Tagging the Bard:
Evaluating the accuracy of a modern POS tagger on Early Modern English corpora

Paul Rayson[1], Dawn Archer[2], Alistair Baron[1], Jonathan Culpeper[1], Nicholas Smith[1]
[1]Lancaster University
[2]University of Central Lancashire
paul@comp.lancs.ac.uk, dearcher@uclan.ac.uk, a.baron@comp.lancs.ac.uk, j.culpeper@lancs.ac.uk, nick@comp.lancs.ac.uk

Abstract

In this paper we focus on automatic part-of-speech (POS) annotation, in the context of historical English texts. Techniques that were originally developed for modern English have been applied to numerous other languages over recent years. Despite this diversification, it is still almost invariably the case that the texts being analysed are from contemporary rather than historical sources. Although there is some recognition among historical linguists of the advantages of annotation for the retrieval of lexical, grammatical and other linguistic phenomena, the implementation of such forms of annotation by automatic methods is problematic. For example, changes in grammar over time will lead to a mismatch between probabilistic language models derived from, say, Present-day English and Middle English. Similarly, variability and changes in spelling can cause problems for POS taggers with fixed lexicons and rule-bases. To determine the extent of the problem, and develop possible solutions, we decided to evaluate the accuracy of existing POS taggers, trained on modern English, when they are applied to Early Modern English (EModE) datasets. We focus here on the CLAWS POS tagger, a hybrid rule-based and statistical tool for English, and use as experimental data the Shakespeare First Folio and the Lampeter Corpus. First, using a manually post-edited test set, we evaluate the accuracy of CLAWS when no modifications are made either to its grammatical model or to its lexicon. We then compare this output with CLAWS' performance when using a pre-processor that detects spelling variants and matches them to modern equivalents. This experiment highlights (i) the extent to which the handling of orthographic variants is sufficient for the tagging accuracy of EModE data to approximate to the levels attained on modern-day text(s), and (ii) in turn, whether revisions to the lexical resources and language models of POS taggers need to be made.

1. Introduction

Annotation of corpora, that is "the practice of adding interpretative [and/or] linguistic information to a corpus" (Leech, 1997: 2), can be applied at a number of levels and by a variety of manual and automatic techniques. It is worth noting that, as grammatical and semantic information has a more explicit presence in the language form, the process of adding *linguistic* annotation is easier to automate than the process of adding *interpretative* information: put simply, there are formal traces which a computer can use to work out the category in which to place a chunk of language. For example, words ending '-ness' are likely to be nouns, as are the word(s) following (though not necessarily immediately) the word 'the'. Complex probabilistic rules can be built up in order to achieve fairly accurate tagging. Social and pragmatic categories

such as those presented in Archer and Culpeper (2003), on the other hand, are far less likely to be recoverable from the language form, and thus must be added manually: Archer and Culpeper (2003) have developed a system for adding social and pragmatic information to a text (in respect to the speaker and addressee's age, gender, role and status) utterance by utterance. In contrast, interpretative information (such as social information about the speakers or information about the provenance of a written text) is usually given in file headers of corpora such as the British National Corpus (BNC).

The automatic annotation of historical corpora is also problematic: with standardised spelling, any given lexeme can be retrieved by the computer with ease, since each lexeme normally has just one associated word-form. However, in the Early Modern English period, including the period in which Shakespeare was writing, each lexeme had word-forms characterised by variable spellings (see, for example Osselton, 1984). For example, for a computer to retrieve all instances of the word 'would' (a form that, in Present-day English, matches the lexeme one-to-one), the computer would have to match an array of forms including: *would*, *wold*, *wolde*, *woolde*, *wuld*, *wulde*, *wud*, *wald*, *vvould*, *vvold*, and so on. This means that any automated tagging program will fail utterly, as such programs rely - amongst other things - on matching words against lexicons. Even regularised editions of Early Modern English texts[1] present potential problems for analysis, such as morphological variants (e.g. 'tellest', 'telleth'), grammatical variants (e.g. 'ye', 'thou', 'thine'), orthographic oddities (e.g. 'wing'd' instead of 'winged', the lack of an apostrophe for the s-genitive, capitalisation practices), and archaic/obsolete forms (e.g. 'becalmed'). Naturally, these difficulties will have a knock-on effect for any subsequent corpus processing tasks, such as the creation of word frequency lists, concordances, collocation statistics, and n-grams. However, our focus in this paper is investigating the problems caused for automatic corpus annotation tools and, in particular, part-of-speech taggers.

