

NEH Workshop White Paper
Genealogy of Texts and Ideas:
Looking Back and Forth through Early English Books Online

The Workshop Findings

Our EEBO data visualization workshop took place February 9-12, 2017. The full list of participants and their credentials is attached. The purpose of the workshop was to examine the potential for how humanities scholars can use the EEBO text transcriptions which are free from copyright restrictions. We believed bringing together scholars from a variety of disciplines would provide us a broad scope to anticipate the variety of research questions the EEBO data might address.

One of the issues in organizing the workshop involved balancing the desire to have as many participants as possible against the familiar problems which arise when groups become too large. Since our objective was collecting information from participants rather than delivering it to participants, we chose to limit the workshop to about 20.

Computer Skills and Programming

One main tension among the participants was whether data mining can be effective for non-programmers. A good bit of the discussion the first and second days revolved around how programming skills influence a researcher's ability to both query data and understand the results of their queries. Everyone agreed that data analysis by programmers would probably be more effective, more accurate, and more reliable. However, by the end of the workshop, we developed a general consensus that properly developed tools could provide non-programmers using big data some insights that programming scholars might not consider. If non-programmers are an important target user base, then a key aim of the tool to be developed must be an easy to use interface that does not require special training, and that maximizes the use of visual interfaces for capturing and presenting a wide range of data patterns and adaptability to creative and unforeseen uses. We elaborate on this point under Making DTA Easy below.

Spelling Variation and Corpus Preparation

Another significant issue discussed was spelling variation. This special issue for the EEBO corpus requires special tools to address and a particular understanding of the problems and best practices solutions for spelling variations. Briefly, spelling did not become normalized in English until around 1700, so every word might be spelled in a variety of ways. Although there are some easily identified regular variants, some words in the EEBO text are simply unique and/or inscrutable. Like covfefe, even area experts reading with context find it impossible to discern meaning for some of the words in the EEBO corpus.

After some deliberate discussions about the unique challenges the EEBO corpus presents, the group did agree that properly preparing the corpus for digital textual analysis requires significant expertise. Because corpus "curation" requires choices and tradeoffs in setting up

things like automated spelling variant tools and part-of-speech taggers, various implementations of a set of corpus preparation tools can change DTA results. Tools like Voyant are built for scholars who prepare their own smaller corpora, and offer little chance of replicability. Should others want to build on a previous scholar's DTA work, having the exact same corpus is necessary for the validity of the statistical results.

We agreed in principle that the idea of "locked down" corpora would benefit scholarship in the long term. "Locked-down" corpora would be "published" as dated, numbered versions. New versions would be created from the successful iterations of the corpus preparation tools used to produce the previous versions. We fully expect that the corpus prep tools available a decade from now will be a great improvement over the tools today, and that means that the corpora we use will evolve. To validate and follow up on DTA scholarship in the future, we will need access to the exact versions scholars used in their work.

Anticipating the Needs of Future Scholars

Probably the most daunting task the workshop faced was trying to map the full breadth and depth of scholarly inquiry into the EEBO corpus. Each participant's perspectives and biases both broadened and limited the discussion and at the end the result was far more limited than we had hoped. During the discussions someone suggested polling many existing EEBO users to find out from them precisely how they imagine DTA could benefit their research.

Making DTA Easy

As a nascent field, DTA offers the ability to look at texts as never before. Big Data perspectives are so different from close reading that interpreting the results can be far from intuitive, even for experts. Different approaches yield radically different kinds of results/findings, and scholars' interpretations of the results of any given DTA search or analysis can also be very different. Computational linguists have been dealing with these kinds of analytics for decades, but from within the silo of their discipline. Adapting existing tools to multidisciplinary inquiry and novice use requires two distinct yet related innovations: a simple and easy user interface and clear, effective data visualizations.

User interfaces can be some of the most confounding barriers users encounter when adopting new technologies. Because even the basic functions of DTA might be unfamiliar to humanities scholars, building an effective DTA tool for them requires creating an interface that is simple to navigate and easy to understand. Although our workshop participants acknowledge the difficulty and complexity of this goal, we agreed that concentrated efforts in this direction are warranted. We spent time discussing the different ways scholars think, specifically addressing the set of brilliant scholars who focus in the humanities precisely because they are not computer savvy or mathematically minded. We decided that creating tools that can bring many more of these colleagues into the DTA fold is worthwhile.

