Discovering clines of variation: the location and analysis of non-canonical forms in general reference corpora

Gill Philip

University of Bologna – Italy

Keywords: categorisation, fixedness, units of meaning, variation, word-play

1. Introduction

As the growth of large corpora and the resulting move away from the serendipitous collection of language examples makes available more and more data from which language norms, and deviations from them, can be identified, studies into phraseology continually reaffirm the fact that "so-called 'fixed phrases' are not in fact fixed" (Sinclair 1996:83). Yet in spite of this recognition, the discovery of non-canonical forms in corpora is still generally considered to be a matter of good fortune (Moon 1998:51). Further to this, the fact that the data which can be extracted is heavily restricted to that which is entered into the search query is cited all too often as one of the drawbacks of corpus-related research into the phenomenon (Moon 1996:252; Deignan, 1999:197). After all, how can one search for something without knowing what that something is? This paper presents procedures for the location of non-canonical "fixed" phrases, and puts forward some arguments in favour of their inclusion in phraseological analysis.

2. Searching for variant forms

Corpus query syntax varies from program to program, but a factor shared by the majority of applications is that search procedures are rarely exploited to the full. Although tagging allows fairly abstract searches to be carried out in the search for grammar patterns (see, for example, Hunston & Francis 1999), the retrieval of lexical variation is all too often limited to word forms and lemmas rather than phraseological structures (Moon 1996; Cicogni & Coffey 2000:550). But this need not be the case: serendipity can be replaced by the systematic; chance by trial and error.

This paper demonstrates a method for the retrieval of variant forms from computer corpora (here, the Bank of English), demonstrating how key-words, wild cards and phrasal structure all contribute to the identification of underlying schemas which can in turn be translated into search queries. It is worth noting that the same principles are relevant to the use of advanced queries in Internet search engines, with the result that the linguist is able to extend and verify his/her findings on a much larger, if less homogeneous, data set – a factor of considerable importance when dealing with multi-word strings, where the location of sufficient examples for analysis can often pose a problem.

3. The linguistic value of phraseological variation

If finding the data is in itself a barrier to the study of variation, so too is the fairly low esteem in which it is held in language description. Perhaps surprisingly, canonical forms of idioms and other figurative phrases are actually quite uncommon in language corpora and are, as a general rule, outnumbered by their corresponding non-canonical forms. Despite this observation, non-canonical forms are still considered to be exceptions to the norm. They are peripheral to the interests of lexicography and foreign language teaching because they can ultimately be reduced to a canonical form, and it is this underlying canonical form, not its variants, which requires documentation and learning.

3.1. The semantic effects of variation

At the far extreme of the scale of variability lie puns and other types of word-play. Unlike their less showy counterparts, these attract considerable interest on the part of linguists. The many layers of meaning that they involve are held in place by an underlying canonical structure which appears to generate an infinite variety of novel utterances – see Partington (1998:121-143) on the exploitation of fixed phrases in newspaper headlines; see also Moon 1998; Philip 2003. In the study of word-play, novel utterances are typically contrasted with their canonical relatives, all but ignoring the relationship that they have with the common-or-garden variation created on the fly by language users. A large swathe of language is overlooked as a result if it being neither normal nor exceptional enough to merit attention. Yet it is precisely to this little-studied middle-ground that phraseology scholars can turn in order to discover more about language and the use that its speakers make of it.

The average language user can and does manipulate conventional structures to the particular communicative situation, but contrary to what our intuition might suggest, this does not necessarily implicate that double meanings, humour, irony or other deliberate textual effects are created. What the study of variant forms shows is the extent to which learned forms can be adapted to make them contextually appropriate, whether this be in order to disambiguate, specify or reinforce meanings.

