DIENSTAG, 14. MÄRZ 2017, 15.00 UHR

QUANTITATIVE ANALYSES OF WORDS, THEIR COMBINATION(S), AND THEIR CO-OCCURRENCE WITH CONSTRUCTIONS

Stefan Th. Gries (Santa Barbara)

Over the last 20 years, linguistics as a discipline has changed considerably. On the one hand, linguistics has become increasingly quantitative in nature to the point where published studies routinely approach even qualitative-seeming topics with statistical methods; on the other, linguists seem to be using corpus data more often. The simultaneity of these trends is not surprising: corpus data are inherently frequentist/quantitative in nature, which means that corpus analyses has a natural tendency to be at least somewhat quantitative.

In addition to this, however, quantitative approaches have also sparked newer theoretical developments (e.g. usage-based linguistics) or new methodological paradigms (such as the widespread adoption of statistical methods in natural language processing at the expense of exclusively rule-based approaches).

In this talk, I will be concerned with quantitative methods in the exploration of the lexicon or, more globally, the constructicon; given that terminological choice, the talk will be from the perspective of the family of usage-based approaches as exemplified in the works of scholars such as Goldberg, Ellis, Bybee, and others and assume a lexis-syntax continuum. In particular, I want to draw attention to how different kinds of frequency data derived from corpora can and should inform our exploration of the lexicon/constructicon.

In the first part, I will discuss how simple corpus-derived statistics can aid in the exploration of phonological aspects of the lexicon, specifically phonaesthemic characteristics of simple words as well as phonological similarity relations in idiomatic expressions can be uncovered with relatively simple statistical methods such as frequencies and conditional probabilities derived from corpora.

In the second part, I will move to slightly more advanced methods based on work concerned with the identification of multi-word units.

Specifically, I will compare two recent recursive algorithms that attempt to find multi-word units in unannotated corpora, one based on co-occurrence frequency, one on measures of association. Four small case studies (two rating studies, a comparison with tagging protocols, and language acquisition data) suggest that the latter appears to be useful to some extent and do better than the former and point to the importance of using the right statistical tools in such studies.

The final part is concerned with the question of association measures and their application to combinations of words as well as combinations of words and

(more) syntactic constructions/patterns. I will discuss aspects and problems of previous work and will propose four dimensions of information that will hopefully inform future work on the lexicon/constructicon that utilized corpus-based co-occurrence data.