CORPUS SEMANTICS

Michael Stubbs

The term "corpus semantics" is a shorthand way of referring to a combination of data, methods and theory: empirical observational evidence from large computer-readable corpora is used to formulate hypotheses about meaning. The data-intensive methods depend on the ability to store and search large text collections for lexical and grammatical patterns. One of the most significant findings has been the pervasiveness, in all kinds of texts, of predictable collocations, and a major theoretical puzzle has been how best to model multi-word units of meaning.

Computer technology was used in the 1960s and 1970s to do preliminary work on word frequency and on collocations (Kučera & Francis 1967, Sinclair et al 1970/2004, Allén 1975). However, it was only by the mid-1980s that computers became powerful enough to make realistic contributions to the construction of the first corpus-based dictionary (Cobuild 1987), and only from around the mid-1990s, that the widespread availability of large corpora and user-friendly software made the data and methods available to scholars with no expertise in computing.

Allén, S. et al (1975) *Nusvensk frekvensordbok baserad på tidningstext. (Frequency dictionary of present-day Swedish based on newspaper material).* Stockholm, Almqvist & Wiksell.

Cobuild (1987) *Collins Cobuild English Language Dictionary.* (Editor in Chief J. Sinclair) London: HarperCollins.

Kučera, H. & Francis. W. N. (1967) *Computational Analysis of Present-Day American English.* Providence, Brown University Press.

Sinclair, J., Jones, S. & Daley, R. (1970/2004) *English Collocation Studies: The OSTI Report.* (ed.) R Krishnamurthy. London: Continuum. [Originally mimeo 1970.]