VosPos:
A project for Virtual Orthographic Standardization and Part of Speech Tagging of Early Modern English texts

Draft Report, November 20, 2006

By Martin Mueller

Introduction and Summary

VosPos sounds like something from a five-year plan in the later years of the Soviet Empire. But the ugly jingle draws attention to the close relationship of two intricately related procedures: virtual orthographic standardization and part-of-speech tagging. You cannot apply part-of-speech tagging to a text in which the 'same' word is spelled in different ways. 'Who wants part-of-speech tagging in the first place?' is a reasonable retort to that concern. But it is also the case that you cannot apply virtual orthographic standardization to many words unless you have part-of-speech tagging. And that takes us to the major goal of VosPos: you want to enable a modern reader to put in a word in its modern spelling and return all or most occurrences of that word, however it is spelled. You want to put in 'jealous' or 'jealousy' without having to worry about the fact that there are about fifty different spellings of the noun and adjective in the 12,000 texts of the Text Creation Partnership released so far:

> gelous | ialouse | ialousies | iealious | iealous | iealousie | iealousies | iealously | iealouslye | iealousnesse | iealousy | iealousye | iealousyes | iealovs | iealovsie | iealovsies | ielous | ielousie | ielousies | jalousie | jealosies | jealous | jealous' | jealouse | jealoused | jealouses | jealousest | jealousey | jealoushes | jealousi | jealousi' | jealousic | jealousie | jealousie' | jealousied | jealousies | jealousiy | jealousle | jealously jealousoe | jealousse | jealousses | jealousy | jealousy' | jealousye | jealousyes | jealousys | jealovs | jealovsie | jelousies | jelousy

VosPos wants to make Early Modern English texts 'tractable' for modern computers. From that perspective, part-of-speech tagging is the real goal. It opens large archives from earlier centuries to sophisticated forms of modern text mining. Virtual orthographic standardization is a prerequisite for achieving that goal and an interim stage in the process. From the perspective of the modern reader who wants to look up this or that, this interim goal may well be the most useful part of the project: it lets you query old texts as if they were modern, and you do not have to worry about regular expression, wild card searches, or other procedures to capture spelling variance about whose nature you are uncertain to begin with. But even for the reader who has no use for part-of-speech tagging itself, this final stage of the project carries the benefit of making virtual orthographic standardization significantly more accurate. With spellings like 'bee', 'doe', 'donne' or 'then' you have to know the part of speech before you can assign the standard spelling. It is of course true that many of these words are not likely to be the source of frequent look-ups.

In this report I talk about what I have done so far but also focus on what remains to be done. This has been one of those projects where you discover half-way that you should really have done it differently in the first place. But there did not seem to be a good way of figuring this out at the time.

The scope of the project

The project operates under the auspices of the Center for Library Initiatives of the CIC Universities and is being funded by a grant from Proquest and subscriptions from the CIC universities as well as a number of other universities. The project is being carried out at Northwestern University. I have been chief cook and bottle washer, with some student help, and very significant support from Academic Technologies.

The initial focus of the project was the 8,500 texts released by spring 2005 in the Text Creation Partnership project. These are texts published between 1470 and 1700. They are digital transcriptions of the microfilm page images now available through Early English Books Online (EEBO). Over the course of the project several Chadwyck-Healey archives to the TCP set, in particular English Poetry, Early English Prose Fiction, Eighteenth Century Fiction, and Nineteenth Century Fiction. There were several reasons for doing so, chiefly the fact that a collection of texts from the eighteenth and nineteenth centuries would provide a substantial word list in more or less standardized spelling. But adding texts as I went along has also been a source of confusion.

The current aggregate of texts includes approximately 650 million word occurrences, and there are about 2.5 million distinct spellings. How many of these spellings need to be mapped to standard spellings before you can say that the texts are fully searchable through standard spellings? Another way of asking the same question is to think of a book as a sequence of pages each of which has the same number of words randomly blotted out. If you think of an average page as having about 400 words would five randomly blotted words seriously interfere with its readability? It would be a rare page where five words make a crucial difference. Thus it seems fair to say that you can declare success if 99% of word occurrences are mapped to standard spellings. You can almost certainly get there by mapping less than half of the distinct spellings.

