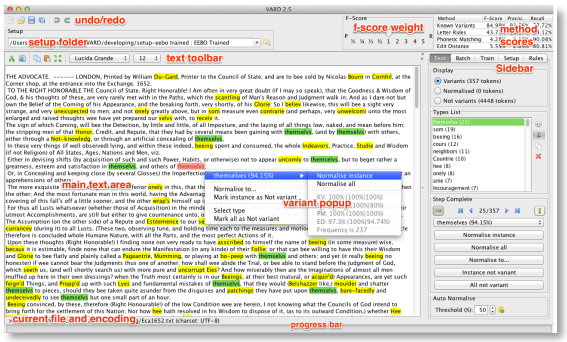
Large-scale Time-sensitive Semantic Analysis of Historical Corpora

Paul Rayson, Alistair Baron, Scott Piao and Steve Wattam
UCREL research centre, School of Computing and Communications,
Lancaster University, UK

Previous attempts to apply automatic semantic analysis to Early Modern English (EmodE) corpora 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. What is required is 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. In this software demonstration, we will show the new Historical Thesaurus Semantic Tagger (HTST) which uses the full set of categories, thematic codes and USAS tags. The user can also enter the date of a text and the software employs dating information in the thesaurus to help it choose more appropriate categories. In addition, given the links from the Historical Thesaurus to the entries in the Oxford English Dictionary (OED), we are able to draw on further information from the senses, definitions and example sentences in the OED in order to assist in the ranking of contextually appropriate thesaurus and thematic codes.



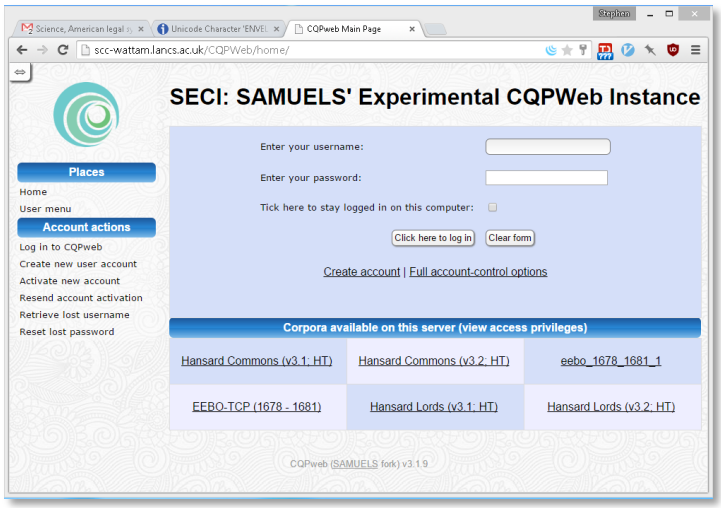
A second significant challenge in the application of corpus and computational linguistics methods to EmodE corpora is historical spelling variation which has been shown to significantly affect their accuracy and robustness *(Archer et al, 2003; Rayson et al, 2007; Baron et al, 2009)*. Following the development 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 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 @scottpiao @StephenWattam

The final challenge in this enterprise is the large-scale and heterogeneous nature of the corpora and metadata. In particular, we have indexed the transcribed portion of the Early English Books Online[2] (EEBO-TCP), over one billion words. The latest release from the Text Creation Partnership (TCP) in February 2015 contains 53,830 transcribed books. The second corpus tagged comprises 200 years of UK Parliamentary Hansard consisting of over 7 million files (≈1 billion words). Running such collections through a pipeline consisting of historical spelling variant normalisation (VARD), part-of-speech tagging (CLAWS), semantic tagging (USAS and HTST) followed by indexing (in tools such as CQPweb and Wmatrix) requires significant computational resources. The software demonstration will show how we have been able to overcome all three of these challenges, demonstrate how accurate such processes are and signpost directions for future work. The HTST is available to use via a web page form at: http://phlox.lancs.ac.uk/ucrel/semtagger/english and our search software is linked at: http://ucrel.lancs.ac.uk/samuels/

TOKEN	LEMMA	POSTAG	SEMTAG1	MWE	SEMTAG2	SEMTAG3
S_BEGIN	NULL	NULL	Z99	0	04.10 [Unrecognised];	04.10 [Unrecognised];
The	the	AT	Z5	0	04.03 [null];	04.03 [Grammatical Word];
cat	cat	NN1	L2 M3	0	03.10.12.02.12.01-08 [0.94736842] [.of cat]; 01.02.04.13.09.02.12-01 [1.00000000] [.types of]; 01.02.06.16.07.04-09 [1.00000000] [.member of family Pimelodidae/common cat-fish];	Y12a07a [Skin with hair attached/fur]; B20t [Particular food plant/product]; B22j [Fish];
sat	sit	VVD	M8 C1 P1 G1.1 G2.1 M6 A9+	0	[MWE] 01.02.08.01.22.08-13 [1.00000000] [Cook .burn/catch on bottom of cooking pot]; 01.05.08.09-06 [1.00000000] [Not moving .remain as opposed to go];	B24t07 [Cooking]; E08i [Absence/privation/cessation of movement];
on	on	II	Z5	0	[MWE] 01.02.08.01.22.08-13 [1.00000000] [Cook .burn/catch on bottom of cooking pot]; 01.05.08.09-06 [1.00000000] [Not moving .remain as opposed to go];	B24t07 [Cooking]; E08i [Absence/privation/cessation of movement];
the	the	AT	Z5	0	04.03 [null];	04.03 [Grammatical Word];
mat	mat	NN1	H5 O2	0	03.02.07.03.09.14-03 [0.93750000] [.mat]; 03.11.04.13.16.15-14 [0.93750000] [.mat]; 03.02.07.03.09.10.01-02 [0.94444444] [.table mat];	Q06f05m [Floor-covering]; Z08v11 [Bowls/bowling]; Q06f05i [Household linen];
.	PUNC	YSTP	PUNC	0	NULL	NULL [];
S_END	NULL	NULL	Z99	0	04.10 [Unrecognised];	04.10 [Unrecognised];



Acknowledgements This work took place in 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.

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.

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.

[2] http://www.textcreationpartnership.org/tcp-eebo/

p.rayson@lancaster.ac.uk @perayson @al586 @scottpiao @StephenWattam