The Lemma and Database Design: Redesigning Representative Poetry Online, Lemmatizing Lexicons of Early Modern English, and Envisioning the Lemmatic Web

Scholarly and Research Communication
VOLUME 3 / ISSUE 2 / 2012

Marc R. Plamondon
Nipissing University

Abstract

This article argues for the usefulness of the lemma as the base element for constructing large databases of texts for digital textual analysis and for providing a new hypertextual reading experience. Support for this is based on the author's experience designing and developing two major Web initiatives: Representative Poetry Online and the Lexicons of Early Modern English. The basic database features of the two websites are delineated, but the latter website, in particular, is described with a view toward showing the importance of a shift away from envisioning the database as constructed upon word entries to one constructed upon lemmata.

Keywords

Database; Lemma; Lemmatic web; Text analysis; Language-based systems; Reading environments; Lexicons.

Marc R. Plamondon is a Professor in the Department of English at Nipissing University, 100 College Drive, North Bay, ON, Canada P1B 8L7. Email: marc.r.plamondon@gmail.com .

CCSP Press
Scholarly and Research Communication
Volume 3, Issue 2, Article ID 020123, 7 pages
Journal URL: www.src-online.ca
Received August 17, 2011, Accepted November 15, 2011, Published August 15, 2012

Plamondon, Marc R. (2012). The Lemma and Database Design: Redesigning Representative Poetry Online, Lemmatizing Lexicons of Early Modern English, and Envisioning the Lemmatic Web. *Scholarly and Research Communication, 3*(2): 020123, 7 pp.

The INKE Research Group comprises over 35 researchers (and their research assistants and postdoctoral fellows) at more than 20 universities in Canada, England, the United States, and Ireland, and across 20 partners in the public and private sectors. INKE is a large-scale, long-term, interdisciplinary project to study the future of books and reading, supported by the Social Sciences and Humanities Research Council of Canada as well as contributions from participating universities and partners, and bringing together activities associated with book history and textual scholarship; user experience studies; interface design; and prototyping of digital reading environments.

The *Lexicons of Early Modern English* (*LEME*) website (2012) was first published by the University of Toronto Libraries and the University of Toronto Press on April 12, 2006, but has its origins in the digitization and encoding of lexicons, which began in the late 1980s. My own involvement with *LEME* began in 2002, and I witnessed and had a significant part to play in the great changes the database of lexicons underwent. As a whole, the evolution of *LEME* reveals a move away from centering upon the dictionary-type word entry to an increasing reliance on the lemma as the basic unit within the large database. The lemma, the canonical form of a word or phrase used to represent the word or phrase and all its inflections (much like the headwords in a dictionary or encyclopedia), is a unit not usually implemented as the basic unit within the design of a database of texts. Anecdotal evidence from the challenges and changes to the *LEME* and *RPO* websites reveals the potential for altering the hypertextual reading experience by having all texts built upon a database of lemmata.

The *Representative Poetry Online* (*RPO*) website (Lancashire, 2011) underwent a major redesign between 2000 and 2003. The website was originally a collection of static HTML files put together by the editor Ian Lancashire, reflecting his love of poetry and honouring the work of his former colleagues in the English Department of the University of Toronto, who had edited and published the original print editions of *Representative Poetry*. Hosted by the University of Toronto Libraries under the immediate supervision of digital librarian Sian Meikle, after the website had grown past its original, more modest beginnings, the editor, foreseeing more growth, decided that the static webpages were becoming unwieldy and a better approach to developing the website was needed. I was hired to help automate, at the very least, the construction of the various indices to the poems. The result was a dynamic website built upon a relational database wherein all the poetic data and metadata are stored. Poems and associated biographical and bibliographical information can now be entered through a Web interface, using strict and not-so strict templates; the indices are generated automatically and a last-line index was added to the website; a keyword search feature was added, allowing user searches through the poetic content for the first time; and poems were served out on an on-demand basis.

This new model, the proud child of my first attempt at designing a relational database and co-ordinating server-side and client-side Web programming, worked very well – for a few months. On January 24, 2003, America Online (AOL) highlighted *RPO* as an interesting website, probably for the upcoming Valentine's Day. As a result, the webserver, on which *RPO* and other University of Toronto websites were housed, received thousands of hits per second, more than it could handle, and AOL users brought down the University of Toronto Libraries' main webserver. During the four hours that followed, the

Plamondon, Marc R. (2012). The Lemma and Database Design: Redesigning Representative Poetry Online, Lemmatizing Lexicons of Early Modern English, and Envisioning the Lemmatic Web. *Scholarly and Research Communication, 3*(2): 020123, 7 pp.

