Rosamund Moon, Cobuild & The University of Birmingham

Data, Description, and Idioms in Corpus Lexicography

Abstract

This paper considers the interaction between theory, data, and lexicographical description, with particular reference to English idioms. It concentrates on one aspect of idioms, that of form and variation. Their variability is very evident in corpus data, but is underplayed in theory and under-represented in dictionaries. The paper looks at specific recurrent types of variation, as evidenced in a large corpus of current English. It then considers the lexicographical consequences.

1. Introduction

One of the challenges in corpus lexicography is to reconcile the conflicting demands of theory, data, and dictionary description. Theories constructed to explain the systems underlying a language are only valid if they account for the phenomena observed in data and are not disproved by further data and counterexamples. But even when theories are proved robust and adequate, this does not necessarily result in something which can be described comprehensively and conveniently in dictionaries. In dealing with idioms, the reconciliation of theory, data, and description becomes even more of a delicate task. This paper explores one aspect of idioms in English, that of form and variation, in order to see how far it is possible to reconcile the diverse constraints of satisfying the system, being true to the data, and producing communicatively and intellectually satisfactory descriptions in dictionaries.

In this case, the base theory is relatively simple. Idioms can be taken here as multi-word units which are non-compositional and typically metaphorical or lexicogrammatically ill-formed in the broadest sense: that is, their meanings cannot be derived from the meanings of their constituent words and morphemes. (This, of course, oversimplifies the term and concept idiom. More detailed discussion can be found in, eg, Fernando 1978, Gläser 1988, and Makkai 1971.) Idioms are fossilized units, restricted collocations with specialized and idiosyncratic meanings, which lie outside the general grammar of the language: see Radford, who talks of sets or classes of anomalous expressions (1988: passim), and Harris, who talks of "a finite learnable stock of 'idiomatic'

material" outside the rules of the language system (1991:43). Zgusta (1971) is a rare case of someone who addresses the problems created by idioms inlexicography.

In practical lexicographical terms, idioms and other kinds of fixed expression are units which must be covered to some extent in general dictionaries, but are usually relegated to subordinate parts of entries and articles, with cursory definitions in monolingual dictionaries, and translations or brief glosses in bilingual dictionaries.

2. Idioms and data

These are the starting-points, but what does the data show? Studies of corpora and other kinds of text demonstrate conclusively that idioms often do not have fixed forms, and are formally unstable. This is a very simple, observable fact, and very important. There are immense repercussions.

In a paper at Euralex '92, I mentioned some preliminary findings of a study of fixed expressions in an 18 million-word corpus of English – the Oxford Hector Pilot Corpus. (This corpus was the object of a research project undertaken by Oxford University Press and Digital Equipment Corporation's Systems Research Center in Palo Alto, California: see Atkins (1992) and Glassman et al (1992).) I reported in particular on frequencies, typologies, and pragmatic functions. The fuller findings of my study supported the distributions and tendencies described in this provisional report. However, something else became very evident. 40% of the 6700 fixed expressions I examined had 'canonical' variations. That is, they regularly varied in form. While my study looked at several kinds of fixed expression, not just idioms, the variation phenomena were spread fairly evenly across typological categories. More recently, I have been looking at idioms in the 200+ million-word Bank of English at Cobuild, and this same phenomenon of variation emerges equally clearly.

Of course, many idioms are apparently invariable, apart from regular inflection for person, tense, or number. For example:

bite the bullet
rain cats and dogs
spill the beans

a red herring
a wet blanket

There may even be further restrictions. For example, corpus data shows that *rain cats and dogs* is used in continuous aspect rather than perfective, and there is no evidence of its passivizing. But in many other cases, there are regular variations. These can be broadly categorized, according to the kind of variation observed, and major groupings are described below.

