Corpus Pattern Analysis

Patrick Hanks
§Brandeis University and *Berlin-Brandenburg Academy of Sciences
*Berlin-Brandenburgische Akademie der Wissenschaften,
Jägerstrasse 22-23, Berlin 10117, Germany.
hanks@bbaw.de

§Department of Computer Science,
Volen Center for Complex Systems,
Brandeis University,
Waltham, MA 02454, USA.
patrick@cs.brandeis.edu

Abstract
Evidence from large corpora shows striking patterns of word use in natural language, the details of which are only now beginning to be adequately recognized and studied. These patterns of usage can be analysed and applied in lexicography as a way of deciding what counts as a lexical meaning distinction and of showing how different meanings are associated with different uses of a word. This has major implications for dictionaries, as well as for lexicons used in computational natural language processing, but lexicography has been slow to respond to the challenges presented by the data. After a discussion of different kinds of corpus evidence and analytic procedures in corpus lexicography, the paper presents a new project of corpus-driven lexicographic analysis of English.

1. Introduction

Corpus Pattern Analysis (CPA) is a new technique for mapping meaning onto words in text. It is based on the Theory of Norms and Exploitations (TNE, see Hanks forthcoming (a) and (b)). TNE in turn is a theory that owes much to the work of Sinclair and Halliday on the lexicon (e.g. Sinclair 1966, 1987, 1991; Halliday 1966), to the Cobuild project in lexical computing (Sinclair, Hanks, et al. 1987), and to the Hector project (Atkins 1993; Hanks 1994). Some recent work in American linguistics (Jackendoff 2002) has complained about the excessive 'syntactocentrism' of American linguistics in the 20th century. TNE offers a lexicocentric approach, with opportunities for synthesis, which will go some way towards redressing the balance.

The focus of the analysis is on the prototypical syntagmatic patterns with which words in use are associated. Patterns for verbs and patterns for nouns are different in kind. Noun patterns consist of a number of corpus-derived gnomic statements, into which the most significant collocates are grouped and incorporated. Verb patterns consist not only of the basic 'argument structure' or 'valency structure' of each verb (typically with semantic values stated for each of the elements), but also of subvalency features, where relevant, such as the presence or absence of a determiner in noun phrases constituting a direct object. For example, the meaning of ***take place*** is quite different from the meaning of ***take his place***. The possessive determiner makes all the difference to the meaning.

No attempt is made in CPA to identify the meaning of a verb or noun directly, as a word in isolation. Instead, meanings are associated with prototypical contexts. Concordance lines are grouped into semantically motivated syntagmatic patterns. Associating a 'meaning' with each pattern is a secondary step, carried out in close coordination with the assignement of concordance lines to patterns. The identification of a syntagmatic pattern is not an automatic procedure: it calls for a great deal of lexicographic art. Among the most difficult of all lexicographic decisions is the selection of an appropriate level of generalization on the basis of which senses are to be distinguished. For example, one might say that the intransitive verb ***abate*** has only one sense ('become less in intensity'), or one might separate ***storm abate*** from ***political protest abate,*** on the grounds that the two contexts have different implicatures. That is a simple example, but in more complex cases (e.g. the verb ***bear***) patterns are indispensible for effective disambiguation. ***Bearing a heavy burden*** is a pattern that normally has an abstract interpretation in English (as opposed to, say, ***carrying a heavy load***), and the meaning is associated with the prototypical phrase, which is quite different in turn from ***I can't bear it.*** In section of this paper, I show examples of how very specific expressions can be generalized while preserving the character of the pattern.
In CPA, the 'meaning' of a pattern is expressed as a set of basic implicatures (e.g., for the verb ***file***: "If you ***file*** a law suit, you are acting as the plaintiff and you activate a procedure by which you hope to obtain redress for some wrong that you believe has been done to you"). For other applications, it may be expressed as a translation into another language, or a synonym set (e.g. "activate, start, begin, lodge".)

