The Nameless Shakespeare

Martin Mueller
Northwestern University
martinmueller@northwestern.edu

Abstract
This article describes the origin, principles, practices, and potential of The Nameless Shakespeare electronic edition as of December 2004.

KEYWORDS: Electronic edition, text-analysis, tagging, data retrieval, interface.

1. The idea of the digital surrogate and its query potential

A play script is a poor substitute for a live performance. Plays are meant to be seen and heard in a live theatre rather than read.[1] But however paltry a surrogate the printed text may be, for some purposes it is superior to the “original” that it replaces. The printed copy has a query potential that can be explored through reading and opens our eyes to aspects of the play that would be harder or impossible to extract from a performance.

The advantages of the surrogate extend to different versions of a printed text. If I owned a First Folio, I would not throw away a modern edition of the text -- not merely because I would be afraid to damage a valuable object by too much thumbing of its pages but because the modern surrogate lets me do some things better and faster.

We are now in the midst of a new and major stage in the “allographic journey” of texts or their migration from one medium to another. The last major phase in that journey was the transition from manuscript to print. It was, by the standards of its day, a remarkably thorough and speedy migration. In 1450 there were no printed books. By the end of the sixteenth century, an entire cultural heritage had been transcribed. To be read meant for the most part to be read as a book.

The allographic journey from print to a digital medium has been an even more rapid process. There are transcription projects, like ARTFL or the TLG, that go back to the 1960’s. The process did not gather speed until the eighties, but since then it has accelerated at a phenomenal rate. We may predict that within ten or certainly twenty years the parts of our cultural heritage that are actively kept will be in digital form. Printed books will certainly not go away, but for a document to be in circulation will mean for it to be digitally available in some form. Or to put it differently, for a document not to be accessible digitally will be a handicap in its competition for attention.

The digital surrogate still is, and is likely to remain for quite a while, a very poor substitute for a book. It is a very poor tool for reading, but it is an excellent tool for finding stuff in a text or across a body of texts. Print culture over the centuries developed its own finding tools, but these are generally inferior to digital finding tools, and the larger the text, the greater the advantage of the digital surrogate.

In planning an electronic edition of a document originating in print culture it is essential to maximize the query potential of the digital surrogate and to think of it as a tool that lets new kinds of readers pursue questions that it would be very difficult or impossible to answer with the help of a printed text. Maximizing that query potential has been the chief goal of The Nameless Shakespeare, which I describe in the following paragraphs.

The Nameless Shakespeare is part of a larger project called WordHoard and funded by the Mellon Foundation. The WordHoard Project applies to highly canonical literary texts the insights and techniques of corpus linguistics, that is to say, the empirical and computer-assisted study of large bodies of written texts or transcribed speech. In the WordHoard environment, such texts are annotated or tagged by morphological, lexical, semantic, prosodic, and narratological criteria. They are mediated through a "digital page" or user interface that lets scholarly but non-technical users explore the greatly increased query potential of textual data kept in such a form.

It is a basic assumption of WordHoard that new kinds of historical, literary, or broadly cultural analysis will be supported through the forms of data access that are made possible when literary texts are treated in the manner of linguistic corpora. Deeply tagged corpora of course support more finely grained inquiries at a verbal or stylistic level. But more importantly, access to the words of a text at such microscopic levels also lets you look in new ways at the imaginative worlds created by those words.

WordHoard consists of the following components:

- A set of guidelines for tagging text corpora to make them suitable for processing in the WordHoard environment

- The Nameless Shakespeare
- The WordHoard software distribution that includes:
    - A relational data model and associated software for ingesting WordHoard documents
    - A Web-based “digital page” that mediates the most commonly sought information through a standard browser
    - A query tool that lets users explore the full query potential of a deeply tagged text archive
    - A collaborative environment to support team-based inquiries into particular archives
    - A WordHoard instance running at Northwestern University providing free and universal access to three text archives, including Early Greek epic, Chaucer, and Shakespeare.

WordHoard is a project of Academic Technologies at Northwestern University in collaboration with the Northwestern University Library and the Center for Computer Research on Language at Lancaster University (UCREL). The project leaders are Martin Mueller (Department of English and Classics) and Bill Parod (Academic Technologies).

