# Balancing form and function in corpus research

Serge Sharoff
University of Leeds

## 1. Introduction

In my short remark to the original discussion I wanted to draw attention to a problem which is rarely discussed in corpus linguistics. Nevertheless it is important to understand it to avoid possible false inferences. It is also related to issues discussed in classic philosophy of mind.

The problem concerns the relationship between the form and function (what lies behind a simple statement "X means Y") and the inventory of meanings (the "storage" for all the meanings of an individual language or sublanguage). In the following two sections I will present my perspective and link it to the position expressed by Edmund Husserl in his project of phenomenology as rigorous science, primarily what is known as "Ideas II" (Husserl 1952).

## 2. Form and function in corpus linguistics

Corpora store words, while researchers in corpus linguistics are more interested in meanings that lie behind the uses. This means that our queries producing concordance lines and collocation lists are based on forms, whereas we have to interpret the results of queries using our own intuition. This results in certain subjectivity of our analyses, ultimately limiting Sinclair's famous claim "trust the text". I will refer to this situation as the Form-In-the-Query-Function-In-the-Mind (FIQ-FIM) model. The problem is aggravated by decontextualisation of our analyses. A word is normally used with a certain purpose in the context of its entire text, while in a concordance line this context is reduced down to something like 60 characters. As a result, to interpret a pattern we need to rely on our intuition more than in the case when we analyse an individual text.[1] The situation with collocation lists is even worse: the variety of uses is reduced to statistics of word bigrams or trigrams.

*International Journal of Corpus Linguistics* 15:3 (2010), **419–424**. DOI 10.1075/ijcl.15.3.14sha
ISSN 1384–6655 / E-ISSN 1569–9811

One of the common direct objects of *visit* in the BNC is *union*, but we need to interpret it as *the Soviet Union*, while the same *union* as the collocate of *join* is more likely to be interpreted as *trade unions*.

In a sense, the FIQ-FIM situation is not too distressing, as it reflects the semiotics of interaction. Human communication is always based on interpretation. There is no way of putting an idea from the head of the speaker to that of the listener other than by encoding it in some (linguistic) form, which needs to be interpreted by the listener, who assigns some (not necessarily the same) meanings to the forms of the utterance (Sharoff 2006). Given that the listener has access to the same set of linguistic resources and the context of situation (with obvious variations in the history of their language acquisition), the result of interpretation normally matches the intentions of the speaker. This gives some hope that our linguistic analysis of concordance lines and collocation lists can be also accepted by the community of researchers, but a new community of researchers in 2050 might challenge our trust in the evidence obtained from our concordance lines.

A possible counterargument to the FIQ-FIM model concerns the possibility of having more than word forms in a corpus. Corpora can be annotated with information on lemmas, part-of-speech tags, shallow parsing, as in the Sketch Engine (Kilgarriff et al. 2004), or a full parse, as in ukWaC (Ferraresi et al. 2008). However, any annotation adds an interpretation (Leech 1993), which needs to be taken with a certain degree of scepticism. In practical terms, automatic annotation is useful, but it comes with limitations with respect to its accuracy as well as the decisions made in designing the annotation tagset or the tagger.

The impressive accuracy of 96.7% reported for POS tagging for English (Brants 2000) still means that there is a mistake per every 30 words on average, and even this level of accuracy is misleading, as it has been achieved with tests carried out on the same text type as the training corpus, e.g. *The Wall Street Journal* (for Brants 2000), while the taggers are normally applied to a much wider range of texts. In the end, the accuracy of TnT on random web texts drops down to 92.7%, and even to 85.7% for some genres, as reported in (Giesbrecht & Evert 2009).[2] The accuracy for other languages is often a bit lower, e.g. 95% for Russian (Sharoff et al. 2008), partly because of the greater number of POS tags in a morphologically rich language (about 500), partly because of the importance of long-distance dependencies, which are not captured in a trigram model.

Apart from the reliability, POS tagging and lemmatisation produce interpretations which are not necessarily compatible with the analyses we want to make. For example, is *reduced* in *reduced fat* an adjective or a participle? How about *unwanted*? Is this a participle (with the lemma "unwant") or an adjective? We can prefer either choice (depending on our goals), but the POS tagger annotating the corpus makes a choice before we start our research. Lemmatisation and tokenisation are also not theory-neutral; lemmatisation unifying words to the lower case could destroy unusual creative spelling, e.g., *uNioN*. Given that there are no spaces between orthographic words in Chinese, tokenisation depends on the assumptions made by the tokeniser, e.g. 联合国安全理事会 ("the UN Security Council" in one word) or 联合国 ("UN") 安全 ("security") 理事会 ("council").