2. Corpus Data: Written, Spoken, and Other

Nowadays, it is inconceivable that a major new dictionary should be planned without making some kind of use of corpus data. Indeed, many older dictionaries, compiled in the days before large corpora became available, have been extensively revised in the light of corpus evidence. Only in America are major dictionaries still published without taking account of —or even claiming to take account of —corpus evidence. Relevant questions now are: what kind of corpus and what kind of use?

Corpus evidence is so plentiful that it is now possible to be selective about it (and its relevance to the project in hand) before starting out on a new lexicographical enterprise. A catchphrase in computational lexical analysis of the 1980s was, "More data is better data." In those days it was considered premature by some computational linguists and speech engineers to ask, "What kind of data?" (although that did not stop other people asking it). Now, twenty years on, in the age of the Internet, we can afford the luxury of being more selective. It seems reasonable to suggest that, for principled as well as practical reasons, general lexicographic projects should focus on the evidence of corpora of printed, published non-specialist texts, and that other corpora (spoken data, chat-room data, email data, domain-specific corpora, etc.) should not be thrown into the general melting pot, but should be built into separate subcorpora which can be related systematically to the evidence of printed texts. The first principled reason for this has to do with the competence/performance distinction. It is easy to dismiss evidence from spoken utterances that do not conform to a theoretical prediction as so-'performance errors'. But a text that has been written down, checked, presented for publication, and re-checked is more likely to represent someone's deliberate

utterance. It is therefore less easy to dismiss its idiosyncrasies as performance errors. If idiosyncrasies are found, some other explanation, consistent with the utterer's linguistic competence, must be sought for them. As we shall see, there is ample evidence to suggest that writers do not merely rehearse their basic linguistic competence, but also sometimes exploit that competence in order to say new and interesting things, or to say old things in a new and interesting way. The distinction between norm and exploitation is essential for effective lexicographic processing of corpus evidence.

Cognitive linguists and others wishing to study how human beings go about exploiting their linguistic competence in order to make meanings (often cooperatively, in conversation with others) will, obviously, want to study the evidence of spoken corpora. Here, some important distinctions must be made. Spoken corpora, as currently conceived, can be problematic. In the first place, some texts currently included in so-called "spoken corpora" are not really spoken at all. If a person writes down a text and then reads it out as a lecture, or if a TV news editor prepares an autocue which a newsreader then reads out, the result can hardly be classified as a genuinely spoken text. The mere fact of spoken utterance is irrelevant. Therefore, as corpus linguists have pointed out, what is required for investigation of the processes of making meanings is a corpus of undiluted transcripts of spontaneous, unscripted conversations. Such a goal, while fascinating, is far removed from the usual run of lexicographic goals. It is hard to see any role for lexicography in interpreting the hesitations and maintenance strategies of normal conversation.

Even more confusing are corpora of chat-room data and emails. Here, there is a genuine role for lexicography, albeit a minor one, in interpreting the myriad new conventions (mostly acronyms and abbreviations) that are constantly emerging. This role is similar to the role that might be played by lexicography in interpreting the conventional abbreviations of small ads in newspaper columns. A particular problem in analysing the language of emails and chat-rooms is that it is never quite clear how spontaneous they are. Some people (the present writer included) agonize long and hard before pressing 'send', often obsessively polishing and rewriting crucial passages and thus obliterating any trace of spontaneity. Others splurge out their stream of consciousness without a moment for second thoughts. Most people, no doubt, are somewhere in between. This means that it is hard for the analyst to know exactly what is spontaneous and what is not.

A different problem arises with specialized domain-specific texts. Of course, every text is about something, and to that extent it is domain-specific. But a great deal of written language nowadays is devoted to specific scientific, technological, business, and sporting topics. Domain-specific conventions fade imperceptibly into the common core: there is no hard and fast dividing line. There is a gradual cline from technical to general register, with no very obvious cut-off point. The uses of ***take*** in examples such as 1 and 2 are both domain-specific and part of the general convention of English.