RPO website was temporarily disabled, the webserver was brought back online, and Sian Meikle and I worked frantically to convert the *RPO* website from one where poem and poet pages are generated on demand to one where the pages are statically available at all times. (At first all poems were delivered with a single ColdFusion script; afterwards, they were all contained within their own HTML file.)

One of the probable reasons that the webserver could not handle the increased load – the current webserver for *RPO* would probably have had no trouble on that fateful day – was the choice of basic unit for the division of the poems into the database. Had the poems been saved as singular entities in free-form text fields within the database, they could probably have been extracted and served out with the accompanying apparatus with less difficulty. The central decision when first planning the *RPO* database was the choice of the basic poetic unit for storing the poems. There were four choices: the poem as a whole, the stanza, the line, and the word. I rejected the first two quickly: having the poem as a whole as a unit of data offered little advantage over the original, static *RPO* website. While I am suspicious that many if not most poetry websites today still retain the poem as a whole as the base unit of data, I believed that more could be done with using a smaller unit. The choice of the stanza did not seem like a substantial enough improvement over that of the poem as a whole.

Choosing the poetic line as the base unit of data seemed to me and my supervisor, Alan Darnell, like the best option. I was unfamiliar with the Text Encoding Initiative (TEI) and its guidelines at the time, but have since found out that TEI supports my decision, though its statement is from the perspective of textual markup: "The fundamental unit of a verse text is the verse line" (Text Encoding Initiative Consortium, 2008, p. 132). As a structural unit of poetry, the line corresponds well to the grid framework typically used to visualize relational database tables. The final possibility, the word as base unit, was more intriguing. It was rejected on the assumption that it would lead to an overly complex and consequently impractical database and accompanying set of programming files. While the decision to reject this option was no doubt the correct decision at the time, to this day I wish I could have pursued this approach to poetic structure, though it probably would have caused a server crash much earlier than the few weeks before Valentine's Day.

Using the poetic line as the base unit of data in a relational database, however, led me to rather easily create what became the first Web poetry search feature that could display results in the keyword-in-context (concordance-style) format. It also allows for some as yet undeveloped features, such as the identification and categorization of poems based on their stanzaic structure and line lengths. And because the indentation of lines is not stored as an integral part of the line but as a characteristic of the line, the varying styles and patterns of indentation can be easily factored into or omitted from poetic line and stanza analysis. The choice of poetic line as base unit seemed like the best choice for the possibility of developing tools for the textual analysis of the corpus of poems. A database of poetry where the basic unit of data is the word, however, would have allowed for a greater depth of poetic analysis. Not only would it make calculating the frequency of the occurrences of words much easier, but it would also have allowed for the calculation of the probability of the occurrences of words in line segments: i.e, whether a word is more likely to occur in the head, middle,

or tail position of the line. When the database of words includes syllabic divisions, pronunciations, and lemmata, the possibilities expand for a greater degree of true poetic analysis: that is, of rhythm, metre, rhyme, and sound structure. The project did not call for such things at the time, and resources, both of finances and hardware, made the decision to construct the poetry database upon the individual words impractical.

When I began working on the *Lexicons of Early Modern English*, a similar question became the first issue to tackle: what is the basic unit of the lexicons for their storage in a relational database? The answer did not present itself as needing much contemplation; in fact, the editor, Ian Lancashire, suggested that the basic unit of the lexicons upon which we were to build analysis tools is the word entry. This also corresponds well with TEI encoding guidelines, which define the elements <entry> and <entryFree> as the basic building blocks for the markup of dictionaries (Text Encoding Initiative Consortium, 2008, p. 255). The three people responsible for the general functioning of the website and its associated data – Ian Lancashire, Sian Meikle, and myself – had no trouble envisioning a website and database centred on the word entry.

