2016

evoText: A New Tool for Analyzing the Biological Sciences

Grant Ramsey
KU Leuven

Charles H. Pence
Louisiana State University, cpence@lsu.edu

Studies in History and Philosophy of Biological and Biomedical Sciences 57 (2016) 83–87

Contents lists available at ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

Studies C essay

evoText: A new tool for analyzing the biological sciences

Grant Ramsey [a], Charles H. Pence [b,*]

[a] KU Leuven, Institute of Philosophy, BE-3000 Leuven, Belgium
[b] Louisiana State University, Department of Philosophy and Religious Studies, Baton Rouge, LA 70803, USA

ARTICLE INFO

Article history:
Received 5 April 2016
Accepted 16 April 2016

Keywords:
Text analysis
Digital humanities
Software
Philosophy of science
Philosophy of biology
History of science
History of biology
Biology

ABSTRACT

We introduce here evoText, a new tool for automated analysis of the literature in the biological sciences. evoText contains a database of hundreds of thousands of journal articles and an array of analysis tools for generating quantitative data on the nature and history of life science, especially ecology and evolutionary biology. This article describes the features of evoText, presents a variety of examples of the kinds of analyses that evoText can run, and offers a brief tutorial describing how to use it.

1. Introduction

What is the nature of science and how has it changed over the past 150 years? What drives scientific change, and what accounts for when and why scientists give up cherished views to adopt new ones? There are a number of distinct approaches to answering questions of this kind, representing divergent ways of understanding the nature and status of particular sciences—or of science in general—arising from a variety of disciplines. There are traditional philosophical approaches, focusing on the conceptual structure of science, asking questions like *what is a scientific explanation?* or *what distinguishes science from pseudoscience?* There are sociological approaches, focusing on questions like *what do scientists actually do on a day-to-day basis?* and *why do scientists make the decisions that they do* (*is it in pursuit of truth*, *power*, *status*)*?* There are also historical approaches, detailing moments in the history of a science to say what happened and why. Each approach has its benefits and limitations. One trade-off between them pits specificity or depth against generality or breadth. Traditional historical approaches, as well as the philosophy of particular sciences (biology, physics, etc.), tend to be narrow in focus, detailing a specific moment within a particular scientific discipline, while general philosophy of science often sacrifices specificity for broader application. It is understandable that many approaches opt for depth rather than breadth: attempting breadth generally brings about superficiality and bias. It is difficult to be representative when confronted with a body of information too large to absorb, and safer to be comprehensive in a small area.

With the digitization of the scientific literature, a new way of engaging with science is beginning to emerge. We, for the first time, have access to digital repositories of hundreds of millions of pages of scientific, philosophical, and historical text. These repositories open up opportunities to examine, on a broad scale, how science happens. Unfortunately, however, there is a lack of tools for studying this corpus of texts in a rigorous way that would produce statistically significant results. Much of the focus has been on producing ways of discovering articles (e.g., Google Scholar) or digitizing material that has not previously been available (e.g., the Darwin Online project (van Whye, 2002) or the Einstein Papers Project (2014)).

We introduce here a powerful new tool, evoText, which provides a window into science that will allow for broad scale quantitative study of the science journal literature—it is thus a way of retaining breadth without becoming superficial or subject to the biases inherent in manually working through large sets of texts. By combining algorithms developed in the sciences and digital humanities with a corpus of science journals, evoText will allow for the study of science in a way not heretofore possible.

In this article, we show what makes evoText distinct from other digital tools, and we describe the corpus of journal articles it

* Corresponding author.
E-mail address: charles@charlespence.net (C.H. Pence).

http://dx.doi.org/10.1016/j.shpsc.2016.04.003

contains and the analysis tools it offers. We then offer an example of the kinds of analyses that evoText can support.

