Developing an automated semantic analysis system for Early Modern English

Dawn Archer, Tony McEnery, Paul Rayson, Andrew Hardie
Lancaster University

Abstract

As reported by Wilson and Rayson (1993) and Rayson and Wilson (1996), the UCREL semantic analysis system (USAS) has been designed to undertake the automatic semantic analysis of present-day English (henceforth PresDE) texts. In this paper, we report on the feasibility of (re)training the USAS system to cope with English from earlier periods, specifically the Early Modern English (henceforth EmodE) period. We begin by describing how effectively the existing system tagged a training corpus prior to any modifications. The training corpus consists of newsbooks dating from December 1653 – May 1654, and totals approximately 613,000.words. We then document the various adaptations that we made to the system in an attempt to improve its efficiency, and the results we achieved when we applied the modified system to two newsbook texts, and an additional text from the Lampeter Corpus (i.e. a text that was not part of the original training corpus). To conclude, we propose a design for a modified semantic tagger for EmodE texts, that contains an 'intelligent' spelling regulariser, that is, a system that has been designed so as to regularise spellings in their 'correct' context.

1. Introduction

As reported by Wilson and Rayson (1993) and Rayson and Wilson (1996), the UCREL semantic analysis system (USAS) has been designed to undertake the automatic semantic analysis of PresDE. Given our interest in corpus-based approaches to EmodE (see, for example, Archer 2003; McEnery forthcoming),[1] using the system on historical data is an obvious extension to our work. We thus began to explore whether a system designed for contemporary English can be re-trained to enable the semantic analysis of EmodE texts. In this paper, we provide our initial findings by reporting on work in progress in the form of our application of USAS to a training corpus. The training corpus consists of newsbooks[2] dating from December 1653 – May 1654, and presently totals approximately 613,000 words. An additional 200,000 words of newsbook data is being held in reserve by us for use as a test corpus in future research.

We begin by explaining how the USAS system works, and provide an indication of its 'typical' success rate on contemporary data. We then introduce our EmodE data. This is followed by an explanation of our experiment, the first stage of which involved:

> (i) Tagging 61,065 words of the training corpus *prior to any modifications* (e.g. spelling normalisations). This enabled us to determine how effectively the existing system tagged EmodE data.
> (ii) Summarizing some of the difficulties that the existing USAS system experienced.
> (iii) Documenting the various adaptations that we made to the system at this time, in an attempt to improve its efficiency.

We then report on the second stage of our experiment, namely, applying the modified system to two of the newsbook texts from the training corpus, The *Perfect Diurnall* (issue 215: 16th Jan. 1654) and *Mercurius Politicus* (Issue 188: 19th Jan. 1654), and a third text taken from the 'Economy' domain of the Lampeter Corpus (The *Advocate*, 1652). Our reason for drawing on a text that is not part of the newsbook corpus is so that we can begin to assess the extent to which the system (at its present developmental stage) can automatically tag EmodE texts in general rather than newsbooks in particular. Finally, we propose a design for a semantic tagger for EmodE texts which contains an 'intelligent' spelling regulariser, that is, a system that is able to regularise spellings in their 'correct' context.

[1] The Linguistics Department at Lancaster University has been/is also involved in the construction of corpora covering the EmodE period. These include the *Corpus of English Dialogues 1560-1760*, compiled by Merja Kytö and Jonathan Culpeper of the universities of Uppsala (Sweden) and Lancaster respectively, and the *Corpus of Newsbooks* (see section 3 for description)

[2] 'Newsbooks' are current affairs pamphlets that were published (or intended to be published) weekly.

1.1 Existing historically based research

Little work has been undertaken on the syntactic and/or semantic annotation of non-contemporary English corpora. Some exceptions to this include Kytö and Voutilainen (1995), the *Penn-Helsinki Parsed Corpus of Middle English*, second edition, the *Parsed Corpus of Early English Correspondence*,[3] and the *Historical Thesaurus of English* (Kay 1994). Each of these bar the *Historical Thesaurus of English* undertake a morpho-syntactic analysis of historical data. Kytö and Voutilainen (1995), for example, have adapted/updated the English morphosyntactic lexicon of the *English Constraint Grammar Parser* (ENGCG) by applying it to a number of texts from the EmodE section of the Helsinki Corpus (e.g. 1500-1710). As its name implies, the *Penn-Helsinki Corpus of Middle English*, second edition (PPCME2) also draws on the Helsinki Corpus but, in this case, Middle English prose text samples have been annotated to allow users to search for words, word sequences and, importantly, syntactic structure.[4] The annotation scheme for the corpus was devised by Anthony Kroch and Ann Taylor, and follows the basic formatting conventions of the Penn Treebank (Santorini 1990). Ann Taylor is also involved in the *Parsed Corpus of Early English Correspondence* (PCEEC) project, the aim of which is to produce a tagged and parsed version of the Corpus of Early Engligh Correspondence (CEEC) by 2005. In contrast to the above, the *Historical Thesaurus of English* (HTE) involves semantic analysis. Indeed, the HTE team are developing a comprehensive semantic listing of the vocabulary from Old English to the modern period, which researchers will be able to search automatically. The system of classification they have adopted is based on a modified taxonomy consisting of three major divisions: (I) The World (e.g. the physical universe, plants and animals), (II) The Mind (e.g. man's mental activities), and (III) Society (e.g. social structures and artefacts). These divisions are sub-divided, in turn, into numbered hierarchical categories.[5]

