evoText: Digital Humanities Meets Evolutionary Biology

Charles H. Pence*

*University of Notre Dame, Program in History and Philosophy of Science (charles@charlespence.net)

Digital Humanities

Since the founding of the *Philosophical Transactions of the Royal Society* in 1665, the primary outlet for the dissemination of scientific results has been the journal. For those interested in studying long-term trends in the sciences, such literature is indispensable. Its bulk, however, is formidable. The journal *Science* has published some 186,000 articles since its inception in 1880, and the JSTOR archive holds nearly 6.5 million individual journal articles.

How do we navigate this vast amount of literature? The development of computational tools in the digital humanities has made this possible.

Archiving the Evolutionary Sciences

We have begun to construct an archive of journal articles in the evolutionary sciences – biology, anthropology, evolutionary psychology, human development, etc. Currently, our archive contains around 500,000 articles, from the following publications:

- *Nature*
- *Science*
- *Q Rev Biol*
- *Am Midl Nat*
- *Am Nat*
- *Evolution*
- *Ecology*
- *Ecol Monogr*
- *Behaviour*
- *Am J Phys Anthropol*
- *Am J Hum Biol*
- *Evol Anthropol*
- *Am J Primatol*
- *J Zool*
- *Genes Brain Behav*
- *Evol Dev*
- *Ethology*
- *Am Anthropol*

Tools for Analysis

Available tools for analysis are an active area of development in evoText. Currently, we have deployed a highly advanced search engine, capable of performing complex full-text searches including boolean operators ("darwin OR huxley"), wildcards ("*fish" or "wom?n"), text stemming ("evolution" matching "evolutionary"), fuzzy matching (words similar to a given term), and proximity searching (two terms within N words of one another).

Once a set of results has been obtained, it can be saved as a "dataset," and various forms of analysis or export can be run on the dataset. Currently, datasets may be exported in a variety of bibliographic formats, and the members of a dataset can be graphed by year. For single-document datasets, a complete list of term frequencies can be queried.

Visit evoText!

http://evotext.crc.nd.edu/

Example: Evolution's First Century

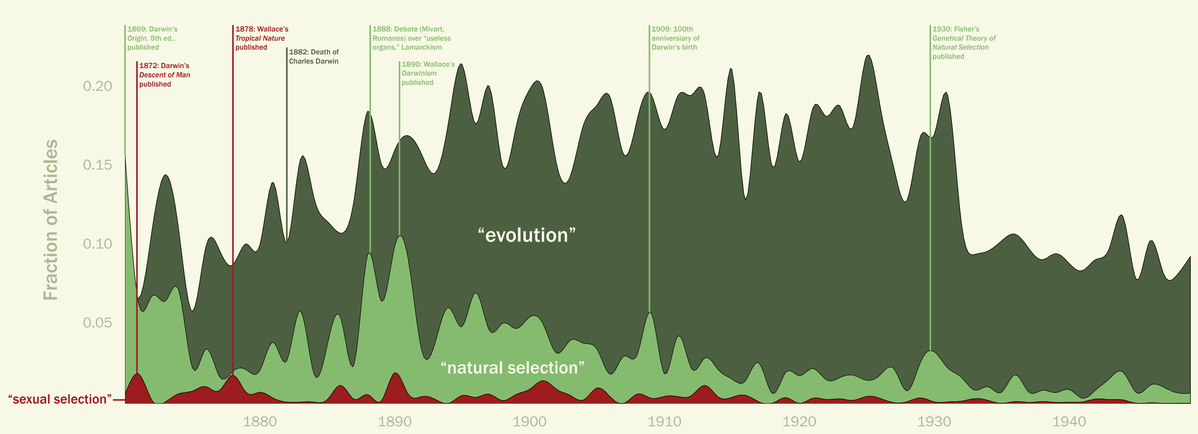

Figure 1: An example of the kind of analysis evoText offers. This dataset is sampled from the first 80 years of the journal *Nature,* including searches for the terms "evolution," "natural selection," and "sexual selection."

