Why do we need pattern dictionaries (and what *is* a pattern dictionary, anyway)?

Patrick Hanks and Jane Bradbury

Abstract
After a lifetime in lexicography, the first author reached the alarming conclusion that words don't have meaning. Does that mean that dictionaries are useless? No, far from it. We argue that, strictly speaking, the neat numbered definitions listed in dictionaries can be regarded as presenting **meaning potentials** rather than meanings as such. Meanings, we say, are events—events activated in a process in which context acts on the meaning potential of each word or phrase that is used. Ordinary dictionary users find dictionaries useful because they can use common sense to supply contextual information that the dictionary does not give explicitly. Computer software for NLP (natural language processing), on the other hand, has little or no common sense to draw on, and so is often baffled by problems of word meaning. Language learners are somewhere in between: some aspects of "common sense" are language-specific; others are universal. Work in recent decades on pattern grammar (e.g. Francis, Hunston and Manning 1996, 1997; Hunston and Francis 2000) and on construction grammar (e.g. Goldberg 1995, 2006) has shown that contributions to the meaning of utterances come from grammatical constructions as well as from individual words. Construction grammarians point out that the meaning of a sentence such as "she slept her way to the top" is quite clear; it is something like: she got a senior job by having sex with powerful men. However, this meaning cannot be deduced from a concatenation of the meanings of the individual words in the sentence. Instead, it is associated, at least in part, with the whole sentence, i.e. the construction as a whole. They argue, with interesting consequences for lexicography, that meaningful constructions such as this are pervasive in ordinary language. But just as a reductionist approach to words (treating words as if they were building blocks in a child's Lego set) is insufficient for an understanding of meaning in language, so also syntactic analysis of grammatical constructions tells only part of the story. Somehow, ways and means need to be found for expressing the conditions under which different meaning potentials of a word are activated. In this article we shall suggest some of the ways in which this can be done. In short, we shall present a case for including much more information about phraseology as well as meaning in dictionaries.

1. Do words have meanings?

What's the meaning of ***blow***? It could refer to what the wind does, or a bitter disappointment. Or it could be something you do with your fist, your nose, a whistle, or even a lot of money.

The *Macmillan English Dictionary for Advanced Learners* (Macmillan, 2002) offers more than fifty potential meanings of the verb ***blow***, including phrasal verbs and idioms. There are another eleven potential meanings for the noun. Out of context, it is impossible to know which of these meanings is being activated; but given some context, things start to become clearer. Here are some sentences from the British National Corpus:

1. Use a fan to ***blow*** air through a screened doorway from the egg room or other work area into the main poultry house.
2. Arbroath has been dealt another jobs ***blow***. The engineering firm of Giddings and Lewis is to make 50 workers redundant.
3. Officials said unidentified saboteurs also used a dynamite-packed petrol tanker to ***blow up*** a bridge near the town of Mostar.

In example 1, the infinitive marker ***to*** designates a verb, while use of the noun ***fan*** before the verb and ***air*** after it suggests that in this instance ***blow*** is being used to denote the process of air being moved around by a machine.

The determiner ***another*** in example 2 shows that ***blow*** in this sentence is a singular noun, so ***jobs*** (being a plural noun) must be being used as a modifier, while ***blow*** is the head of the nominal group. The only possible interpretation here is that ***blow*** is being used to mean some kind of disappointment (an interpretation that is elaborated in the sentence that follows). In example 3, the occurrence of ***up*** after ***blow*** narrows down the possible meaning, and the object that follows the verb, ***a bridge***,

Patrick Hanks is Professor in Lexicography at the University of Wolverhampton, where he is directing a corpus-driven project in computational linguistics, the output of which is the on-line *Pattern Dictionary of English Verbs* (PDEV). He has developed a lexicographical technique called Corpus Pattern Analysis, used for PDEV and other projects.
He is also a Visiting Professor at the University of the West of England in Bristol, where he is the lead researcher on a project investigating family names in the UK (FaNUK).
From 1990 to 2000 he was Chief Editor of Current English Dictionaries at Oxford University Press. Before that (from 1971 to 1990) he was chief editor of English dictionaries for Collins publishers, and worked closely with John Sinclair on the first edition of the *Collins Cobuild English Language Dictionary* (1987).
After taking early retirement from OUP in 2000, he held teaching and research posts in the USA, Germany, and the Czech Republic. His monograph *Lexical Analysis: Norms and Exploitations* (MIT Press, 2013) presents a corpus-driven, lexically-based theory of meaning in language.
patrick.w.hanks@gmail.com

confirms that here we are talking about a physical object being destroyed rather than a person losing their temper or inflating a balloon.