As the wide range of systems for semantic field analysis demonstrate, there is currently little consensus on what constitutes an 'ideal' semantic annotation system (Wilson and Thomas 1997). Consequently, the reader will not be surprised to learn that the semantic tagset that we utilise in our pilot study is different from that of the HTE (see section 2, following, for a description of the semantic tagset, and an indication of its 'typical' success rate). Our purpose in engaging in semantic field analysis is also different. We hope to develop a means of tagging EmodE corpora automatically, with the goal of automated discourse analysis (following Wilson and Thomas 1997) as opposed to constructing a reference thesaurus.

2. The USAS system

The initial tagset of the USAS system was loosely based on Tom McArthur's *Longman Lexicon of Contemporary English* (McArthur, 1981), but has since been considerably revised in the light of practical tagging problems met in the course of previous research. The tagset is arranged in a hierarchy with 21 major discourse fields expanding into 232 category labels. The following table shows the 21 labels at the top level of the hierarchy.

Table 1 The top level of the USAS system

A General and abstract terms	**B** The body and the individual	**C** Arts and crafts	**E** Emotion
F Food and farming	**G** Government and public	**H** Architecture, housing and the home	**I** Money and commerce in industry
K Entertainment, sports and games	**L** Life and living things	**M** Movement, location, travel and transport	**N** Numbers and measurement
O Substances, materials, objects and equipment	**P** Education	**Q** Language and communication	**S** Social actions, states and processes
T Time	**W** World and environment	**X** Psychological actions, states and processes	**Y** Science and technology
Z Names and grammar			

[3] See http://www.eng.helsinki.fi/varieng/team2/1_2_3_cooperation.htm.

[4] See http://www.ling.upenn.edu/mideng/ for more information.

[5] See http://www2.arts.gla.ac.uk/SESLL/EngLang/thesaur/thes.htm for a list of the taxonomy's 'headings'.

The existing USAS system involves two stages. In the initial stage, a part-of-speech tag is assigned to every lexical item or syntactic idiom in the text by the CLAWS part-of-speech tagger (Garside, Leech and Sampson 1987; Garside and Smith 1997), using probabilistic Markov models of likely part-of-speech sequences. Experiments with contemporary texts have consistently shown that CLAWS assigns part-of-speech tags to words in the text with about 97% accuracy. In the second stage, the output from the part-of-speech element is then fed into the semantic tagging suite, SEMTAG (Rayson and Wilson 1996). Experiments with contemporary texts have shown that SEMTAG assigns a semantic tag or tags to each lexical item in the text with about 92% accuracy (Rayson 2001).

3. A description of the training corpus

We have restricted our training corpus to the same text type, e.g. newspapers dating from 1653 to 1654. We decided to concentrate on the more recent EmodE period (i.e. 1640+) so that the problems caused by distance in time would be potentially minimised (cf. Kytö and Voutilainen 1995: 29).[6] The newsbooks are drawn from the Thomason Tracts, a collection of documents published in London between 1640 and 1661. The texts have been encoded in an SGML format conformant with the Text Encoding Initiative (TEI) guidelines, as part of a British Academy-funded project at Lancaster University.[7] The corpus totals 0.8 million words and is drawn from the period March 1653-May 1654. It includes established titles (e.g. *Mercurius Politicus*), titles which begin and end in the period (e.g. *Perfect Diurnal Occurrences*) and titles which had been running for some time but ended in this period (e.g. *Mercurius Democritus*). A breakdown of the training corpus derived from the full corpus is provided below.

Table 2 Breakdown of corpus (to date)

Title of Newsbook	Start date	No. of texts	Total no. of words (to date)
A Perfect Account	Dec 1653	12	31,815
Faithful Scout	Jan 1653	7	6,153
Mercurius Democritus	Nov 1653	4	8,462
Mercurius Nullus	Mar 1654	2	2,320
Mercurius Poeticus	Mar 1654	1	1,950
Mercurius Politicus	Feb 1654	23	118,667
Moderate Intelligencer	Feb 1654	1	3,436
Perfect Diurnall Occurrences	May 1654	6	16,946
Perfect Occurrences	Feb 1654	4	27,851
Proceedings of State Affairs	Feb 1653	16	95,229
The Grannd Politique Post	Mar 1654	1	3,495
The Perfect Diurnall of some passages	Mar 1653	19	133,381
The Politique Informer	Feb 1654	2	6,913
The Politique Post	Feb 1654	19	55,876
The true and Perfect Dutch-Diurnall	Jan 1653	9	21,687
The Weekly Intelligencer	Dec 1653	24	68,225
True and Perfect Informer	Dec 1653	3	7,697
Total		153	612,903

4. First stage of experiment – applying USAS to 25 newsbooks

When we initially explored the possibility of developing an EmodE semantic tagger, we were faced with two possible options. The first involved starting from scratch, using historical data to develop semantic categories. The second involved using the existing resources in the USAS system and adapting them to annotate EmodE data. There are benefits and disadvantages in using either approach. For example, a benefit of the first approach is that the data drives the categories. A disadvantage is the length of time that it takes to develop semantic category systems. Conversely, a benefit of the second option is that the existing resource has been successfully used to tag a variety of contemporary text-types and, in principle, should be able to work in a similar way with EmodE texts.A disadvantage is that the existing resources may need extensive amendment to be applied to the EmodE period. It may also be the case that the semantic

[6] Kytö and Voutilainen (1995: 29) found that ‘the orthographic conventions and morphology of many texts from the third subperiod of Early Modern English in the Helsinki Corpus (1640-1710) approach Present-day English’. We believe the same is true of our newsbook texts.

