Reassessing the Canon: “fixed” phrases in general reference corpora

Gill Philip

University of Bologna – Italy

Abstract

This paper sets forth the argument for revisiting fixed phrases in the light of the knowledge that their fixedness is not necessarily something to be taken for granted. It focuses on the location and analysis of variant forms in general reference corpora. Existing phraseological structures, including collocational frameworks, idiom schemas and semi-prepackaged phrases, are introduced by way of background before a procedure for retrieving non-canonical forms of fixed expressions in general reference corpora is presented. Some implications relating to the study of variant forms are presented, along with suggestions for future research directions.

Keywords: class-inclusion hypothesis, collocational framework, fixed phrases, idiom principle, idiom schema, multi-word expressions, open-choice, units of meaning, variation, word-play

1. Introduction

Although phraseology focuses primarily on phrase building, from word to collocate and beyond, this paper sets out the case for phrase deconstruction. At first glance this might appear to be the antithesis of what phraseology is about. Why dismantle multi-word expressions when so much effort has gone into identifying their most typical realisation – the canonical form[1]?

One reason is that corpus linguists have consistently noticed that canonical forms are not particularly common in language corpora and, crucially, they tend to be outnumbered by non-canonical variants and exploitations. These non-standard forms generally receive little consideration from linguists unless they happen to appear in a particularly eye-catching guise, for example in journalistic and advertising word-play.

Variations of, and deviations from the canonical form are more than simply ‘alternative renderings’ of the ‘same’ basic expression. Their existence fills a semantic and pragmatic need – that of personalising and making one’s language use relevant to the situation in which it is uttered. This paper presents some of the ways in which variant forms of fixed expressions have been classified in existing literature (2), and how they can be retrieved from general reference corpora using off-the-shelf corpus query applications (3). The linguistic value of variant forms is discussed in 4, where particular attention is paid to the evaluation of word-play relative to variation in general (4.1), and the insights that the observation of variant forms can provide in the study of semantic productivity (4.2).

2. Recurrent and non-recurrent forms in language corpora

With the growth and increased availability of large, general reference corpora, phraseology scholars have been able to move away from the manual, serendipitous, collection of citations. The information that can be gleaned from studying multiple representations of an expression on a page of concordances is not only more detailed than that derived from smaller amounts of data supplemented by intuition; for many years, corpus data has been providing evidence that "so-called 'fixed phrases'[2] are not in fact fixed" (Sinclair 1996: 83). Yet although linguists are aware of this fact, very little attention seems to have been given to the variant and anomalous forms except when they create stylistic effects such as puns, irony and humour (for example, Partington 1998: 121-143).

So why has less marked variation been overlooked or cast aside? The fact of the matter is that unmarked variation is seen to be of limited interest because it concerns non-recurrent forms, and as such it is generally believed to contribute little to the description of the canonical form. Canonical forms are linguistically important because of their regularity and consistency. Variants, on the other hand, are embellishments, additions or reductions whose meaning and function are ultimately dependent on, and reducible to, the canonical form from which they are derived. Insofar as corpus studies are concerned, there is a degree of incompatibility between a methodology which favours the identification of recurrent patterns, and the search for non-recurring variants. One of the most commonly-cited drawbacks of corpus-related research into phraseological variation is that "you find what you look for: search tools will only match the pattern sought. An over-restricted search for *a wolf in sheep's clothing* will not find *a sheep in wolf's clothing*" (Moon 1996: 252). Commercial corpus query software is designed for the extraction of lexically- and syntactically-related forms making it difficult to imagine how to retrieve variants based on semantics rather than grammatical structure.[3] And so the location, or rather the discovery, of non-canonical forms in corpora is still generally considered to be a matter of good fortune (Moon 1998: 51), arrived at more by accident than by design. After all, how can one search for something without knowing what that something is?

