From 500 passages to 50,000 books: Creating and using a large-enough historical corpus

Mark Liberman
University of Pennsylvania

For well over a century, scholars have been using historical corpora in systematic studies of morphological and syntactic change. Recently, the development of large parsed corpora has made an enormous quantitative improvement in the efficiency of such research, thereby changing the field qualitatively as well. However, an important problem remains.

Thus the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) has 1,737,853 words in 448 texts (between 75 and 16,593 words per text, or 3,879 on average), covering the 210 years from 1500 to 1710. This is 8,275 words per year, or 82,750 words per decade on average. Despite the heroic labor involved in creating this resource, there remains a problem with sample size for those phenomena that don't occur at rates high enough to give a clear signal with reasonable temporal resolution at this scale. To take a trivial example, if we were interested in the history of that-elision in the VERB (that) S construction, and wanted to look particular verbs, we would have to face the fact that forms of e.g. believe with sentential complements occur at rates of about 200 per million words, for an expected count of about 16 examples per decade in this collection, while forms of decide with sentential complements occur at a rate of about 20 per million, giving an expected count of less than two examples per decade.

And the size of the sample in sources is at least as much of a problem as the size of the sample in words. There is considerable source-specific variation; each source is of course unique to a particular time period; and 448 sources over 210 years is an average of only 21 sources per decade. As a small indication of how much variation different sources can introduce, I note that the Second Quarto edition of Hamlet (1604) has the famous exchange:

*Ham.* Madam, how like you this play?
*Quee.* The Lady doth protest too much mee thinks.

whereas the First Quarto (1603) has

*Ham.* Madam, how do you like this play?
*Queene.* The Lady protests too much.

This is not an entirely isolated difference: Q1 Hamlet has 8 instances of "do you VERB" in questions, while Q2 Hamlet has just 3; Q1 has 12 instances of third-singular verbs in -eth, while Q2 has just 5. These differences presumably don't reflect a linguistic change between 1603 and 1604, but rather the influence of
different copyists or editors. When we have wholly different works by different authors, the variation of course is even greater: among works published in the year 1604, for example, it’s easy to find authors who use 3rd-singular *eth* almost exclusively, and others who almost exclusively use *es* instead.

What can we do about these sampling problems? While making larger historical treebanks on the current plan is a good idea, it’s not practical to achieve the required scale by this route. What I suggest instead is something like the following:

* Create a large, open, digital collection -- at least 200 books per year from (say) 1600 through 1850, and as much as is available from 1475 to 1600;
* Ensure that the texts and metadata are accurate;
* Provide tools that make it easy to create appropriately balanced samples for adequate testing of specific hypotheses.

Because a collection on this scale will comprise several billion words from 50,000-100,000 sources -- at least 2,000 sources per decade -- most sampling problems should be avoided. But is such a collection practical, and can scholars really use this type of unanalyzed collection for serious linguistic research? In this presentation, I’ll discuss a plausible series of steps for creating a resource of this kind, and propose a method for efficiently using it for research in historical morphology, syntax, and semantics. While this approach depends on tools that do not now exist, the required algorithms are well established.