[7] We would like to thank the British Academy for supporting this project (British Academy reference SG-33825).

categories that work for PresDE do not apply fully to EmodE. The best way to determine whether the existing semantic categories are applicable is to apply the USAS system to some historical data, of course. Consequently, we chose the second option.

The initial stage of the experiment involved submitting a selection of the texts (i.e. 25) to WMATRIX (Rayson 2003), a web-based corpus processing environment, and applying the corpus annotation tools, CLAWS and SEMTAG. We then viewed the results to determine the success rates for the tagging, paying particular attention to the Z99 category. SEMTAG assigns the Z99 category to lexical items when the matching procedure (at the semantic level) fails. In contemporary texts, this tends to be because a word has been misspelled or because a lexical item has not yet been included in the lexicon. It is worth noting that we concentrated on the Z99 category because we envisaged that it would serve as an indicator of how many word forms of EmodE were unknown to a lexicon developed on PresDE. We also envisaged that the number of Z99s in historical texts would be significantly higher than the number of Z99s in contemporary texts. To our surprise, however, the difference was relatively minimal (at least in relation to the newsbooks). Table 1.3 (below) provides a breakdown of the Z99s that occurred in each newsbook title in the small test set. Notice that the number of Z99s is between 4.09% and 9.42%. With contemporary texts, the percentage tends to fluctuate around the 3-8% mark.

Table 3 Breakdown of Z99s in each text

Text	No of texts	Total no of words	Total no of Z99s	As a percentage
The true and Perfect Dutch-Diurnall	9	21,696	888	4.09
Mercurius Politicus	4	16,694	821	4.92
Perfect Diurnall … proceedings	1	5,236	493	9.42
The Weekly Intelligencer	7	18,011	729	4.05
The Politique Post	4	9,428	487	5.17
total	25	61,065	3,418	5.60

Our utilisation of the (modern) USAS system to annotate EmodE data *prior to any modifications* enabled us to identify the following problems:

Irregular spellings	e.g. spelling variations such as *bee* and *doe* that the system was unable to recognise and therefore tag consistently (either as part-of-speech or semantically)
Morphological inconsistencies	e.g. verb endings such as *–eth* and *–est* for the 3rd person present, and (e)s for the genitive
Archaic/rare terminology	e.g. lexical items such as *becalmed* that are no longer in everyday usage (including abbreviations)

The easiest solution to many of the above 'problems' is to update (or, one might say, 'backdate') the existing lexicon to include all archaic words and spelling variations/misspellings. By way of illustration, the USAS system assumed that a word final 's' on an unknown word form in a nominal position signalled a plural noun, and therefore assigned 'hors', a spelling variant of 'horse', with an NN2 part of speech tag, before assigning it to the Z99 category. By including the variant spelling, the USAS system will be able to assign the correct part-of-speech (NN1) and semantic tag (i.e. L2: 'Living creatures generally') in the future.

As we considered different possibilities further, we became concerned that 'backdating' the existing lexicon may adversely affect the efficiency of the existing USAS system on contemporary texts. By way of illustration, the 'backdated' lexicon could potentially assign the wrong semantic tags to 21st century texts, because of an inability to distinguish between archaic and current senses of a particular lexical item. While this problem may be overcome with separate lexicons, a more difficult problem was presented by variant spellings that overlap with words already in the lexicon. Take the verb 'be', for example, which because of the variant spelling 'bee', potentially overlaps with the lexical item that denotes a flying insect. Including 'bee' as both a noun and verb leads to difficulties - preliminary investigations suggested that the existing USAS system treats every instance of 'bee' as a singular noun, in spite of (i) its position in the sentence, and (ii) both possibilities being accounted for in the lexicon).

We decided that the best solution to the first problem (the meaning of words changing over time) was to split the lexicon between PresDE and EmodE and to include

(i) ‘New’ lexical items found in the newsbooks in the existing lexicon, *only if* their use and meaning was similar to that in PresDE.

(ii) Other ‘new’ lexical items in a EmodE lexicon, which will run alongside the existing lexicon.

It is worth noting that many of the Z99 items (see Table 3 above) were spelling variants that were repeated in and across the various texts. This was especially the case with proper nouns (i.e. *Charls, Lilburn, Montross, Norwey*). This meant that the number of Z99s that needed to be inputted (i) was less than Table 3 suggests, and (ii) in practice, decreased as we checked more newsbook texts.

The problem of spelling variants potentially overlapping with other words was more complex. Our solution was to develop a spelling regulariser and a component containing template rules that can deal with irregularities ‘intelligently’ (morpho-modifier). The rest of this paper largely reports on work undertaken when developing the first of those components. We begin, however, by explaining why we believe both components – a spelling regulariser and a morpho-modifier – to be necessary (see section 4.1).

