Historical corpus linguistics 2.0.
From language community to individual and from the individual researcher to team-based research

Peter Petré & Enrique Manjavacas

We present the novel corpus and analytic tools developed for the project the *Mind-Bending Grammars* project, which aims at relating longitudinal grammatical change to change across the lifetime of adult speakers. Historical linguistics typically has treated language change as taking place in the language community, stating that change is conventionalized innovation. However, every language community consists of individual speakers, whose mental grammars also undergo change when the community as a whole is changing. Until recently, available corpora were simply not big enough to allow for any quantitative research on this individual level. This has changed with the advent of massive digitalization projects such as *Early English Books Online*. *Mind-Bending Grammars* builds on these digitalization projects and is currently developing a 125 million word corpus which spans well over a century. Unlike existing corpora, however, this corpus consists of 50 independent corpora, each representing the collected works of a single prolific individual. With an average of about 2.5 million words per individual, each such subcorpus is substantially bigger than the seminal Helsinki Corpus, which covered a whole millennium.

While much of the material was already digitally available, designing this corpus brought a number of both old and new challenges with it. First, only individuals whose collected works surpass half a million words were eligible. With preliminary EEBO word counts & estimates from ECCO, this still resulted in more than 1300 candidates. This set was then narrowed down on the basis of criteria such as evenly distribution of words across the lifetime, balance between religious and literary authors, connection to the London intellectual elite, spread over four generations etc. In the end, not many more than the intended 50 remained. The technical preprocessing as well, including conversion of xml, (re)formatting of metadata, tokenization and indexing, will be discussed.

In addition to the corpus design, which truly brings historical corpus linguistics to the level of the individual, the project also takes a new stance on corpus annotation. The project team will look at eight case studies of syntactic change in the same corpus. The concordancing software is built in such a way that team members will see each other’s annotations appear in real-time in the corpus. Also, the software gives a customizable set of notifications whenever a new analytic category is inserted by a user or a user has annotated data. This enables efficient discussion of difficult-to-analyze cases. All annotation also ends up as xml-code in the preprocessed corpus files. This means, first, that a whole range of things can be easily calculated, such as for instance clustering effects of particular constructions, and, second, that all analyses remain centrally associated with the corpus, and are not disconnected (and so lost for other researchers) in an excel sheet or standalone concordancer.

We will briefly demonstrate the corpus and software (in its current state), and show with a brief case study, on the grammaticalization of *be going to*, how the era of ‘Big data’ enables corpus linguistics to ask and answer research questions that were unreachable until now.