1. … transactions where other EC companies took over UK companies
2. Six minus three. … You take the smaller digit from the larger digit.

3, 4, and 5, however, are more highly technical. It can be argued that they have no place in a small general dictionary's account of the meaning and phraseology of the verb ***take***.

3. The company took a charge of $10 million against inventory …
4. A fleet under vice-admiral Sir George Byng was sent to take station off Dunkirk.
5. … a more than useful total on a pitch already taking spin.

For reasons such as these, it is not desirable to include highly technical texts in a general corpus. Obviously, jargon such as that in 3, 4, and 5 can be found in any large general corpus, but distinguishing between domain-specific texts and general texts can at least prevent the jargon becoming overwhelming. Domains have their own conventions, and these must be recognized lexicographically as such in a systematic way.

3. Interpreting the Evidence

A central task for 21st-century lexicographers is to provide a descriptively adequate account of how the words of a language are actually used (or, for historical lexicographers, how the words were used in the past and how conventions of usage have changed over time). This task has been seriously neglected up till now, for a variety of reasons. Until 15-20 years ago, the main reason was lack of evidence, coupled with scepticism as to whether the patterns existed at all. There simply was not enough data on any given word for lexicographers to be able to say how a word is normally used. Now, the problem is lack of generally accepted theoretical guidelines for recognizing and representing the patterns. Corpus evidence shows that most words are strongly associated with a very small number of particular patterns of usage, but that the number of possible usages of each word is extremely large. Lexicographers are reluctant to abandon the notion that their duty is to represent all possible uses of a word, rather than normal and typical uses.

Consider the verb ***hazard***. Corpus-based dictionaries generally show ***hazard a guess*** as an example of use of this verb; some pre-corpus dictionaries do not even mention it. But not even the corpus-based dictionaries mention that (if the corpus evidence is to be believed) over 40% of British usage and nearly 80% of American usage prefer this noun as the direct object. How is the user to know that this collocation is specially privileged? ***Guess*** is the prototypical direct object of the verb ***hazard***, and most other uses of it (*hazard an opinion, hazard a conjecture, hazard a definition*), including uses as a reporting verb, derive their meaning by analogy to this pattern. Facts such as these are of the greatest importance both for natural language generation (both by computer and by foreign learners) and for natural language understanding. If a reader or computer encounters "hazard an [UNKNOWN]", it is more likely that [UNKNOWN] is some kind of guess than anything else. Dictionaries fudge issues like this, by trying to allow for rare but conceivable possibilities, as opposed to the evidence that exists. Many dictionaries still try to construct definitions as if they were statements of necessary and sufficient for word meaning, innocent of the fact that philosophers of language from Wittgenstein to Putnam have long since exploded the notion.

You can't blame the lexicographers: they are governed by market forces, and the market has been conditioned over the centuries to expect definitions, in terms of "all and only" and "necessary and sufficient". What is needed is a culture shift in every aspect of the language learning, language teaching, and language using community (i.e. everybody), so that people accept that word meaning is governed not by necessary and sufficient conditions but by the much more powerful process of analogy to a prototype (Fillmore 1975; Hanks 1994). Every time a learner asks "Can you say X in English?" the teacher should formulate the answer in terms of what is normal rather than what is possible.