The contrasting meanings of ***blow*** in these examples illustrate that many words do not have meaning in isolation: rather, we are forced to say that they have *meaning potential*. We need to examine the context, and in particular the *collocations* of a word, to realize this potential and identify a unique meaning.

2. How do collocations shape meaning?

Collocations are co-occurrences of words near each other in any given text or (at a more general level) they are pairs or sets of words that *typically* co-occur in many texts. One of the most important finding of corpus linguistics has been that (while the number of *possible* co-occurrences of words is in principle infinite) the *actual* number of frequently recurring collocations associated with any given content word in any language is comparatively small. Collocations can be measured and processed lexicographically. Very often, they yield unique interpretations of words that, in isolation, have more than one potential meaning.

The idea that collocation is key to meaning is not new. The first edition of the Cobuild dictionary was accompanied by a book of essays by the lexicographers (Sinclair, 1987). In a chapter entitled 'The analysis of meaning', Rosamund Moon draws attention to the relationship between collocations and meanings:

> Collocation … frequently reinforces meaning distinctions …. The noun *gap* has four main meanings: a physical space, an interval of time, a deficiency, and a discrepancy. Each of these has a distinctive set of collocates. The physical space sense collocates with *mountain, teeth, in,* and *between*. …. The interval of time sense collocates particularly with *year* and *of* …; the deficiency sense collocates with *fill, record*, and *in* …; the discrepancy sense collocates with *close, poor, rich, widen, bridge, trade, generation, narrow, reduce,* and *between*. …
>
> Arguably, the only way to make distinctions in meaning or use within the major delexical verbs such as *have, give,* and *take,* is to split according to the type of object collocate. A further area where collocation supports – or enforces – meaning distinctions is that of verbs and the animate/inanimate identity of subject and object, or valency patterning.

In another chapter, 'The Nature of the Evidence', Sinclair observes:

> Our initial assumption, that the words are distributed at random, is false.

He goes on to illustrate this with a discussion of corpus evidence for the distribution of collocations of the verb *set*, which has since been much quoted. Church and Hanks (1989 [1990]) used it as a basis for their work on statistical analysis of corpus data.

By 1998, after a further ten years of corpus analysis and growth of the Birmingham Corpus into what was to become 'the Bank of English', Sinclair had moved on to declare that "many, if not most meanings, require the presence of more than one word for their normal realization", and to argue that "patterns of co-selection among words … have a direct connection with meaning". Nowadays, data from large corpora, extending to billions of words of text, confirm that word use is highly patterned. It is these phraseological patterns that give readers and listeners the contexts they need to activate the meaning of words. However, despite the initiative of Cobuild, patterns of word use in English and other languages have still not yet been satisfactorily identified or explained. In particular, more information about collocations needs to be given. Foreign learners in particular need to be given much more information than is customary in standard dictionaries about the normal phraseology with which each sense of each word is associated. Thanks to the technology of corpus linguistics, it is now possible to represent such phraseology systematically, although some variations may be expected, depending on the actual corpus and statistical measure(s) used to identify salient collocations.

2.1 Valency and collocation

Valency in language defines the number of syntagmatic arguments that go with a word. For example, the verb ***shower*** in *he showered* has a valency of one; in *he showered the dog* it has a valency of two; in *he showered her with gifts* it has a valency of three. It is sometimes difficult to distinguish between an optional adjunct and an adverbial argument. For example, few people would claim that *he showered her with gifts every day* has a valency of four. 'Every day' is a time adverbial which does not attach itself specifically to the verb *shower*. Instead, systemic grammarians prefer to say that time adverbials normally attach themselves to the general concept 'event verb', rather than affecting the meaning of any one specific verb.

For effective sense disambiguation, information on both collocations and valency is needed. More often than not, the relevant collocations are in a particular syntagmatic relationship with the target word. Hanks (2012) discusses the example of the verb *shower* in more detail: one sense

