**MILNER — DREaM**

<2 images>

**DREaM: Distant Reading Early Modernity**

Milner, M., Sinclair, S. and Wittek, S.

Our proposed paper will provide an overview of the theory and methodology driving the creation of Distant Reading Early Modernity(DREaM), a digital humanities project that has made a massive corpus of early modern texts amenable for use with macro-scale analytical tools. Key focus areas include the technical challenges deriving from non-standardized spelling, the philosophy of our tutorial program, the argument for our approach to the early modern archive, and the potential benefit to early modern scholarship of distant reading techniques.

**From Microfilm Library, to EEBO, to EEBO-TCP, to DREaM**

The foundational work for DREaM began in 1934, when Eugene B. Power used parts of two movie and still cameras to create one of the world’s first microfilm bookcameras, a device he used to photograph thousands of texts in British libraries (Anderson and Power, 1990). In 1998 Power’s microfilm library became the basis for Early English Books Online (EEBO), a database that comprises the images for some 125,000 texts from 1475 to 1700, and has profoundly expanded the horizons of early modern research.1 To date, approximately one-third of the documents on EEBO are available as transcribed, full-text editions. Researchers for the EEBO Text Creation Partnership(EEBO-TCP) are currently working to transcribe the remaining 85,000 documents, which are as yet only available as digitized microfilm images.2

Although completion of the transcription work is still at least 10 years in the future, the prospect of a full-text library of all documents from the first 225 years of English print points to the need for some careful re-thinking about the relation between scholarship and archival sources. As it now stands, the EEBO-TCP corpus amounts to 8.02 gigabytes of XML-encoded text and contains nearly 45,000 documents, for a grand total of well over a billion words (1,155,264,343 by our count). Confronted by the sheer expanse of a corpus several magnitudes larger than anything one could hope to read in a lifetime, early modern scholarship must now work to incorporate digital methodologies that enable a bird’s-eye view of large corpora, an approach that Franco Moretti has dubbed ‘distant reading’ (Moretti, 2007). DREaM has begun the work of making such a view possible.

Unlike EEBO, DREaM enables batch downloading of custom-defined subsets rather than obliging users to download individual texts on a one-by-one basis. In other words, it functions at the level of ‘sets of texts’ (sometimes called *worksets*) rather than ‘individual texts’. Examples of subsets one might potentially generate include ‘all texts by Ben Jonson’, ‘all texts published in 1623’, or ‘all texts printed by John Wolfe’. A user-friendly interface makes subsets available as either plain text or XML-encoded files, and gives users the option to automatically name individual files by date, author, title, or combinations thereof (this file naming flexibility can be useful when interoperating with other tool suites).

The ability to generate custom-defined subsets is important because it allows researchers to explore the early modern canon with distant reading techniques, and to capture otherwise intractable data with visualizations such as graphs, charts, or other forms of graphic representation. On this note, another key feature of DREaM is that it allows users to transfer specially tailored subsets directly to the analytic interfaces of Voyant Tools (voyant-tools.org), a suite of textual visualization tools that collectively constitute the leading platform for open-access digital humanities research.3 In fact, DREaM is actually implemented within Voyant Tools (version 2.0, not yet released, which provides much better support for very large text collections). DREaM thus provides a compelling example of a bridge between massive full-text repositories (that typically provide faceted searching) and more specialized analytic and visualization environments. By enabling simple transference between the EEBO-TCP archive and Voyant, DREaM has significantly expanded the range and sophistication of technologies currently available to researchers who wish to gain a broad sense of printed matter in early modern England.

Notably, however, DREaM does not aim to replace EEBO, or to supplant conventional forms of research. Rather, our goal is to simply add a new item to the scholar’s toolbox, and to increase transferability between distant reading methodologies and more fine-grained forms of analysis.

**Negotiating the Complexities of Non-Standardized Spelling**

Standardized spelling had yet to emerge in early modernity: writers had the freedom to spell however they pleased. To take a famous example, the name ‘Shakespeare’ has 80 different recorded spellings, including ‘Shaxpere’ and ‘Shaxberd’. As one might imagine, variance on this scale presents a serious challenge for large-scale textual analysis. How is it possible to track the incidence of a specific word, or group of words, if any given word could have an unknown multiplicity of iterations?

To address this problem, we enlisted the assistance of VARD 2, a tool that helps to improve the accuracy of textual analysis by finding candidate modern form replacements for spelling variants in historical texts.4 As with conventional spellcheckers, a user can choose to process texts manually (selecting a candidate replacement offered by the system), automatically (allowing the system to use the best candidate replacement found), or semi-automatically (training the tool on a sample of the corpora).

After some preliminary training, we ran the TCP-EEBO corpus through VARD using the default settings (auto normalization at a threshold of 50%). Rather than using the ‘batch’ mode—which proved unreliable for such a big job—we wrote a script that normalized the texts on a one-by-one basis from the command-line. This process took about three days on a commodity machine. VARD normalized 80,676 terms for a grand total of 44,909,676 changes overall.

