Historical Corpus Linguistics: Methods and Applications
Saarland University, Department of Computational Linguistics & Phonetics, 16-17 June 2016

Magnus Huber

Abstract

Introducing the Old Bailey Corpus, version 2.0

This presentation introduces version 2.0 of the Old Bailey Corpus (OBC 2.0), a corpus of speech-related texts covering the 1720-1913 period and based on the proceedings of London's central criminal court (http://www.oldbaileyonline.org). With more than 26 million words of spoken Late Modern English, OBC 2.0 is almost twice as large as version 1.0. The corpus is hosted at the CLARIN-D Fedora Commons Repository of Saarland University and is freely accessible for non-commercial research (https://fedora.clarin-d.uni-saarland.de/cqpweb/obc).

Even though scribes, printers, publishers and the constraints of the printed medium did of course act as linguistic filters, the fact that the proceedings of the Old Bailey were taken down in shorthand makes them a reasonably close representation of what was said in the courtroom. The OBC thus offers the rare opportunity of analyzing speech-related texts in Late Modern English, a period that has been neglected both with regard to the compilation of primary linguistic data and the description of the structure, variability, and change of English.

A particular strength of the OBC is that it includes a very high number of speakers and therefore constitutes a fairly representative sample of spoken, rather formal Late Modern English in the courtroom setting. Every speaker turn has detailed annotation for sociobiographical (gender, social class, age), pragmatic (role in the trial) and textual variables (the shorthand scribe, printer and publisher of individual Proceedings). The combination of sheer size and rich annotation make the OBC the largest diachronic collection of spoken English with this detail of utterance-level sociolinguistic annotation. The corpus is a valuable resource for multivariate analyses in historical linguistics in general, and historical sociolinguistics and pragmatics in particular, enabling researchers to correlate linguistic change and structural variability with the social context.

The presentation will give an overview of the compilation and structure of the OBC and illustrate its linguistic potential by way of example analyses like the development of relativizers.