

Corpus-driven Lexical Analysis: Norms and Exploitations in Word Use

Patrick Hanks

Research Institute of Information and Language Processing,
University of Wolverhampton, UK
and
Bristol Centre for Linguistics,
University of the West of England, UK

It is a truism that meaning depends on context. Corpus evidence now shows us that normal contexts can be summarised and indeed quantified, while the creative exploitations of normal contexts by ordinary language users far exceed anything dreamed up in speculative linguistic theory. Human linguistic behaviour is indeed rule-governed, but in recent years, corpus analysis (e.g. Hanks 2013) has shown that there is not just a single monolithic system of rules: instead, language use is governed by two interlinked systems: one set of rules governing normal, idiomatic uses of words and another set of rules governing how we exploit those norms creatively. Types of creative exploitation include (among others):

- using anomalous arguments to make novel meanings
- ellipsis for verbal economy in discourse
- metaphors, metonymy, and other figurative uses for stylistic effect and other purposes

Traditional dictionaries do a good job of listing the many possible meanings of words. But they do a poor job of reporting phraseology and an even worse job of associating different meanings with phraseological patterns. Moreover, all too often, they list a creative use that happens to have been noticed by a lexicographer as if it were a conventional norm, with resultant confusion, for example:

- A riddle does not mean a hole made by a bullet (but OED says it does).
- To newspaper does not mean to work as a journalist (but Merriam Webster says it does).

The idiom principle formulated by the late John Sinclair (1991, 1998) argues that many meanings depend for their realization on the presence of more than one word. The Pattern Dictionary of English Verbs (PDEV; <http://deb.fi.muni.cz/pdev/>; publicly available, but note that it is work in progress) implements this principle by associating meanings with patterns rather than with words in isolation. A pattern consists of a verb and its valencies (otherwise known as clause roles or arguments). Each argument is populated by an open-ended set of lexical items and phrases, which share a semantic value. This means that different senses of a verb can be distinguished according to the semantic values of its arguments. Thus, executing an order and executing a plan go together and are distinguished from executing a dictatorthese are two different meanings of the same verb, activated by different collocates, even though all three examples have identical syntax. Sinclairs idiom principle can be usefully compared with the theory of construction grammar (Goldberg 1995). An important different is that the Sinclairian approach is empirically well founded: it is corpus-driven. It does not rely on the speculative invention of evidence, which has been shown to be methodologically unreliable. PDEV is likewise rigorously corpus-driven. Every verb (and in due course, every predicatorincluding predicative adjectives) has been or will be analysed on the basis of corpus evidence. Each entry in PDEV has the following components:

- A set of syntagmatically distinct patterns (the phraseological norms)
- An implicature (i.e. the meaning and context) for each pattern
- A set of corpus lines illustrating normal uses of each pattern
- Comparative frequencies of each pattern of use of each verb, showing which patterns are most frequent
- A smaller set of corpus lines illustrating creative exploitations
- A shallow ontology of nouns and noun phrases

The CPA shallow ontology serves as a device for grouping together noun phrases that distinguish one meaning of a verb from another.

References

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

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

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, John. 1998. The lexical item. In Edda Weigand (ed.), *Contrastive Lexical Semantics*. John Benjamins.