Usage based grammar and rarely-used structures: What is adequate data?

Peter Uhrig
FAU Erlangen-Nuremberg

It is commonplace for many cognitive approaches to grammar to state that they are usage-based. While the use of the term has broadened over the years, the basic idea remains, as defined by Langacker: "Substantial importance is given to the actual use of the linguistic system and a speaker's knowledge of this use" (1987: 494). This usually includes a strong focus on storage (see for instance Goldberg (2006)) and often a very prominent role of frequency (see for instance Bybee (2007, 2010)). The two major empirically sound paradigms for data gathering – experimental and corpus research – both have their specific advantages and drawbacks so that for many research questions one of the two is the obvious choice (see however Mukherjee (2004) and Gries (2011) for strong arguments in favour of corpus use).

The present paper will illustrate what problems researchers are faced with in the case of low-frequency constructions. For instance, if we want to find out if a certain verb or adjective allows for a certain type of clausal subject, whether extraposed or not, one may consult a corpus or ask native speaker informants, and none of the two methods seems to be inherently unsuitable for the job. Thus if we want to know if *illegal* takes a *that*-clause subject, we can craft a sentence and submit it to native speaker informants:

(1) It was **illegal** that she was fired just because she was pregnant.

Here, three native speakers were interviewed and two accepted it, so one may be inclined to treat the sentence as acceptable from an experimental linguist's perspective (if the tendency also holds for a larger group of informants).

If we turn to the corpus – here, parsed corpora of the *Treebank.info* project (Uhrig/Proisl 2011) totalling roughly 1.5 billion words were used – we find the structure exactly once:

(2) It is **illegal**, degrading and inhuman that prisoners in Ireland must carry out slops from their toilet cells each morning. (news)

We can conclude that the structure practically never occurs (one may even discard the example above, where one could attribute the acceptability to the co-ordination of *illegal* with *degrading* and *inhuman*), while the competing *to*-infinitive structures occur roughly 2,000 times in the same dataset. From a corpus linguist's perspective, one may thus be inclined to treat the structure as unacceptable (see Stefanowitsch (2006) for how to identify significantly absent structures).

Since 1.5 billion words correspond to roughly 60 to 150 years of linguistic experience of a native speaker (60 following Dąbrowska's (2004: 19) estimates, 150 according to Aston/Burnard (1998: 28)), it is highly unlikely that a native speaker informant has ever read or heard the adjective *illegal* used with a *that*-clause extraposed subject before and it follows that cognitively no entrenchment of such a structure is to be expected. Accordingly, in this particular case, our experiment obviously did not measure usage at all. The problem will become even more challenging in the case of items for which all competing constructions are relatively rare.

The paper will present further evidence to show that researchers relying on one of the two major research paradigms may benefit from double-checking their results with the help of the other to arrive at better data and more reliable data interpretation particularly in the case of low-frequency data.

References

Aston, G., and L. Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press.

Bybee, J. 2007. *Frequency of Use and the Organization of Language*. New York: Oxford University Press.

Bybee, J. 2010. *Language, Usage and Cognition.* Cambridge: Cambridge University Press.

Dąbrowska, E. 2004. *Language, Mind and Brain: Some Psychological and Neurological Constraints on Theories of Grammar.* Edinburgh: Edinburgh University Press.

Goldberg, A. 2006. *Constructions at Work.* Oxford: Oxford University Press.

Gries, S. 2011. Corpus data in usage-based linguistics: What's the right degree of granularity for the analysis of argument structure constructions? In *Cognitive linguistics: convergence and expansion*, eds. M. Brdar, S. Gries and M. Žic Fuchs. Amsterdam/Philadelphia: John Benjamins, 237-256.

Langacker, R. W. 1987. *Foundations of cognitive grammar.* Vol. 1: *Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Mukherjee, J. 2004. Corpus data in a usage-based cognitive grammar. In: *Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, ed. K. Aijmer and B. Altenberg. Amsterdam: Rodopi, 85-100.

Stefanowitsch, A. 2006. "Negative Evidence and the Raw Frequency Fallacy." *Corpus Linguistics and Linguistic Theory* 2, no. 1:61–77.

Uhrig, P., and Proisl, T. *Treebank.info*. Details and access available at <http://treebank.info>.