A defence of semantic preference

Gill Philip
University of Macerata
g.philip.polidoro@gmail.com

1 Introduction

Less striking than collocation, less enticing than semantic prosody, it would be fair to say that semantic preference is the most neglected component of Sinclair's (1990) "extended unit of meaning". This paper intends to reignite interest in semantic preference, discussing some of the issues concerning repetition of words *vs* repetition of ideas, and making a case for the continued (or renewed) practice of consulting KWIC concordances in addition to – or indeed instead of – the more sophisticated and speedy computational tools which are available to the corpus linguist.

2 Collocation

Collocation in its received sense is the co-occurrence of two word forms at least twice in the data examined. Collocations can of course involve more than two words (e.g. "tall, dark and handsome", "single white female"), and "word form" is often extended to encompass some or all forms of the lemma ("serial killer/s", "double whisky/whiskies"). Strictly speaking, different word forms tend to collocate differently from one another (compare "naked eye" and *"naked eyes"), and when they do share collocates, the meaning expressed can be surprisingly different (compare "ruddy cheeks" and "ruddy cheek"). But when we talk of a "collocation" we implicitly include all the acceptable variants and exclude the unacceptable ones.[1]

Being by definition a visible and countable phenomenon, collocation lends itself to automation. A collocation famously "stares you in the face just as it is" (Firth 1957: 14). What this means in terms of the KWIC concordance is that collocations repeated down a page are identifiable as blocks separated by an invisible line: the white space of word boundaries. In terms of decontextualised collocations listings, collocates of a given node can be listed in descending or alphabetical order of statistical significance. What I would like to stress here is that it is a simple computational task to retrieve repeated strings of characters and determine the collocations present in text data. It is a little less simple, but still unproblematic, to retrieve variants of character strings, and therefore flesh out the detail of those collocations. Semantic groupings are another kettle of fish. It is much less simple, and decidedly problematic, to retrieve repeated ideas which may or may not be represented with repeated character strings.

[1] How exactly we do so is something of a mystery, and beyond the scope of the present paper to discuss.

3 Disclaimer

Despite having mentioned the automation of data extraction, the focus of this paper is not to investigate or provide an overview of the state of the art of semantic tagging. The very considerable progress that has been made in this area over the past decade is taken as given and I do not intend to belittle the bewildering complexity that semantic annotation entails. What I do intend to dwell upon is a phenomenon which has emerged in parallel with computational advances: the virtual disappearance of the KWIC concordance in corpus linguistics journals, book series and even at conferences such as this one. The thrust of my argument is that the increasingly sophisticated tools which the average corpus linguist has at his or her disposal are lulling linguists into a false sense of security. If collocations can be extracted automatically, it seems, there is no longer any need to count and measure the data by hand.

While it is clear that "the whole point" of using data in electronic format and running it through concordancing software is to introduce an amount of automation to the analysis, it is also true that it was not intended that the computer should be doing all the analysis. Collocations listings and profiles serve a particular purpose within particular types of language study; but they do not tell the full story and they have to be used as an *aid* to analysis, not a *substitute* for it. This is especially true when the corpus in question is not a general reference corpus but rather a collection of texts which are being analysed using corpus linguistics tools.

4 Semantic preference

Semantic preference is the stepping stone which makes it possible to progress from the concrete realities of collocation to the abstract perception of semantic prosody. Semantic prosody is undeniably more attractive a category in corpus linguistics studies: although counting hits on Google Scholar is a crude measure to use, it is interesting to compare the 1750 hits for "semantic prosody" with 1050 for "semantic preference", not just because of the numerical difference, but also because "semantic prosody" is only used

within corpus linguistics (over 70% of the hits also feature "collocation") yet is 66% more frequent than "semantic preference", despite this latter term having currency throughout the cognitive and linguistic sciences (41% of the hits also feature "collocation") and therefore being the more widely-used of the two. For those scholars whose interest lies in semantic prosody – in pragmatics, in evaluation and in connotation – sketching out the semantic preference is a means to an end rather than worth studying in its own right. Yet for collocations enthusiasts, semantic preference is put together by grouping the recurrent collocates – those extracted by the software – which inevitably means that detail is being lost.

5 Semantic preference in corpora

Why is semantic prosody worth bothering with, then? Insofar as large general reference corpora are concerned, collocation profiles may indeed suffice. But increasingly a corpus is a small collection of texts which are being subjected to corpus-assisted analysis, usually in addition to "manual" analysis; and here semantic preference becomes important. The reason is simple: the shorter the text, the lower the number of collocates extracted via statistical measures, and the lower the frequency of any collocations that are found. A lack of lexical repetition is held to be a feature of good writing. The inevitable corollary is that although word forms may not be repeated, it does not follow that certain notions are not being reiterated in the text: they are simply expressed with different words.