The same procedure can be used to distinguish different senses of words. Again, a simple example: consider the verb ***toast.*** There are at least two senses of this verb: (1) "cook food by exposure to a grill or fire" and (2) "raise one's glass and drink in honour of someone or something". What the dictionaries do not say is how to recognize the difference. It is assumed that human beings know this. In fact, since food is something, the definition "raise one's glass and drink in honour of something" could in theory apply to food as well as to any of various other things. It is *possible* to imagine talking about toasting a piece of bread and meaning raising one's glass and drinking in honour of a piece of bread. There is nothing in the syntax to stop one. But no sane person would say such a thing. This is of course absurd – if you already know the meaning of the verb. But if you are a foreign learner or a computer program, it is potentially confusing. What is needed for non-native speakers and computational applications alike is a dictionary that lists prototypical lexical sets: bread, muffin, and bagels on the one hand and people (including bridesmaids and each other), people's successes and victories, their health and their future if alive, and their memory if dead, on the other hand. An important question for corpus analysis is how to represent such lexical sets. On the one hand, it seems perfectly sensible simply to list two or three members of the set of foodstuffs that are normally toasted, forming the core of an analogical computer program that uses cluster analysis to pick up other toastable foodstuffs (crumpets, for example) and distinguish them from bridesmaids and victories, thus distinguishing between the two meanings of the verb toast. A list of prototypical words is more useful for this purpose than a semantic type such as [[Food]], in part because there are many foods that are never toasted. On the other hand, it would obviously be impossible to list prototypical [[Event]]s, or to specify the prototypical words and names that denote people. These can only be represented as semantic types, not as paradigm sets.

With all this in mind, a new kind of dictionary is being developed at Brandeis University in Waltham, Massachusetts: one that focuses on usage, rather than meanings. The aim is to create an inventory of prototypical usage patterns by corpus pattern analysis (CPA). The patterns represent the normal syntagmatic behaviour of each word, but they are semantically motived. That is to say, if a word has more than one meaning, we ask how anyone can tell one meaning from another. The answer usually (but not always) lies in the immediate context surrounding the word in question.

The methodology of CPA is to extract from the corpus a concordance for each target word, scan it to get a general overview of the word's behaviour, then select a random sample of between 200 and 1000 concordance lines for detailed analysis. In the course of detailed analysis, concordances lines are sorted into groups that have approximately the same

meaning and similar syntactic structures. Semantic values are given for the arguments or valencies of the target word in each group. Methodological discipline requires that every line in the random sample should be classified. The classifications are:

Norms
Exploitations
Alternations (e.g. 'achievements' alternating with 'people')
Names (***Midnight Storm*** is the name of a racehorse)
Mentions (to mention a word is not to use it; the syntagmatics are different)
Mistakes (learned is sometimes mistyped as leaned)
Unassignables.

Exploitations include metaphors and other non-normal uses, e.g. metonymy, as 6. Normally, one takes documents, not information out of a filing cabinet.

6. She was standing at a filing-cabinet taking out more information

A powerful aid in doing CPA is the "Waspbench Word Sketches" program of (Kilgariff and Tugwell 2001). This exploits the concept of mutual information as a measure of statistical significance (Church and Hanks 1989), listing the words that are most associated (in terms of statistical significance) with the target word in different clause roles (subject, verb, object, adverbial, etc.). It does not, however, group together the different clause roles into contrasting sense groups (a process called 'triangulation' in Church et al. 1994). So for example it shows that both patient and woman are significant direct objects of the verb treat, while both with respect and with antibiotics are prepositional phrases significantly associated with treat. What it does not do is to group these to show that 'these patients are treated with antibiotics' activates one meaning of treat, while 'as sisters and daughters women are treated with respect' activates a different meaning of treat. Elucidating the relationship between syntagmatic patterns and activated meanings is one of the goals of CPA.

4. Verb Entries in CPA

The procedure for corpus pattern analysis of verbs is very different from that for nouns, and the results look quite different too. We start with verbs because the verb is the pivot of the clause and there is some reason to believe that the patterns for many nouns will start to fall into place semi-automatically (i.e. with the aid of an interactive computer program) once the verbs have been correctly analysed.

Verb patterns for English can be stated in the form of free text (exploiting the heavy reliance of English on word order to distinguish subject, object, and adverbial), but for purposes of practical use in natural language computing they are slotted into a template with provision for at least following clause-role components:

Subject
Object
Subject-Complement
Object-Complement
Adverbial
Clausal

Provision is also made for significant sub-valency features: for example, the presence or absence of a determiner can radically change the sense (compare an event took place with someone took someone else's place). Examples of verb patterns are shown in Figure 1.