The requirements of *LEME* in preparing period dictionaries, word lists, glossaries, and other lexical texts for the database or even for markup, however, dictate a rejection of a traditional notion of dictionary entry structure. Lancashire defends the decision not to adopt fully TEI encoding guidelines for lexicons: "TEI guidelines for encoding modern dictionaries do not well serve the experimental structures employed in early lexicons." Among other reasons, Lancashire points to the non-traditional (for modern dictionaries) relationship between the headword and the explanation: "The post-lemmatic segment of most Early Modern English dictionaries seldom held definitions as we know them" (Lancashire, 2006, p. 46). Lancashire argues that early modern English dictionaries are best understood as not constructed on the headword-definition model ubiquitous in modern dictionaries and the basis for TEI encoding guidelines: "Principal *LEME* elements are the word-group (for example, alphabetical or topical headings), the word-entry, and its two nested subelements, the 'form' and the 'explanation'. The encoding suggests a bilingual dictionary. *LEME* form and explanation are not headword and definition, as they would be today, but two equivalent units" (Lancashire, 2003, p. 13). As such, *LEME* gives greater weight to the lexical information contained within the explanation part, or the post-lemmatic segment, of a word entry than might otherwise be expected. Thus, headword entries, while searchable on their own, are not the primary way into the lexical information contained in the database. The headwords themselves are sometimes not found in the lexicons as regularized lemmata, but occur in inflected forms; they also occur with non-standardized spellings. Headwords sometimes also occur in the post-lemmatic segments (or what would traditionally be identified as definitions). Thus the importance of the headwords, as they occur in the texts, is diminished and the words that receive editorial lemmatization, those that are actively being described, defined, or highlighted in some way, whether occurring in the headword segment or in the explanation segment, possess greater significance for the reader and thus for the database.

All lexicons in the *LEME* database are encoded, but the encoding does not adhere strictly to either TEI or XML guidelines. The XML-like encoding is designed to minimize the effort in preparing the lexicons and to maximize the information that

Plamondon, Marc R. (2012). The Lemma and Database Design: Redesigning Representative Poetry Online, Lemmatizing Lexicons of Early Modern English, and Envisioning the Lemmatic Web. *Scholarly and Research Communication, 3*(2): 020123, 7 pp.

can be extracted from them, with a view towards saving the information within the relational database. The word entry table with the associated word entry form table and word entry explanation table originally constituted the central tables within the database: these are the tables that contain the text that we assumed would be the principal target of keyword searches. As of 2009, there are 112 fully-analyzed lexicons in the database, containing over ten thousand pages of original text and 354,921 word entries. With the encoded lexicons already "processed," with their word entries dissected and divided into the appropriate database tables, the information needed by a user's search can quickly be retrieved and the data delivered. As we realized the importance of the lemmata, the lemma database table gained in importance; the links between the word entry tables and the lemma table contain much of the power of the website as it currently exists. The lemmata can be searched separately (identified on the website as the "Modern headwords search," currently only available to subscribers of *LEME*) and as part of the general keyword searches. The database currently contains 264,122 distinct lemmata as identified in the 112 fully-analyzed lexicons. With the number of distinct indexed spellings contained within these lexicons at 404,898, the ratio of distinct lemmata to distinct spellings is greater than 65%, representing a significant effort in modernizing and regularizing the vocabulary put to use by the lexicons.

The lemmata, editorially identified directly into the markup of the lexicons, are arguably the basis upon which the greater part of the value of the website rests. Originally conceived as being important for looking up significant words through a standardization of spellings and word forms, the lemmata became a way to link word entries from different lexicons when they contain significant lexical information in common. For example, the word *hamlet* currently appears as a lemma for six word entries; a cursory examination of the six reveals that John Cowell defined it first in 1607 and that Henry Cockeram's definition from 1623 is identical to John Bullokar's definition of 1616. The word *love* as a noun occurs as a lemma ten times in a form position and four times in an explanation position; as a verb, *love* occurs as a lemma nine times in a form and five times in an explanation. Needless to say, the word occurs within the texts with different spellings. Counts of the number of times a particular lemma occurs as a significant component of a word entry within the period can be compiled, offering insight into not just the frequency of these words but also into the lexicographers' attitudes to certain words over others. Such counts can reveal, for example, whether Latinate or Anglo-Saxon words occur more frequently as lemmata in the word entries, and whether one type or the other appears more frequently in the form or the explanation positions. The lemmatic links between word entries can also be used to trace borrowings (or stealings) of word entries by one lexicographer of another and to trace the evolution of word definitions. These last are among the approaches Lancashire (2003) has undertaken in his research into early modern lexicography, using the website.