Standardization and lemmatization

VosPos is about standardization rather than modernization. A form like ‘louyth’ could be modernized to ‘loves’, but here it is standardized to ‘loveth’. The project assumes the survival in modern English of standard archaic forms. The King James Bible is a good example of a standardized archaic text. To put it differently, I have not modernized a spelling where the spelling points to genuine morphological difference. ‘Telle’ becomes ‘tell’, but ‘tellen’ stays ‘tellen’, and so forth.

Lemmatization is the process of mapping different word forms to the standard form in which the word appears in a dictionary. In modern English, nouns have a zero form and an –s form. Verbs have a zero form (love), an –s form (loves) an –ed form (loved), and an –ing form (loving). In Chaucer’s English you find an –en form, which by the time of Spenser has become and archaism, -st and -edest forms, which slowly disappear from the middle of the seventeenth century on, but survive in archaic and jocular usage, and an –eth form that is syntactically identical with the –s form. Adjectives have a –ly form, and short adjectives have –er and –est forms. When you lemmatize an English word, you group all these different forms under the zero form.

It is sometimes easier to lemmatize than to standardize. In modern English, the plural and genitive are orthographically distinguished: mother's vs. mothers. The apostrophe as a genitive marker is rare before 1700, and forms like 'mothers' cannot be orthographically disambiguated without prior part-of-speech tagging.

The assumption behind lemmatization is that most users for most purposes will use the lemma as a search term, just as they do with a dictionary. A "lemmatize search" is what most people associate with a dictionary look-up.

What is a spelling and what is it a spelling of?

I began this project with the naïve sense that most of the spellings represented English words and that the main task would be to map such English spellings to their standard form. I discovered very soon that this is not the case. There are lots of foreign words (mainly Latin). There are lots of names, and it is not always easy to assign them to a particular language. There are abbreviations that are not really words, and there are spellings that have something wrong with them, either because they were typographical errors to begin, were wrongly transcribed in the digitizing process, or were marked as partly illegible. Probably no more than half the spellings are spellings of English words.

Foreign words

About half a million or 20% of the spellings in the combined archives (TCP-CH) are foreign. There are close to 400,000 Latin words, and another 100,000 divided between French (~50,000) and a host of other languages, including Italian, Spanish, Dutch, German, Hebrew, Norse, and Greek. These spellings account for about 2% of word occurrences. The distribution of foreign words led to the decision to record Latin and French in categories of their own and lump all other foreign words together under a miscellaneous category.

English words are not always easy to tell apart from Latin or French words. If you standardize words that end in –ente or –entes by dropping the last 'e', you may turn standard Latin spellings into English words. You could ignore the overlap of English and foreign words since it affects less than 1% of word occurrences, but I have spent a considerable amount of effort trying to identify foreign words as accurately as I could without resorting to context. This is an error prone business, and there are many words that while clearly not English could be any combination of Latin, French, Italian, or Spanish. This is a particular problem with words ending in –o. These may be Latin datives or Italian or Spanish nominatives, and without context it is impossible to assign them confidently, unless you invented an 'extended Latin' category, which might not be a bad thing to do.

Filtering out foreign words has the potential benefit of identifying stretches written in other languages. This may be of benefit with the TCP texts, where some scholars may find it useful to study the range of Latin reference in theological or other texts.

Spellings of foreign words have not been standardized or lemmatized. It may be desirable to do this for Latin, and it would not be particularly difficult to do, since Latin spellings in earlier texts are for the most part highly standardized and there are programs for parsing Latin. But this is a task for a later time and can be done separately.

In the current phase the Scottish dialect has been treated as a form of English. This decision needs revisiting. It is not obvious whether Scottish texts or texts with a lot of Scottish dialects (e.g. Scott's novels) could or should be standardized. After all the spellings in Scott are themselves fairly standardized representations of dialect. Replacing 'gae' with 'go' is a form of translation rather than modernization or standardization.

Similar problems arise with Irish or with the representation of various dialects in American literature. The current lists are inconsistent in this regard.

Names

About a quarter million spellings represent names of one kind or another. In all periods of English, proper names are overwhelmingly capitalized, although in texts before the middle of the sixteenth century you will find a fair number of lower case names. On the other hand, before 1800, nouns and many other words are capitalized as well. Thus sentence-medial capitalization is not a good name indicator per se. On the other hand, if a spelling occurs more than four times and is always capitalized, it is probably a name. There are also a number of suffixes that are strong name indicators. Case frequencies and suffixes let you filter out names quite successfully.