DTA is a statistical endeavor. It is accomplished by using computer algorithms to identify patterns in enormous sets of data. Although some scholars might be comfortable reading through hundreds or thousands of lines of numeric results, obviously charts and graphs make that kind of data easier to digest and to see patterns at a glance that are completely invisible in

other modes of presentation. Some graphic tools are better than others at illuminating certain types of data, and choosing the right type of chart or graph is far from trivial. For this reason we think it best to first approach digitally inexperienced scholars with interactive charts and graphs ideally suited for very specific, limited types of analysis.

Two Case Studies

Word frequencies and collocations have been part of the foundation of linguistics for decades. Although these methods can be easily used by scholars in other disciplines to cultivate useful knowledge, using them effectively can be a challenge. Today's humanities scholars often use word maps to show word frequency in a text or corpus. The more frequently a word is used, the larger it appears on the map. In order to make the image more interesting, the words are often shown in a variety of random colors and orientations. Although the word maps are popular, the random colors and orientation obscure the underlying information. A simple rank list would be more accurate and informative.

Rather than laboring to describe in words how such tools might be used, we created simple mock-ups showing basic interfaces on limited subsets of EEBO data. They are published here and should remain online indefinitely. In this case, a link is worth thousands of words. . .

[A Directed Network Graph](https://textonomer.shinyapps.io/network_graph/) that shows word frequency in the size of the node, and collocation frequency in edge thickness, is much more informative, especially when researchers can change parameters on the fly. Selecting an interesting word as the center of the graph reveals its most frequent, order-dependent, collocations. Selecting an edge¹ can reveal a list of the collocations in context. A tool like this will help humanities researchers expand historical and cultural inquiry far beyond conventional tools. https://textonomer.shinyapps.io/network_graph/

[A Tree Map](https://textonomer.shinyapps.io/treemap/) offers the ability to view salient features of a corpus easily and effortlessly. Tree Maps visually show frequency analysis based on levels of criteria. The first level divides the map into boxes sized proportionately to the frequency of instances with the first level characteristic. For example, if the first feature is 'year of publication', then the largest boxes will represent years with the most publications. Then, each box can be clicked to reveal more proportionally sized boxes for additional features specified. Without such a tool, it can be impossible to visualize interrelated things like the number of texts, number of words, year of publication, authors, publishers, locations, and languages. With a tree map it's easy to see things like which authors published the most words or texts in a corpus, or in a specific year or set of years. <https://textonomer.shinyapps.io/treemap/>

The Workshop Failings - Detailed Requirements for Full Implementation

In our proposal we promised a detailed set of requirements to produce a full implementation of an online toolkit built for the EEBO phase 1 public domain text. This was overly ambitious. It was naive for us to imagine that we could develop a comprehensive set of requirements for both data analysis and user interfaces in a weekend workshop. We determined that a better

¹ the line between nodes that here represents the collocations

plan would interview dozens or even hundreds of currently active EEBO scholars to assemble a wish list of potential visualizations and work from there.

The good news is that in the workshop it became clear that we do not have to start from scratch. For EEBO data analysis, workshop attendee Andrew Hardie has already produced a fast and effective tool that, as far as we were able to determine, is capable of running any big data query we could want. His tool is online and runs on the latest iterations of a locked-down EEBO corpus.

However, there is much work to be done. Because Hardie's robust tool is designed primarily for linguists, it is difficult for others to navigate or to utilize for a range of searches without extensive training. For data visualization and user interfaces, Voyant provides excellent examples of what is available and what can be done. Voyant is not designed for large corpora, but the interfaces can be adapted with some work.

