*. By far the most common of these adverbials is *on*: the combination *urge on* is so close-knit that some would classify it as a phrasal verb. The small set of STEEDS with an adverbial of direction is closely paralleled by a more abstract or metaphorical sense, involving the much larger set of PERSONS with an adverbial of direction: *urging practitioners towards greater involvement, urging on my more sluggardly companions, urged the Party on,* and so forth.

This particular behavioural norm for **urge**, then, consists of two features, composed of two probabilistic lexicosemantic sets and a correlation between them. The verb **urge** complemented by (1) a word denoting a horse as direct object plus (2) an adverbial of direction (e.g., *urged Fontana up the path*) is in

systemic contrast with the verb **urge** complemented by (1) the name of a person or group as direct object plus (2) a *to* infinitive (*urged Stella to blot out the memory*). An intermediate pattern, consisting of the verb **urge** complemented by (1) the name of a person or group as direct object plus (2) an adverbial of direction is often somewhat metaphorical (e.g., *urging practitioners towards greater involvement*).

It is easy to construct more or less plausible-sounding counterexamples, e.g., *She urged Fontana to climb the path.* We cannot, therefore, argue that the correlations mentioned are necessary conditions determining meaning. However, what we can argue is that they represent norms of usage, associated with potential meanings. Moreover, a constructed counterexample may violate some condition of naturalness or textual well-formedness (see Sinclair 1984). Moreover, corpora also contain authentic utterances which violate norms of naturalness, either through performance error or for some rhetorical effect. Naturalness is still not well understood: until the advent of large corpora, there were no adequate means of studying it.

Collocations, too, can present misleading clues to the unwary. Great caution is called for in assigning patterns to interpretations. A simple example is that **urge** collocates regularly with two quite different uses of **on**. On the one hand, we have people urging steeds on, where the particle is intransitive, but on the other hand there is a pattern exemplified as "to urge a course of action on someone", where the particle is transitive and the interpretation is quite different. Collocation with *on*, therefore, is not a sufficient contextual clue for assigning a sense to **urge** in a text.

A Dictionary without Definitions

If we sort out the conventional uses of words in the way suggested here and start by compiling a "dictionary without definitions", we find that, for each word, a tiny number of patterns (in general, not more than a dozen) account for all conventional uses, which constitute a very high proportion of all uses (70 or 80%, often more). The remaining, less conventional uses can be placed in relation to one or more of the major patterns.

Since "conventional uses" in this context is a notion stipulatively defined by explicit criteria derived from corpus analysis, some very ordinary-sounding sentences may end up being counted as non-conventional. For example, the citation

The communiqué urged prudence

is related to the pattern "PERSON urge ATTITUDE" by virtue of the fact that communiqués are utterances by means of which persons express attitudes. This may seem like a painful way of restating the obvious. But we need to say precisely what the conventions of use are before we can say how they are used and exploited to create meanings. Other exploitations are more dramatic and complicated.

This brings us to another paradox. In recent years, it has been fashionable to stress the creative potential of language, due to the endless possibilities for combining and recombining words. The creative potential of language is undeniable, but the concordances to a corpus remind us forcibly that in most of our utterances we are creatures of habit, immensely predictable, rehearsing the same old platitudes and the same old clichés in almost everything we say. If it were not so, language would become unworkable. Humankind cannot bear very much creativity.

Presuppositions and Implications

Each of the lexicosyntactic patterns in which a word participates is associated with a meaning potential (as defined in Hanks 1994). Let us now turn briefly to these. Rather than citing traditional monolingual dictionary definitions, it is an interesting alternative to think in terms of presupposition and implication. In this connection, Anna Wierzbicka's comments on the meaning of **urge** (Fig. 3) are relevant.

Wierzbicka's analyses of meaning are couched in terms of a small number of words which she stipulates as undefinable semantic primitives, hence the rather unusual style. Her discussion, insightful as always, draws attention to several important points about **urge**. In saying that "it doesn't imply that the speaker has power over the addressee", if anything she understates the case; the corpus shows a best-example selectional preference among the direct objects of the active verb for governments and ministers — a far cry from horses.

One conclusion we might draw from Wierzbicka's comments is that a list of presuppositions and implications would be at least as interesting as a list of conventional definitions.