2. The unique nature of evoText

There are many high-quality software packages for analyzing texts, such as SEASR (Ashton, 2011), TAPOR Tools (Rockwell, 2006), MONK (Kumar, 2009), Google's N-gram Viewer (Brants and Franz, 2006; Michel et al., 2011), and JSTOR's Data for Research (Burns et al., 2009). Each, however, is inappropriate for the problems that evoText aims to solve. Several packages—including TAPOR Tools and MONK—require that the user upload texts into the system, making analysis of the size of corpus deployed in evoText (hundreds of thousands to millions of documents) impracticable. Some, such as MONK, require for full capability that the texts be marked up manually in a format like TEI (Ide and Véronis, 1995), which, again, is infeasible for analysis of a corpus as large as ours. Other tools, such as Google's N-gram Viewer and JSTOR's Data for Research, have large corpora of text against which they are deployed, but they have significant limitations. JSTOR's corpus is limited to the journals they happen to have agreements with and is thus unlikely to be a representative sample of all journals. And Google's N-gram Viewer contains only the corpus that the Google Books Project has thus far digitized, limiting the inferences one can make about cultural dynamics from its analysis (Pechenick, Danforth, & Dodds, 2015). No general-purpose tool presently available is optimized for journal articles. The challenges presented by the analysis of millions of small texts (as is the case with journal articles) rather than a much smaller number of considerably larger texts (like books) are unique and significant. Finally, some current programs (such as SEASR or TAPOR Tools) require the user to chain together many smaller analysis steps to perform common data analyses, presenting a usability challenge.

evoText resolves each of these issues. Its corpus contains a vast collection of journal articles, thus not requiring users to upload these texts themselves. Its analysis tools work against plain text, allowing us to add substantial numbers of texts without costly processing or encoding time, and to include specific features to clean OCR text. These qualities, in addition a to user-friendly website, allow users to perform common analyses with a few clicks.

The software powering evoText, called RLetters, is available under the MIT License (Open Source) at https://github.com/rletters/rletters. While evoText will have a corpus of articles curated by us, if a user wishes to analyze a different corpus of articles, they are free to use the RLetters software to accomplish this (Pence, 2016).

3. The journal database

Our journal database currently contains open access content from a variety of *PLoS* journals, and closed access content obtained via text mining agreements as well as partnerships with Nature Publishing Group, Elsevier, and JSTOR. At press time we have more than 400,000 journal articles, but are adding articles on an ongoing basis. The corpus focuses on journals related to evolutionary biology, but is not limited to this topic. Our goal is to be as complete as possible in our collection of evolutionary biology journals, but to include a large array of articles in neighboring disciplines. For example, from JSTOR we include the entire array of journals in their "Ecology and Evolutionary Biology" category as well as "General Science." As the project progresses, we will continue to broaden the corpus.

We are sensitive to the worry that housing the content in this way constitutes, in essence, yet another example of "data siloing" in the digital humanities—the construction of another closed collection of data to which only the evoText maintainers will have full access. At the moment, however, there exists no alternative if one desires to mine more than open-access or public-domain texts. JSTOR, for example, has informed us that a solution in which we store the full text of their articles on our own servers is legally infeasible. This is a recognized problem in textual analysis, of course, as those in charge of closed-access data archives like HathiTrust have repeatedly emphasized (York, 2009). We believe, however, that deciding to analyze only open-access texts is the wrong solution—particularly if publishers can be made to see the demand present for this kind of textual analysis in the scholarly community. We would be glad to work with researchers who would like to negotiate closed-access content agreements similar to our own.

Fig. 1. The relative font sizes in this word cloud represent the frequencies of the ten most common 'evolutionary ___' bigrams in the journal *Nature* during the 2000s decade.

4. The evoText tools

A wide variety of analysis methods are implemented in evoText, and are described here briefly. More detail can be found in Pence (2016), or below in the discussion of our example use of evoText.

Compute Term Frequency. Users can compute term frequency tables for a given dataset, for either single words or multiple-word phrases (n-grams) (modeled after features in Tsukamoto, 2002). These are the most common inputs for other kinds of textual analysis algorithms, meaning that users can easily extract term frequencies and use them to run their own analyses locally if desired.

Co-occurrence and Collocation Analysis. Information may be extracted concerning statistically significant collocations (immediate pairs of words) or co-occurrences (significant connections between words at the sentence, paragraph, section, or article level) (Manning and Schütze, 1999).