2.1. Phraseological skeletons

The change from canonical form to variant, rather than being clear-cut, operates along a continuum. Canonical forms often include grammatical elements which inflect in text, and they may also have alternative lexical realizations. There are several existing studies investigating different types of phraseological frameworks and schemas which incorporate such 'regular irregularities', including Renouf & Sinclair (1991), Francis (1993), and Moon (1998). Each of these scholars focuses on a different type of collocational phenomenon, which can be referred to with the generic term *phraseological skeletons*. The core components of a phraseological unit constitute its skeleton, which is fleshed out by elements with a less central role in creating and maintaining the structure of a phraseological form. The different types of skeleton discussed in this subsection are all illustrated with examples for ease of reference.

Renouf & Sinclair (1991) introduced the concept of *collocational frameworks*. These are defined as "a discontinuous sequence of two words, positioned at one word remove from each other; they are therefore not grammatically self-standing; their well-formedness is dependent on what intervenes" (1991: 128). Collocational frameworks are essentially grammatical collocations with a variable lexical 'slot' intervening, and can be seen as an early stage in the subsequent development of the definition of the *idiom principle* (Sinclair 1991), and the *extended unit of meaning* (Sinclair 1996): the example in Figure 1 (after Renouf & Sinclair 1991: 142) illustrates the way in which the slot-filler, *accident*, serves as an "idiomatic platform" (ibid.) for a series of semantically-related words to the immediate right of the framework, *an … of*; each of these choices in turn would presumably favour a restricted number of collocations, which again limit the successive options, and so on until the outer bounds of the unit of meaning are reached.

an accident of birth
an accident of history
an accident of history or birth
an accident of fate
an accident of post-war politics
an accident of war

Figure 1: Collocational framework

Lexicogrammatical frames (Moon 1998) are a particular type of phraseological collocation that are akin to collocational frameworks in that they are composed of a fixed part and a variable one. The difference between these models is that the variable elements in lexicogrammatical frames must be related, because the resulting clusters of phrases are related not only syntactically, but also semantically. As Moon explains, "[t]here is a common structure which contains a variable slot; the variable element is lexical, rather than grammatical, and the variations found tend to belong to the same semantic set. Because of the similarities in grammatical structure and lexical content, the meanings of the phrases can be said to be roughly synonymous" (Moon 1998: 145-6). A lexicogrammatical frame with the fixed preposition, *beyond*, and variable nouns (after Moon 1998: 39) can be seen in Figure 2.

beyond belief
beyond description
beyond doubt
beyond question
beyond recognition
beyond repair

Figure 2: Lexicogrammatical frame

The type of collocation that Francis (1993) calls *semi-prepackaged phrases* take a further step away from the notion of collocation as word co-occurrence. Instead, she presents collocating semantic sets. Semi-prepackaged phrases are phrases which are understood to be variants of one another, but in which "there is no single lexical item which is essential" (ibid: 144). It is important to stress that this is not the same as semantic preference (Sinclair 1996): in semi-prepackaged phrases, the collocation itself is defined by semantic set, whereas semantic preference is the tendency for a collocation (in the received sense of the term) to co-occur with a restricted range of semantically-related words in the extended co-text: this preference for particular semantic sets contributes towards the definition of the extended unit of meaning, but does not affect the composition of the initial collocation.

the faintest idea
the least idea
the slightest idea
the foggiest idea
the remotest idea
the slightest conception
slightest notion
the foggiest notion

the remotest notion
the foggiest.

Figure 3: Semi-prepackaged phrase

The example in Figure 3 (after Francis 1993: 144) illustrates forms of the semi-prepackaged phrase, *the faintest idea*, and its variant wordings. Despite first impressions, which suggest that the phrase is quite fixed, it can be noted that “the only essential elements are the ‘superlative’ marking of the adjective and the definite article *the* which accompanies it.” (ibid). The method used for the extraction of these variants, in which there is no fixed element, will be examined in 3.