Interface

The evoText interface is designed for easy use on both desktop computers and tablets such as the iPad. A detailed tutorial is available at the evoText website.

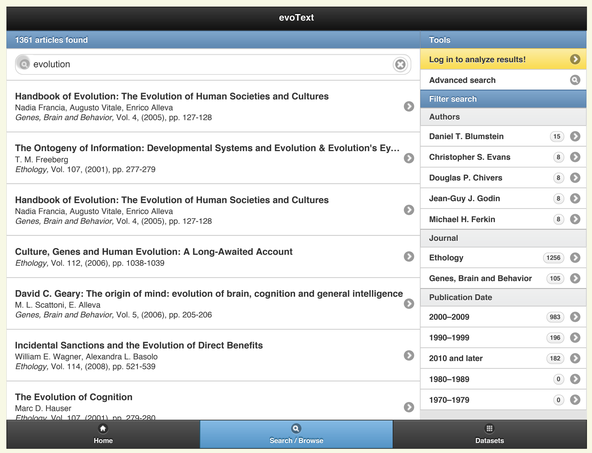

Figure 2: The evoText search interface. Here, a search for "evolution" has been performed. On the left you can see the list of documents available (1,361 hits in our test database). You can proceed to narrow this search by filtering for authors, journals, or publication dates, which are listed in the right-hand column.

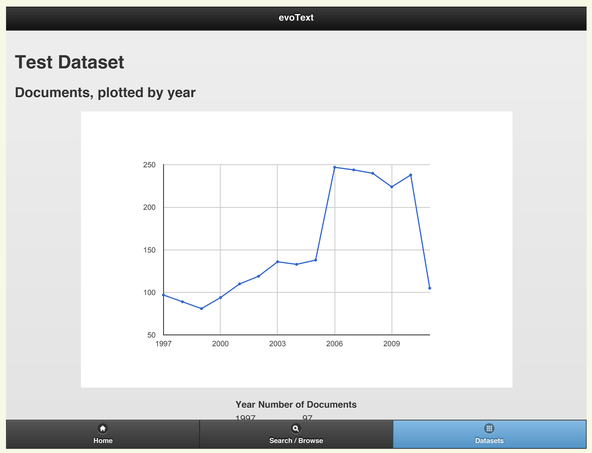

Figure 3: Results from an analysis of an evoText dataset. This graph shows the documents in our test database plotted by year, with tabular data below. This test database consists of the complete run of two journals which began publication within the last several decades.

Future Analysis Tasks

Active development is focused on providing more analysis tasks that can be run on a dataset. In the near future, we plan to support at least the following standard digital humanities analyses:

- Term frequencies and tf*idf (term frequency × inverse document frequency) for an entire dataset, not just a single document
- Most frequently used N-word phrases (N-grams) in a dataset
- Network analysis of word context (a 2-d graph showing how words cluster within documents)
- Extraction, recognition, and mapping of place names found within journal articles

Further, we plan to work on visualization of data using cutting-edge web presentation technologies, especially on tablet devices, including WebGL.

Finally, we aim to continue to increase our archive of searchable journal articles. Many of our collected articles do not yet appear in evoText, as the articles are pending completed optical character recognition (OCR).

Acknowledgments

Many thanks to Grant Ramsey for the initial inspiration to embark on this project. Agustín Fuentes has assisted with work on journal selection and project scope. A University of Notre Dame Institute for Scholarship in the Liberal Arts Exploratory Seminars in Integrative Research Grant has supported the Digital Humanities Working Group: myself, Grant Ramsey, Agustín Fuentes, Eric Lease Morgan, Sean O'Brien, and Matthew Wilkens. Extensive hardware and infrastructure support has been provided by the Notre Dame Center for Research Computing.

RLetters

Would you like to run your own archive of journal articles similar to evoText? The software platform on which evoText is based – RLetters – is open source software and is freely available to researchers at `http://charlespence.net/rletters/`. It requires a web server (with Ruby on Rails, powering the front-end) and a Solr search server (powering the backend), both of which are easy to install and configure. If you'd like to set up an RLetters installation using your own data, send me an e-mail!