Abs-54

Alistair Baron (Lancaster University), Paul Rayson (Lancaster University), and Dawn Archer (University of Central Lancashire)

Dealing with spelling variation in historical corpora: Using VARD to standardise spelling variants from the EmodE period

In papers presented at the previous corpus linguistic conferences (Archer et al. 2003; Rayson et al, 2005, 2007; Baron and Rayson, 2009), and elsewhere (Baron et al, 2009) a strong case has been made for the need to provide standardised versions of EmodE corpora alongside the original published editions. Without standardisation of spelling, commonly-applied corpus linguistics methods such as frequency and key words analysis and part-of-speech and semantic tagging are much less accurate. For example, it has been estimated that around 60% of word types and 35% of word tokens in a variety of corpus samples dating from 1500-1600 are historical variants. Although standardisation could be viewed as a spell checking or translation problem, it has been shown that existing tools such as Microsoft Word are unable to cope with the variety of variants in historical corpora. Our solution is the VARD (Variant Detector) software which allows corpus compilers and users to standardise spelling in corpora before the text is used for corpus analysis.

This session will be a two-hour hands-on workshop using the VARD software with a mixture of presentations and significant audience participation. We will begin with two short presentations each of 20 minutes. The first presentation will be by Anu Lehto from the University of Helsinki team who have recently published the corpus of Early Modern English Medical Texts (EMEMT, see Taavitsainen and Pahta, 2010); she will describe the team's experience of training and applying the VARD tool to develop a standardised version of the EMEMT corpus. The second presentation will be given by the workshop organisers and present an overview of the methods used in the VARD software and the accuracy of the tool.

The remainder of the workshop time (1 hour 20 minutes) will be devoted to hands-on exercises with the workshop participants using the software directly to manually standardise corpus samples. We will provide samples from Early English Books Online but workshop participants will be invited to bring their own corpora to test. As VARD can also be used to deal with other forms of spelling variation, such as in SMS (Tagg et al, 2010), participants would be welcome to bring texts from other sources. The hands-on component will include four main parts. First, familiarisation with the VARD user interface. Second, step-by-step training on the standardisation procedure itself. Third, the participants will independently standardise a selection of corpus samples. Finally, guidance will be given on how to tailor the linguistic resources and rules within VARD to improve the accuracy of the standardisation procedure.

By the end of the workshop, participants will understand how to use the VARD software to standardise spelling variants in EmodE corpora, how to export both original and standardised versions for use in other corpus linguistic software and how much training is required for their own corpora. Participants will be provided with copies of our previous studies on standardising historical corpora, a copy of the VARD software for academic use and a user manual. Since no computer labs are available at the conference venue, participants should bring their own laptops.

References

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

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. *Anglistik: International Journal of English Studies*, 20 (1), pp. 41-67.
Rayson, P., Archer, D. and Smith, N. (2005). VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. In proceedings of the Corpus Linguistics 2005 conference, July 14-17, Birmingham, UK. *Proceedings from the Corpus Linguistics Conference Series on-line e-journal* 1 (1). ISSN 1747-9398.
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.
Taavitsainen, I. and Pahta, P. (eds.) (2010). *Early Modern English Medical Texts: Corpus description and studies*, Benjamins, Amsterdam.
Tagg, C., Baron, A. and Rayson, P. (2010). "I didn't spel that wrong did i. Oops": Analysis and standardisation of SMS spelling variation. In *ICAME 31 Abstracts*, 108–109, Gießen, Germany.