Article:
Deignan, AH (2015) MIP, the corpus and dictionaries: what makes for the best metaphor analysis? Metaphor and the Social World, 5 (1). pp. 145-154. ISSN 2210-4070

https://doi.org/10.1075/msw.5.1.09dei

MIP, the corpus and dictionaries: what makes for the best metaphor analysis?

Alice Deignan

University of Leeds

MacArthur explores the important issue of dictionary use as a core part of the metaphor identification procedure, MIP, proposed by the Pragglejaz group (2007), and developed by Steen et al (2010) as MIPVU. The point which MacArthur takes issue with is its use is in helping to decide the 'basic' meaning of a lexical unit. This is part of Step 3, which consists of identifying the 'contextual' meaning of a lexical unit, that is, its meaning in the text under examination, and whether it has a more 'basic' contemporary meaning. If a more basic meaning is found, the procedure then moves on to consider the two in relation to each other. MacArthur cites the Pragglejaz definition of basic meanings, which tend to be "more concrete […] related to bodily action […] more precise […] and historically older" (2007, p. 3).

In fact, Pragglejaz do not explicitly describe the use of the dictionary in Step 3, though they use it in Step 1, identifying lexical units. However, they mention its use in a discussion of one of their examples, and elsewhere dictionary use seems to have become the norm for this step; Krennmayr discusses the application of MIP in the examination of a sizeable corpus (2008, p. 103), and writes of using the dictionary "heavily" for Step 3. MIPVU uses dictionaries to identify basic meanings, and Steen et al are unambiguous about its importance: "A meaning cannot be more basic if it is not included in a contemporary users' dictionary" (2010, p. 35).

I am one of the ten parents of MIP, so I naturally support the use of dictionaries as described in the Pragglejaz article. However, I am also an offspring of Sinclairian lexicography, and my true preference, in an ideal world, is the direct, unfiltered use of corpora to inform decisions about meaning. This is much easier in practical and technical terms than it ever has been. A huge amount is now available for free online, some of it having become available since Pragglejaz published their article introducing MIP in 2007. The British National Corpus (BNC, a corpus of approximately 100m words compiled in the early 1990s) has been freely available for

some time, and is now also easy to access and use through several gateways, including Brigham Young University. (The Corpus of Contemporary American English (COCA) can also be accessed from this site, and the free software allows for comparisons between the BNC and COCA.)

In this article, I show how I use a corpus to help me to decide whether a word has one or more meanings, and, if the word has more than one meaning, what kind of relationship there is between each meaning. Following MacArthur, and Dorst and Reijnierse (this issue) I investigated say and its inflections. I used corpus data from the BNC, analysed with standard, largely manual, corpus analytical techniques. The software I used was Sketchengine (Kilgarriff et al, 2014; http://www.sketchengine.co.uk). I then go on to discuss corpus-based dictionaries as an alternative to this direct use of corpus data.

Like Fiona MacArthur, I implicitly assumed that when say is used to denote an action that is clearly writing, it is being used metaphorically. Of course, neither a corpus nor a dictionary can ultimately either support or refute that assumption. They can only provide data that as an analyst, I need to consider as objectively as possible, and, in this case, trying to ignore my gut feeling that there is a mapping WRITING IS SPEAKING. I now describe this process and my findings.

Sketchengine shows there to be 318, 800 citations of say and its inflections says, saying and said in the BNC, that is, 2842.5 occurrences per million words of running text. My first step in this kind of analysis is to use the automatic tools available to look at a word's collocational profile. Sketchengine offers raw frequency, T-Score, Mutual Information (MI) and LogDice. In my experience, looking at collocates is often helpful in identifying fixed expressions, which can then be analysed as a group. This makes for a better analysis of the group, and can speed up the overall analysis. Examples from recent analyses that I have done are with see and light. The collocational profile for see shows that pleased and surprised are significant; going to the concordance data for these collocations shows that this us due to the frequency of the expression *[BE] pleased/ surprised/ to see that […]*. Also frequent are modal verbs, in phrases expressing interpersonal negotiation, such as I can see that/ you will see that. Other frequent collocates are lexis associated with text, such as page and chapter. Frequent collocates for light are red, cigarette, bright,

