From correspondence to corpora:
A seminar on digital processing of historical letter compilations
15 November 2013, University of Helsinki

Abstracts of letter project presentations

Mikko Hakala & Minna Palander-Collin (University of Helsinki)

The normalized version of the CEEC: Dealing with spelling variation in a corpus of historical letters

Our presentation deals with the problem of spelling variation in corpus research. We will share our experiences of normalizing the spelling in the Corpus of Early English Correspondence with VARD 2, a tool for dealing with spelling variation (semi-) automatically, and demonstrate the impact of normalization on the results of methods such as keyword and cluster analyses.

Samuli Kaislaniemi (University of Helsinki)

"Sorry! Your corpus is not representative ... but does it matter?"

The Corpus of Early English Correspondence (CEEC) is a socially stratified corpus of English personal letters 1400–1800, and we the compilers take pride in its representativeness. But is it really possible that the CEEC contains no less than 10% of all surviving letters written by women in Tudor England? If true, this suggests that 'representativeness' is a skewed metric.

This talk is an attempt to address the complex and problematic relationship between the archival record (the manuscripts and documents that survive), the "edited truth" (what has been published), 'representative' corpora, and the original manuscript reality of Early Modern English women's letters. My aim is to raise questions about the very fundamentals of the digital resources we create.

Kirsi Keravuori (Finnish Literature Society)

The Correspondence of Elias Lönnrot: A digital scholarly edition

The Finnish Literature Society is preparing a digital scholarly edition of Elias Lönnrot's letters (1828–1884), a correspondence contributing both to the study of Finnish cultural history and of the development of Finnish as a written language. The digital edition comprises of 2 500 letters or drafts written by Lönnrot and 3 500 letters he received. The correspondence has now been digitized and for the most part transcribed. The first phase of the on-line publication, about 1800 private letters written by Lönnrot, is estimated to begin at

the end of 2014. The edition will include facsimiles and transcriptions of the letters and some commentaries.

The bilingual (Finnish/Swedish) correspondence between Elias Lönnrot and his extensive network of friends and associates is a rich source material for the work and mentality of the Finnish scholarly community, their language practices and their endeavors to develop the Finnish language and for the study of 19th century epistolary culture.

The Biographical Centre of the Finnish Literature Society is in charge of the publication project. A demo version of the Correspondence of Elias Lönnrot is available on-line at http://elias.finlit.fi/lonnrot/.

Mikko Laitinen (Linnaeus University):

LALP: Letters of artisans and labouring poor in late modern England

This presentation discusses the collection process of an electronic corpus of lower-order petition letters, *Letters of Artisans and the Labouring Poor, c. 1750–1835* (LALP). I will first briefly describe the content and the set-up of the corpus and will then shift to new research questions that lower-order materials like LALP raise. I will specifically focus on developing literacies among (often inexperienced) writers and look into a continuum of skills that are related to petitions as a genre.

Anni Sairio (University of Helsinki)

From manuscripts to corpus coding: A learning curve

I will talk about the Bluestocking Corpus, which I have compiled of eighteenth-century manuscript letters, and the challenges in carrying out this 250,000-word solo project. What kinds of goals should a corpus compiler set for a small-scale, work-intensive project? What makes sense, and what is enough? What have been the biggest obstacles and challenges?

Tanja Säily (University of Helsinki)

Progress in POS tagging the *Corpus of Early English Correspondence Extension* (CEECE)

Parsed Corpus of Early English Correspondence (PCEEC) has yielded interesting results on variation in part-of-speech frequencies from Late Middle English to Early Modern English (Säily et al. 2011), its 18th-century extension has not yet been POS tagged. In this presentation, I will discuss our decision to use CLAWS (Garside & Smith 1997) to tag the CEECE, the requirements this sets to the corpus and how we are dealing with them. For instance, while we are using the standardised-spelling version of the CEECE (Palander-Collin

& Hakala 2011), there are still a number of issues that the tagger cannot handle and that need to be corrected manually. Furthermore, the COCOA-format parameter coding and the text-level coding of the CEECE present their own problems, which we are working on in collaboration with one of the developers of CLAWS, Paul Rayson.

References

CEECE = *Corpus of Early English Correspondence Extension*. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Modern Languages, University of Helsinki. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/

Garside, Roger and Nicholas Smith. 1997. "A hybrid grammatical tagger: CLAWS4." In Garside, Roger, Geoffrey Leech and Tony McEnery (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, 102–121. http://ucrel.lancs.ac.uk/claws/

Palander-Collin, Minna and Mikko Hakala. 2011. "Standardized versions of the *Corpora of Early English Correspondence*." Corpus Resource Database (CoRD). http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/standardized.html

PCEEC = *Parsed Corpus of Early English Correspondence*, tagged version. 2006. Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

Säily, Tanja, Terttu Nevalainen and Harri Siirtola. 2011. "Variation in noun and pronoun frequencies in a sociohistorical corpus of English." *Literary and Linguistic Computing* 26(2): 167–188.