Plan for Phases and Completion

We believe that the most cost effective, useful, and doable plan of action would be to start with a broad survey of existing EEBO users, to get a clear idea of how humanities scholars are interacting with the data base to produce new knowledge, and exactly what kinds of search parameters and visualization functions would be most valuable to them. The next task is to develop an API (Application Programming Interface) for the Hardie tool, based on these desiderata, while pursuing the most general functionality possible to provide for open-ended uses that have not yet been conceived. Then we would use that API to populate data for Voyant-like interfaces, whose development would also be guided by our results from investigation of users. In this way an easy-to-use integrated and comprehensive tool can be completed for a modest investment of time and money. It would build on the thousands of hours already invested in the EEBO corpus and Voyant tools, and accomplish what has not been done before.

Code Repository

All of the code written for this project can be found at <https://github.com/Textonomer/MockUps>.

Other Records

Brief bio's for the workshop participants are included below.



Allison Mickel is an archaeologist and a lecturer in the Program in Writing and Rhetoric at Stanford University. She received her PhD in anthropology from Stanford in 2016 and her BA from The College of William and Mary in 2011. She uses network analysis, primarily, to examine the processes of knowledge production in archaeological excavation. By analyzing the texts produced during archaeological fieldwork in a number of ways-- structurally, topically, and emotively-- she maps out the information transmission that occurs between people over the course of archaeological work, and ultimately how facts about the past are formed through both the creation of excavation documents, through labor organization paradigms, and through interpersonal interactions.



Andrew Hardie is a Reader in Linguistics at Lancaster University, and currently serves both as Chair of Lancaster's UCREL centre, and as Director of the ESRC Centre for Corpus Approaches to Social Science. His main research interests are the theory and methodology of corpus linguistics; corpus-based descriptive and theoretical grammatical analysis; the languages of Asia; and applications of corpus methods in the humanities and social sciences. He is also a lead developer of the *Corpus Workbench* analysis software, and the creator of its online interface, CQPweb. He is the author, with Tony McEnery, of the book *Corpus Linguistics: Method, Theory and Practice* (2012).



Anne Chao is a modern Chinese historian currently engaged in three different DH projects. She is using network analysis software to plot the social networks of the man who founded the Chinese Communist Party, Chen Duxiu. She is also using text-mining tools to detect shifting meanings in Chen's thoughts and writing. The second project involves creating a biographical database with a colleague on the historical figures of Late Qing, Early Republican China (ca. 1890-1920). Finally she is the manager of the Houston Asian American Archive, an online repository of life stories of the Asian Americans in Houston. She is also an Adjunct Lecturer in the Humanities at Rice.



Anupam Basu is an Assistant Professor of English at Washington University in Saint Louis. He works at the intersection of literature and computational analysis, drawing on emerging techniques to make vast digital archives of early modern print more tractable for scholars. The ngram and collocation browser at <http://earlyprint.wustl.edu> makes some of the data behind his research accessible to a wider audience. His current book project on crime and social change in Tudor and Stuart literature explores the popular representation of criminality, poverty, and vagrancy in the period.



Benjamin Brochstein is a former network admin; database programmer, designer, and admin; CPA; and civil litigator (yes, civil litigator is an oxymoron) re-purposed into academia. His 2008 research into an extraordinary set of 17th century healing narratives focused his attention into the EEBO transcription project and its potential as a data source for statistical analysis in humanities research. He began programming computers in high school in the late 1970's and has maintained a strong interest in information technology for the subsequent four decades. He is currently the project director of this NEH funded workshop. Benjamin's BHAG is to create a DTA tool so simple and intuitive that practically any humanities researcher can take a "fishing expedition" through any corpus with very little effort.



Chad Shaw. "My main research interests are systems biology and the analysis of large scale genomic data. My laboratory has done extensive research in the use of genomic annotations to enhance analysis of microarray experiments. We have developed a variety of web-based software including tools for Gene Ontology analysis as well as an interactive system for exploration of protein-protein interaction networks. We have also developed tools for web-based visualization and sharing of gene expression data. We analyze primary microarray data sets from all array platforms including expression arrays, genome content arrays (aCGH), microRNA arrays, and chromatin arrays with an expertise in data pre-processing and normalization. Our methodological work involves statistical considerations for use of annotations in analysis of very large scale genomic data."