3. Types of variation

3.1 Lexical variations

Firstly, there are lexical variations: two or more realizations of what can be considered the same idiom semantically and pragmatically. The variable or substituting items are often but by no means always broadly synonymous. Examples attested in The Bank of English include:

add fuel to the fire
add fuel to the flames

fill the bill
fit the bill

go to earth
go to ground

hit the roof
hit the ceiling

scent blood
taste blood

They frequently involve differences between British and American English:

burn your bridges (AmEng & BrEng)
burn your boats (BrEng)

have green fingers (BrEng)
have a green thumb (AmEng)

hold the fort (AmEng & BrEng)
hold down the fort (AmEng)

kick your heels (mainly BrEng)
cool your heels (mainly AmEng)

too big for your boots (BrEng)
too big for your breeches/britches (AmEng)

Often, one of the variations is relatively infrequent, although the idiom form is still unstable:

a can of worms
(a bag of worms)

put two and two together
(add two and two together)

not have two pennies to rub together
(not have two halfpennies to rub together) (BrEng)
(not have two nickels to rub together) (AmEng)

A related group can be described as 'focussed'. Here, the variations have slight shifts in meaning which are predictable from the regular meanings of the varying words, or are different in terms of emphasis or register:

cut your cloth according to –
cut your coat according to your cloth

keep your cards close to your chest
play your cards close to your chest

jump through hoops
go through hoops

pull someone's chain
yank someone's chain

sit on the fence
be/stay on the fence

These can in turn be related to idioms where one component word is replaced with another word more relevant to the immediate topic and contextual situation. For example, *hang up your boots* means 'to retire from an activity, typically football or another sport'; variations replace

boots with another artifact which represents the particular activity from which someone is retiring. Examples in The Bank of English include *aprons* for cooks and cleaners, *gloves* for boxers, *handbags* for ladies-in-waiting, and so on. Is *hang up your boots* indeed fixed enough to be considered the canonical form at all? The question will be discussed further below.

3.2 Systematic and syntactic variations

The second group includes cases where the variations are analysable. Cowie et al (1983:xxxiii–iv) draw attention to groups of variants which involve some notion of possession or attribution:

have (an/no) axe to grind
(with/without) an axe to grind

have a finger in the pie
with fingers in the pie

keep a straight face
with a straight face

give someone the nod
get the nod

Others undergo structural transformations of different kinds:

kick someone in the teeth
a kick in the teeth

turn the screw(s) on someone
tighten the screw(s) on someone
a turn/twist/tightening of the screw(s)

circle the wagons
pull the wagons in a circle

give someone a bloody nose
get a bloody nose
bloody someone's nose

let the cat out of the bag
the cat is out of the bag

In some cases transformations, often involving inversion, result in new lexical items altogether:

blow the whistle
whistle-blowing
a whistle-blower

break the ice
ice-breaking
an ice-breaker

make someone's toes curl
toe-curling

3.3 Other types

Antonymous idioms can also be seen as quasi-systematic, although it may not be possible to predict how they will be realized:

off the record
on the record

have all your marbles
lose your marbles

keep your cool
lose your cool

have a monkey on your back
get the monkey off your back

hold the purse strings
loosen the purse strings
tighten the purse strings

A few idioms have regular exploitations, originally jocular, but now recurring consistently in data as institutionalized variants:

every cloud has a silver lining
every silver lining has a cloud

a wolf in sheep's clothing
a sheep in wolf's clothing

call a spade a spade
call a spade a shovel

Another type comprises truncations of longer idioms or proverbs. The shorter forms are commoner, but the longer forms are still used and implied in context:

it's an ill wind
it's an ill wind that blows nobody any good

it's the (last) straw that breaks the camel's back
the last/final straw

a silver lining
every cloud has a silver lining

swings and roundabouts
what you lose on the swings, you gain on the roundabouts

More extreme and problematic cases are where sets of realizations appear to realize a single idiom semantically but contain any of several synonyms or co-hyponyms, not fixed words (cf. the case of *hang up your boots* above):

a kick up the backside/arse/rear end/bum/bottom (mainly BrEng)
a kick in the butt/ass (mainly AmEng)
a boot up the backside/etc (mainly BrEng)
to kick/boot someone (up) the backside/etc (mainly BrEng)

rose-coloured spectacles/glasses
rose-tinted spectacles/glasses
view/look at something through/with rose-tinted spectacles

wash your dirty linen/laundry in public (mainly BrEng)
air your dirty laundry/linen in public (mainly AmEng)
do your dirty washing in public (BrEng)

wash/air your dirty linen/laundry
wash/air your linen/laundry in public
dirty washing/linen/laundry