Even when the texts in question are not particularly short, repetition may be absent, or it may be absent at certain (potentially) crucial points in the text. This is true in literary texts, where again repetition is avoided as a matter of good style, but may also be used deliberately in order to fix concepts in the reader's mind. Taking J.K. Rowling's seven-book *Harry Potter* series as an example, the physical attributes of the characters are described in repeated formulaic chunks which undergo little if any modification over the course of the 198 chapters, e.g.:

"greasy black hair"= Severus Snape

"pale, pointed face" = Draco Malfoy

"red eyes like slits" = Voldemort

However, no narrative can survive on formulaic language alone. More subtle forms of reiteration are used to create impressions in the reader's mind, the lack of lexical repetition preventing the reader from being able to pinpoint where his or her interpretation stems from. And it is in such places in a text that semantic preference takes precedence over collocation, and the use of KWIC concordances becomes essential. The semantic preference is built up by observing and grouping single instances of words with similar meanings, or which in context appear to form a coherent group, e.g. "dishonesty".

had not answered honestly
guilty secrets
lies
The Life and Lies of Albus Dumbledore

Figure 1: HP7, Ch2 "dishonesty"

Once the semantic preference identification procedure gets under way, it becomes apparent that determining similarity is not always straightforward. Bottom-up semantic groupings are rather more complex than top-down ones, and can be unpredictable. In some cases, formal semantics prevails, in others, there is sufficient sharing of attributes for group membership to be considered, (c.f Hanksand Ježek 2008). Sometimes meaning distinctions merge. In Figure 2, "fall" literally refers to Dumbledore's plummet from a high tower, but simultaneously refers to his death (it is a euphemism for death, but also a metonym in this context).

moments after Dumbledore had fallen
moments after Dumbledore fell, jumped, or was pushed
right after Dumbledore had died,
R- right after Dumbledore ...
you said after Dumbledore ´s funeral
four weeks after Dumbledore ´s mysterious death

Figure 2: HP7 "after Dumbledore's death"

6 Who's afraid of the KWIC concordance?

A call to re-evaluate semantic preference necessarily involves a call to resuscitate the KWIC concordance as an essential and fundamental part of corpus data analysis. (Re)turning to KWIC concordances compels the analyst to reconnect with the original text(s) in the corpus, to engage with " text" as well as "data", and to remember that linguistics is not about data extraction, but about how language works.

References

Firth, J.R. 1957. "A Synopsis of Linguistic Theory 1930-55". *Studies in Linguistic Analysis*. Oxford: Basil Blackwell. 1-32.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge (Mass.): MIT Press.

Hoey, M. 2005. *Lexical Priming: a new theory of words and language*. London and New York: Routledge.

Hanks, P. and Ježek, E. (2008). "Shimmering lexical sets". *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitat Pompeu Fabra. 391-402.

Sinclair, J.M. 1996. "The search for units of meaning". *Textus* 9: 75-106.

Automated semantic categorisation of collocates to identify salient domains: A corpus-based critical discourse analysis of naming strategies for people with HIV/AIDS

Amanda Potts
Lancaster University
a.potts@lancaster.ac.uk

1 Introduction

In modern corpus linguistics, bigger is often better. We now have access to reference corpora containing billions of words, and individual researchers routinely collect *ad hoc* corpora of millions or hundreds of millions of words for specific purposes. This technological advancement is a blessing and a curse; while larger corpora contain more examples of both frequent and infrequent patterns to study, the sheer volume of results is often prohibitive to detailed qualitative analysis. This issue is of particular significance to analysts who are interested in combining the power of corpus linguistic tools with the rich scholarly tradition and interdisciplinary flexibility of additional theories emphasising qualitative analysis.

I here demonstrate one method of exploring the representation of social actors in a large opportunistic corpus. Using a corpus-based critical discourse analytical approach with a strong focus on automated semantic tagging of collocates, I compare construal of *AIDS/HIV patients, victims, sufferers,* and *carriers* in a 161-million-word corpus of American newspaper texts from 1981-2009.

2 Theoretical frameworks

In recent years, several 'schools' of research have found that combining methodological elements of corpus linguistics with a discourse analytical theory "helps researchers cope with large amounts of textual data, thus bolstering…empirical foundations, reducing researchers' bias and enhancing the credibility of analyses" (Mautner 2009: 138), and this synergy is at the centre of a rapidly developing field of research.

A major strength of the corpus linguistics approach to discourse analysis is increased variety and representativeness owing to large but governable sample size. Using statistical measures, Mautner suggests that linguistic 'norms' can be accurately represented, and