4.1 The need for a spelling regulariser and morphological component

As Schneider (2002) notes, there are several ways of dealing with spelling variants. We might adapt a software program, such as the ENGCG, so that it can cope with the idiosyncrasies of earlier English texts (cf. Kytö and Voutilainen 1995). Alternatively, we might develop a program, such as ZENSPELL, that normalizes spellings (cf. Schneider 2002). Which route one takes will depend on one’s linguistic interests and ultimate goal. As our ultimate aim is the development of a semantic/part-of-speech tagger for EmodE texts, we originally opted for the route of ‘normalisation’. However, it quickly became obvious that this route alone would prove insufficient.

The spelling regulariser consists of two components, namely, a Perl search and replace script and a list of terms, which ‘matches’ a spelling variant to its ‘normalised’ equivalent. In simple terms, the search and replace script searches for occurrences of the spelling variants contained within the list, and, when found, replaces them with an SGML ‘reg’ tag spelling (the part of speech/semantic taggers are SGML aware). Thus, ‘addes’ should be replaced by ‘<reg o= “addes”>adds’. We do this so that the original spelling is retrievable (because of it being encoded in the corpus markup).

It is worth noting that, whilst we wish to provide an automatic spelling regulariser, our present system is perhaps best described as *pseudo*-automatic, as it can only regularise spelling variants that have been manually included in the list. The list itself was generated by two means. Firstly, a close (i.e. manual) examination of the 25 texts mentioned above (in particular, the items given a Z99 semantic label). Secondly, the use of a program to identify – and isolate - the spelling variants that had been manually marked up in a further 40 of the newsbook texts. We report on the accuracy of the spelling regulariser in section 4.3 (following).

The morpho-modifier is designed to capture those instances when simply regularising the spelling proves problematic; for example, the actual spelling variation might already occur in the existing lexicon, but relate to another part-of-speech or semantic category (*be* and *bee*). The aim of the morpho-modifier, then, is to identify significant (i.e. potentially *problematic*) sequences of text, and apply some specified annotation to that text (i.e. a template rule). In the case of *bee*, for example, we are seeking to develop a rule that treats the spelling variant as a verb (infinitive or base form) when (i) preceded by a general preposition or a modal auxiliary, and (ii) followed by an article or the past tense/past participle form of a lexical verb. The rule for recognizing ‘bee’ as an infinitive will look something like the following (for an explanation of the template rule format we are proposing to adopt see Fligelstone *et al* 1996):

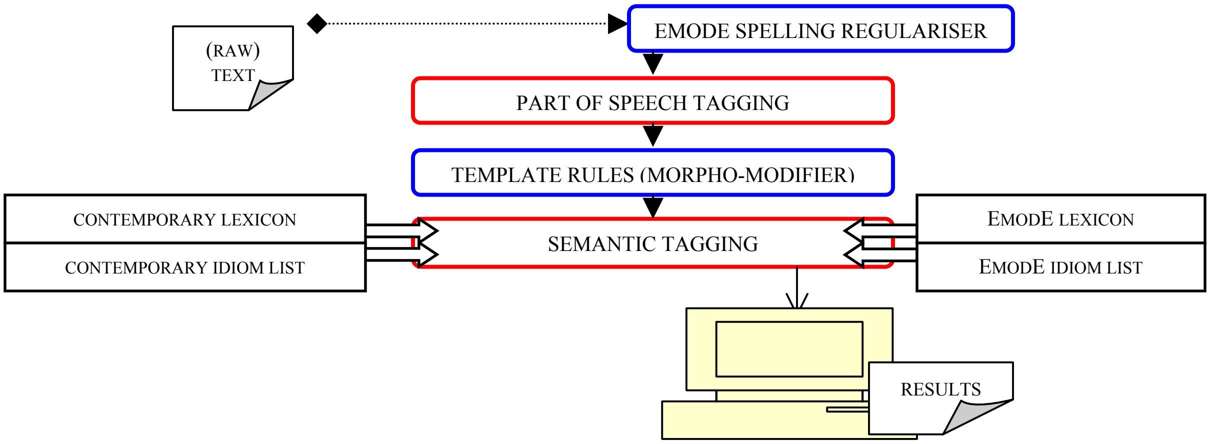
II {bee} (RR*n) VV*

Notice that we have allowed for the possibility that a number of adverbs (RR*n) may occur before the past tense/past participle form of the lexical verb. We also use a wild card (*) so that the rule will match any of several strings (i.e. any form of adverb and the past tense or the past participle of the lexical verb).

4.2 The initial structure of the EmodE tagger

Figure 1.1 (below) provides a visual representation of the initial system. Notice that we incorporated the spelling regulariser at the front-end of the system, before part-of-speech tags are assigned, and the template rule element after the part-of-speech tagger, so that any necessary part-of-speech adjustments can be made prior to the text being semantically tagged.

Figure 1 Diagram Depicting Early Modern English (EmodE) Tagger System



In section 4.3 (following), we highlight the extent to which one of these elements, the spelling regulariser, improved the accuracy of the part-of-speech tagging component of the system when we applied it to three EmodE texts. We then go on to describe modifications that we made to the above structure to ensure even greater accuracy.