Jane Bradbury is Research Associate in Lexicography at the University of Wolverhampton, where she is working on a corpus-driven project in computational linguistics, the *Pattern Dictionary of English Verbs*. She is also a consultant for Oxford University Press, working on the BBC/Oxford Children's Writing Corpus, as well as writing and editing educational resources.
Jane began her career as a lexicographer in 1989, working for Cobuild at the University of Birmingham on a range of projects including the *Collins Cobuild Student's Dictionary*, the *BBC English Dictionary*, the *Collins Cobuild English Grammar*, and the *Collins Cobuild English Guides*, of which she was series editor. From 1997, she worked as a freelance lexicographer whilst training as a teacher of English in order to explore the possibilities of using corpora to develop resources for L1 students and teachers of English, and to explore children's use of language. She is now keen to apply Corpus Pattern Analysis to these areas of interest.
jane.bradbury@mail.com

of this verb (broadly, 'wash the body under flowing water') can be clearly distinguished from other senses because it is intransitive and has a valency of one; however, other senses are less easy to separate on the sole grounds that they have the same number of arguments. For example, it is insufficient simply to report that *shower someone with praise* is transitive and has a valency of three. *Shower someone with rocks*, *shower someone with praise*, and *shower someone with gifts* all have a valency of three, however they have different meanings. To disambiguate these meanings effectively, we must look to both the syntagmatic patterns and the collocations (*rocks*, *praise*, or *gifts*). The point is that all three of these nouns are regular collocates of the verb ***shower***: the different collocates activate different senses of the verb, which need to be explained specifically in dictionaries. Moreover, the different arguments correlate with one another: thus, an *explosion* can *shower* people or locations with *debris*, but no sentences have been found in which an *explosion* showers them with gifts or praise. This general approach to correlating arguments in order to get at the meaning is called triangulation.

Hanks (2013, chapter 5) shows that most meanings of most verbs and other words denoting events work in this way.

- ***Firing*** a person from a job has a different meaning from ***firing*** a bullet from a gun.
- ***Filing*** a lawsuit in a law court denotes activation of a process, whereas ***filing*** papers in a filing cabinet denotes cessation of active use of those papers.

In this paper, we propose that corpus evidence should be analysed by triangulation to group all normal uses according to their valency and syntax, for only then can a well-founded attempt be made to explain the meaning.

3. Why has no one made a pattern dictionary before?

The need for a dictionary that identifies and reports on patterns of syntax and collocation was established by Sinclair *et al*. in the 1980s (in the Cobuild project), and yet still no satisfactory pattern dictionary has been completed. This is because until very recently there was insufficient corpus data to provide an empirical basis for a reliable pattern dictionary. Let us look a little more deeply at the example of *shower*.

We have established that to disambiguate senses effectively, it is not enough to separate by valency alone. The next step is to look at patterns of adverbials and complementation, followed by patterns of collocation. Here, some delicate decisions must be made by the lexicographer. For example, it is clear that *showering someone with presents* is different from *showering someone with praise*. This is because *presents* are (normally) physical objects, whereas *praise* is an eventuality activated by a person's speech or actions. But *rocks* are physical objects too, so should *showering with rocks* and *showering with presents* be lumped together in the same pattern, or be split and dealt with separately?

A similar problem arises with *shower with praise* and *shower with abuse*. Both *praise* and *abuse* are eventualities activated by a person's speech or actions; do they belong in the same pattern?

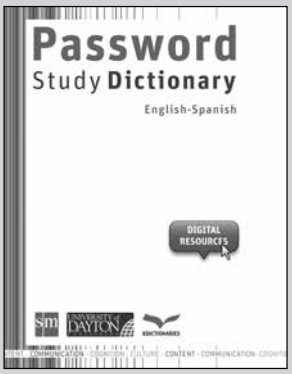
When you start to throw lexical items into the mix along with valencies, the decision as to whether to lump or split becomes difficult, but this is precisely the task that we are ambitiously undertaking.

4. Disambiguation of Verbs by Collocation

The Disambiguation of Verbs by Collocation (DVC) is an AHRC-funded project based in the Research Institute for Information and Language Processing at the University of Wolverhampton. The project aims, by doing Corpus Pattern Analysis, to establish an inventory of normal phraseological conventions, or ***patterns***, for English verbs. For each of these verbs, an initial sample of 250 corpus lines is analysed and tagged to show which pattern they are typical of; this sample size is doubled where a verb is identified as having 10 patterns or more, and doubled again if the total number of patterns reaches 20. Each pattern is linked to a set of tagged corpus lines.

The key objective of DVC is to identify normal usage, or phraseological *norms*. A useful by-product of identifying these norms is that it draws attention to authentic uses of verbs which are not norms but which are one-off deliberate *exploitations* of established patterns, for example for literary or humorous effect. Exploitations (Hanks 2013) are deliberate irregularities in language use, which do not form part of a pattern and must be ruled out as lexicographical evidence.