What the three phraseological skeletons discussed so far have in common is an increasing tendency towards idiomaticity: even the apparently innocuous grammatical framework *a…of*, once its lexical slot is filled, finds its meaning possibilities restricted. Lexicogrammatical frames are defined by a freer type of collocational regularity, which involves a fixed element which collocates with a range of words belonging to the same semantic set, resulting in the formation of a series of loosely synonymous expressions. This is further extended semi-prepackaged phrases, where the collocation involves no fixed element, both slots being filled by members of a restricted semantic set, and a cluster of related, roughly synonymous expressions is generated.

This brings us to *idiom schemas* (Moon 1998), which again are defined by semantic similarity rather than word-form co-occurrence.

one sandwich short of a pic-nic
several cards short of a full deck
a few gallons shy of a full tank
two beanshoots short of a spring roll
a bishop short of a chess set
several hatstands short of a cloakroom
one number short of a logarithm

Figure 4a: Idiom schemas (i)

shake in one’s shoes
quake in one’s shoes
shake in one’s boots
quake in one’s boots
shiver in one’s boots
quake in one’s Doc Marten’s
quake in one’s size 11s

Figure 4b: Idiom schemas (ii)

Idiom schemas “share an underlying metaphorical conceit and their lexicalizations are drawn from sets of co-hyponyms.” (1996: 252). The schemas that Moon illustrates appear to be a sub-type of lexicogrammatical frame, in that there tends to be a base structure which supports the variant forms. The examples provided in Figures 4a and 4b demonstrate the schemas for *one [component] short/shy of a [whole]* (ibid: 252) and *[tremble] in one’s [footwear]* (Moon 1998: 161) respectively, where the words enclosed in square brackets represent the variable slots.

2.2. Capturing the variety of non-canonical forms

While the kinds of schematic representation outlined in 2.1 are adequate for classifying variant forms from a lexicographic point of view, they leave little room for the inclusion of variants which, though exploiting the underlying conceit, do not adhere to the typical phraseological patterning (e.g. *the pan calling the kettle blackbottom*, related to *the pot calling* *the kettle black*: see Appendix; see also Moon 1998: 170-177). Conceit-based exploitations are recognisably related to the canonical form in much the same way as variant forms of semi-prepackaged phrases are appreciably 'the same but different', so it may not always be helpful to treat exploitations separately from schemas.

One way to include exploitations alongside schematic representations of idiomatic and figurative phraseology is to focus on the key components of the idiomatic 'theme' (Philip 2000): these are typically the most salient elements, and they may be syntactic or lexical. Sometimes the recognition of the canonical phrase is triggered by a core collocation (e.g. *red rag*), at other times a combination of salient words and grammatical elements have the same effect (e.g. *like a* [[colour] [fabric]] *to a* [NP]).[4] In contrast with the phraseological models discussed above, the reduction of phrases to key components is not designed for lexicographical description. It aims instead simply to extract as many phraseological permutations as possible in order to paint a comprehensive picture of everyday variation which can then be used as a benchmark for assessing stylistic effects in journalism, literature and translation.

like a red rag to a bull
as a red rag to a bull.
a red rag to the Unionist bull.
the latest red rag from a bullish Beijing
like a red rag to the Euro-sceptics
like waving a red flag in the face of a bull
like waving a red rag at the bull
like putting a "red rag to a bull"
like bulls to a red rag
red flag to a dragon

Figure 5: Idiom theme and variations

Figure 5 shows variation to the idiom *like a red rag to a bull* (after Philip 2000: 231-232); here it can be observed that there is only one invariable element, *red* (no variations to this colour were attested in the corpus, though marked variation may well home in on this component; see Philip 2003: 201-242); the comparative appears to be optional, as indeed is *bull*, which is substituted in a semantically intriguing way: rather than being replaced by co-hyponyms as happens in idiom schemas, the semantic set is *attributive* rather than *taxonomic* (Glucksberg & Keysar 1993: 408-9; see also 4.2). The implications of this are considerable, especially given the current vogue for automated extraction of data from corpora. Attributive semantic sets are not fixed lists and do not conform to traditional notions of semantic relatedness as adopted in thesauri, but instead are determined and interpreted contextually. The variation found in idiomatic and figurative phrases is a combination of grammatical and lexical components, and with the added complication of attributive semantics, it is easy to see why "from a lexicographical viewpoint, they are simply nightmares" (Moon 1996: 252).

