Tagging Early Modern English Medical Texts (1500-1700)

Turo Hiltunen
University of Helsinki
turo.hiltunen@helsinki.fi

Jukka Tyrkkö
University of Tampere
jukka.tyrkko@uta.fi

1 Introduction

The availability of part-of-speech information in corpora improves the precision of corpus queries and facilitates more efficient analysis of syntactic features (Atwell 2008: 505). In addition, it enables the analysis of part-of-speech ratios, which have a variety of uses such as the verification of results from untagged corpora. In historical linguistics, for example, POS-ratios can be used to analyse whether an increase in the frequency of a word is due to an increase in the use of words in the same class (e.g. Mair et al. 2003).

With the automatic taggers available today, it has become relatively easy to tag Present-day English text with a reasonably good accuracy. Things are not as simple when it comes to historical texts, however. The challenges encountered are not only the result of various syntactic and lexical differences between the historical data and Present-day English, but to a considerable extent of orthographic variation. However, Rayson et al. (2007) have shown that it is possible to improve the accuracy of automatic POS-tagging by replacing deviant spellings with their modern variants.

This work-in-progress report reviews the findings of a study that looked at the success rate of automatic POS-tagging with CLAWS4 (Gardside and Smith 1997) when applied to early modern text orthographically standardized using VARD2 (Baron and Rayson 2009), following the procedures described in Lehto et al. (2010). Our data comes from *Early Modern English Medical Texts* (EMEMT), a specialised corpus covering the history of medical writing 1500–1700 (Taavitsainen et al. 2010). We also compare our result to those obtained by Rayson et al. (2007) for two other Early Modern English genres, namely tracts and comedies. The ultimate aim of this work is to produce a POS-tagged version of this corpus to be used in syntactic and phraseological research.

2 Corpus

The EMEMT corpus consists of 462 text samples of varying length, the total word count being ca. 2 million words. The corpus was originally devised to facilitate systematic inquiry into the changing styles of scientific thinking and the textual manifestations thereof, with a particular focus on pragmatic and discourse phenomena. For this reason, no grammatical annotation that would facilitate syntactic research is currently available for EMEMT.

The text extracts in EMEMT contain some mark-up providing information about issues such as the lay-out of texts, typesetters' errors, and figures and symbols omitted from transcriptions. A version with standardised orthography is released as part of the corpus.

3 Process

To analyse tagging accuracy, we first selected eight samples of roughly 1,000 words from the corpus and tagged them manually. The samples span the entire period in focus, as the date of the text is likely to influence the tagger's performance. Following the procedure described in Rayson et al. (2007), we then used CLAWS to tag both the non-standardised and standardised texts and compared the two individually to the manually tagged samples. Before tagging the standardised samples, we made a number of further substitutions to improve the quality of the input; these included expanding abbreviations originally annotated using tildes and equal signs (e.g. `<y~>` or `<y=t=>` for *that*), which had not been standardised in the published version of EMEMT.

4 Results and discussion

The analysis showed that the standardisation of spelling improves the tagger's performance considerably. The effect was particularly large for two texts from the 16th century (*Treasure of poor men* (1526, anonymous) and Gratarolo's *Health of magistates and students* (1574)), where CLAWS achieved roughly 80% accuracy (measured as the proportion of words correctly tagged by CLAWS), if the spelling had not been standardised. After the standardisation, the accuracy was over 90%. Compared to these texts, the samples from the 17th century already show less orthographic variation, but standardising their spelling with VARD2 lead to yet more accurate tagging.

These results provide additional support for the view that the spelling of texts from mid-17th century onwards is relatively similar to Present-day English, as has been suggested in earlier studies (Kytö and Voutilainen 1995: 29; Archer et al. 2003: 27). At the same time, they underline the usefulness of automatic spelling standardisation in creating a POS-tagged corpus of medical texts from the Early

Modern English period.

References

Archer, D., McEnery, A. M., Rayson, P. and Hardie, A. 2003. “Developing an automated semantic analysis system for Early Modern English”. In D. Archer, P. Rayson, A. Wilson and A. M. McEnery (eds.) *Proceedings of the corpus linguistics conference 2003*. Lancaster: University of Lancaster. 22–31.

Atwell, E. 2008. “Development of tag sets for part-of-speech tagging”. In A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics. An International Handbook. Volume 1*. Berlin: Mouton de Gruyter. 501–527.

Baron, A. and Rayson, P. 2009. “Automatic standardization of texts containing spelling variation: How much training data do you need?” In *Proceedings of the Corpus Linguistics Conference 2009*. Liverpool: University of Liverpool.

Garside, R. and Smith, N. 1997. “A hybrid grammatical tagger: CLAWS4”. In R. Garside, G. Leech and A. McEnery (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London. 102-121.

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 I. Taavitsainen and P. Pahta (eds.) *Early Modern English Medical Texts: Corpus description and studies.* Amsterdam: John Benjamins Publishing Company. 279-290.

Mair, C., Hundt, M., Leech, G. and Smith, N. 2003. “Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora”. *International Journal of Corpus Linguistics* 7(2), 245–264.

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 the Corpus Linguistics Conference 2007*. Birmingham: University of Birmingham.

Taavitsainen, I., Pahta, P., Hiltunen, T., Mäkinen, M., Marttila, V., Ratia, M., Suhr, C. and Tyrkkö, J. (compilers). 2010. *Early Modern English Medical Texts*. CD-ROM. Amsterdam: John Benjamins Publishing Company.