3.2. Factors affecting phraseological variation

In the view taken in this paper, canonical phrases are seen as a particular sub-set of collocational frameworks (Sinclair & Renouf, 1991). This means that they provide a structure which permits the productive variation of some slots while resisting change to others. The most revealing productivity occurs along the paradigmatic axis where terms are substituted not only by members of the same semantic set but also by apparently unrelated terms. This phenomenon can be accounted for by the *Class Inclusion Hypothesis* (Glucksberg & Keysar 1993; Glucksberg & McGlone 1999), which notes that although classes are traditionally considered to be taxonomic, in the case of metaphorical language they are attributive. If the relationship between the substituted term and the canonical one is based on common attributes, then everyday, unmarked variation would appear to be functioning along different lines from those which are intuitively believed to be in operation. This contributes significantly to the difficulties encountered in trying to predict how variant forms will manifest themselves. It also helps to explain why computational models of language find non-canonical forms difficult to account for, whereas most language users find both their interpretation and production unproblematic.

The study of canonical forms alongside non-canonical forms in all their guises also highlights the relationships which hold between phraseological items and their co-textual environments. In fact, the analysis of corpus data demonstrates that the central element of an extended unit of meaning (Sinclair, 1996), typically taken to be a single word, can just as readily take the form of an entire phrase, canonical or otherwise (Philip 2003). The data suggests that although it is possible for a canonical phrase to occur in an unusual co-text, thus triggering contextually-relevant interpretations, it is extremely rare for a non-canonical form of a phrase to occur in a context which differs markedly from the norm (ibid.). When it does occur, it tends to occur in transcribed spontaneous speech when two similar structures are fused during on-line processing, resulting in a "crack" in the phraseological priming (Hoey, 2005:11).

4. Conclusions

It is the unpredictability of non-canonical forms that makes them difficult to extract from a corpus, and this certainly hinders attempts to study them comprehensively. Although this paper shows that their location is not governed by mere happenstance, the manual extraction of the type described here still poses a challenge to automation: automatic extraction requires a degree of stability which non-canonical forms do not always comply with. However the study of variant forms can contribute fruitfully to existing linguistic knowledge bases, providing a semantic dimension that has yet to be explored in corpus studies. This in turn will contribute towards future success in automatic extraction of variant forms from corpora, as well as shaping and refining our understanding of how natural languages really work.

References

Cicogni L. & S. Coffey (2000) A Corpus study of Italian Proverbs: implications for lexicographical description. In *Euralex 2000 proceedings*. Stuttgart: Universität Stuttgart, 549-555.

Deignan A. (1999) Corpus-based research into metaphor. In Cameron L. & G. Low (eds), Researching and Applying Metaphor. Cambridge: Cambridge University Press. 177-199.

Glucksberg S & B. Keysar (1993) How Metaphors Work. In A. Ortony (ed) *Metaphor and Thought.* Cambridge: Cambridge University Press, 401-424

Glucksberg S. & M.S. McGlone (1999) When love is not a journey: What metaphors mean. *Journal of Pragmatics* 31, 1541-1558.

Hoey M. (2005) *Lexical Priming: A new theory of words and language*. London: Routledge

Hunston S. & G. Francis (1999) *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia: John Benjamin.

Moon R.E. (1998) *Fixed Expressions and Idioms in English: A Corpus-Based Approach.* Oxford: Clarendon.

Moon R.E. (1996) Data, Description, and Idioms in Corpus Lexicography. In *Euralex '96 Proceedings.* Gothenburg: Göteborg University, 245-256.

Partington A. (1996) *Patterns and Meanings: Using Corpora for English Language Research and Teaching.* Amsterdam and Philadelphia: John Benjamin.

Philip G. (2003) *Connotation And Collocation: A Corpus-Based Investigation Of Colour Words In English And Italian*. PhD Thesis. Birmingham: The University of Birmingham.

Philip G. (2000) An Idiomatic Theme and Variations. In Heffer, C. and H. Sauntson (eds*) Words in Context: A Tribute to John Sinclair on his Retirement.* ELR Monograph 18. Birmingham: The University of Birmingham, 221-233.

Sinclair J.M. (1996) The Search for Units of Meaning. *TEXTUS* IX (1), 71-106.

Sinclair J.M. and A. Renouf (1991) Collocational Frameworks in English. Reprinted in Foley J.A. (ed) (1996), *J.M. Sinclair on Lexis and Lexicography.* Singapore: Unipress. 55-71.