2. How the Nameless Shakespeare was made

The Nameless Shakespeare bears its current name because we have not been able to think of a better one. It is at a minimum a good enough modern spelling edition to serve as the basis for the features the digital surrogate will contain beyond an accurate text. We (a group of scholars and programmers at Northwestern, working with people at The Perseus Project and at Lancaster University) built The Nameless Shakespeare on the Perseus transcription of the Globe Shakespeare, the standard text of the late nineteenth century and a major stage in a textual tradition that still survives in such modern texts as the Riverside Shakespeare, Bevington’s edition, or the Arden Shakespeare. These texts are in virtually complete agreement with each other and with the Globe Shakespeare on the copy text used for each play, and, if one stands a few feet away from the passionately contested minutiae of Shakespearean editing, they do not differ a great deal in their treatment of cruxes or choice of variants.

We used various techniques of digital collation to identify textual differences between the Perseus transcription of the Globe Shakespeare and other digital Shakespeare texts. These procedures generated a list of some 50,000 textual differences. After filtering out differences that were merely orthographic, there remained about 5,000 locations where a textual variance could be said to have some lexical, morphological, or prosodic implication.[2] Dealing with these variants was my work, and in carrying it out I was greatly helped by the excellent transcriptions of the quarto and folio texts in the Internet Shakespeare Editions. It so happened that the Perseus transcribers of the Globe text had added the TLN numbers used by Hinman to keep track of the lines in the folio. Thus it took only seconds to go from any passage in the Globe to the corresponding passage in the Internet Shakespeare transcriptions, which used the TLN numbers not only to identify folio lines but also the corresponding passages in the quarto texts.

I had some very simple decision rules to resolve variants. If a reputable modern editor had kept a reading in the copy text (whether folio or quarto), I would choose that reading. If the editors of the Arden, Riverside, and Bevington texts all departed from the copy text, I would go with the choice of the majority.[3] If they disagreed among themselves, I would use my own judgment. The result is an eclectic and not especially principled text, but it can lay some claim to being a text that faithfully preserves the lexical, morphological, and prosodic habits of the copy text. It differs non-trivially from the Globe Shakespeare, whose Victorian editors thought, perhaps rightly, that the quarto and folio texts needed some editing with regard to grammar, foreign words, and prosody.

Internal orthographic variants, especially with regard to word boundaries and hyphenation, were standardized. The resultant text makes no claim to advance the state of knowledge about editing Shakespeare. It merely claims to be a good enough basis for the addition of several layers of morphosyntactic, semantic, and narratological tagging. The added value of The Nameless Shakespeare consists in the fact that every word occurrence is annotated in the manner of a linguistic corpus.

The morphosyntactic layer of annotation is based on the CLAWS tag set developed at Lancaster University and used for the tagging of the British National Corpus. CLAWS uses algorithms that look at two words at a time and assign a given word to one of about 150 categories based on a probability table. Developed for modern prose, CLAWS is about 97% accurate. For Shakespeare, the automatic output had an error rate on the order of 5%. Through various rounds of error checking, we have reduced this rate to somewhere between 0.5% and 0.7%.

We adjusted the CLAWS tag set to account for the second person singular past and present of verbs, and we added some additional distinctions obscured by the CLAWS tagger. Take a phrase like “my loving lord.” Here ‘loving’ is morphologically a present participle, but it functions as an adjective. The CLAWS tag (AJ0) captures the syntactic function but throws away the morphological information. We introduced tags like VVG-AJ and VVG-NN, which will let users look for participial forms and distinguish between verbal, adjectival, and nominal or gerundial uses.[4] The text is also lemmatized; that is to say every word occurrence (‘loves’, ‘loving’) is referred to a headword or dictionary entry form. Lemmatization distinguishes between ‘love’ as a verb and as a noun.

For the semantic tagging of the text, we used the USAS tag set also developed at Lancaster University. Like CLAWS, USAS is a tool designed to deal with modern English. It assigns every word in a text to one of about 250 categories, which are subsets of 21 main categories. Dawn Archer and Paul Rayson at Lancaster University demonstrated that it can be used quite effectively for seventeenth-news books. On the basis of their success, a group of students at Northwestern under the direction of Dawn Archer worked for two months in the summer of 2004 supplementing and correcting the output of an initial semantic tagging run of the Shakespeare corpus.