2. Background

Many existing historical corpora are available for the Early Modern English period: e.g. the Lampeter Corpus (1640-1740), Corpus of English Dialogues (1560-1760), the Helsinki corpus (Old, Middle and Early Modern English), and the Archer corpus (1650-1990, sampled at 50-year periods). In addition, vast quantities of new searchable textual material are being created in electronic form through large digitisation initiatives currently underway: see, for example, the Open Content Alliance[2], Google Book Search[3], Early English Books Online[4]. These initiatives are largely focussed on historical or more recent out-of-copyright material. As well as image-based digitisation, transcription and OCR-scanning techniques are being used to produce text-based materials. Increasingly, researchers will carry out linguistic analysis and complex retrieval tasks on this historical data (Nissim et al, 2004).

Part-of-speech (POS) tagging is perhaps the most common form of corpus annotation: grammatical annotation can be useful in situations where you want to distinguish the grammatical functions of particular word forms or identify all the words performing a particular grammatical function. For example, you may be interested in the usage of the word 'house' as a verb (as in 'the specimens are now

[1] For example, the Nameless Shakespeare (Mueller, 2005)
[2] http://www.opencontentalliance.org/
[3] http://books.google.com/
[4] http://eebo.chadwyck.com/home

housed in the British Museum'). The past participle of 'house' (i.e. 'housed') can be tagged 'housed_VVN', the –ing participle 'housing_VVG', the –s form of the lexical verb 'houses_VVZ', and so on. If you want *all* the verbal forms of 'house' without doing separate searches, some wildcards will do the trick in principle, e.g. 'hous*_V*'. In practice, the exact query syntax may be determined by the kind of corpus you are interrogating and the kind of software you are using (e.g. some programs have difficulty with the underscore). Alternatively, one might wish to see what kinds of word perform particular functions and their respective frequencies. Grammatical annotation could be used to retrieve all the verb forms in a particular text, and then one could compare these with the verb forms of other texts.

POS tagging software has been under development at Lancaster University since the early 1980s. The software in question, CLAWS (Garside and Smith, 1997)[5], works on the basis of:

(1) a lexicon, including words (or multi-word units) and suffixes and their possible parts of speech, and
(2) a matrix containing sequencing probabilities (e.g. the likelihood that following an adjective there will be a noun), which is applied to each sentence to disambiguate words that could be several parts-of-speech, potentially.

CLAWS achieves 96%-97% accuracy on written texts, and a slightly lesser degree of accuracy on spoken texts (Leech and Smith, 2000).

Early grumbles about corpus annotation from a few (historical) corpus linguists seemed to be based on assumptions that the annotations were damaging the integrity – even purity – of the text, and this argument is still voiced for modern data (Sinclair 2004: 191). This is not a serious objection, as one can have multiple copies of a corpus and just annotate one of them, or a computer can easily hide most forms of annotation. However, most historical linguists working with corpora would recognise the value of grammatical annotation. It would enable one to track usage of particular grammatical categories (nouns, complementisers and passives, for instance), and their associated contexts, across time. Fourteen years ago, Kytö and Rissanen (1993: 1) wrote:

> At present, the usefulness of our corpus is diminished by the absence of grammatical tagging. This means that all searches must be based on words, or their parts or combinations. Programs suitable for tagging historical text material are being developed in various parts of the world, and we hope to start applying these programs to some subsections of our corpus in the near future. It is obvious, however, that equipping the entire corpus with even a fairly simple grammatical tag system would, with our present resources, be a matter of years of hard work.

Here they were talking about the Helsinki corpus, which includes Old, Middle as well as Early Modern English. In fact, from that time to the present, there has been much progress in creating POS tagged and grammatically parsed corpora. Kytö and Voutilainen (1995), for example, have back-dated the English morphosyntactic lexicon of the English Constraint Grammar Parser (ENGCG) before applying it

[5] CLAWS is the Constituent Likelihood Automatic Word-tagging System. More information, including the full C7 tagset used here, can be found at: http://ucrel.lancs.ac.uk/claws/.

automatically to tag a number of texts from the EmodE section of the Helsinki Corpus (e.g. 1500-1710). Durrell et al. (2006) have also adopted the approach of adding historical variants to the POS tagger's lexicon, in their application of the TreeTagger for POS annotation of the GerManC corpus. In addition, Anthony Kroch and Ann Taylor have released syntactically parsed versions of historical corpora, following the basic formatting conventions of the Penn Treebank (Santorini 1990): see, for example, The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE), The York-Helsinki Parsed Corpus of Old English Poetry, the Penn-Helsinki Corpus of Middle English, second edition (PPCME2) and the Parsed Corpus of Early English Correspondence (PCEEC).

It is worth noting that, in all these cases, parsing has involved a great deal of manual post-editing. Perhaps as a consequence of this, tagging efforts have been focused on the earliest periods of English (the entire Old English corpus amounts to about 3 million words), corpora consisting of limited samples, or specific genres. Automated yet accurate procedures would, of course, allow for the tagging of far greater quantities of linguistic data.

