Using an historical semantic tagger as a diagnostic tool for variation in spelling

Dawn Archer and Paul Rayson,
Dept. of Linguistics and Dept. of Computing,
Lancaster University

We are presently re-developing the UCREL automated semantic analysis system (USAS) for present-day English, so that it will automatically tag texts from the 16th, 17th and 18th centuries. The present USAS system initially assigns every word of a given text (or texts) a part-of-speech tag, and then uses that part-of-speech information, as well as a list of single lexical items, a list of multiple-word-units, contextual rules and a template list to assign semantic field codes. Early Modern English (henceforth EmodE) differs in some significant ways from Present-Day English, of course, not least in terms of its vocabulary, morphology and spelling. This has necessitated that we redesign the modern system, so that the historical tagger includes:

- Additional single lexicon and multiword expression (MWE) dictionaries, which capture terminology that is specific to or has undergone semantic change since the Early Modern Period, and
- Two regularisers. The first regulariser is context-independent, that is variant spellings such as 'aulnage' and 'aulneage' are replaced by their 'normalised' form (in this case, 'alnage'), using a simple search and replace perl script. The second regulariser uses contextual information to amend morphological inconsistencies (such as (e)s for the genitive).

During the process of our research, we have sought to assess the extent to which spelling varies by not only century, but also author and possibly genre in the EmodE period. In this paper, we report some of our findings. In particular, we catalogue the different patterns of variation within a selection of text-types from the 16th, 17th and 18th centuries. The text-types include newsbooks, 'economy' and 'law' texts taken from the Lampeter Corpus, fictional texts such as *Gulliver* and *Tristram Shandy* and play texts (in particular, the work of Shakespeare).