Compare Difference Between Datasets. The Craig Zeta algorithm (Burrows, 2006; Craig and Kinney, 2009) can compute the difference between datasets, showing which words, if found in a random article, would be likely to "mark out" that article as belonging to either set.

Compute Term Network. Users can visualize the network of words occurring in the immediate vicinity of a given focal word of interest, an analysis that is useful for determining which words often "travel together" in the literature (He, 1999).

Extract Proper Names. Proper names (of persons, locations, organizations, and so forth) found in journal articles can be extracted. This analysis can be useful to detect locations of field research, organizational networks, etc. (Manning et al., 2014).

Graph by Publication Date. Users can graph the publication dates of a dataset, which is particularly useful if the dataset contains only those articles that match a complex search.

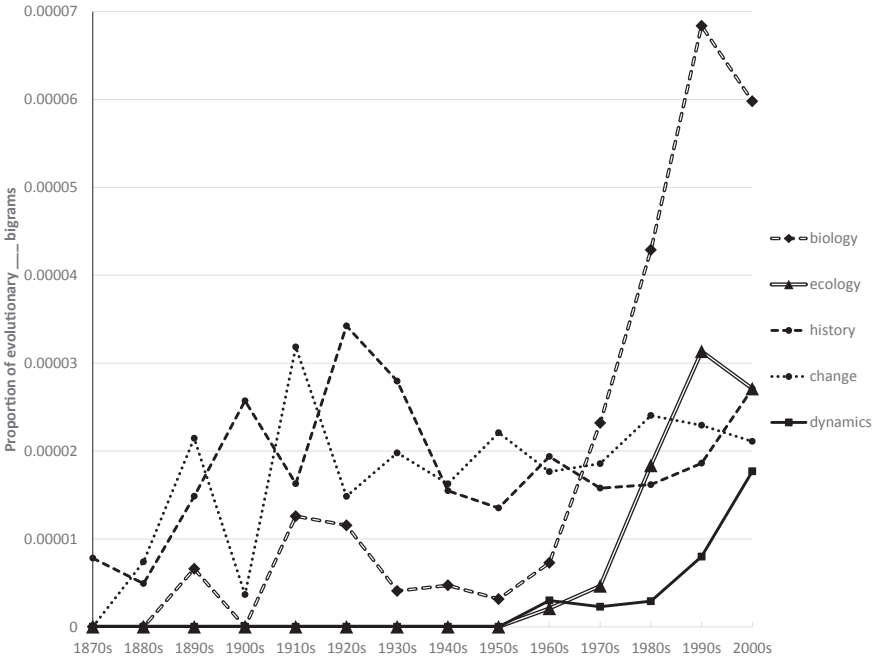

Fig. 2. The frequency of 'evolutionary ___' bigrams in the journal *Nature* from the 1870s through the 2000s (measured as the proportion of all bigrams from each decade).

Export Citations. Lastly, a dataset can be exported in a variety of citation formats to a user's citation manager, including EndNote and BibTeX.

5. An example of an evoText analysis

To show the power of evoText, let's consider an example.[1] For the example, we investigated what the word 'evolutionary' modifies across journals and across time. We began by determining the ten most frequently occurring 'evolutionary ___' bigrams (pairs of words) from the journal *Nature* during the decade of the 2000s. Starting with the most frequent, they are biology, ecology, history, change, dynamics, processes, time, process, genetics, and theory. See Fig. 1 for a word cloud representing the frequency of these bigrams (and see the Appendix for information about how this figure was generated using evoText).

Given this distribution of bigram frequencies, we might wonder how they came to be. When, for example, did we begin to use 'evolutionary biology' at such a high frequency? To examine how these frequencies changed over time, we plotted the frequency of the bigrams (the number of occurrences divided by total number of bigrams), all the way back to the 1870s (see Fig. 2). To reduce clutter, we focused only on the five most frequent bigrams. In this graph, we are able to see how and when these bigrams became as frequent as they now are. Fig. 2 shows some interesting trends. Not unexpectedly, 'history' and 'change' have remained fairly constant over the history of the journal. From its beginning, with the publication of Darwin's *On the Origin of Species* (1859), evolution has always been understood as a historical science concerning organic change. The data from *Nature* confirm these as core elements of our understanding of evolution throughout its history.