Our efforts in this emerging research area centre on the VARD, or Variant Detector, tool, described previously in Rayson et al (2005, 2007). The VARD is designed to assist users of historical corpora in dealing with spelling variation, particularly in EModE texts. The tool, which is under continuing development, uses techniques derived from modern spell-checkers to find candidate modern form replacements for variants found within historical texts.

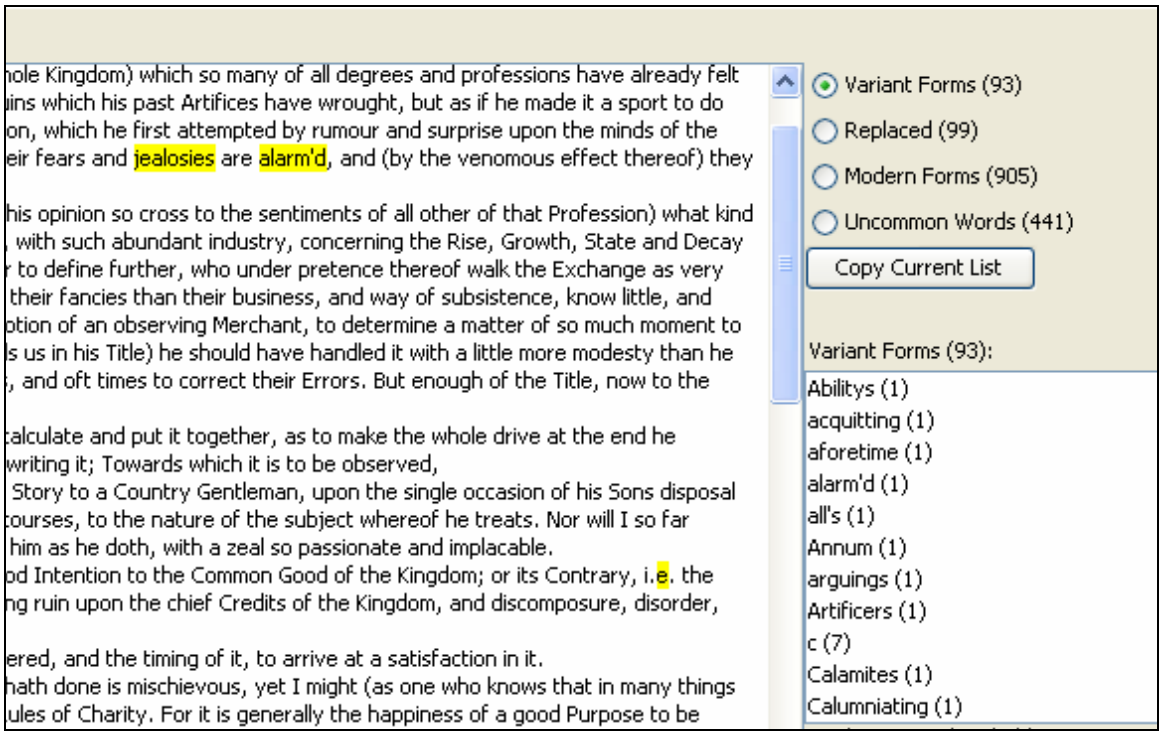

Figure 1: Screenshot showing the grouping of words by VARD

The current version of the tool scans through a given text and marks any words which are not in its modern English word list as potential variants. In a recent addition, the tool also separates out any words which it considers uncommon into a separate list. This process is based on word frequency in the British National Corpus (BNC) (Leech et al. 2001). The search for uncommon words is the first step in finding

variants which are graphically the same as a modern word form; future work will involve using context-sensitive information to find more of these variants. Once the text has been processed, the words found within are divided into 3 groups: variant forms, modern forms and uncommon words. Another group exists for replaced words; this is initially empty and populated during the later editing phase. Each group can be selected and the words within the group highlighted in the text. This is shown in Figure 1.

The VARD produces a list of candidate replacements for each variant form found. To do this, it uses (a) a manually created list of variants mapped to modern form replacements, (b) the SoundEx algorithm, and (c) letter replacement heuristics. Each candidate is then given a confidence measure based on which methods were used to find it, as well as the Levenshtein Distance between the variant and the replacement. The candidates are then ranked by this confidence measure and offered to the user for consideration. If the confidence measure for two candidates is the same, the replacement with the higher frequency (within the BNC) is preferred.