The mapping of names to standard spellings is much more problematic. The adjectives 'greene' and 'browne' map to 'green' and 'brown', but the names 'Greene' and 'Browne' do not. In this release no effort has been made to map proper names to standard spellings. For historical and legendary characters this is clearly desirable, but this is a task for a later release.

In the current release, the original spelling of a name has been propagated to the modern column even though many spellings are manifestly not modern.

Greg Crane and the Perseus Project have had a significant interest in neo-Latin, and several of their initiatives are based on making it more accessible. There are several million words of Latin scattered through the texts in the TCP. It may be an interesting research project to do something with them. If so, the identification of potentially Latin words is a first step towards extracting Latin sections.

Abbreviations, errors, and word fragments

The TCP texts have thousands of abbreviations, typically in footnotes. In the English Drama archive speaker names are often abbreviated. So far, I have only tried to identify character strings that are not words. I have not resolved abbreviations. It would be a good practice (recommended strongly by people in biomedical text processing) to expand abbreviations and map them to the words they stand for. There appear to be about four million occurrences of them, which is four times the collected works of Shakespeare.

The TCP texts were transcribed from xerox copies of microfilms, and transcribers were instructed to mark words they could not read. There are half a million spellings that contain one or more characters marked as missing. In many cases it is possible to establish the missing character with complete certainty.

The TCP texts also have a lot of typographical errors. The most common mistake is the confusion of ‘f’ and long s. Spellings like ‘affign’ can be confidently resolved to ‘assign’. It is not possible to tell from the digital text whether the error is the printer’s or the transcriber’s. Errors of this kind probably run in the low ten thousands, and many can be corrected.

There are also a fair number of word fragments in the TCP texts—non-words that are clearly not abbreviations, but part of a word where either the printer or the transcriber made an error in word separation.

The CH texts appear to be a lot cleaner than the TCP texts. But the accumulated debris of various kinds in the TCP-CH archives makes it harder to estimate the percentage of words that have been mapped. As many as five million character strings (almost 1% of word occurrences) may represent unmappable non-words.

Making sense of the spellings

In this first phase of the project I have focused on finding out something about as many spellings as possible, even if the something does not amount to a standardized spelling. Bismarck once said that laws and sausages should never be watched in the making, and this is probably true as well of the very ‘heuristic’ procedures by which I arrived at my results. The best that can be said of them is that they cleared the way for proceeding more systematically in the second phase with a larger, completely fixed, and more systematically tokenized data set.

The English Poetry Databases includes Old and Middle English texts. I excluded these from my calculations and restricted the project to spellings attested since 1470 or the beginning of print culture in England. The result of these labors is a database with the following columns for each ‘word’:

1. The spelling
2. The standard word form if it is a possible English word
3. The lemma with which the word is associated

4. Total count in all collections
5. Count of upper-case spellings
6. Count of lower-case spellings
7. Count across 50-year ranges
8. Is it a possible English word?
9. Is it a possible Latin word?
10. Is it a possible French word?
11. Does it belong to some other language?
12. Is a proper name?
13. Is it an abbreviation
14. How long is the word
15. Does it contain non-ASCII characters
16. The reverse spelling of the word

The reverse spelling of a word is an extraordinarily simple and helpful tool for finding out patterns because words are much more clearly defined at the end and sorting words by reverse spelling reveals many regularities.

Working with these data in the environment of a Microsoft Access database, I have come up with these results:

1. ~350,000 words with some 500 million occurrences have been identified as English and mapped to modern spellings. About a third of them are mapped to themselves, but that provides some information because you need to know of a spelling whether it is standardized or not.
2. 31 spellings with 113 million occurrences have been mapped as punctuation
3. Some 1,300 spellings appear to be abbreviations and account for four million word occurrences.
4. About 320,000 spellings represent Latin words and account for 8 million word occurrences. The Latin corpus in these texts (mainly TCP) is eight times as large as the plays of Shakespeare.
5. Some 380,000 spellings represent names and have about 18 million occurrences. But 100,000 of the spellings are also classified as English or Latin words. If you filter out those out, there are some 280,000 name forms with some four million occurrences.

What about the unmapped spellings? There remain close to 1.5 million spellings about which at this point I know nothing. They add up to about 4.5 million word occurrences. To this you should add some 450,000 hyphenated forms that have close to two million occurrences.