3. Searching general reference corpora for variant forms

Given the difficulties to be encountered in defining types of variation, it becomes clear why corpus searches generally produce only a limited range of non-canonical citations. The degree of semantic productivity to be encountered in non-canonical forms can be surprisingly complex and unpredictable, and the retrieval of variants becomes well-nigh impossible if no fixed element can be defined. However, if the search criteria insist on there being certain fixed elements present, any examples which do not contain these word forms (but which contain other key components, or use phonologically or graphically similar forms) will remain hidden. Tagging – both grammatical and semantic – can aid the process, but do not resolve the problem entirely. While both probabilistic algorithms and expert linguists are able to predict likely variants, the reality of variation as revealed in corpus data (as can be seen in the Appendix) often results in the formation of *ad hoc* semantic classes which evade prediction. So the corpus user has to define a search strategy which will maximise the retrieval of a useful data set.

For most corpus linguists, this means carrying out a single, wide-reaching and general search which is subsequently refined. What few attempt is to combine the results of successive related searches before embarking on the refining and selecting procedure. One such method was used by Francis (1993) for retrieving semi-prepackaged phrases.

> "I concordanced *idea* and found this meaning with the adjectives *faintest*, *least*, *slightest*, *foggiest* and *remotest*. I then concordanced all these adjectives in order to find head-nouns combining with them to form the same meaning, and found *conception* and *notion*." (ibid: 156)

The resulting data was then combined into a single file, from which a *stepped concordance* – a concordance with no single, invariable node – was produced (ibid: 144).

Cignoni & Coffey (1998, 2000) adopted a different approach to extract idiom and proverb variants from the untagged Italian Reference Corpus (Bindi et al. 1991), which involved "making searches for one or more key words for each idiom and subsequently editing out irrelevant material with a word-processor" (Cignoni & Coffey 1998: 292). Manual editing is quite feasible with a corpus of this size (15 million words), but the larger the corpus, the more laborious the selecting becomes.

Philip (2000, 2003) developed a procedure for retrieving variants of idiomatic phrases in the 450 million-word Bank of English. Given the size of the corpus, simple key-word searches would have been impossible to edit manually, so it was essential to formulate a series of searches that would be inclusive of all potentially relevant data, yet restrictive enough to exclude as much 'noise' as possible. The results of the searches were collated, then edited with a PC concordance package to eliminate duplicates and irrelevant concordances (Philip 2003: 127-129).

1a	red+1,5bull	[*red* followed by *bull*; five word window]
1b	bull+1,5red	[*bull* followed by *red*; five word window]
2a	rag+1,5bull	[*rag* followed by *bull*; five word window]
2b	bull+1,5rag	[*bull* followed by *rag*; five word window]
3a	flag+1,5bull	[*flag* followed by *bull*; five word window]
3b	bull+1,5flag	[*bull* followed by *flag*; five word window]
4	to+a+bull	[*to a bull*; no words intervening]
5	red+rag	[*red rag*; no words intervening]

6	red+flag	[*red flag*; no words intervening]
7	like+a+1,5to+a	[*like a* followed by *to a*; five word window]

Figure 6: Search procedures[5] for *like a red rag/flag to a bull*

The queries shown in Figure 6 incorporate various lexical and phrasal elements of the canonical form, and exploit the maximum number of unspecified words (the ‘five-word window’) permitted between search terms in order to include as much data as possible.