A user can make their way through a text, right-clicking on any highlighted variant to be presented with the list of candidate replacements. Clicking on an offered replacement changes the variant to the modern form selected. It is important to point out that the variant form is never discarded. In the output file it is placed inside an XML tag, and juxtaposed with the modern replacement form, as follows:

my very <replaced variant="lippes" found by="ps" replacementType="pr" ed="2">lips</replaced> might freeze to my teeth

The process of replacing a variant is shown in Figure 2. As can be seen, the tool displays how it arrived at the confidence measure by indicating which methods were used to find the replacement (shown with ticked options in the pop-up window).

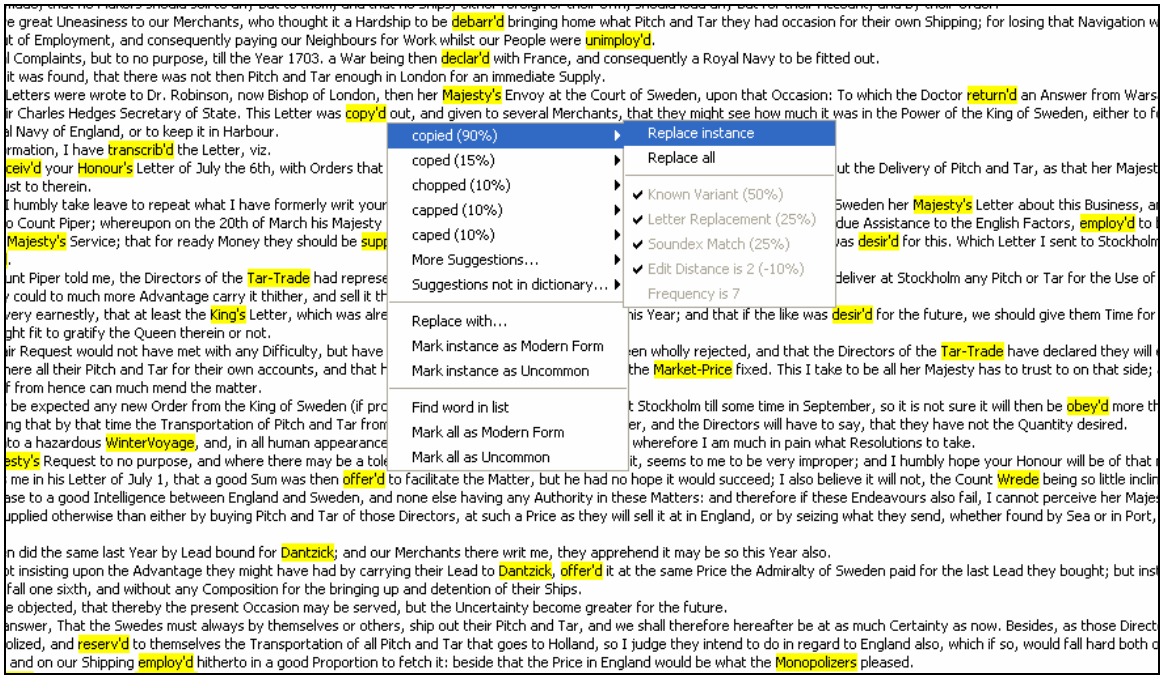

Figure 2: Screenshot showing a user selecting a replacement for a variant

As well as manually selecting and discarding the variant forms, the user can instruct the tool to automatically replace any variant form with its highest ranked candidate replacement. The user can also provide a threshold confidence measure, which is the minimum score the candidate replacement must reach for it to be used. By using this feature with a relatively high threshold, the user can automatically replace the most common variant forms, thereby saving a substantial amount of time.

VARD now has a much larger dictionary than that used for previous experiments. The decision was taken to increase the size of the dictionary (currently 26,071 entries) as, previously, many modern words were being incorrectly marked as variants. The increased size of the dictionary has resulted in some problems, however. Variant forms such as *bee, doe* and *wee* (*be, do* and *we*) are no longer marked as variants, because the enlarged dictionary now includes other lexemes that happen to share these wordforms as their "standard" form". This, in part, led to the introduction of the uncommon words group as described above. A user can scan this list for potential variant forms. It is better to have such words marked as uncommon rather than variants, as it could be the case that *bee, doe* or *wee* is the correct modern form. In our future work, we will focus on the use of context-based template rules to disambiguate such "real-word" variants in historical texts.[6]

A further function recently added to the tool, which was not available for previous experiments, is the ability to join two words separated by white space. This allows the user to deal with cases where a variant form of a modern word is split into two separate words (e.g. *to morrow*, *some one*, *your selfe*), and words split over line breaks. Words split over two lines are fairly frequent in Early Modern English, as this is one strategy that compositors used to avoid "ragged right".