DVC also allows the calculation of the relative frequency of each norm for each verb, shown as a percentage. An account is given of the meaning (semantic and pragmatic – we do not distinguish) associated with each phraseological norm, using a shallow ontology of semantic types. The DVC ontology is based on lexicographical need rather than received scientific theory. For example, there is no place in the ontology for a semantic type 'mammal', because there are no verbs in



PASSWORD Study Dictionary English-Spanish

Ediciones SM
Madrid, Spain
April 2013
Editorial coordination: Yolanda Lozano and Concepión Maldonado
Flexicover, 1088 pages
ISBN: 978-84-154-7867-6
From the series: **Kernerman Semi-Bilingual Dictionaries**

English that select all and only mammals as arguments. On the other hand, as we shall see below, there are plenty of verbs that select 'horse' as an argument.

Case study: *harness*, verb

The DVC accounts for the normal patterns for ***harness***, verb, as follows:

1. 5% **[[Human]] harness [[Horse]]**
   [[Human]] puts harness on [[Horse]] in preparation for riding or driving it, or getting it to pull a cart, carriage, or plough
2. 95% **[[Human | Institution]] harness [[{Eventuality 1 | Entity 1} = Resource]] (to [[Eventuality 2 | Entity 2]])**
   [[Human | Institution]] makes use of [[{Eventuality 1 | Entity 1} = Resource] (in conjunction with [[Eventuality 2 | Entity 2]]) for some purpose

In Pattern 1, the lexicographer faces a dilemma that is a typical issue in DVC research. Prototypically, it is horses that get harnessed, but (as it happens) only 50% of the BNC citations for this pattern involve horses. The remaining 50% involve harnessing other animals: the British National Corpus (BNC) gives us the following examples of animals other than horses that get harnessed:

- dogs (huskies, for pulling sledges)
- oxen
- bullocks
- deer
- donkeys
- reindeer
- camels
- mules

When a speaker or writer talks about ***harnessing*** a bullock, reindeer, or mule, this is not a linguistic exploitation for effect; they are literally talking about the act of putting one of these animals into a harness in order to ride it, drive it, or get it to pull a cart etc. DVC must account for this regular alternation for the benefit of both language users and NLP applications. Therefore, it might be better to state Pattern 1 as **[[Human]] harness [[Horse | Animal]].**

However, if [[Animal]] is given as an argument alternation of this pattern, the scope is too broad, as it could be taken as implying that it is normal to harness cats, primates, and cows, which is not correct. On the other hand, as we have seen, stating [[Horse]] alone is over-restrictive, appearing to rule out dogs, bullocks, oxen, etc. The answer to this apparently irresolvable dilemma is that, whatever semantic type (or set of types) is chosen, it is really only a form of shorthand, encapsulating a set of lexical items that are prototypical in this slot. Semantic typing is helpful as far as it goes, but it is possible to put too much weight on the type, as opposed to the actual lexical items that 'populate' the semantic type.

The DVC Ontology places the semantic type 'Animal' in a hierarchy, as follows:

Animate
  Human
  Animal
    Horse
    Dog
    Cat
    Primate
    Cow
  Bird
  Insect
  Fish
  Snake
  Spider
  Cetacean

Given this ontological set, by choosing the type [[Animal]] as an alternate for [[Horse]], the lexicographer can signal that it is normal for other types of living creatures to be put into a harness (though not birds, insects, fish, or cetaceans, which are separate semantic types, associated with distinctive sets of verbs).

Pattern 2, which refers to the non-literal ***harnessing*** of abstract resources in order to use them, would once have been considered an exploitation:

[[Human | Institution]] harness [[{Eventuality 1 | Entity 1} = Resource]] (to [[Eventuality 2 | Entity 2]])

However, DVC has discovered that this pattern now accounts for 95% of uses of ***harness***, verb, in this corpus: a clear example of an exploitation becoming a norm. It will be interesting to compare the relative frequencies of these two patterns in other corpora.

The example below shows a one-off exploitation of ***harness***:

Perot wants to take us all back in time and *harness* us behind mules!

The writer is not suggesting that people will literally be forced to wear harnesses and pull carts behind mules: most readers will work out that this is a metaphorical extension of Pattern 1, with the intended meaning that Perot would treat people as no better than beasts of burden, valued for their physical strength only. However, in

Fortunes of National Cultures in Globalisation Context

The international conference on Fortunes of National Cultures in Globalisation Context was held by the Faculty of Eurasian and Oriental Studeis at Chelyabinsk State University (CSU, Russia) on 4-5 April 2013. It embraced a wide range of topics concerned with cultural and globalization phenomena, including preservation of languages and national identities, linguistics, cognitive linguistics, international cooperation, etc. Considering that any culture represents an interdisciplinary object of research, the issues of ethnology, political studies, history, psychology, linguistic and cultural studies, philosophy, semiotics and lexicography were all discussed.