traffic, green, dark, flashing, blue, shed. These immediately suggest interesting combinations such as red light area, light a cigarette, give the green light, coming in on a blue light (i.e. in an emergency vehicle), shed light on and others. Fascinating questions arise for the metaphor researcher, such as: is red light used metaphorically as an antonym of green light? How frequent is the metonym blue light? Is its recent use as a verb becoming established enough to be evidenced in the corpus? What variants are there on shed light on, and is it ever literal? A happy day can be spent among corpus data investigating these and other questions, and with the luxury of corpora in the billions, such as the Oxford English Corpus, there is enough data to examine several hundred naturally-occurring citations of each collocation.

However, say does not seem to offer such interesting possibilities. A quick look at its collocates, using the various measures available with Sketchengine, shows that the verb is unusually bereft of any lexical collocations that would suggest phraseology or idiomaticity. Measures which favour unusual collocations, such as MI, give prosecuting and spokeswoman, then proper names as top collocates. These are significant statistically because in themselves they are infrequent word forms in the corpus, but they are clearly not of interest as indicators of distinct senses, or fixed expressions. Examination of citations of the collocations confirms this.

The top ten collocates identified using T-score are, in descending order: I, the, he, that, to, it, you, was, and, a. These are all delexical words, and the list is not very different from what a standard frequency list for a corpus looks like. The top 10 most frequent words in the BNC are: the, of, and, to, a, in, to, it, is, was (Leech, Rayson and Wilson, 2001); I, the most significant collocate of say using T-Score, and the most frequent overall, is the 11th most frequent word in the BNC. Manually, I later identified the fixed expressions: *It's h*ard/ difficult/ impossible to say; *It's t*rue to say; but neither is frequent enough relative to other uses for its component lexis to appear in lists of collocates identified using statistical measures. To me, this lack of lexical collocates suggests the possibility—albeit inconclusively at this stage-- that say is almost delexical, making it a hard candidate to study for metaphor. This is similar to the conclusion reached by Dorst and Reijnierse (this issue), through the application of Steen's (2008) theoretical model. I agree with them that this means that such metaphoricity as say may have is perhaps of little interest.

I then analysed a random sample of 1000 citations by hand, attempting to identify the main meanings, and in particular, whether they referred to speaking or writing. In a small number (fewer than 10 citations, or 1%), say is a noun, in expressions such as have your say/ have a say [in]. In a similar number of citations, it can also be an adverb, meaning 'to take an approximate example', for example in the citation:

> … what positions are achievable by turning, say, just two faces F and R.

In 19 citations, it is clearly metonymic; the subject of the verb is a company or a government body. I have not taken the exact frequencies of these occurrences very seriously, because it seems very likely to me that they will be heavily influenced by the genre make-up of the corpus. News texts for example, might tend to include more such metonymic uses than conversation.

Setting aside these citations, I then examined the remainder, the majority, to decide whether speaking or writing was intended. I found that in fact, it is often difficult to tell from the citations. For example, in the concordance line:

> ...Journal of Epidemiology, the researchers said ' It is possible that…'

said appears to refer to speaking; it is followed by a direct quotation. Given the mention of a journal though, I expanded the concordance, which gives:

> Writing in the International Journal of Epidemiology, the researchers said 'It is possible that lung function like height is a marker of the effects of growth and development of childhood environmental factors such as nutrition and infection.'

This seems to indicate that said refers to a message communicated in writing. The reverse was the case for the following concordance citation:

> …assess the role of certain nutrients', he said. Radioactive dust …

I examined the context prior to said, which gives:

> More and more doctors realise that diet can influence disease, William Pryor, professor of chemistry and biochemistry at Louisiana State University, USA,

> told the conference. The National Cancer Institute was sponsoring more than 30 studies to assess the role of certain nutrients, he said.

Here, I took the reference to a conference to indicate that speaking, rather than writing, was referred to.