Meaning

I say: you should do X

I assume that you may not want to do it

I don't want to stop saying this because I want to cause you to feel that you have to do it

I say this, in this way, because I want to cause you to do it

I don't want you to think about it for a long time

Discussion

Urging is an attempt to get the addressee to do something.

Unlike *ask* and *request*, it doesn't imply that the speaker is seeking a benefit for himself.

Unlike *order* and *command* it doesn't imply that the speaker has power over the addressee.

[It is] pressing and forceful.

The speaker perceives or anticipates unwillingness on the addressee's part.

There is usually some sense of urgency. The speaker wants the addressee to respond and to respond **now**. Unlike the case of *command*, however, it is not necessarily an external action which the speaker wants. Rather, it is a psychological response.

Figure 3. Wierzbicka's account of "urge" (extracts)

Another (which she herself would like us to draw) is that **urge** has only one basic sense, not five or six as shown in standard dictionaries. This is a controversial claim which may not be resolvable by appeal to any objective criterion. Word meanings may be expressed as an increasingly delicate hierarchy of generalizations. The decision whether to lump or split senses — i.e., how to divide up the sets of collocates with which a word occurs — may in the end be no more than a matter of differences in the analysts' tastes.

Contextually Determined Default Interpretations

How does **urge** differ from its near synonyms? Wierzbicka contrasts it with **ask, request, order,** and **command**. Other near synonyms include **incite** and

Lemma: **incite, incites, inciting, incited**

One pattern with three variations:

1.1 [PERSON *or* SOMETHING] incites [PERSON] to-INF [DO [BAD]]

1.2 [PERSON *or* SOMETHING] incites [PERSON] to-PREP [ACTION *or* ATTITUDE [BAD]]

1.3 [PERSON *or* SOMETHING] incites [ACTION *or* ATTITUDE [BAD]]

lexical items in the set [DO [BAD]] found in this context include: *rebel, revolt, go on strike, assassinate, be naughty, break the law, commit [CRIME], go shoplifting, [VP] illegally, breach [NP]*

lexical items in the set [ACTION [BAD]] found in this context include; *riot, arson, debauchery, discord, dissention, denunciation, hatred, crime, lewdness, murder, trouble, unrest, violence, revolution, demonstration.*

Figure 4. A behavioural profile of "incite"

encourage. Figure 4 shows a behavioural profile for **incite**, together with examples of its indirect objects and infinitive-verb complements, as found in the British National Corpus. It will readily be seen that there is an intensional criterion uniting all the uses illustrated, namely that what you incite people to is an action and it is bad. This preponderance imposes a default interpretation on neutral terms. If I report that John incited Barry to speak to Astrid, the default interpretation is that by speaking to her, Barry broke some rule or did something nasty to Astrid. There is, of course, nothing intrinsically nasty in the semantics of **speak**, and it is not a necessary condition of inciting that it be bad. Rather, the implication that speaking, in this context, is disobedient or bad is inherited, in the absence of anything to the contrary, from the default interpretation of **incite**.

You *can* incite people to do good things, but it is now more normal to incite them to do bad things. 'Twas not ever thus. The most modern citation for **incite** in OED, dated 1880, is:

> A certainty, and an overflowing gladness in the heart, which are capable of inciting to heroic deeds.

This use will probably strike the modern reader as old-fashioned or unconventional, but almost certainly not as wrong or ungrammatical. The English word appears to have acquired its "bad" axiological value comparatively

recently (within the past century or so), so that Murray's definition of **incite** ("to urge or spur on; to stir up, animate, instigate, stimulate") is no longer wholly adequate.

There is clearly some sort of hierarchy in the language which allows one default interpretation to override another. It is not clear what the hierarchy is or how it works. What is allowed to override what? Are the hierarchies different for each lexical item, or are there general principles involved? They cannot be a simple matter of semantic density, as proposed by Wilks (1975):

> Word sense and structural ambiguity in natural language will always, in any system, give rise to alternative competing structures, all of which can be said to 'represent' whatever chunk of natural language is under examination. What I mean by 'preference' is the use of procedures, at every level of the system, for preferring certain derived structures to others, on the basis of their 'semantic density'.

Consider the following (made-up) examples:

1. The doctors did not treat all the patients in this hospital effectively.
2. The doctors did not treat all the patients in this hospital with respect.