Plenary speakers included Nikolay Alefirenko (Belgorod State University), Konstantin Averbukh (Moscow City Pedagogical University), Vera Budykina (CSU, Conference Organizer), Olga Felde (Siberian State University), Zan Hun Ge (Capital University of Education, Beijing), Igor Golovanov (Chelyabinsk Pedagogical State University), Ilan Kernerman (K Dictionaries, Tel Aviv), Valery Kuznetsov (Moscow State Linguistic University), Elvira Sinetskaya (Institute for Oriental Studies of the Rusian Academy of Sciences, Moscow), and Svetlana Ter-Minasova (Lomonosov Moscow State University and President of the National Association of Teachers of English in Russia).

▶

other corpora, we may find that

[[Human 1]] harness [[Human 2]]

has become established as a pattern in its own right in certain domains or in a more recent time-frame than that of BNC.

5. Conclusion

The *Pattern Dictionary of English Verbs* (PDEV) represents a new development in lexical analysis, based on careful empirical analysis of a corpus. We hope that it will take its place alongside other innovative approaches such as FrameNet in accounting for words and meanings. It represents only one of many possible approaches to identifying and explaining patterns of word use and the connection between such patterns and their meanings. If it is successful, PDEV can function as a set of 'seed' patterns for semi-automatic expansion over much larger sets of data, including domain-specific corpora, corpora of children's language, historical corpora, etc. We do not claim that it is possible that any pattern dictionary could account for all and only the meanings of words in any natural language. "All and only" represents a theoretical goal that was exploded as unrealistic and distorting for natural-language research (including lexicography) during the second half of the 20th century. Instead, the aim now is to represent prototypical usage and associate it with prototypical meaning.

PDEV is work in progress and is in the public domain. It can be accessed at http://deb.fi.muni.cz/pdev/.

Although it is still only work in progress, we urge you to explore it. Comments and feedback are invited.

References

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*, 26–29 June 1989. University of British Columbia. Revised in *Computational Linguistics* 16 (1), 1990.

Francis, Gill; Hunston, Susan; and Manning, Elizabeth. 1996. *Collins COBUILD Grammar Patterns 1: Verbs*. London & Glasgow: HarperCollins.

Francis, Gill; Hunston, Susan; and Manning, Elizabeth. 1997. *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives*. London & Glasgow: HarperCollins.

Goldberg, Adele. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Hanks, Patrick. 2012. 'How People Use Words to Make Meanings: Semantic types meet Valencies'. In Alex Boulton and James Thomas (eds.). *Input, Process and Product: Developments in Teaching and Language Corpora*. Brno: Masaryk University Press.

Hanks, Patrick. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Hunston, Susan, and Francis, Gill. 2000. *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam & Philadelphia: John Benjamins.

Moon, Rosamund. 1987. 'The analysis of meaning' in Sinclair, John (ed.).

Rundell, Michael (ed.). 2002. *Macmillan English Dictionary for Advanced Learners*. Oxford: Macmillan.

Sinclair, John (ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD Language Dictionary*. London & Glasgow: Collins ELT.

Sinclair, John. 1998. 'The Lexical Item' in E. Weigand (ed.) *Contrastive Lexical Semantics*. Amsterdam & Philadelphia: Benjamins.

Sinclair, John, Patrick Hanks, and others (eds.). 1987. *Collins COBUILD English Language Dictionary*. Glasgow: Collins.

▶

There were 7 parallel sessions, namely: Theoretical and Methodological Aspects of National Culture Studies in Various Paradigms of Knowledge; National Spiritual Culture: Traditions and Innovations; Cross-Cultural Communication, Cross-Cultural Competence and Globalisation; The Dialogue Between Cultures: West, East and Russia; National Mentality Representation in the Modern Information Globalisation and Preservation of National Cultures in Literary Perception; Lexicography, Terminology Banks and National Identity.

The latter session concerned national and cultural aspects of lexicography and problems associated with the formation of corpora and databases for dictionaries. New tendencies in lexicographic practice were discussed and compared to the Russian tradition of dictionary compilation. These included papers on 'Professional communication in terms of globalization' (Averbukh), 'Terminology system of higher education of Russia: National identity or harmonization?' (Budykina), and 'Lingua franca, mother tongue, and pedagogical lexicography: Developing a global dictionary series for learners' and a masterclass on 'The current status, changes and prospects in the dictionary world' (Kernerman).

The conference proceedings, comprising 800 pages, are the issue of these discussions.

Vera Budykina
vbudykina@gmail.com