> The following citation was also ambiguous between speaking and writing:
>
> …disappearing regularly every year' one source said. 'There are indications that some peregrines..

The expanded citation is:

> A number of peregrine falcon nests on the mainland have already been raided to supply the illegal falconry trade to the Continent and the Middle East. 'There are peregrine eggs disappearing regularly every year,' one source said. 'There are indications that some peregrines are stolen by foreigners…

In this case, I thought that it is simply impossible to be certain whether speaking or writing is intended. I decided probably speaking, but I am not confident of this.

In fact, in over 90% of citations, say appears to refer to speaking; there were only 26 of the 1000 where I was able to determine clearly that writing was referred to. I could tell this either because the subject was human but the context indicated that the medium was written, or because the subject was a written text, such as report, statement, manual, or instructions.

I wrote above that I had tried to approach my data ignoring my belief that there is a mapping from SPEAKING to WRITING. However I had still not approached the analysis without any assumptions; I had assumed that there would be a distinction somewhere in the data between speaking and writing, because I believed that this was actually a difference that mattered. At this point though, I started wondering whether either the writer of the text, or the reader, cares or minds whether it was speaking or writing that actually took place. Perhaps say simply means 'communicate', and is unspecific as to medium. In terms of meaning distinctions then, we are looking at not two meanings but one single unspecified meaning that can be applied in various contexts.

My experience of examining corpus data for polysemous words is different from examining these citations for say. In my experience, whatever ambiguity might exist between different meanings of a word in the abstract, there is little or none once naturally-occurring citations are examined. For example, grasp is generally agreed to have at least two meanings, a literal meaning, of 'take hold', and a metaphorical meaning, 'understand quickly', a realisation of a conceptual mapping between understanding and holding in the hand. In the BNC, the most significant collocates of the verbal form of grasp (including its inflections, using logDice) are as follows, in descending order of significance: nettle, firmly, wrist, arm, tightly, significance, incapable, essentials, mettle, opportunity, hand, elbow, hands, failed, scruff, wrists, meaning, complexities. Grasp the nettle is a fixed idiom, and examination of the citations shows that mettle is a mis-spelling and/ or mis-pronunciation of nettle (though apparently so frequent that it may become a legitimate variant of the idiom). Grasp an opportunity exemplifies a different metaphorical sense of grasp. The remaining significant collocates are clearly associated with either the literal meaning or the metaphorical sense of 'understand'. In other words, it is possible to disambiguate meanings to some extent even before citations themselves are examined, just on the basis of collocates. Once citations are examined, it is immediately apparent which meaning is intended, and it is never necessary to resort to further context. This is strikingly different from the picture for say, where the collocates tell us virtually nothing, citations are very often ambiguous, and even full context often does not show whether a 'literal' or 'metaphorical' sense is intended.

For some citations of say, it is not only difficult to say whether writing or speaking is intended, but the whole question seems completely beside the point. For example:

> Much the same can be said about sentence parsing.
>
> Tritium has a 12.3 year half-life, which is the same as saying that 5.5 per cent of its atoms decay each year.
>
> I get very cross when people say that he wasted his talent.
>
> The uncertainty principle says that only one of these measurements can….

> People living on the proposed site say their future is now more uncertain than ever.

For most of these, whether speaking or writing is intended is irrelevant; the utterance is about what was communicated, not how. Indeed, most of these citations of say probably do not refer to a single act of communication, but to a message that was communicated a number of times, quite possibly in different ways. The last citation, for instance, might refer to 'people living on the proposed site' making both spoken communications **and** written ones, through a range of text types.

There are three citations in the sample in which a survey 'says' something, including:

> Three out of 10 aged over 15 admit to it, with half believing marijuana should be legalised. One in four has taken Ecstacy, while LSD users have doubled to 35 per cent, says a survey by jeans company Wrangler.

