Renaissance and Reformation
Renaissance et Réforme

Lexicons of Early Modern English (LEME). Database

Heather Froehlich

Volume 42, Number 2, Spring 2019

URI: https://id.erudit.org/iderudit/1065130ar
DOI: https://doi.org/10.7202/1065130ar

See table of contents

Publisher(s)

Iter Press

ISSN

0034-429X (print)
2293-7374 (digital)

Explore this journal

Cite this article

Froehlich, H. (2019). Lexicons of Early Modern English (LEME). Database. *Renaissance and Reformation / Renaissance et Réforme*, *42*(2), 167–171. https://doi.org/10.7202/1065130ar

the usability of the site, some users such as I would like to know more about the project. The history of attempts to catalogue incunabula, for example, is an interesting one that reaches back into the eighteenth century. The history of the *ISTC* itself is intriguing as well. Lotte Hellinga's "Ten Years of the Incunabula Short-Title Catalogue (*ISTC*)" in *Bulletin du bibliophile* 1 (1990), 125–32, starts the story, and that story will clearly continue into the foreseeable future.

ROBERT E. BJORK
Arizona State University

Lancashire, Ian, gen. ed.
***Lexicons of Early Modern English* (*LEME*). Database.**
Toronto: University of Toronto Library and University of Toronto Press, 2019. Accessed 6 March 2019. leme.library.utoronto.ca/.

Lexicons of Early Modern English (hereafter *LEME*) is a historical database of lexical items that presents a survey of vocabulary early modern English people would have encountered. Covering 1,139,993 words in total, *LEME* offers a user 60,891 modernized English headwords to explore, all of which are culled from language-learning resources including lexical encyclopaedias, monolingual and multilingual dictionaries, hard-word glossaries, spelling lists, and other forms of lexically valuable treatises such as grammars and specific literary texts. Designed as a collated reference hub for lexical data, *LEME* is broadly presented as an index, with the ability to search for a range of variables.

LEME began in 1992 as a prototype and has been in existence in one form or another since approximately 1996. Ian Lancashire serves as general editor, with a not-insignificant army of collaborators including database programmers, research assistants, and text entry firms working with him. They are all named in the documentation regarding the project available in the header menu. It would also be remiss to discuss this project without acknowledging ample support for nearly thirty years from the University of Toronto Libraries' technical services department and the University of Toronto Press as its long-term publisher, representing an enormous amount of institutional and infrastructural buy-in. The labour and collaboration underlying these sorts of partnerships are essential to the long-term success of any digital project.

As its introduction states, *LEME* "is not a period dictionary like *The Middle English Dictionary* or the yet unrealized Early Modern English period dictionary," but rather a survey of vocabulary which was circulating in early modern England in both printed and manuscript form between 1475 to 1755.[1] The editors of *LEME* are careful to present it in contrast to resources like the *Oxford English Dictionary* (*OED*; oed.com), which is wise. Designed primarily as a monitor corpus for the history of the entire English language, from Anglo-Saxon to contemporary English, the *OED* does not account for lexical and language-learning documents as much as it accounts for examples of language in use. Thus, rather than make the rhetorical argument that English is an ever changing, ever expanding language as the *OED* does, *LEME* seeks to present a survey of vocabulary that someone alive between 1475 and 1755 could realistically be expected to encounter in some kind of language-learning context.

The resource represents two tiers of data on offer, named as "Analyzed" and "Unanalyzed," but best considered as "edited" and "unedited." Analyzed (or edited) texts have undergone a TEI-compliant XML encoding process to make them highly searchable across a range of variables.[2] This is all driven by an extremely flexible markup scheme: the editors of *LEME* developed relevant markup features that account for a range of formulations of each individual lexical entry. As an example, each language for multilingual entries is individually annotated and linked to the headword provided by the material object. Other metadata encoded in the markup allow for searches which include textual features such as citations, scribal corrections, damage, a category called "doubtful" (for uncertain readings), editorial corrections, editorial additions, parts of speech, and pronunciation. A full list of all metadata is provided in the Help documents.

Meanwhile, unanalyzed (or unedited) texts offer significantly less metadata and searchability, in large part due to their accessibility through *Early English Books Online* and its associated Text Creation Partnership initiative. The expectation appears to be that a user can follow up with a copy of the volume

1. leme.library.utoronto.ca/help/intro#scope.

2. eXtensible Markup Language (XML), a commonly used format for encoding text in a machine-readable way, is often paired with standards developed and maintained by the Text Encoding Initiative (TEI, tei-c.org).

in question as a series of images or through a full text search.[3] That said, it is strongly implied that the presence of an unanalyzed text represents a placeholder form until it is available in a fully edited and therefore analyzed form. When a text gains "analyzed" status, users can download the full associated entry as a plain-text-format, TEI-compliant file. This tiered system can feel unnecessarily complicated for the end user, but also serves as a valuable reminder that *LEME* continues to grow as a resource.

All of the entries in *LEME* are searchable using simple search strings, though more advanced searches such as regular-expression, proximity, keyword-in-context, and Boolean searches can also be performed. Users can search by specific lexical work or across all available lexical works. More specific searches can be performed, as the markup schema for Analyzed Texts allows users to account for variables such as date, type of lexical work, author, title, Wing/STC catalogue number, genre, and/or subject. Search results produce output for headwords, explanations, sub-headwords, sub explanations, and cross references, all of which are accessible as durable URLs for citation elsewhere. Each individual search function is clearly identified under the drop-down menu "search."