The evolutionary biology, ecology, and dynamics bigrams, on the other hand, all take off in frequency around 1950. This timing corresponds with the period following the synthesis of genetics and Darwinism, in which evolution became a discipline of its own, with a society and a dedicated journal. The first annual meeting of the Society for the Study of Evolution occurred in 1946 and the first volume of the society's journal, *Evolution*, was published the following year (Smocovitis, 1994). Therefore, it makes sense that referring to evolutionary biology and evolutionary ecology would be infrequent before the crystallization of evolutionary biology as an independent discipline, but increasingly common afterward. The term 'evolutionary dynamics' takes off around the same point, though it has a slower increase than biology and ecology. Its increase might in part be attributed to it being a more technical term than 'change', and it is interesting to note that 'evolutionary change' decreases during the same period that 'evolutionary dynamics' increases. This may be purely coincidental, but it may also point to a replacement of 'change' with 'dynamics'.

If we were to further investigate this increase in 'dynamics', we might wonder whether the journal *Nature* is representative of broader trends in evolutionary science. A plausible initial hypothesis would be that because of *Nature*'s status as a general science journal, it would be slower to introduce the term and would use it less frequently. As an initial test of this hypothesis, we could compare the frequency of use of 'evolutionary dynamics' in *Nature* with its usage in a more specialized journal like *Evolution* (measured as the number of articles containing the bigram divided by total number of articles for each year). When we do so, we find that the bigram takes off in roughly the same year in both journals and has a similar pattern of increasing frequency (see Fig. 3). With some allowance for noisy signal, the pattern of use of 'evolutionary dynamics' appears to be broadly similar in both journals, suggesting that we are seeing a field-wide change in terminology. Of course, the absolute frequency in *Nature* is lower by more than an order of magnitude, but this can be attributed to the fact that *Nature* publishes on a wide array of topics.

6. Conclusion

evoText is a tool for historians, philosophers, scientists, and any others who wish to gain insight into the nature and history of

[1] All data used to generate these figures is available online, with 10.6084/m9.figshare.3180220.

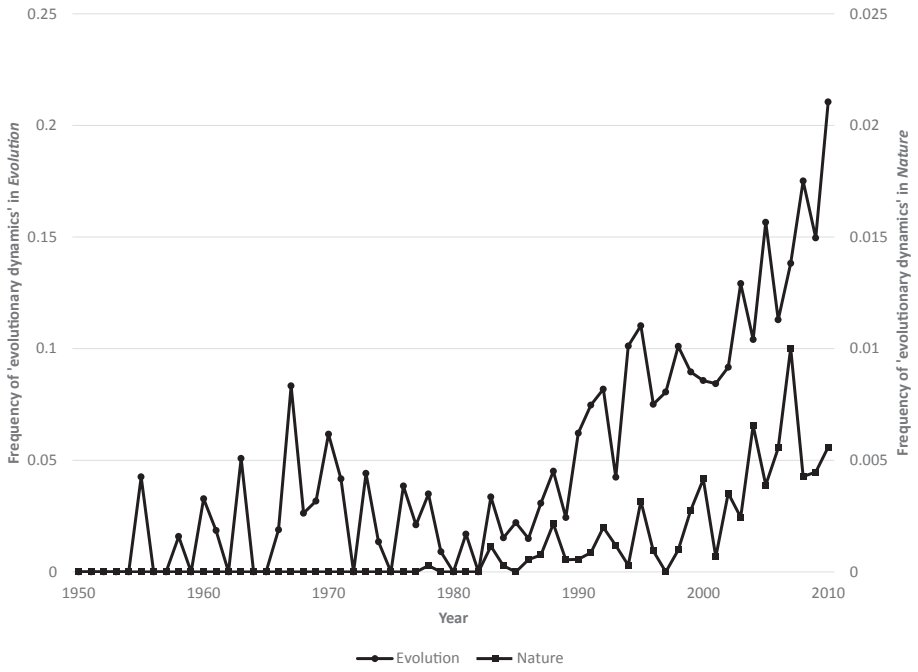

Fig. 3. The frequency of 'evolutionary dynamics' bigrams in the journals *Evolution* and *Nature* from 1950 through 2000 (measured as the proportion of documents from each year that contain at least one occurrence of the bigram).

