THE GEOLOGICAL SOCIETY OF AMERICA
GROUNDWORK
Furthering the Influence of Earth Science

Charting the Geosciences with Google Ngram Viewer

Danita S. Brandt, *Department of Earth and Environmental Geosciences, Michigan State University, East Lansing, Michigan 48824, USA, brandt@msu.edu*

INTRODUCTION

Frequency of mention in books can be used to trace the evolution of a discipline, from the first recorded use of the word or phrase to its current standing, as measured by the number of books that include the phrase. Ngram Viewer, the tool developed by a team at Google Books (Michel et al., 2011) places a database ("corpus") of >500 billion words at the disposal of its users (http://books.google.com/ngrams). Here I describe how this tool can be used to examine patterns suggested by qualitative ideas about the intellectual development of the geosciences. An example of the Ngram Viewer output is given in Figure 1.

N-GRAMS

An N-gram is a contiguous string of *n* items from a given sequence of text or speech. A 1 gram (also known as a unigram) is a string of characters uninterrupted by a space, e.g., "trilobite" or "3.14159." An N-gram is a sequence of 1 gram, e.g., "trilobite extinction" (2 gram or bigram), and "Michigan State University" (3 gram or trigram). N-grams are used by computer scientists and computational linguists for text mining and natural language processing (Jurafsky and Martin, 2014). Google Books, a service of search-engine giant Google Inc., has amassed a database of more than 25 million scanned books. From this resource, a subset of over five million books, chosen for the quality of their optical scan and metadata (e.g., date of publication), comprises the corpus of Google Ngram Viewer. Currently, Ngram Viewer is restricted to a maximum word string length of *n* = 5 (five-grams), and counts only N-grams that occur at least 40 times in the corpus. The data consist of books published from the 1500s to 2000, and includes children's literature, trade, and other books but no journal articles. The full data set is available at www.culturomics.org and www.ngrams.googlelabs.com.

CAVEATS TO USING THE CORPUS

Problems with the unfiltered use of the Google Books corpus are well-documented, including errors introduced during optical scanning and entering metadata (Nunberg, 2009). Pechenick et al. (2015) described limits to inferring cultural and linguistic evolution from the Google N-gram corpus, including the problem of the burgeoning number of scientific texts since 1990, which skews the results toward academic usage of N-grams and is therefore less reflective of cultural context. However, if the user's purpose is to trace the history of a scientific discipline rather than a cultural phenomenon, as the purpose is here, the bias Pechenick et al. (2015) described skews in a constructive direction. Because the database consists of books only, rather than journal articles, N-gram results might lag the intellectual development of a discipline.

APPLICATION TO THE GEOLOGICAL SCIENCES

Ngram Viewer is useful for suggesting testable hypotheses by identifying correlations. Two important caveats to keep in mind when using Ngram Viewer are, as in any analysis, correlation does not necessarily indicate causation, and, as with any online resource (Wikipedia, for example), Ngram Viewer provides a starting point to stimulate further investigation, not an end in itself. Here, in approximate chronological order, are three examples of Ngram Viewer searches drawn from geological topics chosen to illustrate the potential and the limitations of these data. Search terms and phrases (the N-grams) are enclosed in quotes.

The frequency of the unigram "geology" shows an increase at 1830, coincident with publication of the first volume of Charles Lyell's *Principles of Geology*. Volume one was followed by volumes two and three in 1832 and 1833, respectively. The N-gram frequency chart supports the hypothesis that Lyell's books contributed to an increase in the frequency of the unigram "geology"; the conclusion that Lyell's work had a major impact on the growth of geology is supported independently by historians of our discipline (Rudwick, 2010).

N-gram frequency of "micropaleontology" reached a maximum in the early 1950s, coincident with that decade's "petroleum" boom, and reflects the well-documented connection between microbiostratigraphy and petroleum exploration (Haq and Boersma, 1998). However, not all possible correlations are easily tested using Ngram Viewer; an attempt to chart the N-grams "micropaleontology" and "petroleum" on the same graph returned a display in which the line tracing the frequency of "micropaleontology" was indistinguishable from the *x*-axis; the frequency of the N-gram "petroleum" swamped "micropaleontology." The corpus is also sensitive to N-gram size and word order; the trigram "extinction of trilobites" successfully returned results; a query for "trilobite extinction" returned no N-grams. Although Ngram Viewer does not allow for easy comparison of N-grams with wildly different occurrence rates, this obstacle can be overcome by downloading and replotting the Ngram Viewer data using programs such as *R*.

Cause-and-effect is suggested by the graph of "geosynclines" and "plate tectonics" (Fig. 1). The graph traces the displacement of the older "geosynclines" paradigm for explaining crustal tectonics by the emergence of "plate tectonics." The dramatic shift from "geosynclines" to "plate tectonics" occurred in the mid-1970s, as plate tectonic theory supplanted the pretectonic explanation of crustal dynamics and made its way into textbooks. The apparent causal connection between the

GSA Today, v. 28, doi: 10.1130/GSATG348GW.1. Copyright 2018, The Geological Society of America. CC-BY-NC.