David J. Birnbaum is Professor and Chair of the Department of Slavic Languages and Literatures at the University of Pittsburgh. He has been involved in the study of electronic text technology since the mid-1980s, has delivered presentations at a variety of electronic text technology conferences, and has served on the board of the Association for Computers and the Humanities, the editorial board of *Markup languages: Theory and practice*, and the Text Encoding Initiative Council. Much of his electronic text work intersects with his research in medieval Slavic manuscript studies, but he also often writes about issues in the philosophy of markup. In July 2017 he will be directing a three-week NEH Institute in Advanced Topics in the Digital Humanities at the University of Pittsburgh entitled "Make *your* edition: models and methods for digital textual scholarship".



Lateefat Alabi. "I'm a current senior in the Material Science and NanoEngineering and Statistics departments at Rice but the STEM track is a recent development. My documented interests have been in European, World, and Art History. The summer of 2016 was spent working at HP using readily available sentiment algorithms to analyse consumer reviews and I look forward to transferring what I learned in that experience here."



Erez Lieberman Aiden received his PhD from Harvard and MIT in 2010. After several years at Harvard's Society of Fellows and at Google as Visiting Faculty, he became Assistant Professor of Genetics at Baylor College of Medicine and of Computer Science and Applied Mathematics at Rice University. Dr. Aiden's inventions include the Hi-C method for three-dimensional DNA sequencing, which enables scientists to examine how the two-meter long human genome folds up inside the tiny space of the cell nucleus. In 2014, his laboratory reported the first comprehensive map of loops across the human genome, mapping their anchors with single-base-pair resolution. In 2015, his lab showed that these loops form by extrusion, and that it is possible to add and remove loops and domains in a predictable fashion using targeted mutations as short as a single base pair. Together with Jean-Baptiste Michel, Dr. Aiden also developed the Google Ngram Viewer, a tool for probing cultural change by exploring the frequency of words and phrases in books over the centuries. The Ngram Viewer is used every day, by millions of users worldwide.



Ido Machol is Lead scientific programmer, Baylor College of Medicine, Texas "For many years, I have worked as a software team leader, exercising daily problem solving, in software development and requirements deciphering. I learned that in order to deliver a quality product on time, a good planning of activities is required. Leading development teams for complex systems at small and big companies, taught me a lot about personal communications, human interactions and human-machine interactions. As a group leader, I was concerned about the quality and applicability of the software we made, always looking for new creative methods to accomplish ever demanding new tasks. I now combine these past skills together with natural sense of curiosity and enthusiasm, to optimize complex bio informatics problems into robust genomic pipelines



John Mulligan is a lecturer at the Rice University Humanity Research Center (HRC). Since earning a B. A. in Bates College in 2006 (with a minor in mathematics) and his Ph. D. in English Literature at Brown University in 2016, he has taught English literature and experiential learning courses at Rice under the HRC's Public Humanities initiative, funded by the Andrew W. Mellon Foundation. He writes on the aesthetics of abstraction in Romantic-era literature and science, and uses digital media to experimentally engage with historically specific problems of representation and communication in the humanities and sciences. His writing (on William Blake and Isaac Newton) has been published in Oxford's *Notes & Queries*, and one of his digital transmedia experiments (on Andreas Vesalius' *Fabrica*) has been publicly installed in the Texas Medical Center Library.



Joseph Dexter (A.B. Princeton) is a Ph.D. candidate in the Department of Systems Biology at Harvard Medical School and (with Pramit Chaudhuri) co-founder and co-director of the Quantitative Criticism Lab. His research in biology focuses on analysis of metabolic and signaling networks and employs a range of computational and experimental techniques, including mathematical modeling, machine learning, *in vivo* imaging, and microfluidics. In addition to his quantitative work on intertextuality and stylometry with QCL, he maintains an active research agenda in more traditional areas of classics and has published on ancient theatre, Latin epic, and reception studies.



Michael Barlow received his PhD in Linguistics from Stanford University and is Associate Professor in the Applied Language Studies and Linguistics Department at the University of Auckland in New Zealand. Dr. Barlow has created several text analysis programs including concordancers MonoConc and ParaConc and a collocation extraction program, Collocate. A recently developed program, WordSkew, is designed to apply corpus analysis techniques while at the same time taking note of the structure of texts.



