University of Cologne

Philosophische Fakultät

CCLS

COLOGNE CENTER OF LANGUAGE SCIENCES

CCLS: Stefan Th. Gries (UC Santa Barbara)

Towards more and more independent dimensions in corpus linguistics

Monday, 6 December 2021, 2:00 p.m.

For the most part, corpus linguistics is an inherently quantitative statistical discipline, in only in the trivial sense that any corpus-linguistic study will depend in some way or other on the frequencies with which certain 'things' occur and co-occur in (parts of) corpora. These frequencies are then used in a variety of ways, either directly or as input to compute other corpus-linguistic statistics such as dispersion, association (e.g., collocations or colligations/collostructions), keyness, and others. However, most such research in the past has been too low-dimensional, which can surface in two ways: (i) we might not use all the dimensions of distributional information that may be relevant to a certain topic/question; (ii) the dimensions we do intend to cover actually conflate various sources of information in complex ways. In this talk, I will discuss three applications that try to address these issues in different ways and in two parts: Part 1 is concerned with (i) and discusses two applications that attempt to demonstrate the advantages of increasing the number of dimensions we consider; Part 2 is concerned with (ii) and outlines how one might reconsider the role that many traditional corpus-linguistic measures (can) play.

In part 1, the first application is concerned with keyness, i.e. the way in which corpus linguists usually try to identify which words are characteristic for, or key to, a certain topic, register or genre. The traditional way to do so is based on computing for each word an association measure that quantifies how much a word is over represented in a target corpus of interest vs. a reference corpus, but I will outline an improved version of this that also considers the degree to which the words in question are distributed across the target and reference corpora. The second application is concerned with the identification of multi-word units (MWUs, such as according to, up for grabs, on the other hand, ...). Some existing work at least within corpus linguistics has approached this task on the basis of frequency and/or association measures such as MI or the log-likelihood ratio G^2. In this case study, I will discuss an approach to MWUs that, in its current brain-storming version at least, involves no less than 8 dimensions of corpus information and I will discuss the no-too-bad, but also not-yet-great results of this approach.

In part 2, I will discuss some recent work that attempts to evaluate existing corpus-linguistic measures with regard to how much they measure what they pretend to measure. Using both association measures and dispersion measures (AMs and DMs), I will argue that some of the most widely used AMs and most DMs actually do not (just) measure what one would think they do - association between elements and the distribution of elements across a corpus respectively - and will exemplify a proposal to 'cleaning up' these measures so as to increase the validity of these measures and make it possible to better study the role of the constructs they represent (for, for instance, cognitive-linguistic or psycholinguistic studies).