An even more extreme case is the quasi-idiom represented in the following item, which loosely means 'mad, crazy, eccentric, stupid':

one sandwich short of a picnic
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

Here well-formedness of use in discourse requires the speaker/writer to be creative and find a new, amusing variation on the theme.

It is possible to see such groups as idiom-schemas (Moon, 1994:182). They share an underlying metaphorical conceit and their lexicalizations are drawn from sets of co-hyponyms. That, however, is a lexicological rationalization; from a lexicographical viewpoint, they are simply nightmares.

4. The problem of identifying idioms

Variation presents a practical problem for corpus lexicographers and lexicologists. 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*. Corpus research into idioms requires awareness of variability – or the variants will not be found – and it is also a matter of serendipity. Software can help, for example with collocational profiles which foreground potential idiom combinations. These are likely to be most useful where the idioms are fixed or contain relatively low-frequency words, or are themselves very common lexical items. Note, however, that 30% of the idioms in The *Collins COBUILD Dictionary of Idioms* (1995) occur less often than once per 10 million words in The Bank of English.

A second problem lies in identifying the canonical form of an idiom; variability is a matter of interpretation. For example, *have an axe to grind*. There is an immutable core *axe to grind*, in restricted collocation after a preceding word such as *have* or *with*, and a premodifier such as

an, any, or *no.* Is the idiom a notional *have an axe to grind,* with variant realizations, or is it an unvarying string *axe to grind* with restrictions on its collocations? Intuitively, *axe to grind* does not seem to be a meaningful unit, but others may disagree.

A third problem is that it is possible to interpret groups of variations as realizing parallel idioms, rather than variations of single idioms. This may well be the case with

be left holding the baby (BrEng)
be left holding the bag (mainly AmEng)

throw someone to the lions
throw someone to the wolves

eat humble pie (BrEng)
eat crow (AmEng)

where there is a reasonably substantial lexical and metaphorical difference, although they mean roughly the same. However, it does not seem right to take the same line with an idiom such as *(wash/air your) (dirty) linen/laundry (in public).*

There is clearly a continuum between these types. In dictionary praxis, variations are more likely to be treated as independent items if their lexis is very different. But the theoretical position is far from clear. Note that theorists such as Rose (1978) and Ruhl (1978, 1989) both identify common underlying patterns in groups of idioms or fixed phrases: system rather than disorder or arbitrariness.

5. The repercussions for dictionaries

The data requires theory to account for variation: sometimes predictable, sometimes motivated, sometimes arbitrary. The evidence is too strong to ignore. How can dictionaries deal with it?

Notions of typicality and prioritization are required here. Large corpora provide evidence of which variations occur and in what proportions. This in turn feeds into dictionary entries, where it is possible to indicate 'normal' or 'stereotypical' variations, as opposed to more marginal ones.

This is easier in specialist dictionaries than in general ones. For example,*The Collins COBUILD Dictionary of Idioms* (1995) deliberately set out to cover idiom variations in depth. Major variations are given as

alternative headphrases, explicitly built into definitions, and indexed. Where necessary, secondary definitions are given to explain particular variations. In contrast, minor variations are mentioned less prominently. In extreme cases, statements such as "this expression is often varied" are made. Examples illustrate the range of variations found.

The phenomenon of variation complicates placement of entries, at least in paper dictionaries, which are organizationally tied to the alphabetical sequence. Users may be unaware that the form they are looking up is non-canonical, and so they may fail to realize where in the dictionary they should look. The only solution here is extensive, careful indexing or cross-referencing, and the problem will largely disappear in purpose-built electronic dictionaries.