What is the sense of *treat* in these two sentences? Semantic density assigns the sense "apply medical care or attention to" in both cases: correctly for 1, but incorrectly for 2.

Polarities and Semantic Prosody

Krzeszowski (1990) contrasts truth-conditional semantics with axiological, evaluative semantics. He claims that all lexical items are assessible on an axiological (good-bad) scale. While this sweeping claim goes too far (what is the axiological value of **twig** or **telephone directory**?), there can be no doubt that in the past a disproportionate amount of attention (as far as analysis of language in use is concerned) was paid to the true-false polarity, while other polarities were neglected. Evaluation on the good-bad scale, too, is an important component of lexical analysis.

A similar point is made by Sinclair (1991), who gives examples of what he calls "semantic prosody". For example, the phrasal verb **set in** normally has a negative semantic prosody: subjects of **set in** in his corpus include *rot, decay, malaise, despair, decadence, impoverishment, infection, prejudice,*

rigor mortis, numbness, bitterness, anarchy, and *disillusion.* But these are different kinds of BAD things from the set of BAD things to which people are incited. Bad things that set in are states of affairs, bad for the individual concerned; the bad things to which people are incited are actions, with bad effects for the rest of society.

At least in part, then, it is the negative axiological value or negative semantic prosody that distinguishes English **incite** from English **encourage** and **urge** and French **inciter**. So, if **incite** is axiologically negative, are we to conclude that either **urge** or **encourage** is positive? No; this would be an oversimplification, flying in the face of the evidence. The evidence of the British National Corpus suggests that **urge** and **encourage** have neither positive nor negative semantic prosody, but are neutral. We must look elsewhere for a relevant distinguisher between them.

You can **encourage** children to behave badly, if you are so minded, as idiomatically as you can encourage them to help the elderly (a good thing, I suppose) or to do well at school. On the other hand, if you **urge** children to behave badly, you are doing something much more perverse. This is because the choice of **urge** invokes the presupposition that the person being urged had no previous intention of doing the action in question — rather the reverse — while the person doing the urging feels justified in so doing. A wicked person urging children to behave badly would, therefore, be planting new ideas in their minds, while abusing the sense of moral righteousness implied by the choice of **urge**. Moreover, people who urge tend to lead from the rear: there is no implication that they will participate in the action that they are urging on others. You can urge people to do something heroic or dangerous, and then stand back and watch.

An encourager, on the other hand, tends to lead from the front, or at least, be mixing in there with his addressees. There is a presupposition that, if I encourage you to do something, you are already well on the way to having committed yourself to doing it, without my encouragement. Moreover, while urging and inciting are volitional, encouraging is not. Circumstances, states of affairs, and sunny weather give encouragement, as well as human agents. More importantly, **encourage** is often found in the passive and in the infinitive (and so without any mention of an agent at all). The focus of **encourage** is on the resultant state of mind of the patient, rather than on any action by the agent.

Now, what happens *after* the act of incitement or encouragement? Here again there are differences in the implications. If I say that James **incited** Bill

to do something, I imply that Bill did it. **Encourage** and **urge** have no such implication.

It is not always possible to see any direct evidence for presuppositions and implications in a set of concordances. If the evidence is there in the texts at all, it may take a distant or indirect form (for example, in the continuation of the story in subsequent paragraphs), rather than being visible in syntactic or collocational form in the immediate vicinity of the target word. It may, therefore, be preferable to approach teasing out such implications as a matter of identifying mutual beliefs by the traditional techniques of introspection and comparison of intuitions, rather than through computational analysis of texts. We must beware of allowing our euphoria over having sufficient evidence to observe patterns of linguistic behaviour to blind us to the possibility that there may be other distinctions — distinctions of belief, for example — that are equally important, but must be sought by other means than corpus analysis.

Projecting Meaning Potentials onto Syntax

I have been arguing that the meaning potential of each verb (if not of other word classes) is determined, at least in part, by the totality of the contexts in which it occurs. It is, therefore, necessary to show how the meaning potential of a verb projects onto the syntactic patterns with which it is associated. A finite verb, the pivot of a clause, has certain grammatical structures — subject, object, complement, and adverbial — associated with it, structures which are in turn linked to the verb's meaning. But to account for the different inferences which native speakers draw from the same word in different contexts, a more delicate subcategorization is required.