There is some double and fuzzy counting in these figures, and I purposely give rounded figures. But the overall results are pretty clear. If you work your way through 2.5 million spellings in descending frequency, you probably reach the law of diminishing return with a million. The last 1.5 million spellings account for no more than 1% of word occurrences! This statistic is a little deceptive since so much of the fabric of a language is

made up of function words with low semantic content. Most of the content of a document is probably caught in only 20% of the word occurrences. If you map 99% of word occurrences, the missing one percent will probably consist largely of content words, it would be more accurate to say that you are missing five percent of the significant words rather than one percent of all words.

The spelling lists from the first phase will be used by Chadwyck-Healey in their LION database, and a somewhat earlier version has been used in an experimental implementation of the Philologic database. It is clear from those results that the current lists are quite useful already even if they contain many mistakes and omissions, which await correction in the next round.

The second phase of the project

In the second phase of the project, we will go about things in a much more systematic fashion. While using the spelling lists that resulted from the first phase, we will in other regards start from scratch, and we will not, as we did in the first round, add archives as we went along.

The second phase will have a fixed corpus, consisting of the latest TCP release, and Chadwyck Healey archives for English Poetry (EPD), English Drama (ED), Early English Prose fiction, Eighteenth Century Fiction (ECF), Nineteenth Century Fiction (NCF), and Literary Theory (LT). That is about a billion words.

Secondly, we will take a systematic approach to character issues. The Chadwyck-Healey texts were encoded before Unicode came along and use hundreds of entity references. The TCP texts are available in a UTF-8 version. We will A conversion process for the Chadwyck Healey is underway. We therefore expect to be able to work with text archives that use UTF-8 wherever possible and have a consistent set of entity references.

Thirdly, we will take a more nuanced approach to tokenization, and finally, we will develop a systematic procedure for dealing with hyphenated words.

Tokenization

Tokenization is a cause of many headaches, and the biggest headache is caused by the fact that the many uses of the point are hardly ever disambiguated in digital transcription. The eighteenth-century fiction archive is the one shining exception: sentences are explicitly encoded.

The point is the third most common ‘word’ after the comma and ‘the’: it occurs 31 million times and accounts for almost 2% of all word occurrences. Standard tokenizing routines treat all points the same way, and so did we. But it seems highly desirable to distinguish at least three major functions of the point:

1. Marking a sentence boundary
2. Marking an abbreviation
3. Marking a decimal point

Paul Schaffner has pointed out that there are other uses: it may surround Roman numerals. A single point may play the double role of marking an abbreviation and the end of a sentence. It is not clear whether points can be retroactively disambiguated with sufficient precision, but it would certainly be worth doing if you can at least 80% right. We will try to do so.

Abbreviations in word lists are almost certainly represented with greater clarity if the point at the end is included. And there are obvious benefits to defining sentence boundaries with greater accuracy.

Hyphenation

There are two salient features of hyphenated words. In a corpus of any size, about a third of the words are nonce words in the sense that they occur only once in that corpus. Nonce words probably make up an even larger percentage of hyphenated words. Secondly, hyphenation in nearly all cases combines words that are separately common. Thus it makes sense to approach the standardization of hyphenated spellings by looking at the separate components rather than the compound.

There are no firmly established conventions for representing hyphens in digital texts. Unicode has different code points for the different functions that one expresses on a keyboard by choosing the minus sign. The Chadwyck Healey and TCP texts differ in this regard. The Chadwyck-Healey texts may also be not entirely consistent with each other.

The best way around this is probably to treat every word as a compound of one or more components.

Segmentation by date and genre

The TCP and Chadwyck-Healey texts have elaborate bibliographical that can be used to segment the data by very broad categories of genre and date. With regard to tokenization, for instance, you expect a collection of play texts to have many abbreviations that refer to speakers and few abbreviations that refer to the books of the Bible. With theological texts it is just the opposite. There are likely to be some domain specific heuristics that can be employed with a tolerable degree of extra effort if they promise to deliver markedly greater accuracy.

The importance of disambiguating different uses of the point is very helpfully discussed by Udo Hahn and Joachim in "Levels of Natural Language Processing for Text Mining", *Text Mining for Biology and Biomedicine,* ed. Sophia Ananiadou and John M'Naught (Boston and London: Artech House, 2006), pp. 16-17. Their discussion focuses on the need to identify abbreviations and sentence boundaries—exactly the areas of concern for early modern texts.