## 3. Producing and reproducing meaning

Even after taking care of the fragile link between the form and function, we can run into problems with identifying the set of functions or meanings. A WordNet-like model assumes that all meanings can be enumerated and mapped to all forms. When two forms can express the same meaning, this is a case of synonymy; when one form can express two meanings, this is a case of polysemy. This model is either explicitly taken or implicitly assumed in many cases of corpus research. However, the crucial questions for this model are:

- How do we know which sense inventory is better? WordNet is commonly used in computational research, but its system of senses is considerably different from other machine-readable dictionaries like LDOCE (Procter 1978) or the 1911 version of Roget's thesaurus.[3] Which system is the "right" one?
- How do senses transform in a given context? Semcor (Landes et al. 1998) is an example of a corpus marked with senses taken from WordNet, but there is considerable disagreement between the annotators on whether the synset (synonym set) in Semcor was chosen correctly or not.
- How do new concepts appear? Compare the history of uses of such words like *icon* or *file*. We cannot take any existing set of synsets for granted.
- How do sublanguage communities co-exist and interact? One example is the use of *parallel* texts by translators and computational linguists, with translators assuming that "parallel" texts are texts in different languages having the same communicative function (Bowker 2000). The usage in the translation community has developed independently along a different metaphor ("parallel" means that there is no direct intersection between the two texts), but we can learn and understand 'synsets' used in other communities.

The WordNet model is reasonable for many applications, but the ultimate answer to these questions lies in a more dynamic model which takes into account the process in which meanings are dynamically generated from interpretation of forms in a given situation.[4] A possible computational implementation of this model is closer to the distributional semantics hypothesis: words occurring in similar contexts tend to have similar meanings, so that the similarity class of *experience* is *knowledge*, *opportunity*, *life*, *encounter*, *skill*, *feeling*, *reality*, *sensation*, *dream*, *vision*, *learning*, *perception* (see Sharoff et al. 2006, and further references there). Meanings that are regularly encountered in specific contexts can be "sedimented" into a core, which is evoked when the context is right. When more texts are appearing with usages in other contexts (*icon* as referring to small signs on the computer screen), the set of default associations changes to accommodate the new meaning.

The opposition of the static WordNet model and the dynamic meaning development model relates to the opposition of the static Platonic realm of ideas vs. the Heraclytian flux, which has been known in the philosophy of mind for millennia. One of the instantiations of the second line of thought is Edmund Husserl's phenomenology, which can have many different interpretations, but what is relevant to our research is the concept of the "meaning-endowing acts" which need to accompany any case of interpretation of signs (Smith 1990). The second relevant concept from Husserl's phenomenology is the intersubjectivity of the lifeworld. Our image of the world is recreated in daily experiences of interacting with it, including other living beings. In other words, the reason for the stability (and even "objectivity") of this image lies in the repeated reliable process of meaning constitution, and not in the fact that we have privileged access to the universal Platonic realm of concepts (or WordNet). Getting back to computational linguistics, various distributional similarity methods also take different approaches to interpreting the similarity in corpora, but they often end up with fairly robust "intersubjective" interpretations. For example, the two other methods for producing distributional synonyms for *experience* generated *knowledge*, *skill*, *practice* (in Sketch Engine) and *skill*, *ability*, *knowledge* (in Infomap (Takayama et al. 1999)), which is not too different from the method used by us, which follows Rapp (2004).

## Notes

[1] I agree that we get more evidence from concordance lines than from a single text. No doubt the concordance lines are useful, but my task in this issue is to draw attention to the price we pay for using them.

[2] This accuracy is reported for German using a manually annotated portion of deWaC (Baroni & Kilgarriff 2006), but the reported accuracy for TnT on a German corpus (TIGER) with cross-validation is even higher, 96.9% (Giesbrecht & Evert 2009).

[3] Available at: http://www.gutenberg.org/etext/10681 (accessed April 2010).

[4] Similar criticisms are raised in Kilgarriff (1997).

