Sources of evidence for linguistic analysis

Round table discussion
Leeds University

Is linguistics an empirical science? How reliable are the data on which linguistic analyses and theories are based? These questions are not new, but in light of the disturbing findings of the Reproducibility Project in psychological sciences, the need to revisit them has become more pressing. This round table discussion will start with presentations from three postdoctoral researchers, who will discuss the question of data collection and analysis and the interpretation of linguistic evidence.

The audience is invited to submit questions in advance (by email to philsoc.roundtable@gmail.com), by the 6th of November.

Chair: Prof. Cécile De Cat (Leeds)
Panelists: Dr Aaron Ecay (York), Dr Seth Mehl (Sheffield), Dr Nick Zair (Cambridge)

Big and small data in ancient languages
(Nicholas Zair, Cambridge)

Ancient linguists often have to deal with 'bad' data; in practice, this often means 'small' data, and a particularly fruitful approach has been to view the data through the lens of sociolinguistic theory, especially with regard to multilingualism, language as a marker of identity, and language death. However, there are dangers in using theory to 'fill in the gaps' in our data; for example, we might wonder how relevant modern cases of language death are to ancient linguistic situations, and to what extent disparate pieces of evidence can reasonably be made to fit into a narrative (pre)defined by 'what we expect'. On the other hand, in recent years there has been a great increase in digital resources for ancient languages. These often allow much faster collection and analysis of large amounts of data, but can also pose challenges – not least the danger of making 'bad' data worse. I will discuss issues surrounding sources for and use of ancient linguistic data, providing case studies from ancient Italy.

Corpus semantics: From texts to data to meaning
(Seth Mehl, Sheffield)

Corpus semantics applies quantitative data science techniques to questions of meaning in language. In order to be successful, such research must account for the nature of corpus data and the nature of semantic meaning, and connect the two. In this talk, I explore the nature of linguistic data, the processes for collecting and structuring it, and the possible relationships between corpus data, semantic meaning, and quantitative calculations. Can computers count words to find meaning? In addressing this question, I present examples from my own research on the Linguistic DNA project, which employs computational methods and close reading to model semantic and conceptual change across tens of thousands of texts, and over a billion words, of Early Modern English.

Bridge to nowhere?
Progress and problems in relating syntactic variation and change to syntactic theory.
(Aaron Ecay, York)

Modern formal syntactic theories rely on a crucial methodological assumption: that speakers' mental representations can be accessed and interrogated by way of acceptability judgments of test sentences created by the investigator. When studying extinct language varieties, however, native speakers are not available to provide judgments. And when variable phenomena are studied judgments are no help, since speakers (by and large) accept all variants equally. Corpora, and the quantitative data they provide, are one methodology for studying syntactic variation and change. But how can the data yielded by corpora connect with theoretical analyses?

In order to bridge the gap between the theoretical and quantitative-empirical domains, linguists have developed linking hypotheses. In this talk, I will review several of these, such as:

- The constant rate hypothesis (Kroch 1989)
- The exponential model of morphophonological rules (Guy 1991)
- The variational model of syntactic acquisition (Yang 2000)

After reviewing the models and how they serve as effective linking hypotheses, I'll go on to consider the work that has followed from them. These models, I will argue, have all encountered challenges arising from the increase in available data and quantitative sophistication brought about in recent decades by the computer revolution.

What, then, will happen next at the interface between syntactic theory and quantitative data? Several developments are on the horizon. Firstly, traditional syntactic acceptability judgments have taken a recent quantitative turn (see e.g. Sprouse et al. 2013). Secondly, newly available sources of data, larger by orders of magnitude than what was previously available, have in the past several years been brought to bear on questions of both historical and contemporary variation, uncovering finer variation than was previously assumed to exist (see e.g. Grieve 2012). Finally, new quantitative linking hypotheses are being developed to augment those listed above (see e.g. Kauhanen 2016). These point to a future where the gap between syntactic theory and syntactic variation is narrower, and will be bridged by a common understanding of the processes that produce and regulate variation.

Grieve, J. (2012) “A statistical analysis of regional variation in adverb position in a corpus of written Standard American English.” Corpus Linguistics and Linguistic Theory 8, pp. 39-72.

Guy, G. (1991) “An exponential model of morphological constraints.” Language Variation and Change 3, pp. 1-22.

Kauhanen, H. (2016) “Neutral change.” Journal of Linguistics. (Accepted to appear; available online).

Kroch, A. (1989) “Reflexes of grammar in patterns of language change.” Language Variation and Chance 1, pp. 199-244.

Sprouse, J., Schütze, C., Almeida, D. (2013) “A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010.” Lingua 134, 219-248.

Yang, C. (2001) “Internal and external forces in language change.” Language Variation and Change 12, 231-250.