When Big(gish) Data Goes Online: Cross-disciplinary opportunities and challenges

Jukka Tyrkkö
Linnéuniversitet

The last three decades have witnessed a remarkable increase in the number and volume of linguistic corpora available to the research community. Structured corpora comprising hundreds of millions or even billions of words of data are no longer unusual, and unstructured data sets such as Google Books, which are increasingly used in a very corpus-like manner, can encompass over a hundred billion words. Many of these large datasets are also made available online, and server-side query tools such as CQPweb, SketchEngine and the Brigham Young front-end to MySQL make it easy for anyone to use very large corpora both quickly and efficiently. While these corpora may fall short of criteria used to define 'big data' in some disciplines, the volume of text available is typically far beyond anything a single researcher or a research team could ever hope to process either manually or with the help of rudimentary search tools. However, while online corpora do open up new worlds of discovery, they also typically impose considerable limits to the types of queries available, provide quantitative data in difficult to process and sometimes misleading manner, and generally do not allow the researcher direct access to the underlying full datasets, more often than not for reasons of copyright and publishing agreements.

Although many of these large text collections and corpora were primarily designed with the linguist in mind, scholars from a wide variety of fields within the humanities and social sciences are also increasingly turning to these data sets for both qualitative and quantitative evidence, such as finding illustrative quotes or indications of diachronic trends that support theoretical arguments. Instead of extrapolating arguments from small and necessarily anecdotal evidence, humanities scholars are increasingly open to the idea of studying cultural, societal and political questions using 'big data' and methodologies such as *culturomics* (Michel et al 2011; Nunberg 2010) and *distant reading* (Moretti 2005). As the conceptual and methodological worlds of qualitative and quantitative research collide, the new challenge is how to operationalize joint research endeavors in the most beneficial fashion (see, e.g., McEnery and Baker 2016).

In this paper, I will discuss some of the opportunities and challenges that these large data sets and online interfaces can bring about, drawing examples from a collaborative project involving a team of social scientists and a corpus linguist. Using the *British Hansard Corpus*, a computer-readable, richly annotated edition of British Parliamentary debates (1803-2005), our objective has been to challenge certain claims made in political science about country references in historical political discourse, namely, that references to foreign nation states as examples to be followed only emerged as a major discursive strategy of policy-making around the time of the Second World War (Meyer et al 1997). The 1.6-billion-word dataset, which includes 7.6 million speeches delivered by over 40,000 MPs, is a new kind of historical corpus: not a sample drawn from an amorphous population, but an exhaustive and arguable complete record of a specific well-defined register of language use. Fully annotated both for standard linguistic variables and semantically tagged using data from and the conceptual network developed for the *Historical Thesaurus of the Oxford English Dictionary* and the Samuels semantic tagger (Alexander et al in press), the Hansard corpus has proven extremely useful and informative, but the data has also coughed up various surprises and potential problems, particularly if one were to rely solely on the online interface. In the present paper, the pros and cons of the online version and the standalone corpus are discussed and evaluated with particular reference to their usefulness in cross-disciplinary (digital) humanities projects, where efficient data management and ease of accessibility have to be balanced with the inherent complexity of textual accounts of ideas and concepts.

References

Alexander, Marc and Mark Davies. 2015-. Hansard Corpus 1803-2005. Available online at http://www.hansard-corpus.org.

Alexander, Marc, Fraser Dallachy, Scott Piao, Alistair Baron, Paul Rayson. In Press. Metaphor, Popular Science and Semantic Tagging: Distant reading with the Historical Thesaurus of English'. In *Digital Scholarship in the Humanities (DSH).*

Alasuutari, Pertti, Marjaana Rautalin and Jukka Tyrkkö. Accepted. The formation of interdependent decision-making: The case of British Parliament, 1803-2005. Presentation to be delivered at The Australian Sociological Association conference. Melbourne. 28.11-1.12.2016.

McEnery, Anthony and Helen Baker. 2016. *Corpus Linguistics and 17th-century Prostitution: Computational Linguistics and History*. (Corpus and Discourse). London: Bloomsmury Academic.

Meyer John W, John Boli, George M. Thomas and Francisco O Ramirez. 1997. World Society and the Nation-State. In *American Journal of Sociology*. 103(1): 144–181.

Michel, Jean-Baptiste, Erez Lieberman Aiden et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. In *Science* 331, 176-182 (published online ahead of print: 12/16/2010). <http://www.sciencemag.org/content/331/6014/176.full.pdf>.

Moretti, Franco. 2005. *Graphs, maps, trees: Abstract models for a literary history*. London: Verso.

Nunberg, Geoff. 2010. Humanties research with the Google Books corpus. http://languagelog.ldc.upenn.edu/nll/?p=2847.