4.3 A comparison of the contemporary USAS system and the EmodE tagging system on three texts

We decided that the best way to demonstrate the spelling regulariser component was to initially tag two texts, the *Perfect Diurnall* (issue 215) and *Mercurius Politicus* (issue 188) using both the existing (contemporary) USAS system and the EmodE tagging system, and then compare the results with a third text that was not part of the training corpus (*The Advocate* 1652, from the *Lampeter Corpus*). In this way, we believe our results more accurately reflect the extent to which the system can analyze EmodE texts in general as opposed to those texts on which it has been 'trained'.

Table 4 (below) displays the results we obtained for the two newsbooks texts. An extract from each of the tagged versions is given in Appendix I.

Table 4 Results for *Mercurius Politicus* and *Perfect Diurnall*

	MERCURIUS POLITICUS		*PERFECT DIURNALL*	
	Contemporary WMATRIX system	INITIAL EMODE SYSTEM	Contemporary WMATRIX system	INITIAL EMODE SYSTEM
No of words	4,286	4,270	5,528	5,534
Headline error rate	124 (2.9%)	53 (1.2%)	223 (4.0%)	75 (1.4%)
Variant spellings corrected	-	72	-	148
Variant spellings left uncorrected	123	52	205	75
Variant spellings mis-corrected		1		
POS errors	73	34	97	42
POS errors due to variant spellings	55	30	91	36

Notice that the overall error rate (that is to say, variant spellings and/or part-of-speech errors) is remarkably low when the texts are tagged by the contemporary USAS system (i.e. 2.9% and 4.0%). This suggests that we were right to expect the period after 1640 to be relatively contemporary and that, with amendments, the system will be able to tag historical texts dating from this point with a reasonable level of accuracy.[8] The errors that did occur were due to variant spellings and/or incorrect part-of-speech tagging. Spelling 'errors', for example, were largely due to one or more of the following:

- the insertion of an additional letter[s] (i.e. the addition of *e* in *bee* and *doe*),
- the deletion of an additional letter[s] (i.e. *labors* as opposed to *labours*), and
- the substitution of a letter[s] (i.e. *t* as opposed to *ed* and *i[e]* as opposed to *y*).

Items that were incorrectly tagged included those with –*'d* endings, which the system treated as contractions of a modal auxiliary (VM), and items that signalled their genitive status via an *[e]s* ending. EmodE capitalisation practices also proved problematic, with the result that common nouns were tagged as proper nouns and vice versa.

When we tagged the same texts again using the EmodE version of the software, the spelling regulariser reduced the overall error rate by 1.7% in the case of *Mercurius Politicus* and 2.6% in the case of the *Perfect Diurnall*. Typical mistakes corrected by the spelling regulariser include:

- JJT (possest) → VVN (possessed). Example: '…places which are as yet possessed'
- NN1 (finde) → VV0 (find). Example: '…and find many difficulties'
- NN1 (chiefe, disloyall) → JJ (chief, disloyal). Examples: '…the chief occasion', 'disloyal subjects'

As Table 4 reveals, not all of the spelling variations were altered (see 'variant spellings left uncorrected' column). In part, this is because the design of our system is such that it has to recognise an item as a 'word' before it is part-of-speech tagged. The easiest way of doing this is to separate lexical items using spaces. However, this meant that variant spellings that began or ended a line were not picked up by our system (see section 4.4 below). In addition, examples like 'King' as a personal name (i.e. NP1) were also left unaltered. Such 'errors' are correctable - by including the part-of-speech that they can represent in the relevant lexicon lists (i.e. personal name (NP1), common noun (NN1) and prefixed title (NNB)).

When we repeated the second part of the experiment with the third text from the Lampeter Corpus (i.e. tagged the text using the EmodE tagger), we found that the headline error rate was significantly greater (i.e. 6.3% as opposed to 1.2% and 1.4%; cf. Table 4 above and Table 5 below).

Table 5 Results for *The Advocate* (1652)

Text	No. of words	Headline error rate	Spellings corrected	Spellings left uncorrected	Spellings mis-corrected	POS errors	POS errors due to variant spellings
The Advocate	4,863 (6.3%)	308	227	148		285	172

This is not surprising, of course, as the lexicon was trained on the newsbooks. Consequently, the *Lampeter* text contained variants that had not yet been included in the spelling regulariser list. That said, the overall error rate could be greatly reduced by solving the problem of the system not recognising items which occur at the beginning/end of lines, and thus not altering any variants (i.e. *themselves*, *imploiment*, etc.) that occurred there, even though they may be included in the spelling regulariser list (see section 4.4 below).

