# Quantifying Early Modern English spelling variation: change over time and genre

Alistair Baron and Paul Rayson
Lancaster University

Dawn Archer
University of Central Lancashire

It is widely accepted that texts representative of a variety of genres display a marked degree of spelling variation throughout the Early Modern English (EModE) period, in spite of the English language's gradual standardization between 1500 and 1700 (Vallins & Scragg, 1965; Görlach, 1991; Nevalainen, 2006). Until recently, our knowledge of this gradual standardization of variant spelling forms was largely founded on qualitative studies. In Baron et al (2009a), however, we've been able to report a large-scale study of spelling variation, and its decline, in the EModE period: several corpora from the period were analysed and a steady decrease in the ratio of spelling variants to modern spellings observed, until around 1700 when the orthography had largely become standardised. Our recent research has also shown that this spelling variation has a negative impact on the accuracy of well-defined corpus linguistic methods such as key word analysis (Baron et al, 2009a), part-of-speech tagging (Rayson et al, 2007) and semantic annotation (Archer et al, 2003).

The large-scale, diachronic documentation of spelling variation is now possible because of the VARD and DICER tools. VARD (Baron and Rayson, 2009) enables texts containing spelling variants to be standardised, and, in the process, supplements them with modern equivalents. As well as increasing the accuracy of corpus linguistic tools, this allows for the analysis of spelling variation trends. DICER (Discovery and Investigation of Character Edit Rules) (Baron et al, 2009b), aids this analysis through the examination of these standardization decisions; letter replacement rules which can transform the variant form to its modern equivalent are extracted and a detailed database of these rules and their frequencies is built.

In a previous study (Archer and Rayson, 2004), using earlier versions of these tools, we've been able to confirm obvious patterns such as the interchangeability of <u>/<v> depending on their initial/medial positioning, etc. We've also identified patterns we were not "cued" to find, some of which seem to be specific to particular genres. In addition, our preliminary investigations of 3823 variant forms suggest that genres typically associated with standardization, because of their inherent "prestige" and "power" (i.e. Religious, Political and Law texts) actually displayed more variation than, for example, (some) seventeenth century newsbooks.

Here, we extend this preliminary work by presenting an analysis of various spelling features in the EModE period. We focus on how these characteristics changed over time, and also examine the influence of genre and text type. Some previously observed trends will be quantified for the first time, such as the use in print of *-`d* instead of *-ed* and *-ick* instead of *-ic* well into the 17th century (Nevalainen, 2006: 36), but less obvious and possibly unknown trends uncovered by our analysis will also be investigated. We go on to claim that this type of analysis will not only aid in the development of more precise standardization techniques, but also allow researchers to better understand the orthography found in EModE period texts and perhaps aid insights into why certain spellings existed.

## References

Archer, D., McEnery, T. Rayson, P. and Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.) *Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University, pp. 22-31.

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

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. (2009a). Word frequency and key word statistics in historical corpus linguistics. In *Anglistik: International Journal of English Studies*, 20 (1), pp. 41-67.

Baron, A., P. Rayson, and D. Archer. (2009b). Automatic standardization of spelling for historical text mining. In *Proceedings of Digital Humanities 2009*, Maryland, USA. University of Maryland. pp. 309–312.

Görlach, M. (1991). *Introduction to Early Modern English*. Cambridge University Press, Cambridge.

Nevalainen, T. (2006). *An Introduction to Early Modern English*. Edinburgh Textbooks on the English Language, Edinburgh University Press, Edinburgh.

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 *Proceedings of Corpus Linguistics 2007*, July 27-30, University of Birmingham, UK.

Vallins, G.H. & Scragg, D.G. (1965). *Spelling*. André Deutsch, London.