Automatic Mapping of Spellings

In the second round we will need to have a much more algorithmic approach to the question of how to relate unknown spellings to standard spellings in an authority list. Phil Burns of Academic Technologies has been working on these issues. By my calculation, there are more than a million spellings that have not been mapped. It is a reasonable assumption that the great majority of them are variants that are some 'edit distances' away from known spellings. It would be a great success if a third could be captured by algorithms of one kind or another.

One step that we have not yet taken is to create a full form dictionary of mapped words, that is to say a dictionary of all potentially legal forms of a given lemma (act, acts, acted, acting, acteth, actest, actedest). It is likely that such a dictionary will not only match quite a few spellings but will also move others within a closer edit distance of existing spellings.

The Oxford English Dictionary is an obvious but problematical help in this enterprise. It actually contains far fewer spellings than are currently in my lists, not to speak of the larger lists resulting from tokenization in the second phase. And much in the information kept in was assembled at a time when rigid compliance with database procedure was unknown. Thus it is not easy to extract information from it in a consistent fashion. Nor is it possible to think of headwords as being always standard spellings in the way that a modern user of a VOSPOS interface would expect. On the other hand, there is promising evidence that more sophisticated forms of massaging the data will produce valuable mapping results.

The mapping of names is a very difficult problem. Many names do not map to a single form: Catherine, Catharine, Katharine, and Katherine are equally legal versions of the same name. For geographical, historical, and mythological names there is typically a standard spelling. We have not yet developed procedures for mapping variant names to standard forms. Many can probably be related to external dictionaries.

Ambiguous words

Some early modern spellings cannot be mapped to standardized forms without prior POS disambiguation. Spellings like 'doe', 'bee', or 'then' are good examples of this. This is, however, a more limited phenomenon than might appear at first sight. I have kept a list that so far runs in the low hundreds, and the complete list probably is not much more than 1000 spellings. Ambiguity within English is probably a lesser problem than ambiguity from spellings that are valid words in more than one language.

What deliverables and when?

The progress of the second phase of VOSPOS will depend on work that is currently going on in Academic Technologies at Northwestern and serves a variety of purposes. We are implementing a version of Philologic at Northwestern that will provide access to the set of TCP and Chadwyck-Healey databases in one search space to institutions that have the right of access to all of them. The availability of that Philologic implementation will

be critical to VOSPOS progress. Some checking of data in context is an important part of analysis in the next phase, and Philologic will be much the quickest tool for that.

Phil Burns has spent several months developing a work flow called MorphAdorner that will support virtual orthographic standardization and part of speech tagging. This is likely to be fully operational in early 2007, with documentation and source code available at that time.

When both Philologic and MorphAdorner are in place, work on Phase II of VosPos can start in a serious way. It will take these steps:

1. We will tokenize the entire corpus from scratch based on new tokenizing routines built in MorphAdorner. Tokenizing routines in modern NLP tool kits depend on rules and heuristic largely derived form modern texts. We will do some empirical survey of the corpus to identify practices that are not reflected in modern rules—such as surrounding Roman numerals with points.
2. We will develop a consistent way of dealing with hyphenated forms in different corpora.
3. We will extract from the mapped spellings so far rules or statistical tendencies that will help in identifying unknown words
4. We will provide virtual orthographic standardization and part of speech tagging using a tag set called NUPOS that is based on current best practices but accommodates historical forms and is very explicit in allowing the classification of parts of speech at different levels of granularity from the very coarse (23 tags) to the moderately fine (160 tags). This tag set is described in “NUPOS: a tag set for written English from Chaucer to the present day” (http://panini.northwestern.edu/mmueller/NUPOS.pdf).

Possibilities

Systematic virtual orthographic standardization presupposes a thorough review of spelling practices across different texts and periods. The procedures reveal many typographical or transcriptional errors that can in principle be retroactively applied to a text. For instance, in the 8,500 texts of the TCP there are 500,00 spellings with 2.5 million occurrences where something is missing in the word. There tens of thousands of instances, where you have spellings like “accordi.g” (the point is my symbol for the gap element used to mark uncertainty). In these cases there is no uncertainty: the missing letter can be supplied with complete certainty.

There are many cases where a long ‘s’ has been represented by an ‘f’, and it is similarly possible to correct the error, although in such cases it may not be possible to determine whether the error is the typesetter’s or the digital transcriber’s. But there are thousands of such cases.

Finally, it may be possible to add markers of sentence boundaries to many texts, perhaps in the form of a special <milestone type=”eos”/>. If this works in eight or nine of ten

cases, it is a lot better than nothing. And the tag can easily be ignored by those who do not trust it.