science. It contains a database of hundreds of thousands of articles and supplies tools to run sophisticated algorithms. These algorithms allow researchers to plumb the depths of the sciences, gaining data and insights not heretofore possible.

We hope that you will explore evoText to see how it might shed new light on your research projects. If you have questions about evoText, we encourage you to contact help@evotext.org. The evoText database is currently focused on particular areas of biology (principally ecology and evolutionary biology) as well as general science, though we plan to expand into other areas in the future. Active development continues on evoText—if there are specific subjects, journals, tools, or other features that would be useful to you, please let us know and we will see how we might accommodate your needs.

Acknowledgments

This work was supported in part by U.S. National Science Foundation (NSF) grant #SES-1456573, awarded to Grant Ramsey and Charles H. Pence for the evoText Project, and by the National Evolutionary Synthesis Center, which was funded by NSF grant #EF-0905606. Thanks as well to the University of Notre Dame Hesburgh Libraries for assistance with content licensing, especially Monica Moore.

Appendix. Generating Fig. 1

Fig. 1 was directly generated using evoText, while Figs. 2 and 3 used data generated by evoText in conjunction with Microsoft Excel for graphing assistance. Fig. 1 is the most involved, so we offer here a tutorial for how to generate it. To get started using evoText, visit http://www.evotext.org and create a user account. Your account will allow for the storage of your personal user data, your datasets (about which more below), and your analysis results.

To begin generating Fig. 1, click the "Start a new analysis" button on your dashboard page. This brings you to a list of the kind of questions you can answer using evoText. For our task, scroll to the bottom and select "What's the frequency of word use within a given set of articles?" You'll now be presented with a page of information about the analysis method you've selected. Click "Start."

You now need to provide your analysis with some data. For this figure, we want to search in the 2000s decade of *Nature*. Click "Create another dataset," and you'll find yourself in evoText's search interface. Building this dataset is easy. Start by clicking "2000–2009" on the right-hand side, under "Filters … Publication Date." Now you're only seeing results in the list for articles published in the 2000s. Then click "Nature" under "Filters … Journal." You now have the set of articles that you're interested in (at press time, this was 29,412 journal articles). Save this collection as a *dataset* by clicking the green "Save" button. Give your dataset a descriptive name (like "Nature 2000s"), and click "Create dataset." This dataset is permanent, and in the future you will be able to run more analyses on it by selecting "Link an already created dataset" instead of "Create another dataset" in the "Collect data" window.

We have now returned to the "Create data" window, with your newly created and named dataset in the list of "Datasets for this job." Click "Set Job Options." This is one of the more complicated analysis tools in all of evoText. To generate our word cloud, set the following options:

- Analyze single words or n-grams? Select "N-grams," as we are interested in multiple-word phrases.
- Size of phrases to analyze: 2. This specifies that we want the frequency of bigrams (phrases of two words).
- Number of n-grams to analyze: We're interested in the ten most commonly occurring bigrams, so leave the "Return all n-grams" button unchecked and enter 10 into the field.
- Include only n-grams that contain one of the following words (space-separated): Enter 'evolutionary' (without quotes), and

we will only get information on bigrams that include the word 'evolutionary'.