In respect of the type of errors that occurred, our findings were very similar to the newsbook texts, but more frequent. For example, *e* was frequently added to possessive pronouns such as *me*, *we* and *he* and auxiliaries such as *be* and *do.* As previously explained, the pronouns can easily be corrected via the spelling regulariser, but variants such as *bee* and *doe* require template rules. Another frequent problem in the *Advocate* was the division of possessive pronouns into two parts (i.e. *our selvs*). This type of error can also be amended by including the variant form and a normalised form in the spelling regulariser. Future research, however, must focus upon how stable spelling variation is in the period. Given that variation

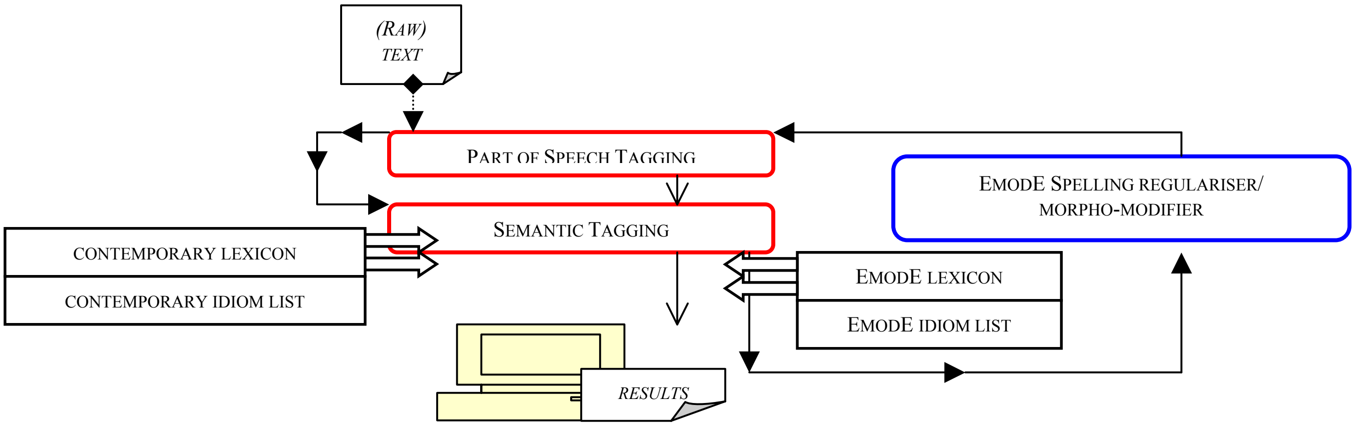
[8] As the modified EmodE tagger appears to be better at tagging than CLAWS is at tagging PresDE (cf. Table 4 and the results we give for modern texts in section 2), it is worth noting that the impressive results are explainable, in part, because they effectively relate to an induced lexicon (see Garside and McEnery 1993 for an indication of the improvements that an induced lexicon can have upon the accuracy of part-of-speech tagging when tagging modern texts).

appears to vary by author and possibly genre, this variation may represent an interesting factor for this research.

4.4 Revised design of the EmodE tagging system

We found that our decision to alter spelling variants could potentially lead to a problem of miscorrection. For example, 'Scots' (NN2) was incorrectly changed to a possessive (i.e. Scot_NN1 's_VBZ) in *Mercurius Politicus*. We had envisaged that such errors would be amended when the text passed through the '*template rules'* component (see Figure 1 above). However, as the existing USAS system appears to cope quite well with the texts, we realised that it would be better to re-structure the EmodE tagger so that the texts are initially part-of-speech and semantically tagged, and then pass through a component that combines regularising instructions and template rules (see Figure 2 below).

Figure 2 Diagram Depicting Early Modern English (EmodE) Tagger System



The strength of such an approach is that the system will be able to use the part-of-speech information to instruct the system when to change a word, and when to leave it unchanged. For example, a noun ending in *(e)s* is most likely to be a genitive (as opposed to a plural) when preceded by a possessive pronoun and followed by a noun, in which case "his kings service" can be normalised to "his king's service" but 'the kings of Europe' can be left unaltered. The modified system will make the change, and then send the text through the part-of-speech tagger again so that it can be re-tagged. A secondary benefit of the restructuring is that the problem of identifying 'words' using spaces is eradicated.

5. Conclusions and future intentions

This paper relates to ongoing work to determine the feasibility of retraining the USAS system to cope with English from earlier periods, specifically the EmodE period. Our initial results suggest that the existing USAS system, including its transition probability matrix, can adequately account for the grammatical features of EmodE. This suggests, in turn, that English grammar has remained quite stable for some 350 years. However, the contemporary system is less successful when attempting to account for EmodE spelling variation and part-of-speech differences. Consequently, we are presently developing an EmodE system that will contain a spelling regulariser and morpho-modifier, as well as PresDE and EmodE lexicons.

We have shown the benefits of one of these elements in this paper– the spelling regulariser. Yet, the procedure we have adopted is currently at best *pseudo* automatic, not least because the program can only search and replace for items included in the spelling regulariser list. The reader should note that we are presently working on the possibility of developing a truly automatic procedure for identifying – and regularising – spelling variants, along the lines proposed by Robertson and Willet (1991, 1993), that is to say, via the use of fuzzy matching algorithms. In the meantime, we are continuing to develop template rules, so that the system can deal with morphological inconsistencies (such as the use of [e]s to mark the genitive) intelligently. We are also investigating the viability of providing more extensive coverage of the possible variants by using the variant facility of the *Oxford English Dictionary*.

The next stage is to apply the EmodE tagger mark II to the remaining training corpus, and a selection of texts from the Lampeter corpus, before undertaking experiments using the semantic categories, using the newsbook test corpus to validate our findings).

8. Bibliography

Archer, Dawn 2003 *The role of the question in EmodE courtroom proceedings: a corpus-based approach*. Unpublished PhD thesis. Lancaster: Lancaster University.

