Plenary-3

Stefan Th. Gries (University of California, Santa Barbara)

Quantitative and exploratory corpus approaches to registers and text types

Following Baker's (2006) set of definitions of *discourse*, in this plenary I will be concerned with discourse$_2$, a notion of discourse that involves "different types of language use or topics" and that is related to notions such as *genre*, *style*, and/or *text type*. With that orientation, I will cover the following three different aspects, which I consider relevant for a large number of corpus-based studies to discourse$_2$, but also to the other ways in which *discourse* is understood.

1. As many others before me, I will briefly address the dichotomy of qualitative vs. quantitative research that is so commonly made in discourse analysis, sociolinguistics, socio-cultural linguistics, to name but a few disciplines. Unlike many others before me, however, my way of addressing this dichotomy will not consist of the usual we-need-both-types-of-approaches-complement-each-other call to arms, but I will attempt to make the point that, in a very trivial sense, qualitative work *is* ultimately based on quantitative work and, thus, needs to take lessons from quantitative methods into consideration.

2. I will summarily discuss a variety of previous case studies with an eye to demonstrate the potential of bottom-up approaches to register-like corpus parts. On the one hand, I will outline ways in which corpus divisions into registers as made by corpus compilers can be tested for their discriminatory power. On the other hand, I will discuss how the study of any linguistic phenomenon in corpora should feature bottom-up analytical steps in order to identify the most promising registers/text types for a particular linguistic phenomenon.

3. In an attempt to extend previous work in this area, I will then exemplify a new idea how such bottom-up approaches can be made more comprehensive. Rather than using only a particular phenomenon in question as a diagnostic for the corpus parts to distinguish, as mentioned above and argued for in previous work, I will explore an approach to identifying homogenous parts in corpora that can include more diagnostics, assign different weights to them depending on which diagnostics are considered (more) important, and use these diagnostics in a cluster-analytic approach (with various ways to follow up and ascertain the discriminatory power of the results).

To the extent possible, I will discuss how such methods reflect and underscore similarities between corpus linguistics in general, register/text type-based studies in corpus linguistics in particular, and psycholinguistic theories. Linguistic elements to be discussed include lexical items, various types of *n*-grams, grammatical patterns/constructions, and argument structure constructions, plus maybe more; data to be discussed are from the BNC, the BNC Baby, the ICE-GB, plus maybe more.