- Exclude any words? Select "Most common words (stop words)." This removes a set of uninformative "stop words" (such as 'the', 'and', 'a', 'of', and so on) from the frequency list. This feature can also be used to exclude a custom list of words inputted by the user.
- Language of text (for stop word list): English. evoText includes standard stop lists for a variety of languages.
- Stem words? No. This option removes endings from words, making, for example, 'evolution' and 'evolutionary' analyze as the same word.
- Text block method: By number of blocks. These options control how we will chop the text into pieces before counting up its words, useful for various algorithms in the digital humanities.
- Number of blocks: 1. This allows us to look at all the documents in the dataset in a single block.
- Split blocks across documents: Checked. We want to get one block that includes *all* the documents in our dataset, not one block per journal article.
- Create a word cloud: Checked.
- Word cloud font and color: Choose as you like! We used the "Vollkorn" font and "Blues" color to generate the word cloud in Fig. 1.
- Show words in the inclusion list in the word cloud? Unchecked. If this box is checked, the word cloud will display 'evolutionary biology', 'evolutionary history', 'evolutionary genetics', etc. Since 'evolutionary' is in the inclusion list—the list of words to include, from above—unchecking this box will remove 'evolutionary' from each entry, giving us a word cloud containing 'biology', 'history', 'genetics', etc., as appears in Fig. 1.

Click "Start analysis job." Jobs in evoText are performed in the background, and you will be e-mailed when a job finishes. Some jobs will take seconds, while more computationally intensive ones can take days. Click "Fetch Results" at the top of the screen, and you will be able to watch your job's progress. When it's done, click the green "Download" button and select "Word Cloud (PDF)" to download your word cloud. And we're done!

References

Ashton, A. T. (2011). Semantically rich tools for text exploration: TEI and SEASR. In *Digital humanities 2011, Stanford, CA* (pp. 270–271).

Brants, T., & Franz, A. (2006). *The Google web 1T 5-gram corpus version 1.1 (LDC2006T13)*. Philadelphia, PA: Linguistic Data Consortium.

Burns, J., Brenner, A., Kiser, K., Krot, M., Llewellyn, C., & Snyder, R. (2009). JSTOR – data for research. In *Research and advanced technology for digital libraries, lecture notes in computer science* (pp. 416–419). Berlin: Springer.

Burrows, J. F. (2006). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing, 22*, 27–47. http://dx.doi.org/10.1093/llc/fqi067.

Craig, H., & Kinney, A. F. (2009). *Shakespeare, computers, and the mystery of authorship*. Cambridge: Cambridge University Press.

Einstein Papers Project. (2014). *The collected papers of Albert Einstein [WWW document]*. URL http://einsteinpapers.press.princeton.edu/.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*, 133–159.

Ide, N., & Véronis, J. (1995). *Text encoding initiative: Background and context*. Dordrecht: Kluwer.

Kumar, A. (2009). MONK project: architecture overview. In *JCDL '09*. Austin, TX: ACM.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System demonstrations* (pp. 55–60). Baltimore, MD: Association for Computational Linguistics.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*, 176–182. http://dx.doi.org/10.1126/science.1199644.

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: strong limits to inferences of socio-cultural and linguistic evolution. *PLoS One, 10*, e0137041. http://dx.doi.org/10.1371/journal.pone.0137041.

Pence, C. H. (2016). RLetters: a web-based application for text analysis of journal articles. *PLoS One, 11*, e0146004. http://dx.doi.org/10.1371/journal.pone.0146004.

Rockwell, G. (2006). TAPoR: building a portal for text analysis. In R. Siemens, & D. Moorman (Eds.), *Mind technologies: Humanities computing and the Canadian academic community* (pp. 285–289). Calgary: University of Calgary Press.

Smocovitis, V. B. (1994). Organizing evolution: founding the society for the study of evolution (1939–1950). *Journal of the History of Biology, 27*, 241–309. http://dx.doi.org/10.1007/BF01062564.

Tsukamoto, S. (2002). KWIC Concordance for Windows: easy access to corpora. In T. Saito, J. Nakamura, & S. Yamazaki (Eds.), *Language and computers, English corpus linguistics in Japan* (pp. 327–340). Amsterdam: Rodopi.

van Whye, J. (2002). *The complete work of Charles Darwin online [WWW Document]*. URL http://darwin-online.org.uk/.

York, J. (2009). This library never forgets: preservation, cooperation, and the making of HathiTrust Digital Library. *Archiving Conference, 2009*, 5–10.