Olivia Vane is a PhD student in Innovation Design Engineering at the Royal College of Art, London. Her research explores how interactive data visualisation maybe used to 'make sense' of digitised cultural collections. She is currently working with partners at the Wellcome Library, the V&A, and the Cooper Hewitt Smithsonian Design Museum. Olivia holds a BA Natural Sciences and an MSci History and Philosophy of Science from the University of Cambridge.



Paul Schaffner is a librarian at the University of Michigan who devoted seventeen years (and counting) to managing the output of the Text Creation Partnership (EEBO-TCP etc.), creating and tweaking the capture guidelines, applying them to innumerable edge cases, munging the associated MARC, swallowing many an objectionable compromise, and managing a distributed staff of cumulatively about a hundred editor. Its faults are his fault. During this time he also spent six years on TEI Council. He came to the library originally in 1997 to manage the transformation of the Middle English Dictionary into the online Middle English Compendium, having spent the previous eight years as an historical lexicographer. This present year, he is working mostly on revisions to the MED, thus coming full circle and returning to his medievalist roots. His technical skills are primitive and old-fashioned (basic scripting, text editing and processing, Perl), and his formal training rapidly rusting: he trained largely as a philologist (BA Haverford (early English); MA Cantab. (Anglo-Saxon, Norse & Celtic), PhD Cornell (medieval studies/philology), MLS Michigan). His bent is toward simplicity (verging sometimes on the simplistic), practicality, and repeatability. His interest in text analysis is mostly in the degree to which it can be made to sort out algorithmically or otherwise the chaos introduced into text corpora by the human and mechanical accidents that gave it birth. In his spare time he collects hand tools, hardware catalogues, and hymn books.



Scarlett Liu is a senior from Rice University graduating in May 2017. She doubles majors in statistics and mathematical economic analysis. During her undergraduate studies, she became increasingly interested in the power of big data. In the fall of 2016, she co-worked with Lateefat Alabi analyzing the text sentiment trend during the English Revolution using text data provided by the site Early English Books Online.



Suzanne Kemmer is Associate Professor of Linguistics and Cognitive Sciences at Rice. She earned her Ph.D. in Linguistics at Stanford University and taught at UCSD before coming to Rice. She is author of the monograph *The Middle Voice* and, with Michael Barlow, editor of the influential volume *Usage-Based Models of Language*. She has published widely in Cognitive Linguistics, Language Universals/Language Typology, Lexical Semantics, Historical Linguistics and English Linguistics. She was an early adopter of the use of large corpora for linguistic research, connecting the basic theoretical tenets of Cognitive Linguistics on the importance of frequency in the learning and use of grammar with the frequency patterns visible in corpora. She served twice as President of the International Cognitive Linguistics Association, has taught in two Linguistic Society of America Institutes (at UCSB and UC Boulder), and was a senior fellow at the Max Planck Institute of Evolutionary Anthropology in Leipzig and also at the Helsinki Collegium for Advanced Studies. At Rice she directed the Neuroscience Program from 2010-2013 and the Cognitive Sciences Program from 2007-2014. Her current research is on the relation of visual perception to linguistic constructions of fictive motion.

Student Assistants



Alex Hayes is a junior majoring in Statistics at Rice University. He is interested in using machine learning techniques to understand human behavior and language. Recently exposed to textual analysis, Alex is looking for research opportunities in natural language processing before heading off to graduate school in statistics next year. In his free time, he enjoys photography, staying in touch by letter, and practicing his Spanish on unsuspecting victims.



Daniel Cohen is a Rice University freshman planning to double major in Cognitive Sciences and Linguistics. He is interested in cognitive linguistics and language endangerment. This is the first conference Daniel has attended, and he is hoping to gain a lot from the experience. In his free time, he enjoys playing basketball and pool, reading, and having heated political debates.



Erin Song is a Sophomore Statistics major at Rice University. She has broad interests in machine learning and data analysis.



Eugene Yeom is a junior studying Linguistics and English at Rice University. She loves linguistics and has taken courses on discourse analysis, corpus linguistics, Middle English literature, and early Shakespeare. She is considering graduate study in linguistics and is currently looking into research opportunities on Korean and Japanese. She loves discussing the Harry Potter series and wants to learn how better to cook for herself.