From earlier tests regularising the spelling of historical texts with the VARD program, we can reasonably expect the error rate to be at acceptable levels. Archer et al. (2003) provide information on POS error rates for two texts dating from 1654, totalling approximately 9,804 words. When the texts were not regularised, 170 POS errors were reported (1.7%), of which 146 were due to variant spellings; when they were regularised, this dropped to 76 POS errors (0.8%), of which 66 were due to variant spellings. However, these figures are not as good, perhaps, as they may sound, since they represent errors reported by the program. It is possible that there were other errors that went unreported. The study we report here includes careful manual checking of all experimental data, not just the errors retrieved. In addition, one might also argue that, by 1654, the standardisation of spelling was well underway, and so the challenge for regularising spelling is less than that represented by our Shakespearean texts.

3. Data for the experiment

Shakespeare's style might be said to vary along two key dimensions: (1) genre (i.e. the traditional categories of comedy, tragedy or history), and (2) period (by this we mean the way in which his style changed significantly during his lifetime; indeed, it is sometimes considered by scholars to contain four phases). Unfortunately, Shakespeare clearly did not have the corpus linguist in mind when he produced his plays, as he did not write all genres in all periods, allowing one to fully explore both dimensions. We have selected plays from one genre, comedies (as classified in the

[6] A related problem is the frequent interchanging of *then* and *than* in earlier stages of English. Again, we intend to apply context-based rules to identify the modern-day equivalent.

First Folio), because that is the only genre that covers his entire writing career (1584-1613). The particular plays are *Taming of the Shrew*, *Love's Labour's lost*, *Merry Wives of Windsor*, *Twelfth Night* and *Tempest*. These plays have been selected so as to evenly span his writing career (though the exact dating of the some of the plays is controversial). However, given that orthography is of particular importance to us, the key date is when they were printed. All plays are from the First Folio printed in 1623 (sourced from the Oxford Text Archive[7]). Note that comedies tend to be amongst the most conversational and speech-like historical data, something which is likely to present POS tagging with further difficulties (e.g. ellipsis).

In addition to the Shakespeare data, we selected three files from the Lampeter corpus to provide a contrast and a further test in the experiment. The Lampeter corpus contains transcribed versions of literature published as tracts or pamphlets in the 17th and 18th centuries. For each of six domains, there are two texts per decade contained in the corpus. In our experiment, one file was selected from three domains: economy and trade (eca1641), law (lawa1643) and science (scia1644). All three files were selected from the earliest decade sampled in the corpus, i.e. 1640s, to provide a greater challenge than that attempted in our earlier work (Archer et al. 2003). Siemund and Claridge (1997: 64) report that the Lampeter corpus "retains the original orthography, punctuation and word divisions".[8]

For each of the texts, we selected 1,000-word samples for analysis. Although this is a relatively small amount, it was felt that this would provide a sufficient sample for estimation of POS tagging accuracy while still remaining manageable for manual checking within the limited time available. The word count feature within Microsoft Word 2003 was used for the selection. We used a random line position[9] in the text to begin the sampling and then selected a minimum of 1,000 running words including up until the following speaker change or end of the sentence. We did not count speaker names and we removed line number markers, which appear in the Shakespeare corpus files. We opted to keep any stage directions in the Shakespeare files to be analysed. Following this process, we selected 5,011 words from the five Shakespeare files and 3,025 words from the three Lampeter files.

4. Evaluation

The experimental set up was as follows. For each 1,000-word sample, we applied the following process, which is represented graphically in figure 3:

1. Apply automatic POS annotation using the CLAWS software with the standard modern-language linguistic resources
2. Manually post-edit the POS tags in the tagged versions resulting from step 1 to produce a gold standard (or baseline) for comparison
3. In parallel to step 2, run the untagged text through VARD, and automatically select modern equivalents over 70% likelihood[10]
4. Apply automatic POS annotation to the texts resulting from step 3

[7] http://ota.ahds.ac.uk/

[8] We also considered using data from the Archer corpus (Biber et al. 1994) in our experiments. However, we found that in this corpus the original orthography is retained for some texts, but normalized (often silently) for others. As such, it is not ideal for our purposes here.

[9] Generated from http://www.random.org

[10] The figure of 70% was selected from earlier trials as a good cut-off point to balance precision and recall

5. In parallel to steps 2 and 3, manually check the untagged text and insert modern equivalents alongside EmodE variants
6. Apply automatic POS annotation to the texts resulting from step 5
7. Compare results pairwise from steps 1 and 2 (to check baseline accuracy), then steps 2 and 4 (to check if there is any improvement in accuracy using automatic VARD), then steps 2 and 6 (to check how much improvement we could achieve using a 'perfect' VARD)

The resulting accuracy rates are displayed in Table 1. Compared to the reported accuracy figures for CLAWS on modern standard British English (96-97%), there is a significant drop in performance on the EModE data. We have listed the gold standard versions without an accuracy figure since they have been manually post-edited and could be assumed to be virtually free from errors.