Surveys in themselves are inanimate, and therefore either the use of say is part of a metonymical utterance, or 'jeans company Wrangler' is a personification. Taking a metonymical interpretation here, there are several possibilities; the metonymy PERSON STANDING FOR COMPANY could be explained as 'a representative of the survey company wrote a press release' or 'somebody who works for Wrangler was interviewed and verbally reported the results of a survey they had commissioned' or by several other plausible scenarios. Which of these is closest to what happened is impossible to ascertain, and it probably does not matter for the communicative purpose of the writer.

To summarise, say has one generalised meaning, to communicate. This can be through speaking or writing. It is sometimes used to summarise the overall message of a series of communications, which can include both speaking and writing. Although only humans can speak or write, the subject of say is often an organisation, metonymically standing for the people associated with it. Say is not therefore a realisation of a mapping of WRITING to SPEAKING. The mapping may exist, but other possible linguistic realisations need to be investigated to confirm this or otherwise.

It will be noted just how much intuition is involved in this kind of corpus analysis. Corpus linguists have not denied the role of intuition in interpreting data: "introspection is dangerous when used as a source of invented evidence, although necessary for the interpretation of real evidence" (Hanks, 2012: 407). However, analysing a corpus is not simply a matter of applying intuitions; a corpus can be misleading if not analysed rigorously. There is a very real danger of finding what you look for and simply not seeing other evidence—a danger that I came close to through my initial conviction that I would find separate groups of citations denoting speaking and writing. Hanks recalls Sinclair describing some linguists' "tendency to use corpora as 'fish ponds' in which to fish for examples supporting their theories, rather than to look and see what is going on" (2008: 220). The data for say could have been treated like this; apparently prototypical examples of speaking and writing could have been found in among the 318,000 citations, and the rest ignored.

As noted above, corpora are generally available these days, but metaphor analysts might not have the time, training or inclination to carry out a concordance analysis for themselves. In this case, a corpus-based dictionary is a good compromise. To write it, highly skilled analysts, in the form of lexicographers, will have spent vast amounts of time studying the data. Most lexicographers of my acquaintance work full time on this and have many years of experience, a profile that few academics, even those who specialise in corpus analysis, can rival.

Dorst and Reijnierse (this issue) point out that the choice of dictionary depends on texts to be analysed. MacArthur expresses surprise at the choice of a learners' dictionary for MIP. I'd argue though that a learners' dictionary is a good choice for several reasons. The first of these is pragmatic: learners' dictionaries were the first to be based on thorough analyses of contemporary corpora from first principles. The market for learning English as a foreign or second Language is huge, globally. The 'engco' model of language forecasting estimates over 1.2 billion second and foreign language speakers of English in the early 2000s, compared to something over 300 million native speakers (Burns and Coffin, 2001). English language teaching materials are a major business, and publishers have invested substantially in their development. The Cobuild project, under the leadership of John Sinclair, exploited this situation, as Hanks writes:

> "For [Sinclair], lexicography was a means to an end…by getting involved with the complex business of dictionary publishing, he could:
>
> (a) get funding for the creation of ever larger corpora (it is hard for us in the age of the Internet to remember how difficult and expensive this was in the 1980s);
>
> (b) encourage the study of lexis as a linguistic level, looking at multiple contexts for each word in order to see what is really going on, rather than accepting speculative theories about what might be going on;
>
> (c) mastermind the creation of dictionaries, grammars, and course books that would help learners get to grips with idiomatic and pragmatic uses of language, as opposed to teaching them word lists and grammatical abstractions." (2008: 220)

As is now well documented, the Cobuild dictionary project led the way for the use of corpus data in lexicography. The first Cobuild dictionary, based on research using the Birmingham University International Language Database, a corpus of 7 million words, was published in 1987; the second edition, based on the 200 million word Bank of English, appeared in 1995. Other learners' dictionaries quickly adopted the use of corpora; 1995 saw the publication of the Oxford Advanced Learners Dictionary (OALD, 5th edition) and the Longman Dictionary of Contemporary English (LDOCE, 3rd edition), both of which used the 100 million word British National Corpus supplemented by smaller corpora, as well as the first edition of the Cambridge International Dictionary of English (CIDE), based on the 100 million word Cambridge Language Survey. Hanks writes that Sinclair "initiated a long and (now) thriving empirical lexical analysis" (2008, p. 220), both in academic research and in lexicography. The 1987 edition of Cobuild was the product of a number of years of research from first principles, detailed in a collection of papers edited by Sinclair (1987). I think it is not exaggerating to see this generation of dictionaries as research tomes in themselves.