Some justification needs to be made regarding the decidedly low-tech aspects of the search procedures illustrated in Figure 6. In the first place, they were devised to be used not only with the Bank of English, but also with corpora which had no tagging, making more sophisticated queries impossible to carry out. The advantage of this is however that they can be used on any corpus or text collection, using even the most rudimentary of concordance packages. They can also be run on Internet search engines, making it possible to use the web to verify findings derived from corpus data, and to provide supplementary data when the corpus in unable to provide sufficient examples for the study of longer phraseological units.

In the course of carrying out multiple searches of this type, it was discovered that some relevant examples featured none of the apparently essential key-words, and others lacked the expected syntactic patternings. This confirms the rationale behind the use of multiple queries to compile a reliable set of phraseological data. The many variants which do not conform to predictable patterns and standard synonyms can be found, but only though a combination of keywords, syntactic frameworks and wildcards over repeated searches.

4. The linguistic value of phraseological variation

If finding the data has in itself been something of a barrier to studying variation, so too is the fairly low status attributed to non-standard forms. 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 (Moon 1998, Cignoni & Coffey 2000, Philip 2003). Yet they are considered to be exceptions to the norm because they are non-recurrent and ultimately reducible to the canonical form.

While there are very sound lexicographical and pedagogical reasons for concentrating on repeated patterns, variant forms reveal a great deal about human linguistic behaviour. Not all variants are deliberate, and not all are marked or ambiguous. Variant forms occupy a very substantial grey area lying between the extremes of the canonical form and the eye-catching puns that induce in us the “smugness effect” (Partington 1996: 140). Yet unmarked variation tends to be overlooked entirely, with marked variation typically being compared and contrasted with the canonical form alone. It is important to incorporate unmarked variation into phraseological description, because it allows marked forms to be judged with respect to other variants as well as with the canonical form from which it stems, as the discussion in 4.1 shows.

4.1. Variation and the open-choice principle

To illustrate the difference that an awareness of typical variation can contribute to the evaluations of marked forms, let us consider the some examples of *the pot calling the kettle black*:

1. The words kettle, black and pot suddenly spring to mind.
2. the pan calling the kettle blackbottom
3. Talk about Mr Pot and Mr Kettle
4. POST CALLING THE KESTLE BLACK?

These examples give some idea of the range of variant forms that are encountered in the Bank of English (see Appendix). Considered individually, and with sole reference to the canonical phrase, each of these examples would be understood to be marked: example 1 exploits the keywords and the underlying conceit, but not the standard structure; example 2 replaces *pot* with the semantically-related *pan*, and elaborates the conceit by replacing *black* with *blackbottom*; example 3 personifies *pot* and *kettle*, and exploits the expression without specifying any other components of the phrase; example 4 maintains the overall phraseological patterning, replacing *pot* and *kettle* with the semantically unconnected but phonologically similar alternatives, *post* and *Kestle.*

How marked are these variants? If compared to the full cline of variation (Appendix) it becomes apparent that, with the exception of example 4, they are little more than instantiations of the variation tendencies associated with this particular idiomatic phrase. Each can be read within the paradigm of its variation type as well as in relation to the canonical form. Taking example 3 as a case in point, the personification of *pot* and *kettle* (Figure 7) can be seen to be one of the tendencies that variations to this idiom follow. Read in this context, then, *Mr Pot and Mr Kettle* is not nearly as marked as it seems to be when evaluated against the canonical form alone.

The case of Pot versus Black Kettle
Hello pot, my name's kettle.
pot, meet kettle.
Pat Pot meets Mariah Kettle.
Talk about Mr Pot and Mr Kettle?
dear pot, yours kettle.

Figure 7: Personification of *pot* and *kettle*

Whereas examples 1-3 are contextualised but not wholly context-dependent, examples 4-7 are true puns as they incorporate allusive and connotative meanings into the interpretation of the variant form. Example 4 substitutes one pair of litigants, *pot* and *kettle* with another pair, the Post Office and a member of the public, as revealed in the subsequent context. And although they fall into the same variation type – colour-term substitution – examples 5-7 are also marked, as the substituted terms evoke connotative meanings which are central to the textual meaning of the variant (*grey* denoting dullness; *schwarz* and *noir* German and French cultural connotations respectively).

