Prerequisites to a corpus-based analysis of EEBO-TCP

Alistair Baron and Andrew Hardie

With Phase 1 containing over 25,000 transcribed texts, approximately 700 million words, and Phase 2 already containing nearly 15,000 texts, over 200 million words (and growing), EEBO-TCP offers an unrivalled resource to corpus linguistic studies of Early Modern English in terms of scale and coverage. However, several preparatory steps are necessary to enable meaningful corpus analyses of the EEBO data. This paper describes these processes and some of the concomitant difficulties, whilst also outlining the analysis methods offered by the corpus analysis tool CQPweb.

Historical corpus linguistics is a well-established research area, and a wide range of corpora have been developed. These corpora have tended to be carefully designed and finely structured, and to focus on a particular genre or text-type; in consequence most are relatively small in size, e.g. the Corpus of English Dialogues at 1.2 million words (Culpeper & Kytö, 2010). By contrast, the scale of the EEBO-TCP data is comparable to the very largest modern corpora now available, e.g. web-crawled corpora of billions of words (Baroni et al., 2009).

Powerful software tools are required to process corpora on this scale. One such tool is CQPweb, a user-friendly web-based interface to the IMS Open Corpus Workbench. CQPweb's analysis functions include concordancing (displays of word occurrences with their immediate context), collocations (statistical analysis of word co-occurrence) and keyword analysis (words significantly more frequent in one set of texts compared to another). To get the most out of these analysis techniques, corpus metadata is needed to allow users to narrow analyses to a particular set of texts or to compare sub-corpora defined by metadata-based filtering. We have been able to extract various metadata fields from the EEBO-TCP headers of each text, most notably the date and place of publication, which allows diachronic analysis. We have also implemented various types of corpus annotation of the EEBO-TCP data, notably part-of-speech tags (grammatical labels assigned to each word) and semantic tags (topic or concept labels assigned to each word). This allows CQPweb analyses to be performed at the annotation level rather than the word level; for example, a researcher could look at which topics are more prevalent in different time periods.

One particular issue with the computational analysis of historical texts is the large amount of spelling variation generally present. We have previously shown that like all Early Modern English corpora, EEBO-TCP contains a large amount of spelling variation (Baron et al., 2009). It has also been shown that this spelling variation has a detrimental effect on the accuracy of various corpus linguistic techniques, e.g. part-of-speech annotation (Rayson et al., 2007) and keyword analysis (Baron et al., 2009). Here we show how the Variant Detector (VARD) tool (Baron & Rayson, 2009) can be used on EEBO-TCP to automatically insert modern equivalents alongside the original word-forms.

The preparatory steps of metadata extraction, spelling modernization and corpus annotation allow significantly more powerful and accurate computational analysis to be performed on EEBO-TCP.

References

Baron, A. and Rayson, P. (2009). Automatic standardization of texts containing spelling variation, how much training data do you need? In M. Mahlberg, V. González-Díaz and C. Smith (eds.) *Proceedings of the Corpus Linguistics Conference, CL2009*, University of Liverpool, UK, 20-23 July 2009.

Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. *International Journal of English Studies* 20 (1): 41-67.

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43(3): 209-226.

Culpeper, J. & Kytö, M. (2010). *Early Modern English dialogues: Spoken Interactions as Writing*. Cambridge University Press, Cambridge.

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In Davies, M., Rayson, P., Hunston, S. and Danielsson, P. (eds.) *Proceedings of the Corpus Linguistics Conference: CL2007*, University of Birmingham, UK, 27th-30th July 2007.