The website database and associated textual markup of the lexicons remain centred on the word entry as a unit: the word entry remains a practical unit for marking, dividing, and even defining the lexicons, and there are no plans to replace it in its importance. But as the value of having the database centred on lemmata rather than word entries became more apparent, Lancashire decided to undertake the lemmatization of full word entries of select lexicons. The relational database container for the lexicons allowed us to pursue this task, which would have been more difficult with marked-up texts:

"Lexical-analysis needs, especially, favour database technology. For example, *every word* in a database word-entry can be lemmatized for retrieval in a standard form" (Lancashire, 2006, p. 52). Lemmatizing the texts, whether in part or in full, requires a great deal of time, concentration, and consistency. Fully understanding this, the editor asked me to produce one or more tools to aid in the task. The result is a semi-automation of the lemmatization process, where words are extracted, transformed as needed, and compared to previous words and lemmata. The pre-existing lemma data, that which had been encoded for the main headwords and important terms in the word entries, are a major source of information that aid in the lemmatization process. As the lemmatization work progresses, tables of correspondences between early modern spellings, modernized spellings, and lemmata are expanded, which are then used for subsequent work. This iterative process means that the work gets progressively easier as more and more lemmatization work is performed; a great proportion of words are automatically lemmatized with little or no input from the lemmatizer (a tool that produces lemma forms of words, for analysis). To date, the process does not use a semantic parser, mainly because a regularization of the spellings was never a first step in the process but an integral part of the lemmatization and because of the great amount of abbreviations and foreign words in the lexicons. However, a semantic parser could make the process of lemmatization much less arduous.

In my work for *LEME* and for my own research into poetic phonology, I have found an increasing dependence upon a database of word forms (spellings) and an increased desire for a database of lemmata and their associated lexemes and a convenient (and accurate) methodology for converting any given written text into a format that can interface with such a database. This is indeed what I had naively envisioned in the spring of 2001 when presented with the task of designing a database for *RPO*. Had I followed my first instinct then, I would have developed not just one of the most useful database-based websites of English poetry, but also one of the most – if not the most – useful databases of poetry for advanced textual analysis. I would have also, possibly, bankrupted the various sources of funding for the project, including Ian Lancashire's research grants.

Reconstructing websites such as *RPO* upon a database of lemmata will open up new possibilities for reading experiences. A more seamless integration with available digital tools and reference sources, such as period dictionaries, modern dictionaries, pronunciation dictionaries, and encyclopedias, will be possible. Linking between texts and external resources is mostly contingent on the designer's whims and ability to foresee the needs of the user. Reading fully lemmatized texts in a hypertextual environment will allow for greater control on the user's part of the selection and activation of tools and resources while reading. Linguistic and literary researchers, and perhaps even historians, sociologists, and psychologists, will have a wealth of raw data to draw upon and analyze. Ideally, a standardized database of lemmata will be available to all, allowing texts and tools to be prepared and developed using these standards. Development in this direction could revolutionize the World Wide Web: where the Semantic Web envisions units of information as virtual objects to be used and juggled, an equivalent Lemmatic Web would be based on the lemma as an object, upon which all texts and all aspects of verbal communication are built.

Plamondon, Marc R. (2012). The Lemma and Database Design: Redesigning Representative Poetry Online, Lemmatizing Lexicons of Early Modern English, and Envisioning the Lemmatic Web. *Scholarly and Research Communication, 3*(2): 020123, 7 pp.

References

Lancashire, Ian. (2003). The lexicons of early modern English. *Computing in the humanities working papers*. URL: http://www.chass.utoronto.ca/epc/chwp/CHC2003/Lancashire2.htm [September, 2011].

Lancashire, Ian. (2006). Computing the lexicons of early modern English. In A. Renouf & A. Kehoe (Eds.), *The changing face of corpus linguistics* (pp. 45-62). New York, NY: Rodopi.

Lancashire, Ian. (2011). *Representative poetry online*. URL: http://rpo.library.utoronto.ca [September, 2011].

Lancashire, Ian. (2012). *Lexicons of early modern English*. URL: leme.library.utoronto.ca [September, 2011].

Text Encoding Initiative Consortium. (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. L. Burnard & S. Bauman (Eds.) URL: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf [September, 2011].