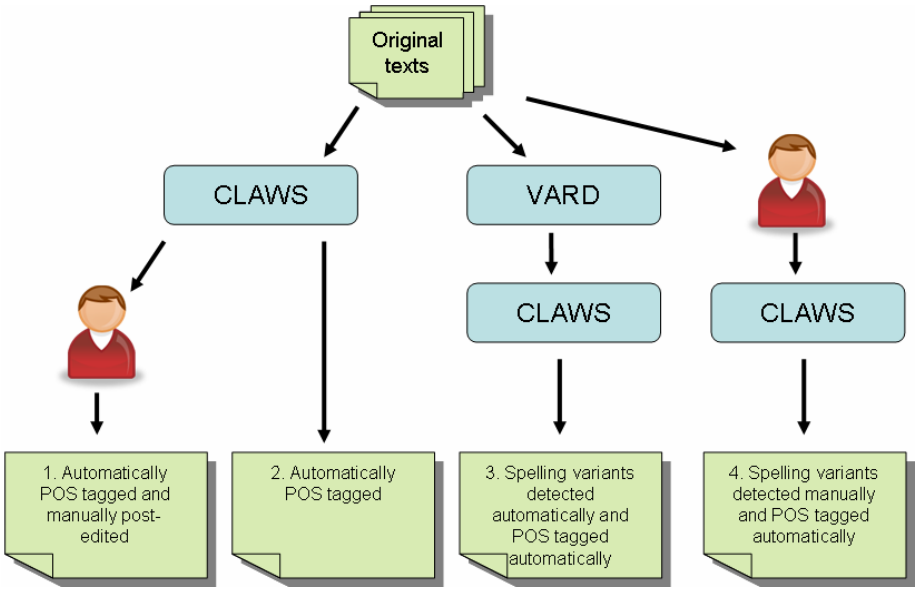

Figure 3: Annotation and manual checking process

	POS tagging accuracy	
	Shakespeare	Lampeter
1. Gold standard (manually post-edited)	n/a	n/a
2. Automatically POS tagged only	81.94%	88.46%
3. Variant spellings automatically modernised	84.81%	89.39%
4. Variant spellings manually modernised	88.88%	91.24%

Table 1: Comparison of POS tagging accuracy

During the process of manual correction of POS tags, we encountered a few instances where it was difficult to accurately determine the correct POS tag. In the example listed in (1), it is debatable whether *Lunaticks* is an adjective or noun. It is attributive in function, and the complement of a singular verb (in parallel with *this is madde*), suggesting an adjective. On the other hand its inflection suggests a plural noun, and only the plural noun usage is listed in the online OED.

(1) Why, this is Lunaticks: this is madde, as a mad dogge.
(Merry Wives of Windsor)

In (2), *Auant* is almost certainly a variant of *avaunt*, but its grammatical wordclass is less clear-cut. Its meaning essentially seems to be connected to the sense 'to be off, go away, depart' (OED *avaunt* v.2); it could be construed as an imperative form of the verb, or as an interjection (OED *avaunt* adv., int.):

(2) Auant perplexitie: What shall we do, If they returne in their owne shapes to wo? (Love's Labour's Lost)

The number of unresolved cases was extremely small, representing fewer than 0.1% of all words. These few instances have been excluded from the experiment. For the most part, it was possible to apply our existing set of guidelines for POS-tagging[11] without difficulty. This lends some support to the widely held view that – at the syntactic level – the English language has not changed substantially since the Early Modern period.

In the case of the Shakespeare data, the unaltered text results in a reduction of CLAWS accuracy to just under 82%. With an automatic modernisation of spelling variants using VARD, this increases by almost 3%. There appears to be a ceiling to the accuracy even with manual modernisation of spellings, with another 4% increase in accuracy to just under 89%. At first glance, we might note that this is still around 8% removed from the reported CLAWS accuracy. However, we should note that, given the difference in style/genre, a more accurate comparison would be to CLAWS' accuracy on modern plays, and legal, scientific and business writings.

Examining the results for the Lampeter data, we see that the initial reduction in accuracy is not as striking as it is with the Shakespeare data. The accuracy rate for running CLAWS over unedited Lampeter data was 88.46%. With an automatic pre-processing step inserting modernised forms, the accuracy rate improves by almost 1%. With a manual process of checking for EmodE variant spellings, we observed another 2% increase in POS tagging accuracy.

There are two clear effects at work here which result in the Lampeter data being less problematic for CLAWS than the Shakespeare samples. First, the Lampeter corpus samples date from at least two decades after the Shakespeare First Folio was printed, and thus are more "standardised". Second, there is the difference in grammatical style in the genres selected from the Lampeter corpus: in general, texts in the latter are written in a form of expository prose that is stylistically very similar to that of modern-day mainstream varieties of text, i.e. the kind of data that CLAWS' language model has been derived from.