So, for example, the meaning of the verb **bank** differs depending on its transitivity. But we also need to say something in the syntax about the semantic type of its subject and object. An aircraft banks (intransitive); people bank money (transitive); a pilot banks an aircraft (also transitive, but the semantic type of *aircraft* is very different from the semantic type of *money*). These two facts (the verb's transitivity patterns and the semantic types of its arguments) determine the way in which we interpret it.

The patterns just mentioned may be exemplified in the following four sentences taken from the British National Corpus. A few relevant fragments of parsing (clause roles) and tagging (wordclasses) are shown, including relevant semantic types.

1. [SUBJ Jani [PERSON]] banked-VBD [OBJ £60,000 [MONEY]] through successful libel actions against Options magazine and the London Evening Standard.

2. . . . [SUBJ she [PERSON]] is believed to have banked-VBN [OBJ £10_million [MONEY]] since being booted out of Downing Street two years ago.

3. [SUBJ The plane [AIRCRAFT]] banked-VBD [NO OBJ], and he pressed his face against the cold window.

4. [SUBJ I [PERSON]] banked-VBD [OBJ the aircraft [AIRCRAFT]] steeply and turned.

To account for sentences such as these, the relevant parts of the accompanying entry for *bank* in a formal dictionary entry, showing how the meaning potential projects onto the syntax, would be something like this:

5. [SUBJ[PERSON]] ___ [OBJ[MONEY]]
= deposit or invest MONEY in a bank or other financial institution for safe keeping

6. [SUBJ[AIRCRAFT]] ___ [NO OBJ]
= raise one wing higher than the other in order to change direction

7. [SUBJ[PERSON]] ___ [OBJ[AIRCRAFT]]
= cause AIRCRAFT to raise one wing higher than the other in order to change direction

Semantic types remain to be identified and listed explicitly, in the form of lexical sets. If an accurate a-priori description of semantic types were possible, then we might expect that the semantic types of a language would be as familiar to us as the well-established part-of-speech classes: verb, noun, adjective, etc. In a sense, they are: joking aside, we all know, informally, that the expression "Bill Clinton" falls into the class PERSON. But a satisfactory formal account of such classes is not yet available. Preliminary empirical work suggests that all a-priori assumptions are suspect. For example, the class PERSON seems plausible enough, but it turns out to be unsatisfactory. As a matter of syntax, it may work better if divided into two classes: defined, on the one hand, by properties which Bill Clinton shares with cats, horses, and monkeys, such as eating, sleeping, and climbing (i.e., the type ANIMAL), and on the other hand by properties which he shares with nations, governments, business organizations, family-history societies, and computers (which we

might call COGNITIVE), namely analysing, negotiating, banking money, making statements, expressing sympathy, and so forth. The details are not at all clear, so for present purposes it is no less unsatisfactory to use PERSON.

Identifying Sets

Identifying just those features of a lexicosyntactic environment that are relevant to making meanings is, as we have seen, no easy task. In the words of Firth (1950):

> We must separate from the mush of general goings-on those features of repeated events which appear to be part of a patterned process.

Moreover, there is no reason to believe that the procedures that help us to use and understand verbs are relevant to, say, nouns or adverbs. Indeed, the interplay of semantics and syntax is different for different verbs. At one extreme, we find verbs which are syntactically complex, although they may be semantically quite straightforward. A case in point is **bother** (Fig. 5). Here, the syntax is quite complicated, but not much hangs on the semantic type of the nouns that occur in the different syntactic roles. Things bother people; people can't be bothered (or don't bother) to do something. It makes little or no difference to the meaning of **bother** what kind of action they couldn't be bothered to do, or what kind of person couldn't be bothered to do it. The meaning of **bother** in "did not bother to do something" and in "couldn't be bothered to do something" is very similar, although the strange thing is that the passive isn't really passive at all: both the active and the passive pattern come down to much the same thing: "did not (or would not) take the trouble to".

Presence or absence of the negative is the most relevant aspect of the interpretation of this verb. It changes the meaning quite dramatically. So, for example, an issue which bothers journalists is not an issue which takes trouble over them. Conversely, "Ava was bothered by my infidelity" is truly passive; it means that my infidelity had a particular effect on Ava (whereas "Ava couldn't be bothered to notice my infidelity" does not mean that my infidelity could not have the effect in question on Ava).