Idioms are always difficult to treat lexicographically. This is not just because of the problems of variation and lexical form. There are other problems presented by idioms: how to convey the meaning and usages of what are essentially context-bound items, with vague or plastic meanings and heavy connotations. All these factors have repercussions for (paper) dictionary design in terms of physical extent. It takes a lot of space to deal with idioms fully and effectively.

6. Conclusions

In conclusion, idiom variations are yet another area where corpus investigation leads to new kinds of requirements and descriptions in dictionaries. Lexicographers and lexicologists must adjust and accept that language cannot always be forced into pleasingly neat systems, and they must devise techniques to reflect this; otherwise, they are being misleading. Ironically, corpora show up the woolliness, indeterminacy, and instability of idioms at the same time as they show up strong phraseological constraints on simplex lexical items, where individual words and meanings are often associated with strikingly regular lexicogrammatical patterns: see for example, Francis (1993), Sinclair (1991). Corpus data therefore leads to a redefinition of lexical units, where the lexicon can be seen to be an agglomeration of loosely, rather than tightly, organized groups.

In terms of dictionaries, then, lexicographers have to fit their descriptions to their data. Because dictionaries inevitably have a normative role, it is important for them to reflect data accurately. For lexicologists, the challenge is to further analyse and identify the systems underlying idiosyncrasies; for lexicographers, to develop better ways of

monitoring and describing them. It is only then that theory, data, and description can be truly reconciled.

References

Atkins, B.T.S. (1992) "Tools for computer-aided lexicography: the Hector Project" in: Kiefer/Kiss/Pajzs (eds.) *Papers in Computational Lexicography. COMPLEX '92 Budapest*, Linguistics Institute pp. 1–59

The Collins COBUILD Dictionary of Idioms (1995) London, Glasgow: HarperCollins

Cowie, A.P., R. Mackin, I.R. McCaig (1983) *Oxford Dictionary of Current Idiomatic English* 2 Oxford, Oxford University Press

Fernando, C. (1978) "Towards a definition of idiom: its nature and function" in: *Studies in Language* 2,3 pp. 313–343

Francis, G. (1993) "A corpus-driven approach to grammar" in: Baker/Francis/Tognini-Bonelli (eds.) *Text and Technology. In Honour of John Sinclair* Philadelphia, Amsterdam, John Benjamins. pp. 137–156

Gläser, R. (1988) "The grading of idiomaticity as a presupposition for a taxonomy of idioms" in: Hüllen/Schulze (eds.) *Understanding the Lexicon* Tübingen, Max Niemeyer. pp. 264–279

Glassman, L., D. Grinberg, C. Hibbard, J. Meehan, L.G. Reid, M.-C. van Leunen (1992) *Hector: Connecting Words with Definitions* (SRC Report 92a) Palo Alto CA, Digital Equipment Corporation Systems Research Center

Harris, Z. (1991) *A Theory of Language and Information* Oxford, Oxford University Press

Makkai, A. (1972) *Idiom Structure in English* The Hague, Mouton

Moon, R.E. (1992) "There is reason in the roasting of eggs: a consideration of fixed expressions in native-speaker dictionaries" in: *Proceedings Euralex-92* pp. 493–502

Moon, R.E. (1994) *Fixed expressions and text: a study of the distribution and textual behaviour of fixed expressions in English* Unpublished PhD thesis: University of Birmingham

Radford, A. (1988) *Transformational Grammar* Cambridge, Cambridge University Press

Rose, J.H. (1978) "Types of idioms" in: *Linguistics* 203 pp. 55–62

Ruhl, C.E. (1978) "Alleged idioms with hit" in: *The 5th LACUS Forum* Columbia SC, Hornbeam pp. 93–107

Ruhl, C.E. (1989) *On Monosemy: A Study in Linguistic Semantics* New York, State University of New York

Sinclair, J.M. (1991) *Corpus, Concordance, Collocation* Oxford, Oxford University Press
Zgusta, L. (1971) *Manual of Lexicography* The Hague, Mouton