The search process produces a list of each lexicon containing a match to a specified search, including author, year, and modern headwords. These results are presented in chronological order (older to more recent). The technical system underlying the delivery of all this material comes from an SQL database,[4] and its search system draws from standards in lexicography, corpus linguistics, and database-development scholarship. Given the sheer volume of variant-spelling material, the linkage of each entry to a standardized headword is truly impressive. A heatmap (see figure 1, below) visualizes density of hits throughout the years of coverage in ten-year blocks representing five years each.

3. The Text Creation Partnership (textcreationpartnership.org) was a large, multi-institutionally funded project to hand-transcribe nearly sixty thousand texts included as part of *Early English Books Online* (*EEBO*). The University of Michigan hosts (quod.lib.umich.edu/e/eebogroup/) the full-text searchable apparatus for accessing Early English Books Text Creation Partnership texts in two phases: Phase I covers the first twenty-five thousand texts, which are fully in the public domain; Phase II covers an additional forty-four thousand texts which will enter the public domain in January 2020.

4. See leme.library.utoronto.ca/help/encoding for additional details.

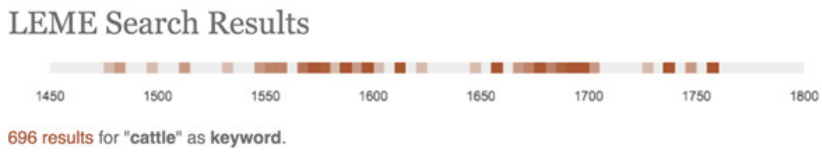

Figure 1: heat map for search results

All search results appear without having to jump through multiple pages with a finite number of entries per page. Although it is not easy to jump to a particular century or decade without scrolling or ardent use of the control-F/command-F "find" command built into web browsers, it is not impossible to navigate.

However, this search functionality may be by design, so that users are forced to skim through potentially historically-relevant items on the way to the result(s) they are searching for. Users can store multiple entries in a "notepad" function which makes their specifically relevant results exportable via email and PDF (though this is less advertised: one must be aware of the "print" function's "preview as PDF" option and then save the file locally). Alternative methods of export could also include copying all results from the notepad into a research notes document. The export feature for each Analyzed text similarly is extremely useful, but also offers a relatively high barrier to entry. Without an awareness of TEI standards and a working knowledge of XML as a classification scheme, it is easy to imagine this aspect of the project—which by all accounts has been set up to be quite the powerhouse—may be less visible to scholars who could otherwise do some remarkably creative things with it.

LEME's analyzed texts perhaps most crucially offer multiple ways to account for the very high level of spelling variation that is regularly found in the period, including modernized headwords and regular expression searches, both of which are very useful for identifying vocabulary in one of many permutated original spellings. If *LEME* did nothing else but offer the ability to curate a list of all available historical variations of early modern spelling for tasks such as scalable text normalization, it would be a massive success. But it does so much more than that: it has plainly been designed for a wide range of potential use cases. This is not something one can necessarily say for all digital projects. Despite the complications the tiered system occasionally

presents, *LEME* also happens to be one of the most robust bibliographies of early modern language-learning documents available, covering 250 works and providing a bibliography totalling 1,400 works under the remit of "lexicons" or "language-learning resources."

A scholar with no interest beyond the bibliographic data for all these language-learning documents will still find *LEME* to be a rich resource as a starting point. As a companion to other linguistic databases such as the *OED* and its Historical Thesaurus, *LEME* offers a way to triangulate contextualization within editorial projects, provides additional details regarding pronunciation for metrical and other voice-based scholarship and practice, and opened the landscape for subsequent projects like VARD.[5] In our post-*EEBO*-TCP data deluge, *LEME*'s focus on material linguistic history is utterly essential for scholars and practitioners of early modern language, variation, and change.

HEATHER FROEHLICH
The Pennsylvania State University

Carolyn W. Nelson et al., eds. *Union First Line Index of English Verse, 13th-19th Century*. Database.
Washington DC: Folger Shakespeare Library, 2009. Accessed 30 June 2019. https://firstlines.folger.edu/.

The Folger Shakespeare Library's *Union First Line Index of English Verse* is one of the fundamental digital tools for the scholar of early modern English verse. The database offers more than 250,000 first lines from English verse in manuscripts from the thirteenth to the nineteenth century (with a focus on 1500–1800), and also some printed verse from 1603 to 1710. The *Union Index* is the result of a compilation of indexes from various major libraries, namely the Bodleian, the British Library, the Folger, Harvard, the Huntington, Leeds, and Yale's Osborn Collection. It also includes metadata from other first-line indexes, including Meredith Sherlock's manuscript sources for Rochester's poetry and Steven W. May and William A. Ringler Jr.'s *Elizabethan Poetry: A Bibliography and First-Line Index of English Verse, 1559–1603* (London: Thoemmes Continuum,

5. VARD is a software programme developed to pre-process early modern English corpora for spelling variation: ucrel.lancs.ac.uk/vard/about/