5. The pot calling the kettle grey?
6. Surely a case of the pot calling the kettle schwarz.

7. It is time the pot stopped calling the kettle 'noir’

The creation of a pun may be considered by some as the reaffirmation of the *open-choice principle* (Sinclair 1991) within phraseological chunks. Certainly, the choice of substituted term appears to be very free indeed. But open choice implies far more freedom than is actually available in this sort of phraseological manipulation, because whatever element is substituted, its meaning is always read in relation to canonical phrase. The new element forces the reader to analyse the phrase both compositionally and non-compositionally, and the overall meaning is a combination of the old phrase and the new, and not a new phrase in its own right (Philip 2003). This type of variation can be described as a *palimpsest effect*. Just as vellum was re-used in medieval times by over-writing the pages of old books –hiding but not erasing the original text – puns constitute a linguistic palimpsest in which the new meaning is written over the old one, but fails to cancel it out completely.

The study of canonical forms alongside non-canonical forms in all their guises highlights the relationship of phraseological items and their cotextual environments. The analysis of corpus data demonstrates that the core of an extended unit of meaning, typically taken to be a single word, can just as readily take the form of an entire phrase, but not necessarily in its canonical form. In fact, various internal parameters are at work in ensuring that a variant retains enough of the canonical form to be recognised as relating to it, but a further factor is external: cotext. Just like single words, fixed phrases too attract colligational features, semantic preferences and semantic prosodies (Philip 2003: 239-40). If the cotext of an innovative use features the norms typically associated with the canonical form, then these norms offset the effects of the internal variation, inducing the reader to relate the variant to its canonical form. In example 8, the cotext provides enough of the expected patterning associated with *green with envy* to ensure that this is the expression that is interpreted alongside the colours of the Irish flag. Interestingly, word-play can be created by locating a canonical form in an atypical cotext, whereby contextually-relevant interpretations merge with the meaning of the (unchanged) phrase. Consider the effect in example 9, where the meaning of *in the pink* (‘happy and healthy’) undergoes forced reinterpretation due to its association with homosexuality.

8. Stunning Miss Ireland Emir Holohan-Doyle wraps our national flag around her – hoping to make her Miss World rivals green, white and orange with envy!
9. Peter Tachell is the author of Europe in the Pink -- lesbian and gay equality in the new Europe.

It is extremely rare to encounter a non-canonical form in an atypical cotext, almost certainly because the proportion of open-choice to idiom principle would be too unbalanced in favour of open-choice, making real-time interpretation very difficult. When this occurs in corpus data, it tends to be found when two similar structures are fused during on-line processing, typically in transcribed spontaneous speech, and represents a “crack” in the phraseological priming (Hoey, 2005:11).

4.2. Variation and the emergence of **ad hoc** *semantic classes*

The variation that occurs in phraseological skeletons often follows unpredictable patterns (see 2.2, 3), and one of the most interesting and potentially important features to emerge from the analysis of variant forms is the phenomenon of *ad hoc* semantic classes.

The *Class Inclusion Hypothesis* (Glucksberg & Keysar 1993), notes that semantic classes are often created attributively, especially in the case of metaphorical and figurative language. Attributive categories differ from taxonomic categories in that the metaphor schema or conceit that is in operation "is used to attribute an organized set of properties to the metaphor topic by projecting onto a target domain, such as *crime*, all of the relevant properties of a source domain, such as *disease*" (Glucksberg & McGlone 1995: 48). If the relationship between the substituted term and the canonical one is based on common attributes, rather than relations of co-hyponomy, then this accounts for much of the non-standard semantics that can be observed in corpus data. It also helps to explain why most language users find little difficulty in interpreting and producing variants such as those in Figure 8, whereas the generation of such sets and the prediction of tendencies in variation, whether computationally or manually, continues to challenge.

black sheep of the family
black sheep of the Mitchell family
black sheep in the Compositae family
black sheep of Britain's financial services
the black sheep of the EU
black sheep of the sporting world.