All of these points relevant to the interpretation of **bother** are at a high level of syntactic structure. At the other extreme lie verbs such as **abandon** (Fig. 6) where the syntactic patterning is extremely simple. Little more can be

1. [PERSON] (not) **bother** to [DO];
[PERSON] (not) **bother** [DOing]

They did not bother to vote. | *If you think I have no chance, I won't bother risking disappointment.*

1.1 [PERSON] **can't be bothered** to [DO]

The calculations are so tedious that theorists can't be bothered to do them.

2. [PERSON] (not) **bother** about [SOMETHING/SOMEONE];
[PERSON] (not) **bother** with [SOMETHING/SOMEONE]

He did not bother about the noise of the typewriter because Danny could not hear it above the noise of the tractor. | *Don't bother about us.* | *The only rocks worth bothering with are 'hard' rocks.* | *Who bothers with such folk?*

2.1 [PERSON] (not) **be bothered** about [SOMETHING]

The only thing I'm bothered about is the well-being of the club. | *I'm not particularly bothered about how I look.*

2.2 [PERSON] (not) **be bothered** with [SOMETHING]

Mrs Sprott didn't want to be bothered with such things.

3. [PERSON] **be bothered** by [PROBLEM];
[PERSON] **be bothered** that [CLAUSE];
[PROBLEM] **bother** [PERSON]

Ava was bothered by my infidelity. | *Poor Philip was particularly bothered by this symptom.* | *Neal Grossman is bothered that the press has dubbed this 'California cuisine'.* | *It is an issue which bothers journalists.* | *all hot and bothered.*

4. [PERSON/ANIMAL] (not) **bother** [PERSON];

She decided not to bother him now with the Russian material. | *I'm sorry to bother you at work.* | *I realize this has been a blow to you, Mr Pertwee: we won't bother you much longer.*

5. [FACT] (not) **bother** [PERSON]

Looks don't bother me; it's personality that counts.

Noun uses:

Changing it would have been too much bother. | *not worth the bother* | *unless one goes to the bother of fetching a ladder ....*

I think you're in a bit of bother. | *a spot of bother that Lloyd George had in 1911.* | *bother boys; bother boots.*

Figure 5. A behavioural profile of "bother"

Lemma: **abandon, abandons, abandoning, abandoned**

COMPLEMENTATION: transitive verb with NP as object, occasionally with PREP 'to':

abandon [PERSON or THING] (to [ABSTRACT [BAD] or CREATURE [WILD [BAD]]])

PRESUPPOSITION: The abandoner formerly had possession, use, or control over the thing abandoned and found it useful or pleasing.

IMPLICATION: The abandoner has relinquished possession, use, or control of it and has gone elsewhere, either a) because of force majeure, or b) because he/she no longer wants/values it, or c) because he/she can no longer look after it.

Types of direct object, with examples:

abandon a *person* (one's wife, children, new-born baby, lover, client, ... a political party, the working classes) or *pet* (dog, cat, ...)

a *vehicle* (which is, of course, a kind of physical object; car, van, lorry, bicycle, motorbike, ship, ...)

any of various other sorts of *physical object* (a gun, hammock, sofa, rhubarb tart, school clothes, piano, luggage, ...)

a *place* (a stronghold, a military position, the road, a house, home, Birmingham, Warsaw, a factory, one's country, a backwater, ...)

an *abstract thing* (one's principles, position, a belief, idea, policy, theory, plan, system; apretence, certainty, hope, caution, commitment, control, ...)

a *procedure* (which is itself a kind of abstract thing: a debate, discussion, experiment, attempt, practice, a career, a way of life, the armed struggle, a race, an event, a match, a meeting, ...)

Figure 6. A wide range of lexical sets: the case of "abandon" (verb senses only)

said than that it is a transitive verb, occasionally with a 'to' prepositional phrase. However, the range of direct objects is remarkably wide. The question arises, to what extent different meaning potentials of this verb arise out of combinations with different semantic types of direct object or indeed individual lexical items (and this is an especially difficult question when there are so many of them). Equally difficult to answer are questions about how the various direct objects should be grouped into different sets in order to identify different meaning potentials. So, for example, it may seem plausible to argue a priori that a ship is a kind of vehicle, as are a car and a van. However, the

inferences that may legitimately be drawn from the use of *abandoned* in fragments 1 and 2 below (for example, inferences about lifeboats and the danger of death by drowning) are quite different from those that may be drawn from 3. (All fragments are drawn from the British National Corpus.)