The numbers denote frequencies in the British National Corpus. Eventually, numbers will be added showing the comparative frequency of each pattern. It is important to emphasize that this is preliminary work in progress; many changes may be expected before completion.

A note on the conventions in figure 1: Double square brackets indicate semantic types. Within square brackets, a semantic role is sometimes indicated after an equals sign. Curly brackets (braces) indicate specific lexical items and are also used for phraseological grouping. Round brackets (parentheses) indicate optionality: items that may not be present at all but which, if present, provide an important clue to the meaning of the verb.

abandon/V BNC FREQ: 4063

1. [[Person]] abandon [[Process]]
2. [[Person]] abandon [[Abstract]]
3. [[Person]] abandon [[Location]]
4. [[Person]] abandon [[Artefact]]
5. [[Person 1]] abandon [[Person 2]]
6. [[Person]] abandon [[Animate = Pet]]
7. [[Person]] abandon [[Process]]
8. [[Person]] abandon [[Self]] {to [[Sensation]]}

abase/V BNC FREQ: 17

[[Person 1]] abase [[Self]] ({before [[Person 2]]})

abdicate/V BNC FREQ: 127

1. [[Person 1 = Monarch 1]] abdicate [NO OBJ] ({in favor of [[Person 2 = Monarch 2]]})
2. [[Person]] abdicate {responsibility} ({for [[TopType]]})
3. [[Person]] abdicate [[Role]]
4. [[Person]] abdicate {from [[Role | Action]]}

abide/V BNC FREQ: 313

1. [[Person]] abide by [[Rule]]
2. [[Person]] {cannot abide} [[TopType]]

NOTE: A number of 17th-century syntactic norms, not listed here, are still used in Legal and Religious (Christian) domains.

accede/V BNC FREQ: 237

1. [[Person1]] accede {to [[Person2]]'s [[SpeechAct]]}
2. [[Person1=Monarch]] accede ({to throne})

accent/V BNC FREQ: 17

1. [[Person={Dancer | Musician}]] accent [[Rhythm]]
2. [[Person=Speaker]] accent [[Language=Spoken]]
3. [[TopType]] accent [[Entity=Visual]]

accentuate/V BNC FREQ: 357

1. [[Event]] accentuate [[Problem | Process | State]]
2. [[Artefact | Shape 1 | Color 1]] accentuate [[Visible Feature | Shape 2 | Color 2]]
3. {[[Person=Musician]] | beat} accentuate [[Rhythm]]

Table 1: Some CPA verb patterns

5. Noun Entries in CPA

Some nouns are of course nominalizations of verbs, and these have a valency structure, mutatis mutandis, similar to those of the equivalent verbs. We shall say nothing more about them here. Here, we are concerned with the prototypical syntagmatic behaviour of noun-y nouns, in particular referring expressions. The word ***storm*** can serve as an example of the patterns for nouns. Conventional metaphorical and idiomatic uses (Table 3) are distinguished from literal uses (Table 2) by their syntagmatics, as explained in Hanks (forthcoming).

WHAT DO STORMS DO?
Storms ***break***
Storms ***blow***.
Storms ***rage***.
Storms ***lash*** coastlines.
Storms ***batter*** ships and places.
Storms ***hit*** ships and places.
Storms ***ravage*** places.
Before it breaks, a storm is ***brewing, gathering,*** or ***impending***.
There is often a ***calm*** or a ***lull before*** a storm.
Storms last for a certain period of time.
A major storm may be associated with a certain year, e.g. ***the great storm of*** [Year]
Storms ***abate***.
Storms ***subside***.
Storms ***pass.***
People can ***weather, survive,*** or ***ride (out)*** a storm.
Ships and people may get ***caught in*** a storm.
WHAT KINDS OF STORMS ARE THERE?
There are ***thunder storms, electrical storms, rain storms, hail storms, snow storms, winter storms, dust storms, sand storms,*** and ***tropical storms.***
Storms are ***violent, severe, raging, howling, terrible, disastrous, fearful,*** and ***ferocious.***
Storms, especially snow storms, may be ***heavy***.
An unexpected storm is a ***freak*** storm.
The centre of a storm is called the ***eye of the storm.***
STORMS ARE ASSOCIATED WITH ***rain, wind, hurricanes, gales,*** and ***floods***.

