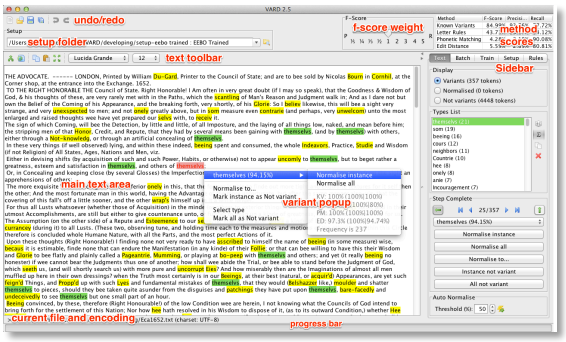
VARDing to modernise spellings in historical texts for improved corpus analysis

Paul Rayson, Alistair Baron and Andrew Moore
UCREL research centre, School of Computing and Communications,
Lancaster University, UK

Early attempts to apply corpus linguistics (CL) and natural language processing (NLP) methods to Early Modern English (EmodE) corpora did not adjust their methods and tools and, for example, have employed existing taxonomies developed for modern corpora such as the USAS tagset *(Rayson et al, 2004)*. However, this fails to account for significant meaning and vocabulary shifts over time, as well as wide-scale historical spelling variation in the corpora which cause problems for existing CL and NLP methods and tools.

To help address the first problem, we require a broad coverage taxonomy combined with historically sensitive meaning categories. The Historical Thesaurus of English (HT),[1] developed at the University of Glasgow over forty years, provides a high-quality semantic lexical database containing 793,742 entries manually classified into 225,131 thesaurus categories arranged in a hierarchical structure. A key challenge is to scale the semantic disambiguation in USAS from a smaller semantic field taxonomy of 232 tags designed for modern English, to that of the HT. A smaller set of four thousand thematic codes devised at Glasgow and arranged at an intermediate level in the hierarchy can also be applied in order to produce semantically tagged output from the Historical Thesaurus Semantic Tagger (HTST) which uses the full set of categories, thematic codes and USAS tags.



The second significant challenge in the application of corpus and computational linguistics methods to EmodE corpora is historical spelling variation which has been shown to significantly affect their accuracy and robustness *(Archer et al, 2003; Rayson et al, 2007; Baron et al, 2009)*. Following the development of the Variant Detector (VARD) software *(Baron and Rayson, 2008)*, this problem can be addressed by inserting modern equivalents which can then be tagged, counted and searched for with appropriate software alongside the original historical variants. We are undertaking a large crowdsourcing exercise which will permit the large-scale manual training of time-sensitive models for matching historical spelling variants. These models can then be applied automatically to our corpora to achieve more accurate results. Moreover, we can make use of variant spelling and dating information in the OED to improve the accuracy and coverage of VARD.

[1] http://historicalthesaurus.arts.gla.ac.uk

p.rayson@lancaster.ac.uk @perayson @al586 @apmoore94

Very little EmodE data is available in corpus form which has been manually VARDed so far. A handful of existing corpora have been made available in original and standardised (or normalised) forms: Innsbruck Letter Corpus, Early Modern English Medical Texts (EMEMT), Corpus of Early English Correspondence (CEEC) and Corpus of English Dialogues (CED). In our VARDsourcing research so far, we have sampled the Early English Books Online (EEBO) Text Creation Partnership (TCP) phase 1 corpus in 25-year periods, creating 10 files of 2,000 words each, all of which have been manually VARDed by three EmodE experts. A committee consisting of Merja Kytö, Jonathan Culpeper, Dawn Archer, Alistair Baron and Paul Rayson, then reviewed these decisions to reach a consensus and to create a gold-standard VARDed corpus. This gold standard will be used to train and evaluate new VARDers who will then VARD another 160,000 words of EEBO-TCP phase 1. All the VARDed data will be made publicly available. You can view the VARDing guidelines and participate in the VARDsourcing challenge by visiting our website (http://ucrel-vardsourcing.lancs.ac.uk/) in order to create an account.

Acknowledgements This research was part of the Semantic Annotation and Mark-Up for Enhancing Lexical Searches (SAMUELS) project (http://www.gla.ac.uk/samuels/) funded by the Arts and Humanities Research Council in conjunction with the Economic and Social Research Council (grant reference AH/L010062/1), January 2014 to March 2015. Lead institution: University of Glasgow. Other partners: University of Huddersfield, University of Central Lancashire, University of Strathclyde, Oxford University Press. International partners: Brigham Young University (Utah), Åbo Akademi University (Finland), and the University of Oulu (Finland). The VARD crowdsourcing experiment is funded by JISC in the UK.

References

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

Archer, D., Kytö, M., Baron, A., Rayson, P. (2015). Guidelines for normalising Early Modern English corpora: decisions and justifications. *ICAME Journal*, Volume 39, May 2015. DOI: 10.2478/icame-2015-0001

Baron, A. and Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. In *proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, 22nd May 2008.

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

Piao, S., Dallachy, F., Baron, A., Demmen, J., Wattam, S., Durkin, P., McCracken, J., Rayson, P. & Alexander, M. (2017) A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech and Language*, vol 46, pp. 113-135. DOI: 10.1016/j.csl.2017.04.010

Rayson, P., Archer, D., Piao, S., & McEnery, A. M. (2004). The UCREL semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop*, Lisbon, Portugal, 2004. (pp. 7-12). Lisbon.

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.