1. Julian Russell was working on the Greek cruiser Oceanos when it went down in the Indian Ocean. He helped man the bridge and co-ordinate the rescue operation after the crew allegedly abandoned ship.

2. And any prudent owner would carry a sufficiency of lifebelts and life-jackets for the passengers and crew. I can even see two lifebelts in front of the bridge. But they haven't done the obvious thing and abandoned ship.

3. Sniffer dogs, brought in to follow the men after they abandoned the vehicle in a ditch, tracked them to Laughterton, near Gainsborough, Lincs.

On the other hand, the inferences that may legitimately be drawn from 4 and 5 (for example, about criminal or antisocial behaviour) are remarkably similar to those that may be drawn from 3.

4. A spokesman said it appeared the Escort XR3i collided with parked vehicles at 9pm on Monday and the driver then abandoned the car.

5. The thieves abandoned the van and its load at the scene and escaped in a Saab car.

It seems then that, in relation to **abandon** at least, **car** and **van** are better members of the set **vehicle** than is **ship**. (The difference is of course underlined by the syntactic oddity — no determiner — of the phrase **abandon ship**.)

This problem of inferencing with the aid of real-world knowledge is only too well known in artificial intelligence and philosophy of language. Although the details of each inference lie outside the scope of corpus analysis, the general possibility of each such cluster of inferences is one factor that the analyst must take into account when grouping corpus uses together as evidence for a word's potential meaning. The problem of grouping is one of relevance. We cannot rule out the possibility that, when described empirically, the relevant set of semantic types for each verb in a language will turn

out to be slightly different from those relevant to every other verb. If this question is investigated empirically, it is likely that at least some gross overlaps will be discovered, e.g. that there are many features common to the direct objects of, say, causative verbs of motion or verbs of perception. These overlaps remain to be established as an empirical fact. One thing is already clear, however: the sets of semantic types are extremely fuzzy. If they can be identified at all, set membership will have to be stated in terms of similarity to a contextually determined prototype.

It is, incidentally, hard enough to group into sets those nouns that actually do occur with reasonable frequency in particular clause roles in relation to each verb, without agonizing about those that do not. Questions such as "Is a hot-air balloon a vehicle?" are not relevant to the corpus analyst working on **abandon** unless two or more examples of abandoned balloons are found.

Notwithstanding the manifold difficulties that lie ahead, the task of accounting empirically for the different paradigmatic sets associated with each verb deserves a high priority for many reasons: distinguishing one word sense from another will improve our understanding of the lexical component of natural language; sharpen up the sense categorizations and the wording of explanations in our dictionaries; and last but not least, contribute to improvements in the many elusive tasks of man-machine interface.

Conclusion

In this paper, I have drawn attention to some of the benefits for lexical analysis of using an electronic corpus by way of evidence. Natural languages are more regularly patterned, syntagmatically, than has been recognized in the past, and corpus analysis can help us to see the patterns for what they are. But the patterns do not spring, untouched by human hand, fully fledged from the corpus. They have to be teased out, often painstakingly and slowly. Procedures have to be developed for distinguishing relevant features from mere noise. Appropriate levels of generalization have to be chosen at every step.

Frequency, too, is important. Just as, in any natural language, a few words are very very frequent while others are quite rare, so also one or two uses of individual words are extremely common, while others are much less signficant. An analysis that gives comparative frequencies, varying perhaps

by text type, will bring added value for users of many different kinds.

Different meanings and different nuances of meaning are dependent to a large extent on the context in which the word is used. The unique contribution that an individual word brings to a text is ordinarily something quite broad and general, not something precise in fine detail. Precision and subtlety of interpretation arise, rather, out of the combinations of contexts in which the word participates. This makes it all the more extraordinary that dictionaries, with their multiple word senses, have paid so little attention to the contextual features associated with different senses. Evidently, lexicographers of the past did not feel that they had sufficient evidence to make the sort of generalizations about words in use that corpus analysts are now beginning to make.

