The extent of spelling variation in Early Modern English

Many researchers have commented on the large amount of spelling variation found in texts from the Early Modern English (EModE) period; however, few have explicitly quantified these levels or investigated any difference in variation between text types, genres, authors or dates. In order to investigate the impact of spelling variation in historical corpora when using automated corpus linguistic tools, this paper details a quantitative survey of the spelling variation contained in various EModE corpora.

The development of the VARD tool (Rayson et al, 2008; Baron and Rayson, 2008), which was designed to automatically standardize spelling variation in historical corpora, has facilitated the exploration of spelling variation more subtly and systematically, producing the analysis presented in this paper. The analysis is also useful for the future development of VARD; increased knowledge about the characteristics of spelling variation will allow for improvements to be made in VARD's capability of processing variants.

Previous studies have shown that spelling variation produces a negative effect on the accuracy of corpus linguistic techniques such as part-of-speech annotation (Rayson et al, 2007), semantic analysis (Archer et al, 2003) and key word analysis (Baron et al, forthcoming). The latter study also showed that higher levels of spelling variation are directly associated with an increased effect on accuracy. Through the quantitative analysis presented in this paper researchers will be able to assess the extent of spelling variation in a particular historical corpus and subsequently the likely effect on automated corpus linguistic tools.

A quantitative survey of spelling variation in different types of text is interesting from a linguistic point of view also. For example, the effect of genre on spelling variation has previously been explored by Archer and Rayson (2004). The VARD tool allows for analysis of the effect of genre, text type, author and date on a much larger scale.

References

Archer, D. and Rayson, P. (2004). Using an historical semantic tagger as a diagnostic tool for variation in spelling. *Thirteenth International Conference on English Historical Linguistics (ICEHL 13).* Vienna, Austria: University of Vienna. (23-29 Aug. 2004).

Baron, A. and Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, 22 May 2008.

Baron, A., Rayson, P. and Archer, D. (forthcoming). Word frequency and key word statistics in historical corpus linguistics. *International Journal of English Studies*.

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.

Rayson, P., Archer, D., Baron, A. and Smith, N. (2008). Travelling Through Time with Corpus Annotation Software. In Lewandowska-Tomaszczyk, B. (ed) *Corpus Linguistics, Computer Tools, and Applications – State of the Art. Palc 2007*. Peter Lang, Frankfurt am Main.