## References

Baroni, M. & Kilgarriff, A. 2006. “Large linguistically-processed Web corpora for multiple languages”. In *Companion Volume to Proceedings of the European Association of Computational Linguistics*, *Trento.* 87–90.

Bowker, L. 2000. “Towards a methodology for exploiting specialized target language corpora as translation resources”. *International Journal of Corpus Linguistics*, 5 (1), 17–52.

Brants, T. 2000. “TnT -a statistical part-of-speech tagger”. In *Proceedings of 6th Applied Natural Language Processing Conference, 29 April — 4 May, Seattle*, 224–231.

Ferraresi, A., Zanchetta, E., Bernardini, S. & Baroni, M. 2008. “Introducing and evaluating ukWaC, a very large web-derived corpus of English”. Paper presented at *The 4th Web as Corpus Workshop: Can we beat Google?, Marakkech, 1 June*, as part of the *LREC 2008 Conference.* Available at: http://clic.cimec.unitn.it/marco/publications/lrec2008/lrec08-ukwac.pdf (accessed April 2010).

Giesbrecht, E. & Evert, S. 2009. “Part-of-Speech (POS) Tagging — a solved task? An evaluation of POS taggers for the Web as corpus”. In *Proceedings of the Fifth Web as Corpus Workshop (WAC5), Donostia-San Sebastián, 7 September,* 27–35. Also available at: http://cogsci.uni-osnabrueck.de/~severt/PUB/GiesbrechtEvert2009_Tagging.pdf (accessed April 2010).

Husserl, E. 1952. *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie, Zweites Buch: Phänomenologische Untersuchungen zur Konstitution*. Edited by M. Biemel. The Hague: Martinus Nijhoff Publishers. [1989. *Ideas II: Studies in the Phenomenology of Constitution*. Rojcewicz, R. & A. Schuwer (transl.). Dordrecht: Kluwer.]

Kilgarriff, A. 1997. “I don’t believe in word senses”. *Computers and the Humanities*, 31 (2), 91–113.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. 2004. “The Sketch Engine”. In *Proceedings of Euralex 2004, Lorient, France,* 105–116. Also available at: ftp://ftp.itri.bton.ac.uk/reports/ITRI-04–08.pdf (accessed April 2010).

Landes, S., Leacock, C. & Tengi, R. I. 1998. “Building semantic concordances”. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press, 199–216.

Leech, G. 1993. “Corpus annotation schemes”. *Literary and Linguistic Computing*, 8 (4), 275–281.

*Longman Dictionary of Contemporary English*. 1978. Procter, P. (Ed.). UK: Longman Group Ltd.

Rapp, R. 2004. “A freely available automatically generated thesaurus of related words”. In *Proceedings of the Fourth Language Resources and Evaluation Conference, LREC 2004, Lisbon*, 395–398.

Sharoff, S. 2006. “How to handle lexical semantics in SFL: A corpus study of purposes for using size adjectives”. In G. Thompson & S. Hunston (Eds.), *System and Corpus: Exploring Connections.* London, Oakville: Equinox, 184–205.

Sharoff, S., Babych, B. & Hartley, A. 2006. “Using comparable corpora to solve problems difficult for human translators”. In *Proceedings of International Conference on Computational Linguistics and Association of Computational Linguistics, COLING-ACL 2006, Sydney,* 739–746. Also available at: http://acl.ldc.upenn.edu/P/p06/p06–2095.pdf (accessed April 2010).

Sharoff, S., Kopotev, M., Erjavec, T., Feldman, A. & Divjak, D. 2008. “Designing and evaluating a Russian tagset”. In *Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008, Marrakech.* Also available at: http://corpus.leeds.ac.uk/serge/publications/lrec2008-msd.pdf (accessed April 2010).

Smith, B. 1990. “Towards a history of speech act theory”. In A. Burkhardt (Ed.), *Speech Acts, Meanings and Intentions: Critical Approaches to the Philosophy of John R. Searle*. Berlin: de Gruyter, 29–61.

Takayama, Y., Flournoy, R., Kaufmann, S. & Peters, S. 1999. “Information retrieval based on domain-specific word associations”. In *Proceedings of PACLING ‘99, Waterloo, Ontario, Canada, June 1999.*

## *Author’s address*

Serge Sharoff
Centre for Translation Studies
University of Leeds
Leeds, LS2 9JT
UK

S.Sharoff@leeds.ac.uk