Tagging historical corpora – the problem of spelling variation

Paul Rayson, Lancaster University
Dawn Archer, University of Central Lancashire
Alistair Baron, Lancaster University
Nicholas Smith, Lancaster University

Spelling issues tend to create relatively minor (though still complex) problems for corpus linguistics, information retrieval and natural language processing tasks that use 'standard' or modern varieties of English. For example, in corpus annotation, we have to decide how to deal with tokenisation issues such as whether (i) periods represent sentence boundaries or acronyms and (ii) apostrophes represent quote marks or contractions (Grefenstette and Tapanainen, 1994; Grefenstette, 1999). The issue of spelling variation becomes more problematic when utilising corpus linguistic techniques on non-standard varieties of English, not least because variation can be due to differences in spelling habits, transcription or compositing practices, and morpho-syntactic customs, as well as "misspelling". Examples of non-standard varieties include:

- Scottish English[1] (Anderson et al., forthcoming), and dialects such as Tyneside English[2] (Allen et al., forthcoming)
- Early Modern English (Archer and Rayson, 2004; Culpeper and Kytö, 2005)
- Emerging varieties such as SMS or CMC in weblogs (Ooi et al., 2006)

In the Dagstuhl workshop we focussed on historical corpora. Vast quantities of searchable historical material are being created in electronic form through large digitisation initiatives already underway e.g. Open Content Alliance[3], Google Book Search[4], and Early English Books Online[5]. Annotation, typically at the part-of-speech (POS) level, is carried out on modern corpora for linguistic analysis, information retrieval and natural language processing tasks such as named entity extraction. Increasingly researchers wish to carry out similar tasks on historical data (Nissim et al, 2004). However, historical data is considered noisy for tasks such as this. The problems faced when applying corpus annotation tools trained on modern language data to historical texts are the motivation for the research described in this paper.

Previous research has adopted an approach of adding historical variants to the POS tagger lexicon, for example in TreeTagger annotation of GerManC (Durrell et al, 2006), or "back-dating" the lexicon in the Constraint Grammar Parser of English (ENGCG) when annotating the Helsinki corpus (Kytö and Voutilainen, 1995).

Our aim was to develop an historical semantic tagger in order to facilitate similar studies on historical data to those that we had previously been performing on modern data using the USAS semantic analysis system (Rayson et al, 2004). The USAS tool relies on POS tagging as a prerequisite to carrying out semantic disambiguation. Hence we were faced with the task of retraining or back-dating two tools, a POS tagger and a semantic tagger. Our proposed solution incorporates a corpus pre-processor for detecting historical spelling variants and inserting modern equivalents alongside them. This enables retrieval as well as annotation tasks and to some extent avoids the need to retrain each annotation tool that is applied to the corpus. The modern tools can then be applied to the modern spelling equivalents rather than the historical variants, and thereby achieve higher levels of accuracy.

The resulting variant detector tool (VARD) employs a number of techniques derived from spell-checking tools as we wished to evaluate their applicability to historical data. The current version of the tool uses known-variant lists, SoundEx, edit distance and letter replacement heuristics to match Early Modern English variants with modern forms. The techniques are combined using a scoring mechanism to enable preferred candidates to be selected using

---

[1] http://www.scottishcorpus.ac.uk/
[2] http://www.ncl.ac.uk/necte/
[3] http://www.opencontentalliance.org/
[4] http://books.google.com/
[5] http://eebo.chadwyck.com/home

likelihood values. The current known-variant lists and letter replacement rules are manually created. In a cross-language study with English and German texts we found that similar techniques could be used to derive letter replacement heuristics from corpus examples (Pilz et al, forthcoming). Our experiments show that VARD can successfully deal with:

- Apostrophes signalling missing letter(s) or sound(s): 'fore ("before"), hee'l ("he will"),
- Irregular apostrophe usage: again'st ("against"), whil'st ("whilst")
- Contracted forms: 'tis("it is"), thats ("that is"), youle ("you will"), t'anticipate ("to anticipate")
- Hyphenated forms: acquain-tance ("acquaintance")
- Variation due to different use of graphs: <v>, <u>, <i>, <y>: aboue ("above"), abyde ("abide")
- Doubling of vowels and consonants –e.g. <-oo-><-ll>: triviall ("trivial")

By direct comparison, variants that are not in the modern lexicon are easy to identify, however, our studies show that a significant portion of variants cannot be discovered this way. Inconsistencies in the use of the genitive, and 'then' appearing instead of 'than' or vice versa require contextual information to be used in their detection. We will outline our approach to resolving this problem, by the use of contextually-sensitive template rules that contain lexical, grammatical and semantic information.

References

Allen, W., Beal, J.C., Corrigan, K.P., Maguire, W. and Moisl, H. (to appear) 'Taming Unconventional Digital Voices: The Newcastle Electronic Corpus of Tyneside English', in Beal, J.C., Corrigan, K.P. and Moisl, H. (eds.) Using Unconventional Digital Language Corpora. Houndmills: Palgrave Macmillan.

Anderson, J., Beavan, D. and Kay, C. (forthcoming): 'The Scottish Corpus of Texts and Speech', Models and Methods in the Handling of Unconventional Digital Corpora, J. Beal, K. Corrigan, H. Moisl (eds.), Houndmills: Palgrave-MacMillan

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.

Culpeper, J. and Kytö, M. (2005). Exploring speech-related Early Modern English texts: lexical bundles re-visited. Presented at the 26th conference of ICAME, University of Michigan, USA, May 2005.

Durrell, M., Bennett, P., Ensslin, A. (2006). Towards a Methodology for Constructing and Annotating Historical Corpora: Tackling Structural and Lexical Variability in Early Modern German Newspaper Texts, *4th Days of Swiss Linguistics Conference*, Basel, Switzerland, November 2006.

Grefenstette, G. (1999). Tokenization. In van Halteren, H, (ed.) Syntactic wordclass tagging, Kluwer, The Netherlands, pp. 117 – 133.

Grefenstette, G. and Tapanainen, P. (1994) What is a Word, What is a Sentence? Problems of Tokenization. In Proceedings of 3rd conference on Computational Lexicography and Text Research (COMPLEX'94), Budapest, July 7-10, 1994, pp. 79 – 87.

Kytö, M. and Voutilainen, A. (1995). Applying the Constraint Grammar Parser of English to the Helsinki Corpus. *ICAME Journal* 19, pp. 23 – 48.

Nissim, M., Matheson, C. and Reid, J. (2004). Recognising Geographical Entities in Scottish Historical Documents. *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*.

Ooi, V. B. Y., Tan, P. K. W. and Chiang, A. K. L.: Analysing weblogs in a speech community using the WMatrix approach. 27th conference of the International Computer Archive of Modern and Medieval English (ICAME) University of Helsinki, Finland, 24-28 May, 2006.

Pilz, T., Ernst-Gerlach, A., Kempken, S., Rayson, P. and Archer, D. (forthcoming) The identification of spelling variants in English and German historical texts: manual or automatic. Literary and Linguistic Computing.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), 25th May 2004, Lisbon, Portugal, pp. 7-12.