A careful check through the list resulted in 373 term normalizations that we found problematic in one way or another. The problematic normalizations amounted to 462,975 changes overall, or only 1.03% of the total number of changes. These results were satisfactory: our goal was not to make the corpus ‘perfectly normalized’ (an impossibility, not least because perfection is debatable in this context), but, more pragmatically, to make it generally normalized, which is the best one can reasonably expect from an automatic process. On this point, it is important to note that VARD encodes a record of all changes within the output XML file, so scholars will be able to see if the program has made an erroneous normalization.

Some of the problematic VARD normalizations seem to have derived from a dictionary error. For example, ‘chan’ became ‘champion’ and ‘ged’ became ‘general’. In other instances, the problematic normalizations were ambiguous or borderline cases that we preferred to simply leave unchanged. Examples include ‘piece’ for ‘peece’, and ‘land’ for ‘iland’. There were also cases where the replacement term was not quite correct: ‘strawberie’ became ‘strawy’ rather than ‘strawberry’, and ‘hoouering’ became ‘hoovering’ rather than ‘hovering’. We fixed as many of these kinks as we could by making adjustments to the VARD training file and running the entire corpus through the normalization process a second time.

Of course, it is not difficult to imagine scenarios wherein a researcher may prefer to work with original spellings rather than normalized texts. With such projects in mind, we have kept both normalized and non-normalized versions of the EEBO-TCP corpus.

**The DREaM Tutorial Program**

As noted above, one of the central objectives of DREaM is to create an interface that will maximize user-friendliness, allowing scholars with a minimal level of technical expertise to quickly and efficiently create subsets tailored for whatever specific research question they wish to pursue. We are building DREaM for our own research, but we also have a much broader pedagogical perspective in mind. To meet this objective, we have launched a pilot tutorial program, currently under way, that will teach scholars how to use DREaM, but will also point to ways in which DREaM could more effectively serve the demands of scholarly investigation.

In a series of tasks that build toward the production of a short case-study report, pilot users must articulate a detailed research question and provide a description of their argument. In addition to establishing a valuable feedback loop for the project, this assignment aims to nudge new users toward a more comprehensive, more practical understanding of how macro-scale textual analysis can complement scholarly practice. The key conceptual challenge, as we see it, hinges on new users’ ability to understand, and learn to negotiate, the gap between distant reading and more conventional means of engaging archival sources.

Our pool of pilot users derives from the membership of our parent project, Early Modern Conversions, a five-year interdisciplinary research initiative that has brought together a team of more than 100 scholars, partners, and graduate student associates from universities in Canada, the United States, England, New Zealand, and Australia.5 Early Modern Conversions provides a propitious testing ground for DREaM because it is at the vanguard of early modern research, and because it entails a rich diversity of disciplinary approaches. Our presentation for DH2015 will report on the results of the tutorial program and on the progress of the project overall.

**Screenshots**



Figure 1. The DREaM interface. Search fields in the middle of the screen enable users to define a subset of EEBO-TCP texts by keyword, year, author, and publisher. Below the search field, an ‘Export’ button opens a dialogue box that offers the option of sending the subset directly to Voyant-tools.org, or downloading it as a ZIP archive. Users may also choose to download subsets as either plain text or XML-encoded files. A drag-and-drop mechanism (bottom) enables automatic naming of files within a subset by date, author, title, or combinations thereof.



Figure 2. A sample subset transferred to Voyant Tools. Beginning in the top left corner, one sees a word cloud representing the frequency of keywords in terms of font size. At a glance, it shows that the highest frequency words in the subset are ‘good’ and ‘come’. Below the word cloud, there is a summary that lists statistics for basic categories such as word count, vocabulary density, word frequency, etc. In addition, the summary lists words that have a notably high frequency for each year: ‘Rome’ and ‘death’ appeared with particular frequency in 1594, while ‘virtue’ and ‘envy’ stood out in 1612. Moving to the bottom left corner, one sees an ordered list of frequencies for each word in the corpus accompanied by a thumbnail graph that tracks the frequency of words over the 40-year delimitation. At a glance, the tool shows a significant spike for the word ‘knight’ in 1624. In the middle of the screen, a ‘Corpus Reader’ tool enables users to drill down into the corpus to examine the context for particular terms.

**Notes**

1. See http://eebo.chadwyck.com.

2. See http://eebo.odl.ox.ac.uk/e/eebo/.

3. See http://voyant-tools.org.

4. See http://ucrel.lancs.ac.uk/vard/about/.

5. See http://earlymodernconversions.com.

**References**

**Anderson, R. and Power, E. B.** (1990). *The Autobiography of Eugene B. Power, Founder of University Microfilms.* UMI, Ann Arbor, MI.

**Moretti, F.** (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, London.