Table 2: Corpus-based profile of ***storm***, noun (literal uses)

A lot of fuss about a comparatively trivial event is described as a ***storm in a teacup***.
Someone who is in trouble is glad to find ***any port in a storm***.
A personality such as an artist, or an artefact such as a work of art or a product may ***take*** a place ***by storm***.
A military force or a military officer may ***take*** a place ***by storm***.
An action may ***cause, provoke, raise, create***, or ***unleash*** a storm.
A bad or unpopular thing may cause ***a storm of protest, controversy***, or ***criticism***.
A successful performance may be greeted by ***a storm of applause***.
Someone who is upset may burst into ***a storm of weeping*** or ***tears***.

Table 3: Metaphorical and idiomatic uses of ***storm***, noun

6. Conclusion

The paper started by arguing that general lexicography should focus on analysis of printed texts with a wide general readership, in order to identify the norms of usage found in deliberate, carefully thought out utterances. Spoken usage and domain-specific usage can then be related to this central body of general norms shared by all members of a speech community. Until now, the focus of lexicography has been on meaning; the paper suggests that a focus on usage – syntagmatics – should preoccupy 21st century lexicographers. CPA is a first step in this direction. The distinctions in CPA are semantically motivated, but the contents consist of summaries of typical patterns of usage, to which meanings, translations, synonym sets, and other data can be appended with little ambiguity.

References

Atkins, B. T. S. 1993. *Tools for computer-aided corpus lexicography: the Hector project.* Acta Linguistica Hungarica, 41.

Church, Kenneth W., and Patrick Hanks. 1989. 'Word Association Norms, Mutual Information, and Lexicography' in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*; reprinted in Computational Linguistics 16:1, 1990.

Church, K., W. Gale, Patrick Hanks, Don Hindle, and Rosamund Moon. 1994. 'Lexical Substitutability' in B. T. S. Atkins and A. Zampolli (eds.), Computational Approaches to the Lexicon. Oxford: Clarendon Press.

Fillmore, Charles J. 1975. 'An Alternative to Checklist Theories of Meaning' in *Papers from the First Annual Meeting of the Berkeley Linguistics Society.*

Halliday, Michael. 1966. 'Lexis as a Linguistic Level' in C. E. Bazell, J. C. Catford, M. A. K. **Halliday, and R. H. Robins** (eds.): *In Memory of J. R. Firth.* Longman. Reprinted in part in 1976 as Chapter 6 of Halliday: System and Function in Language, ed. G. Kress. Oxford University Press.

Hanks, Patrick. 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.

Hanks, Patrick. Forthcoming (a). 'The Syntagmatics of Metaphor and Idioms'. To appear in *International Journal of Lexicography*, March 2005.

Hanks, Patrick. Forthcoming (b). Norms and Exploitations: Mapping Meaning onto Use. MIT Press.

Jackendoff, Ray 2002. *Foundations of Language*. Oxford University Press.
Kilgarriff, Adam, and David Tugwell. 2001. 'WASP-Bench: an MT Lexicographers' Workstation Supporting State-of-the-art Lexical Disambiguation' *in Proceedings of MT Summit VII*, Santiago de Compostela.
Sinclair, John. 1966. 'Beginning the Study of Lexis' in C. E. Bazell, J. C. Catford, M. A. K. **Halliday, and R. H. Robins** (eds.): *In Memory of J. R. Firth*. Longman.
Sinclair, John (ed.). 1987. *Looking Up: an Account of the Cobuild Project in Lexical Computing*. HarperCollins.
Sinclair, John 1991: *Corpus, Concordance, Collocation*. Oxford University Press.
Sinclair, John, Patrick Hanks, et al. 1987. *The Collins Cobuild English Language Dictionary*. HarperCollins.