Closer inspection of the Lampeter data highlighted an additional problem facing automatic spelling regularisation for this period, namely code-switching between English and Latin. At the end of one of the Lampeter text samples there appears a passage of Latin, which CLAWS failed to identify with its 'FW' (foreign word) tag. If we were to remove this passage from the experiment, the fourth accuracy figure reported in Table 1 would have been higher still (93.22%)

Table 2 shows some typical examples of mistagging due to spelling variants, as well as correct tagging in spite of spelling variants.[12]

---

[11] Guidelines for wordclass tagging (based on Present Day English) are available on the UCREL website: http://www.comp.lancs.ac.uk/ucrel/bnc2sampler/guide_c7.htm

[12] POS-tags used in these examples are: APPGE (possessive form of personal pronoun), JJ (adjective), NN2 (plural common noun), VVN (past participle), VVZ (present tense lexical verb, 3sg.), NN1 (singular common noun), PPIS1 (personal pronoun, 1sg., nominative), II (preposition), NNT1 (singular temporal noun).

Example	Comments
with your_APPGE sweete_JJ breathes_NN2 puft_VVN out (Love's Labour's Lost)	*sweete*, *breathes* and *puft* are correctly tagged despite being variants for *sweet*, *breaths* and *puffed* respectively
it is leggs_NN2 and_CC thighes_VVZ (Twelfth Night)	*leggs* (*legs*) is correctly tagged *thighes* (*thighs*) is mistagged as a verb rather than a noun
Be not deni'de_NN1 accesse_NN1 (Twelfth Night)	*deni'de* (*denied*) is mistagged as a noun rather than a verb (past participle); *accesse* (*access*) is correctly tagged as a noun
I_PPIS1, 't is strong (Twelfth Night)	*I* (*aye* meaning yes) mistagged as a personal pronoun
the Razors_NN2 edge (Love's Labour's Lost)	*Razors* mistagged since apostrophe is missing (*Razor's*) from genitive
but your_APPGE selfe_NN1 (Love's Labour's Lost)	*yourself* incorrectly tokenised as two separate lexemes rather than a singular reflexive personal pronoun
Ile_JJ ride_NN1 home to_II morrow_NNT1 (Twelfth Night)	*Ile* incorrectly tokenised as one word, tagged as adjective, *ride* tagged as noun (rather than verb); *tomorrow* incorrectly tokenised as two words: a preposition plus a time noun

Table 2: Examples of mistagging and spelling variants

There were also a large number of cases where CLAWS correctly tagged the word despite it containing some sort of spelling variation. Of the total number of words in the Shakespeare sample, 5.78% were variants which CLAWS tagged correctly. CLAWS is robust enough in these cases to guess the correct tag using its probabilistic model, and heuristics such as common patterns of word endings.

However, as can be seen, for the Lampeter data and to a greater extent for the Shakespeare data, we observed a significant reduction in POS tagging accuracy for both of the historical datasets compared with modern data. We have also shown that a large part of this reduction in accuracy was due to spelling variation. The last example in Table 2 illustrates, moreover, that some of these errors are consequential: it is because CLAWS thinks *Ile* is an adjective (JJ) that it assigns to the word *ride* the tag for singular common noun (NN1). Resolving the first of these errors would naturally eliminate the error on the second.

5. Conclusion

In previous studies that have relied on the automatic detection of problems, we have reported that POS-tagging accuracy (for example in CLAWS) is affected by spelling variants in Early Modern English (see, for example, Archer et al. 2003). In the research reported here, we have carried out a more detailed study of variant spellings with a full manual analysis. We have also extended the previous study to encompass data from the Early Modern English period that is representative of an earlier phase in the process of standardisation of spelling (i.e. the Shakespeare corpus).

In the Shakespeare data, we observed a reduction in POS tagging accuracy from around 96% to 82%. We also showed that insertion of modern equivalent spellings alongside EmodE variants helps to improve the accuracy of POS-tagging such data. The ceiling for such an improvement is around 89% accuracy. The effect of

spelling variation was less marked in the Lampeter corpus samples of expository prose dating from the 1640s. However, there was still a significant reduction in accuracy to around 88.5% with a ceiling for improvement, if all spelling variants are detected, of 93.2%.