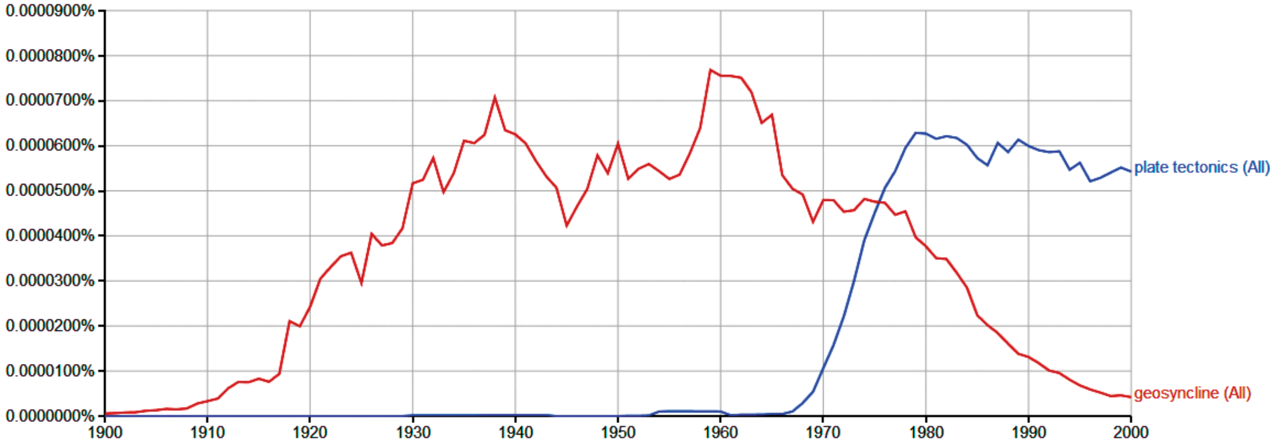

Figure 1. Screenshot of Ngram Viewer chart showing the frequency in the Google Books corpus of the N-grams "geosyncline" and "plate tectonics," from 1900 to 2000. *Y*-axis is frequency of the N-gram in the corpus.

rise of plate tectonics and the fall of geosynclines can be examined more closely by accessing the corpus on which the search is based. In addition to the chart (Fig. 1), Ngram Viewer searches return links to the corpus on which the search is based, binned by year of publication. Clicking on these bins opens a Google search page with links to each publication included in the corpus. The diligent researcher can then sort through the titles and assess the quality of the data on which the Ngram Viewer chart is based.

OTHER USES FOR N-GRAMS IN THE GEOSCIENCES

Charting word frequency trends can contribute to identifying directions for research or investment of resources. In the U.S., a number of Departments of "Geology" became Departments of "Geological Sciences" in the late 1970s and early 1980s (including the department at Michigan State University), mirroring the increase in frequency of the bigram "geological sciences." In 2016, MSU's department changed its name, again, to "Earth and Environmental Sciences," reflecting the increase in frequency of the "Environmental Sciences" bigram, which started in 1990. The N-gram frequency of other geologic disciplines also chart what might be interpreted as evolving priorities, especially in the textbook-rich academic environment: References to "evolutionary biology" now approach those of "paleontology." As frequency of the bigram "evolutionary biology" increased, through the mid-1970s, the Paleontological Society debuted its new journal, *Paleobiology*.

The decisions to change department names, revise course descriptions, and initiate new journals described here were made before there was a Google Books corpus, but these decisions were undoubtedly affected by trends in metrics, like student enrollment and funding priorities, which are now indirectly reflected in that database.

SUMMARY

The output of Google's "shiny new toy for nerds" (Zhang, 2015), Ngram Viewer, is not sufficient to support hypotheses of causality suggested by the correlations it generates, but its accessibility and ease of use can serve an important function in introducing scholars to the possibilities of digital research (Cohen, 2010). The frequency of N-grams through time maps where we have been, and, mindful of the adage, "those who cannot remember the past are condemned to repeat it," history ought not be ignored in identifying trends in support of education, policy, planning, and funding objectives of our discipline.

ACKNOWLEDGMENTS

A.M. Velbel introduced me to Ngram Viewer and was instrumental in the evolution of this manuscript. Three reviewers contributed to a more focused and improved final version.

REFERENCES CITED

Cohen, D., 2010, Initial thoughts on the Google Books N-gram Viewer and datasets, http://www.dancohen.org/2010/12/19/initial-thoughts-on-the-google-books-N-gram-viewer-and-datasets/ (last accessed 10 May 2017).

Haq, B.U., and Boersma, A., eds., 1998, Introduction to marine micropaleontology (2nd edition): Amsterdam, Elsevier, 376 p.

Jurafsky, D., and Martin, J.H., 2014, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd edition): New York, Prentice Hall, 1024 p.

Lyell, C., 1830, Principles of geology, being an attempt to explain the former changes of the Earth's surface, by reference to causes now in operation: London, John Murray, volume 1.

Lyell, C., 1832, Principles of geology, being an attempt to explain the former changes of the Earth's surface, by reference to causes now in operation: London, John Murray, volume 2.

Lyell, C., 1833, Principles of geology, being an attempt to explain the former changes of the Earth's surface, by reference to causes now in operation: London, John Murray, volume 3.

Michel, J.B., Shen, Y.K, Presser Aiden, A., Veres, A., Gray, M.K., Brockman, W., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., and Lieberman Aiden, E., 2011, Quantitative analysis of culture using millions of digitized books: Science, v. 331, p. 176–182, https://doi.org/10.1126/science.1199644.

Nunberg, G., 2009, Google's book search: A disaster for scholars: The Chronicle of Higher Education, http://www.chronicle.com/article/Googles-Book-Search-A/48245/ (last accessed 10 May 2017).

Pechenick, E.A., Danforth, C.M., and Dodds, P.S., 2015, Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution: PLoS One, v. 10, no. 10, https://doi.org/10.1371/journal.pone.0137041.

Rudwick, M.J.S., 2010, Worlds before Adam: The reconstruction of geohistory in the age of reform: Chicago, University of Chicago Press, 648 p.

Zhang, S., 2015, The pitfalls of using Google N-gram to study language, https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-N-gram/ (last accessed 10 May 2017).

Manuscript received 11 May 2017
Revised manuscript received 3 January 2018
Manuscript accepted 7 February 2018