To the corpus analyst, a list of mutually exclusive choices, all given equal weight, for word-sense disambiguation no longer seems an appropriate model. Rather, a model of words in use must show how the totality of the patterns in which each word regularly participates contribute to its meaning on any particular occasion when it is used, with more or less subtle changes of emphasis. The more frequent patterns, of course, have a more profound effect, both at a subconscious psycholinguistic level and at a sociolinguistic level, while infrequent patterns are less influential. We might even speak of "dominant" and "recessive" features in a lexical description, especially as historical corpora are developed, enabling us to measure changes in frequency over periods of time.

Important features of word meaning also include presupposition and implication. Here there is a much more subtle challenge for the corpus analyst. Although it is normally possible to secure agreement among a group of native speakers about the presuppositions and implications associated with individual words, it is by no means clear how such information might be teased out of a corpus by a human analyst, still less computed on the basis of textual evidence.

A great deal of work remains to be done in understanding how meanings are related to uses, a task in which corpus analysis will inevitably play a central role. For example, we need more effective procedures for distinguishing relevant from irrelevant features of linguistic environments. And we need to sort the lexical items that make up the relevant features into lexical sets, using intuitively plausible intensional criteria.

Note

The research on which this paper is based was done in 1992-3 as part of the Digital-Oxford Hector Project, and I would like to thank the staff of Digital's Systems Research Center in Palo Alto, especially Loretta Guarino, Bob Taylor, and Mary-Claire van Leunen for making it possible. I would also like to thank Sue Atkins, Lucinda Coventry, Valerie Grundy, and John Sinclair for comments on earlier drafts of this paper. In addition, I would like to thank Professors Barbara Lewandowska-Tomaszczyk and Marcel Thelen and their fellow organizers of the Łódź-Maastricht Duo Colloquium on Meaning and Translation (1995) for giving me the opportunity to present these ideas publicly and for forcing me to put my thoughts in order. An earlier version of the first part of this paper appeared in the proceedings of that Colloquium.

Bibliography

Church, K. and P. Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics* 16(1).

Church, K., W. Gale, P. Hanks, and D. Hindle. 1991. "Using Statistics in Lexical Analysis". In U. Zernik (ed.): *Lexical Acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Church, K., W. Gale, P. Hanks, D. Hindle, and R. Moon. 1994. "Lexical Substitutability". In B. T. S. Atkins and A. Zampolli (eds.): *Computational Approaches to the Lexicon.* Oxford: Oxford University Press.

Firth, J. R. 1950. "Personality and Language in Society". In *The Sociological Review*, xlii. Reprinted (1957) In *Papers in Linguistics 1934-1951.* Oxford: Oxford University Press.

Firth, J. R. 1957. "A Synopsis of Linguistic Theory 1930-1955". In *Studies in Linguistic Analysis.* Philological Society, Oxford. Reprinted in Palmer, F. (ed.) 1968. *Selected Papers of J. R. Firth.* Harlow: Longman.

Hanks, P. 1994. "Linguistic Norms and Pragmatic Explanations, or Why Lexicographers need Prototype Theory and Vice Versa". In F. Kiefer, G. Kiss, and J. Pajzs (eds.) *Papers in Computational Lexicography: Complex '94.* Research Institute for Linguistics, Hungarian Academy of Sciences.

Krzeszowski, T. 1990. "The Axiological Aspect of Idealized Cognitive Models". In J. Tomaszczyk and B. Lewandowska-Tomaszczyk (eds.) *Meaning and Lexicography.* Amsterdam and Philadelphia: John Benjamins.

Sinclair, J. M. 1984. "Naturalness in Language". In J. Aarts and W. Meijs (eds.) *Corpus Linguistics.* Amsterdam: Rodopi.

Sinclair, J. M. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Stubbs, M. 1995. "Collocations and Semantic Profiles." *Functions of Language* 2.

Wierzbicka, A. 1987. *English Speech Act Verbs.* New York: Academic Press.

Wilks. Y. A. 1975. "A Preferential, Pattern-Seeking Semantics for Natural Language Inference". *Artificial Intelligence* 6.

Zadeh, L. 1965. "Fuzzy Sets." *Information and Control* 8.