The narratological tagging distinguishes between speakers, prose and verse, as well as different kinds of verse (rhyme, couplets, stanza).

A TEI-conformant version of The Nameless Shakespeare, including morphosyntactic and narratological tagging, is freely available for non-commercial purposes.url A version that includes semantic tagging will be released in the summer of 2005.

For purposes of data retrieval the various levels of tagging have also been kept in a relational database environment, which allows for quite granular queries based on the intersection of different tagging categories. As a digital surrogate, The Nameless Shakespeare may be said to “know” each of the followings things, separately or in combination:

1. The precise location of a word occurrence by act, scene, line, position in line, and the canonical citation form.
2. The word form or spelling resident at that location
3. The lemma or lexical item to which the word form

belongs
4. The word type of the lemma by broad categories of noun, adjective, verb, etc.
5. The grammatical state of the word, such as “plural noun”, “present participle”
6. The number of its position in the line
7. Whether it is verse or prose
8. Whether it rhymes or not
9. The identity of the speaker
10. The date range associated with the work in which the word occurs

Implicit in the knowledge about word occurrences is various kinds of frequency information about the count of word forms, lemmas, or word types in individual works, groups of works bundled by chronology or genre, or the entire corpus.

What can a user do with this knowledge? In answering this question consider the difference between “looking up” and “looking for.” Looking up is a familiar activity: you look up a word in a dictionary, a concordance, or an index, and you are given a definition or a set of citations. The act of looking up may be complicated by secondary constraints, such as one word near another (“love” near “death”) or passages in a subset of a corpus defined by date or genre. But looking up always starts from a spelling, even though truncation or wildcard searches may extend the range of spellings that are looked up in a single pass.

While in “looking up” the point of departure is always a spelling, and the result set is always a set of locations, in “looking for” the point of departure may be any combination of criteria, and the result set is a list of previously unknown words that meet those criteria.

The Nameless Shakespeare supports “looking up” queries (although some other electronic Shakespeares do a better job of that), but its real strength lies in the support it provides for finding unknown words that meet specified criteria, such as

1. words found only in Hamlet
2. words found for the first time in Hamlet
3. adjectives used by Ophelia in verse
4. words used in Hamlet and occurring also in another specified play but never or only rarely in the rest of the

corpus

Such lists are very time-consuming to compile from printed sources. They can be generated in seconds through the interface of The Nameless Shakespeare, and they provide very useful starting or supporting evidence for all manner of hypotheses.

It is important to keep in mind the cost of the enabling assumption behind all the data tables. Words are no longer treated as an ordered sequence, but as tokens in a bag to be counted and compared in various ways. Some inquiries are greatly aided by this brutal application of "divide and conquer." Others are not.

3. The interface of the Nameless Shakespeare

The Nameless Shakespeare is currently accessible through an aging interfaceurl. In the context of the WordHoard project we are designing a more sophisticated and generic interface that will support the display and querying of various texts tagged in the manner of The Nameless Shakespeare.

A key feature of the WordHoard interface will be a digital page that puts basic information about a text at hand by using the information that the text may be said to know about itself. The digital page will support the side-by-display of arbitrarily chosen passages. You can do this if you have two copies of a book, but it is a slow business, even assuming that you can get both pages to lie flat and still. In WordHoard, a search will generate a list of words, from which one will normally generate some KWIC display. From the KWIC display one can in turn choose displays of wider contexts and direct them into the left or right window with the hit lines lined up in parallel. Having the lines to be compared within a single field of vision will be an eye-opener as well as a time-saver.

WordHoard will also have a built-in feature that adjusts the reporting of a result set to its size. If a result set is sufficiently small, going directly to a KWIC report will be most helpful to users. If the result set is larger, it is more helpful to add an interim step of a summary report that aggregates data by some set of criteria and makes it easier for users to focus on the subset for which they want more detailed results.

