Stefan Evert (Universität Osnabrück):

Why the British National Corpus isn't a random sample:
Understanding and managing non-randomness in corpus data

Dienstag, 13.01.2009
16:15- 17:45 Uhr
Raum A 702

Abstract:

In order to make proper use of quantitative data obtained from corpora, it is essential to perform a statistical analysis that corrects for chance variation in the frequency counts. Many different statistical methods are available for this purpose, but all of them are based on the assumption that the observed data constitute a random sample from a large population ("the language").

This randomness assumption is rarely tenable for corpus data, most often because the unit of measurement — typically words (e.g. for lexical frequencies and morphological distributions) or sentences (e.g. for syntactic phenomena) — does not coincide with the unit of (random) sampling — typically entire texts or long contiguous excerpts (ranging from 2,000 words in the Brown corpus to 50,000 words and more in the British National Corpus).

In my talk, I will first motivate the need for a random sample analysis of corpus frequency data and review basic statistical procedures such as binomial and chi-squared tests. In the second part of the talk, I take a closer look at the causes and consequences of non-randomness introduced by a mismatch between the unit of measurement and the sampling granularity. The theoretical discussion of non-randomness effects will be illustrated with thought experiments (the "library metaphor") as well as empirical data from the BNC and other corpora.