Figure 8: Attributive semantics in *black sheep of the family*

The semantic productivity occurring along the paradigmatic axis, where terms are substituted not only by members of the same semantic set but also by apparently unrelated terms, is an area of study that is waiting to be explored, and to which corpus analysis can contribute enormously. *Ad hoc* semantic sets remain on the whole something of an unknown quantity. What do they tell us about the ways in which we classify the world around us? Can attributive semantic sets be predicted at all, and if so, how can such knowledge be incorporated into AI and NLP? Up until now, most research in the field has been based on opportunistically-collected and invented examples, and has not been detailed and exhaustive enough to tackle such questions. Using corpus data to study variants means that more examples are available, with the additional advantage that these belong to the same, homogeneous data set.

5. Concluding remarks

This paper has shown that variant forms of fixed expressions can be found in corpora by following quite simple procedures, effectively debunking the myth that their retrieval is governed by happenstance. Non-canonical forms are indeed unpredictable, but they seem to follow tendencies in their variability, suggesting that their apparent randomness is in fact fairly systematic. The types of variation that emerge merit further study as they provide data that is otherwise difficult to access regarding how language users manipulate words and meanings.

The fact that variations tend to follow trends provides tantalising evidence of the idiom principle in operation. As demonstrated in 4.1, changes to the canonical form are necessarily restricted if the meaning value is to be preserved, and non-canonical forms are inclined to occur within a 'canonical' cotext, where the most typical features associated with the canonical form and its extended unit of meaning are all present. This suggests that the phraseology external to the fixed expression shares the role of transmitting meaning, exerting most influence when the internal phraseology is weakened due to variation.

Regularities found in corpus data serve as a benchmark in language description, where they illustrate normal language use. Where then does variation fit? While it is true that the canonical form is more important to document and learn, textual occurrences of fixed phrases are likely to involve variation. This opens up a need for pedagogical and lexicographical descriptions to address fixed phrases from a more inclusive viewpoint, where creativity is considered an integral feature of phraseology. The prospects are enticing.

Notes

1. Throughout this paper, the ‘dictionary citation form’ of a fixed expression is referred to as the *canonical form* (though it should be remembered that established alternative forms can co-exist; see Moon 1998: 122-124). Variants are defined as *marked* if the changes to the canonical form affect the semantic and/or pragmatic meaning conveyed (e.g. puns and word-play), and *unmarked* when the changes cause little or no real change in meaning.

2. The term “fixed phrases” will be used here to include all types of conventional, phraseological chunks such as idioms, metaphors and similes, proverbs, sayings and clichés, bound collocations and binominals. See Moon (1998: 19-25) for a detailed definition of these types.

3. Although there are several NLP applications which incorporate semantic tagging as an aid to identifying possible contenders for multi-word expressions, they do not resolve entirely the problem of locating variant forms: if statistically-based, they “are not accurate for dealing with MWEs of very low frequencies, particularly those occurring only once or twice” (Piao et al. 2005: 379); and if they are dependent on human judgements of lexical and semantic use, they are subject to the same shortcomings that befall manually-entered, trial-and-error corpus query searches – namely that one is unlikely to expect (and look for and find) the unexpected.

4. The analysis of variants makes it possible to sketch out the trends that seem to be followed for individual phrases, but it is much more difficult to predict what these trends might be in the absence of relevant data. To complicate matters further, it is common for several types of variation, whether grammatical or semantic, or involving e.g. rhyme, inversion or truncation, to interact in a single example (Philip 2000: 223).

5. The search routines are defined in the Look Up query language used with the Bank of English; an explanation of the formulae is provided in square brackets.

References

Bindi R., M. Monachini & P. Orsini (1991) *Italian Reference Corpus: General Information and Key for Consultation.* Istituto di Linguistica Computazionale. CNR, Pisa.