For the texts of the Nameless Shakespeare, the WordHoard interface will also add a feature that puts the transcription of the original folio and quarto texts in the Internet Shakespeare immediately at the disposal of the reader. Clicking on any line will take you to the relevant folio or quarto page with the hit line visually marked. This will be a valuable pedagogical tool since even for the novice reader it will provide instant and easy access to the original texts. A pilot version of this feature is already in operation in the current interface.

Frequency is a fundamental property of a word. In listening to conversation we pick up instantly on whether a word is common or rare. Computers are very good at keeping count, and a tagged corpus in a digital environment can easily deliver information about the relative frequency of its parts. An important goal of the WordHoard environment will be to make frequency information accessible to users who are not statistical experts, but might be interested in using various procedures to follow up hunches and test hypotheses.

Finally, the WordHoard environment is also likely to go beyond the "tokens in a bag" approach towards the corpus. In the XML representation of the data, morphosyntactic information is encoded as a part-of-speech attribute of the word element that encloses each separate word, as in <w pos="DPS>my</w> <w pos="VVG-AJ">loving</w> <w pos="NN1">lord</w>. A search tool like XAIRA can look for strings of attributes as well as for words and can retrieve all instances of an adjective followed by a noun. We expect to integrate it or a tool like it into our interface.

The WordHoard interface will consist of one or more Java applications that users will download and through which they will query the information on the server. Pieces of the interface will be released as they become sufficiently stable. A first version of the digital page is likely to be available by the spring of 2005. It remains to be seen whether the final project will consist of one tightly integrated package or of a set of closely related, but independent modules.

Notes

[1] This paper is a revised version of a talk delivered at the CASTA meeting at the University of Victoria in November 2003. It describes the state of The Nameless Shakespeare and the WordHoard project as of December 2004.

[2] I know it is contentious to claim that a variant is "merely" orthographic. The distribution of such merely orthographic variants may tell you about the typesetter and be powerful evidence for inquiries into the genesis of a text or its relationship to other texts. Such inquiries lie outside the research scope of the Nameless Shakespeare, at least for the moment. It is a tempting idea to associate the modern spelling in each word occurrence with the spellings in the quarto and folio texts, thus putting the spelling variance of the original documents into a database environment where it can be easily queried. The excellent transcriptions of The Internet Shakespeare would be a good source for doing something like this, but I have not had the time or patience to do this within the confines of the current project.

[3] Dawn Archer, Tony McEnery, Paul Rayson, Andrew Hardie (2003). Developing an automated semantic analysis system for Early Modern English. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16. UCREL, Lancaster University, pp. 22 - 31. Downloadable from <http:www.comp.lancs.ac.uk/computing/users/paul/public.html>.

[4] To be more precise, the automatic tagging was done with the latest CLAWS set (C7), which has about 150 categories. This was stepped down to the 60 tags of the C5 set, and our modified C5 set has about 100 different tags.

Works Cited

Archer, Dawn, Tony McEnery, Paul Rayson, Andrew Hardie (2003). “Developing an automated semantic analysis system for Early Modern English.” *Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University. Eds. Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. pp. 22 - 31. 6 Dec. 2004 <http:www.comp.lancs.ac.uk/computing/users/paul/public.html>.

Arden Shakespeare CD-ROM: Texts and Sources for Shakespeare Studies. (1997). Thomas Nelson.

The ARTFL Project. 6 Dec. 2004 <http://humanities.uchicago.edu/orgs/ARTFL/>.

CLAWS. 6 Dec. 2004 <http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/>.

The Complete Works of Shakespeare. (1997). Ed. David Bevington. New York: Longman.

Internet Shakespeare Editions (ISE). 6 Dec. 2004 <http://ise.uvic.ca/>.

The Nameless Shakespeare. 6 Dec. 2004 <http://www.library.northwestern.edu/shakespeare/>.

The Riverside Shakespeare. (1974). Ed. G. Blakemore Evans. Boston, MA: Houghton Mifflin.

Thesaurus Linguae Graecae (TLG). 6 Dec. 2004 <http://www.tlg.uci.edu/>.

The WordHoard Project. 6 Dec. 2004 <http://www.comp.lancs.ac.uk/computing/research/ucrel/projects.html#wordhoard>.