Native speaker dictionaries followed rather more slowly: the New Oxford English Dictionary, based on research using the British National Corpus, was published in 1998. The second (2003, revised 2005) and third (2010) editions were

revised and supplemented using the 200 million Oxford English Corpus. Although corpus principles have been adopted by native speaker English dictionaries, this was later, and is less widespread than their use in learner dictionaries. I'd thus argue that learners' dictionaries have a stronger traditional of using corpus data, and an enormous amount of intellectual resource has gone into this.

The other reason for using learners' dictionaries is perhaps even more important, and it concerns who each dictionary is written for and for what purposes. Learners of English need to know about central and typical patterns of the language (Sinclair, 1987). Learners' dictionaries thus devote a large proportion of their space to describing the most frequent words of the language, carefully presenting each of their senses, typical collocates and grammatical patterns, with notes about connotation, style and register. Definitions are illustrated with examples judged to be typical by the lexicographer. This seems to me to be the next best thing to conducting a corpus analysis, and for the researcher in a hurry, perhaps better. In contrast, native speaker dictionaries traditionally cover "hard" or specialist words, and sometimes give some history. Many native speakers use them mainly to check spellings. A native speaker dictionary is much less likely to include a detailed analysis and description of the meanings of say, speak, light, or grasp, because its intended users already know how these words are used. Why would I need to check with a dictionary to see whether say can be used to describe writing, in my first language? I know this. (The last time I used an English dictionary for myself was to check the spelling of minuscule.) I'd argue then that a learners' dictionary gives a truer picture of the way that words are used, from corpus evidence. This is not to criticise native speaker dictionaries; they are doing something different.

Ultimately though, I would argue that the problem starts from the attempt to impose an either/ or distinction on an aspect of communication that is gradable. As Pragglejaz write "we recognize that words, and language more generally, differ in the degree to which they express metaphoricity" (2007, p. 2). There are good reasons for trying to conduct the exercise. Pragglejaz point out that striving for consistency in metaphor identification enables researchers to compare results, and develop a shared understanding of what they are looking at. Attempting to introduce clear-cut distinctions into a field that is characterised by prototypicality is far from unusual in

many areas of scholarship. We are all trying to fit the living language into a set of boxes, and no solution will be ideal.

References

Burns, A. & Coffin, C. (eds.) (2001) Analysing English in a Global Context. London: Routledge.

Dorst and Reijnierse (this issue)

Hanks, P. (2008) The lexicographical legacy of John Sinclair. International Journal of Lexicography, 21 (3), 219-229.

Hanks, P. (2012) The corpus revolution in lexicography, International Journal of Lexicography, 25 (4), 298-436.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014) The Sketchengine: 10 years on, Lexicography 1 (1) 7-36.

Krennmayr, T. (2008) Using dictionaries in linguistic metaphor identification. Selected Papers from the 2006 and 2007 Stockholm Metaphor Festivals, eds. N.-L. Johannesson & D.C. Minugh, 97–115. Stockholm: Department of English, Stockholm University .

Leech, G., Rayson, P. & Wilson, A. (2001) Word Frequencies in Written and Spoken English: Based on the British National Corpus. London: Longman.

Pragglejaz Group (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. Metaphor and Symbol, 22 (1), 1-39.

Sinclair, J. (ed.) (1987) Looking Up: An Account of the COBUILD Project in Lexical Computing. London: Collins ELT.

Steen, G. J. (2008) The paradox of metaphor: Why we need a three-dimensional model of metaphor. Metaphor and Symbol 23 (4), 213–241.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). A method for linguistic metaphor identification: From MIP to MIPVU. Amsterdam & Philadelphia: John Benjamins.