Fligelstone, Steve, Rayson, Paul and Nicholas Smith 1996 Template analysis: Bridging the gap between grammar and the lexicon. In Thomas, Jenny and Mick Short (eds) *Using Corpora for Language Research*. London: Longman.

Garside, Roger, Leech, Geoffrey and Geoffrey Sampson 1987 *The Computational Analysis of English: a corpus-based approach*. London: Longman.

Garside, Roger and Tony McEnery 1993 Treebanking: The Compilation of a Corpus of Skeleton-Parsed Sentences. In Black, Ezra, Garside, Roger and Geoffrey Leech (eds*) Statistically-driven Computer Grammars of English: The IBM/Lancaster approach*. Amsterdam: Atlanta, pp. 17-35.

Garside, Roger and Nicholas Smith 1997. A Hybrid Grammatical Tagger: CLAWS4. In Garside, Roger, Leech, Geoffrey, and McEnery, Anthony (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* London: Longman, pp. 102 – 121.

Kay, Christian 1994 Historical Thesaurus of English: Progress and Plans. In Merja Kytö, Matti Rissanen and Susan Wright (eds), *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora.* Rodopi, Amsterdam,pp 111-120.

Kytö, Merja and Atro Voutilainen 1995. Applying the Constraint Grammar Parser of English to the Helsinki Corpus. *ICAME Journal* (19): 23-48.

McArthur, Tom 1981 *Longman Lexicon of Contemporary English*. Longman Group UK Limited.

McEnery, Tony forthcoming *Swearing in English*. Routledge.

Robertson, Alexander M. and Peter Willet 1991 Digram and trigram matching for the identification of word variants in historical text databases. In McEnery, Tony (ed). *13th Information Retrieval Colloquium, Lancaster 1991.* Chippenham: Antony Rowe Ltd, pp. 12-21.

Robertson, Alexander M. and Peter Willet 1993 Evaluation of techniques for the conflation of modern and seventeenth century spelling. In McEnery, Tony and Chris Paice (eds). *14th Information Retrieval Colloquium, Lancaster 1992.* London: Springer-Verlag, pp. 155-168.

Rayson, Paul 2001 *Wmatrix: a web-based corpus processing environment*. Software demonstration presented at ICAME 2001 conference, Université catholique de Louvain, Belgium.

Rayson, Paul 2003. *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster: Lancaster University.

Rayson, Paul and Andrew Wilson 1996. The ACAMRIT semantic tagging system: progress report. In L. J. Evett, and T. G. Rose (eds) *Language Engineering for Document Analysis and Recognition*, LEDAR, AISB96 Workshop proceedings, pp 13-20. Brighton, England. Faculty of Engineering and Computing, Nottingham Trent University, UK.

Santorini, Beatrice 1990: *Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47*, Department of Computer and Information Science, University of Pennsylvania.

Schneider, Peter 2002 Computer assisted spelling normalization of 18th century English. In Peters, Pam, Collins, Peter and Adam Smith (eds) *New frontiers of corpus research: Papers from the Twenty First International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, pp. 199-211.

Wilson, Andrew and Jenny Thomas 1997 Semantic annotation. In Garside, Roger, Leech, Geoffrey and Anthony McEnery (eds). *Corpus Annotation*. Addison Wesley Longman Limited, pp. 53-65.

Wilson, Andrew and Paul Rayson 1993 Automatic Content Analysis of Spoken Discourse: a report on work in progress. In: C. Souter and E. Atwell (eds). *Corpus Based Computational Linguistics*. Amsterdam: Rodopi, pp. 215-226.

9. Appendix I: An extract of the texts

Mercurius Politicus

A_AT1 party_NN1 of_IO hors_NN2 was_VBDZ sent_VVN thereupon_RT to_TO seise_VVI him_PPHO1 ;_; but_CCB he_PPHS1 quitted_VVD that_CST place_VV0 about_RG 2_MC daies_NNT2 before_RT ,_, and_CC the_AT Countrey_NP1 are_VBR so_RG false_JJ to_II us_PPIO2 ,_, that_CST they_PPHS2 will_VM give_VVI no_AT direction_NN1 which_DDQ way_NN1 to_TO follow_VVI upon_II any_DD such_DA occasion_NN1 Those_DD2 Scots_NN2 that_CST will_VM not_XX rise_VVI with_IW the_AT Highlanders_NN2 are_VBR plundred_VVN by_II them_PPHO2 ;_; the_AT Highlanders_NN2 lay_VV0 Contributions_NN2 upon_II them_PPHO2 ,_, which_DDQ bring_VV0 them_PPHO2 very_RG low_JJ ,_, because_CS they_PPHS2 pay_VV0 likewise_RR toward_II the_AT maintenance_NN1 of_IO our_APPGE English_JJ Army_NN1 ._. …

(Contemporary version of software)