In future work, we intend to continue implementing improvements to the automatic variant spelling detection system (VARD). Crucially, we aim to incorporate contextual rules to detect EmodE variants that are otherwise undetected, since they appear in a modern lexicon (see, for example, 'then' for 'than' and vice versa). We also need to carry out similar experiments on the same data employing higher levels of annotation, e.g. semantic tagging. Our existing semantic tagger is a knowledge-based tool which exploits a large lexicon. As such, it is not as robust as the probabilistic approach used by CLAWS (see Piao et al., 2004). The semantic tagger can not guess the semantic field of a word from its immediate neighbours (using the probabilistic HMM technique) or other surface clues (such as the suffix rules employed in CLAWS).

In addition to improving the accuracy of annotation tools by pre-processing variant spellings, further key considerations are altering the existing taxonomies so that they can be applied in an historical context, by, for example: (i) changing the POS tagsets embedded within our POS taggers to reflect changes in grammar over time (Britto et al., 1999; Kytö and Voutilainen, 1995), and (ii) adapting the sense distinctions of semantic categories and hierarchical structures of modern semantic tagsets to accommodate changes in meaning over time (Archer et al, 2004).

Acknowledgements

The authors would like to thank Jeremy Bateman for his assistance is the manual post-editing of POS tags. The work described here was partly funded during a small research project funded by the British Academy (entitled "Scragg Revisited").

References

Archer, D. and Culpeper, J. (2003). Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics. In: A. Wilson, P. Rayson and A. M. McEnery (eds.) *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Peter Lang: Frankfurt/Main, 37-58.

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, Lancaster University, pp. 22 - 31.*

Archer, D., Rayson, P., Piao, S., McEnery, T. (2004). Comparing the UCREL Semantic Annotation Scheme with Lexicographical Taxonomies. In Williams G. and Vessier S. (eds.) *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004)*, Lorient, France, 6-10 July 2004. Université de Bretagne Sud. Volume III, pp. 817-827.

Biber, D., Finegan, E., and Atkinson, D. (1994). ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers. In Fries, U., Tottie, G. & Schneider, P. (eds) *Creating and Using English Language Corpora*. Amsterdam: Rodopi, pp. 1–14.

Britto, H. Galves, C., Ribeiro, I., Augusto, M., and Scher, A. (1999). Morphological Annotation System for Automatic Tagging of Electronic Textual Corpora: from

English to Romance Languages. *Proceedings of the 6th International Symposium of Social Communication*. Santiago, Cuba, pp.582-589.

Garside, R. and Smith, N. (1997). A Hybrid Grammatical Tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Longman, London, pp. 102 – 121.

Kytö, M. and Rissanen, M. (1993). General introduction. In Rissanen, M., Kytö, M., and Palander-Collin, M. (eds.) *Early English in the computer age: explorations through the Helsinki corpus*. Mouton de Gruyter, Berlin, pp. 1-17.

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

Leech, G. (1997) Introducing Corpus Annotation. In Garside, R., Leech, G., and McEnery, A. (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 1 – 18.

Leech, G. and Smith, N. (2000). *Manual to accompany The British National Corpus (Version 2) with Improved Word-class Tagging*. Accessed 22nd June 2007. http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2postag_manual.htm

Leech, G., Rayson, P., and Wilson, A. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman, London.

Mueller, M. (2005). The Nameless Shakespeare. *Working Papers from the First and Second Canadian Symposium on Text Analysis Research (CaSTA)*. Computing in the Humanities Working Papers (CHWP 34).

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*.

Osselton, N.E. (1984). Informal spelling systems in Early Modern English: 1500-1800. In N.F. Blake and C. Jones (eds.) *English Historical Linguistics: Studies in development*. The Centre for English Cultural Tradition and Language, University of Sheffield, pp. 123-137.

Piao, S. L., Rayson, P., Archer, D., McEnery, T. (2004). Evaluating lexical resources for a semantic tagger. In *proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 26-28 May 2004, Lisbon, Portugal, Volume II, pp. 499-502.

Rayson, P., Archer, D., Smith, N., (2005). VARD versus WORD: A comparison of the UCREL variant detector and modern spellcheckers on English historical corpora. In *Proceedings of Corpus Linguistics 2005*, Birmingham University, July 14-17.

Rayson, P., Archer, D., Baron, A. and Smith, N. (2007). Tagging historical corpora - the problem of spelling variation. In *proceedings of Digital Historical Corpora, Dagstuhl-Seminar 06491*, International Conference and Research Center for Computer Science, Schloss Dagstuhl, Wadern, Germany, December 3rd-8th 2006. http://drops.dagstuhl.de/opus/volltexte/2007/1055/

Santorini, B. (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.

Siemund, R. and Claridge, C. (1997). The Lampeter corpus of Early Modern English Tracts. *ICAME Journal*, 21, pp. 61-70.

Sinclair, J. (2004) (edited with R. Carter) *Trust the Text: Language, Corpus and Discourse.* London and New York: Routledge.