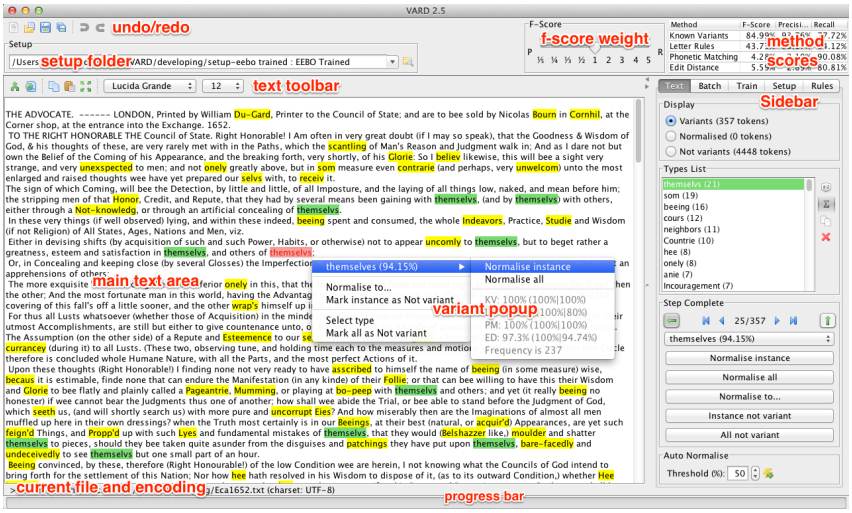
VARD 2.5

VARD is an interactive piece of software designed to assist corpus users in dealing with the problems caused by spelling variation. Spelling variation has been shown to have a negative impact on the accuracy of semantic annotation (Archer *et al.*, 2003), part-of-speech tagging (Rayson *et al*., 2007) and key word analysis (Baron *et al*., 2009). Whilst primarily aimed at normalising Early Modern English spelling variation, VARD has been successfully adapted to deal with other varieties of English spelling variation, and spelling variation found in other languages.

The software assists with manual normalisation by suggesting candidate normalisations for detected spelling variants. As decisions are made by the user, VARD learns how to best normalise the spelling variation in your corpus to the point where it can successfully automatically normalise the entire corpus after training. The output from VARD retains the original spelling form within an XML tag, e.g. <normalised orig="satisfy'd">satisfied</normalised>.



Evaluation with the Innsbruck Letters Corpus (1386–1688) showed that after training, 62% of spelling variant tokens could be automatically normalised successfully, with a precision rate of 95%. Furthermore, VARD has been used to automatically normalise the Early Modern English Medical Texts (EMEMT) Corpus (Lehto *et al*., 2010), with 73% of variant tokens successfully normalised, and precision estimated to be above 98%.

The latest version includes a new cleaner user interface, options for dealing with foreign text tags, greater control of VARD's setup and plenty more besides.

VARD is free to use for academic research, simply request a download on the website: **http://ucrel.lancs.ac.uk/vard**. The website includes a user guide, FAQ, mailing list details, version history and a list of related publications, as well as bug reporting and feedback forms.

DICER

DICER (Discovery and Investigations of Character Edit Rules) is a supplementary tool to VARD which can be used to analyse in greater detail the spelling variation in a corpus. Given a variant string and its VARD normalisation, DICER locates the differences between the two strings and counts how many differences there are, where the differences occur and precisely which characters are changed. Given a number of these pairs, quantities are summed and stored in a database for analysis.

DICER was first developed to find letter replacement rules, with the aim of improving the performance of VARD; the spelling rules produced by DICER can be added to VARD. However, DICER was also found to be valuable in the investigation of spelling trends over time, genres, text-types, authors or any other meta-data available for study.

DICER is a web-based tool freely available for academic research.
Website: http://corpora.lancs.ac.uk/dicer

There are various publicly viewable analyses available to view on the website. If you would like to analyse the spelling variation in your own corpus, please get in touch.

Contact: a.baron@lancaster.ac.uk / @al586

Key References

Archer, D., McEnery, T., Rayson, P. & Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In D. Archer, P. Rayson, A. Wilson & T. Mcenery, eds., *Proceedings of Corpus Linguistics 2003*, 22–31, Lancaster University, Lancaster, UK.

Archer, D., Kytö, M., Baron, A. & Rayson, P. (2015). "Guidelines for normalising Early Modern English corpora: decisions and justifications." ICAME Journal: 39.

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.

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

Lehto, A., Baron, A., Ratia, M. and Rayson, P. (2010). Improving the precision of corpus methods: The standardized version of Early Modern English Medical Texts. In Taavitsainen, I. and Pahta, P. (eds.) *Early Modern English Medical Texts: Corpus description and studies*, pp. 279–290. John Benjamins, Amsterdam.

Rayson, P. and Baron, A. (2011). Automatic error tagging of spelling mistakes in learner corpora. In Meunier F., De Cock S., Gilquin G. and Paquot M. (eds.) A Taste for Corpora. In honour of Sylviane Granger, Studies in Corpus Linguistics, 45. John Benjamins, Amsterdam.

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, 27-30 July 2007.

Reynaert, M., Hendrickx, I., & Marquilhas, R. (2012). Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2. on Annotation of Corpora for Research in the Humanities (ACRH-2), pp. 87-98.

Tagg, C., Baron, A. and Rayson, P. (2012). "i didn't spel that wrong did i. Oops": Analysis and normalisation of SMS spelling variation. In Cougnon, L. A. and Fairon, C. (eds.) SMS Communication: A linguistic approach. Special issue of Linguisticae Investigationes 35:2, pp. 367-388.