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

Cignoni L. & S. Coffey (1998) A Corpus-based study of Italian idiomatic phrases: from citation forms to ‘real-life’ occurrences. In *Euralex '98 proceedings*. Liège: University of Liège. 291-300.

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. Available: http://amsacta.cib.unibo.it

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.

Piao, S.S., P. Rayson, D. Archer, T. McEnery (2005) Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech and Language* 19, 378-397.

Renouf A. & J.M. Sinclair (1991) Collocational Frameworks in English. In Aimer K & B Altenberg (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* London: Longman, 128-143.

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

Appendix

the pot calling the kettle black

```
certainly a case of the pot calling the kettle black. Keatingspeak is now
    it is a case of the pot calling the kettle black # mdash; P. Hudson,
it is not a case of the pot calling the kettle black. <p> McEnroe genuinely
is rather a case of the pot calling the kettle black? The RSPCA used to run
  is this a case of the pot calling the kettle black? Could holidaymakers
  a classic case of the pot calling the kettle black, MEPs were obliged to
 In a prime case of the pot calling the kettle black, 48-year-old Iglesias,
which is a bit like the pot calling the kettle black. <p> As if to answer
 one. <p> It's like the pot calling the kettle black," said parish council
her, because it was the pot calling the kettle black. <p> That Scotland
      Grove. January 12 Pot calling the kettle black? IT is not often that
   </dt> TALK about the pot calling the kettle black # Linfield actually
          would be `the pot calling the kettle black." It's hard to see how
MPs, this really is the pot calling the kettle black." He believes that the
    WVW: Isn't that the pot calling the kettle black? After her time in
     <ZGY> Rather <ZG0> pot calling the kettle black <ZGY> <M0X>
mind, that would be the pot calling the kettle black with a vengeance. And
       </date> Sir: The pot calling the kettle black! Press tells doctors
look like a case of the pot calling the kettle. . . Hannahs make-over has
   Surely a case of the pot calling the kettle schwarz. <p> Germans have
  out of steam. <p> The pot calling the kettle grey? Labour will be a
`It is time the pot stopped calling the kettle 'noir # The scandal is
  happy. I've heard of pots calling the kettle black, but this is more in
    in the kitchen and pots calling the kettle black. I wondered if it was
      because that would be calling the kettle black, but I don't like the
 been for years as well. So we've got a kettle calling the pot black round
      would be a classic example of the kettle calling the pot black. And
     that a case of the Doc calling the kettle black? <h> John, Neil or
 low she says to me the pan calling the kettle blackbottom and I had to
is this a case of the pot-i calling the kettle black # <p> Mahoney laughed
  know, really. This is the pot and the kettle getting together and
 new boss is a useless jerk - a pot and kettle case if ever there was to
 superiority. There is a bit of pot-and-kettle about its outrage. Growth
       North and south are like pot and kettle and neither out-shines the
   out demons and evil spirits. Pot and kettle, or what? KEITH PORTEOUS
 so clever and witty, the words pot and kettle do spring to mind about his
      spent on petrol the phrase, `Pot, kettle and black," springs to mind.
       anything's gone wrong. The words kettle, black and pot suddenly
 awful afternoon, which called to mind `kettle" and `pot", and culminated
wonder, ever heard the words `pot" and `kettle # Take Bruce Anderson, the
 mind you. The case of Pot versus Black Kettle (1927). A BAND in Texas have
 no-smoking area?" Hello pot, my name's kettle. I have a phobia of dirty
  the speaker `sound stupid" (pot, meet kettle). She was sitting next to
  of humility. Talk about Mr Pot and Mr Kettle? Finally I must chide you
    there was a case of dear pot, yours kettle. Mandy didn't bother too
 it," he declared. Pat Pot meets Mariah Kettle. liamfay@clubi.ie <xr> 9108
o move along. </p> <h> POST CALLING THE KESTLE BLACK? </h> <p> SARAH Kestle
```