A_AT1 party_NN1 of_IO <reg o="hors"> horse_NN1 was_VBDZ sent_VVN thereupon_RT to_TO <reg o="seise"> seize_VVI him_PPHO1 ;_; but_CCB he_PPHS1 quitted_VVD that_CST place_VV0 about_RG 2_MC <reg o="daies"> days_NNT2 before_RT ,_, and_CC the_AT <reg o="Countrey"> country_NN1 are_VBR so_RG false_JJ to_II us_PPIO2 ,_, that_CST they_PPHS2 will_VM give_VVI no_AT direction_NN1 which_DDQ way_NN1 to_TO follow_VVI upon_II any_DD such_DA occasion_NN1 Those_DD2 <reg o="Scots"> Scot_NN1 's_VBZ that_CST will_VM not_XX rise_VVI with_IW the_AT Highlanders_NN2 are_VBR <reg o="plundred"> plundered_VVN by_II them_PPHO2 ;_; the_AT Highlanders_NN2 lay_VV0 Contributions_NN2 upon_II them_PPHO2 ,_, which_DDQ bring_VV0 them_PPHO2 very_RG low_JJ ,_, because_CS they_PPHS2 pay_VV0 likewise_RR toward_II the_AT maintenance_NN1 of_IO our_APPGE English_JJ Army_NN1 ….

(EmodE version of software)

The Perfect Diurnall of some passages

Thus_RR much_RR I_PPIS1 have_VH0 conceived_VVN it_PPH1 fit_JJ to_TO shew_VVI you_PPY ,_, that_CST I_PPIS1 might_VM undeceive_VVI you_PPY in_II some_DD opinions_NN2 ,_, upon_II which_DDQ I_PPIS1 find_VV0 you_PPY ground_VVD your_APPGE Arguments_NN2 in_II your_APPGE Letter_NN1 to_II the_AT Gentlemen_NN2 of_IO Badgenoth_NP1 :_: As_CSA first_MD you_PPY look_VV0 upon_II your_APPGE Commonwealth_NN1 (_(as_CSA you_PPY call_VV0 it_PPH1)_) as_CSA one_MC1 firm_NN1 and_CC fixt_VV0 Government_NN1 ,_, whilest_JJT the_AT Lord_NN1 knowes_NN2 ,_, there_EX is_VBZ at_II this_DD1 instant_NNT1 no_AT such_DA thing_NN1 in_II England_NP1

(Contemporary version of software)

Thus_RR much_RR I_PPIS1 have_VH0 conceived_VVN it_PPH1 fit_JJ to_TO <reg o="shew"> show_VVI you_PPY ,_, that_CST I_PPIS1 might_VM undeceive_VVI you_PPY in_II some_DD opinions_NN2 ,_, upon_II which_DDQ I_PPIS1 find_VV0 you_PPY ground_VVD your_APPGE Arguments_NN2 in_II your_APPGE Letter_NN1 to_II the_AT Gentlemen_NN2 of_IO Badgenoth_NP1 :_: As_CSA first_MD you_PPY look_VV0 upon_II your_APPGE Commonwealth_NN1 (_(as_CSA you_PPY call_VV0 it_PPH1)_) as_CSA one_MC1 firm_NN1 and_CC <reg o="fixt"> fixed_JJ Government_NN1 ,_, <reg o="whilest"> whilst_CS the_AT Lord_NN1 knowes_NN2 ,_, there_EX is_VBZ at_II this_DD1 instant_NNT1 no_AT such_DA thing_NN1 in_II England_NP1 ,_,

(EmodE version of software)

The Advocate

I_PPIS1 Am_VBM often_RR in_II very_RG great_JJ doubt_NN1 (_(if_CS I_PPIS1 may_VM so_RR speak_VVI)_) ,_, that_CST the_AT Goodness_NN1 &; Wisdom_NN1 of_IO God_NP1 ,_, &; his_APPGE thoughts_NN2 of_IO these_DD2 ,_, are_VBR very_RG rarely_RR met_VVN with_IW in_II the_AT Paths_NN2 ,_, which_DDQ the_AT scantling_NN1 of_IO Man_NN1 's_GE Reason_NN1 and_CC <reg o="Judgment"> judgement_NN1 walk_VV0 in_RP ;_; And_CC as_CSA I_PPIS1 dare_VV0 not_XX but_CCB own_VV0 the_AT Belief_NN1 of_IO the_AT Coming_NN1 of_IO his_APPGE Appearance_NN1 ,_, and_CC the_AT breaking_NN1 forth_RR ,_, very_RG shortly_RR ,_, of_IO his_APPGE Glorie_NP1 :_: So_RR I_PPIS1 <reg o="believ"> believe_VV0 likewise_RR ,_, this_DD1 will_NN1 bee_NN1 a_AT1 sight_NN1 very_RG strange_JJ ,_, and_CC very_RG <reg o="unexspected"> unexpected_JJ to_II men_NN2 ;_; and_CC not_XX <reg o="onely"> only_RR greatly_RR above_RL ,_, but_CCB in_II <reg o="som"> some_DD measure_NN1 even_RR <reg o="contrarie"> contrary_JJ (_(and_CC perhaps_RR ,_, very_JJ unwelcom_NN1)_) unto_II the_AT most_RGT enlarged_JJ and_CC raised_JJ thoughts_NN2 <reg o="wee"> we_PPIS2 have_VH0 yet_RR prepared_VVN <reg o="our selvs"> ourselves_PPX2 *_FU with_IW ,_, to_TO <reg o="receiv"> receive_VVI it_PPH1 ._.

(EmodE version of software)