2024

How'd They Make That?: Analyzing the Ingredients that Produce a Google Ngram Chart

Laura Baker
Abilene Christian University

How'd They Make That?: Analyzing the Ingredients that Produce a Google Ngram Chart

Laura Baker, Abilene Christian University, Librarian for User Experience and Assessment, bakerl@acu.edu

Nutrition Information

Google Ngram is an online tool that graphs the relative frequency of words or phrases that appear over time in the millions of books digitized by Google. It is often used by digital humanists to analyze the popularity of topics, to track the rise and fall of words throughout the ages, and to explain changes in culture and history. Google Ngram is free and requires very little training to use, making it a popular tool for many. It is astonishingly convenient. Anyone can create an Ngram chart in a matter of seconds. But what are we really getting?

This recipe takes a deep dive into how Google Ngrams work. We will explore how Google Ngram charts are produced, what factors influence the results, and what conclusions may be drawn from the output.

Learning Outcomes

- Become familiar with Google Ngram as a text analysis tool.
- Learn to construct a Google Ngram.
- Realize the strengths and limitations of the Ngram corpora and draw conclusions from the chart.

- Develop critical thinking skills that can apply to other types of text analysis in digital humanities.

Number Served

15-20 participants

Cooking Time

Appropriate for two 50-minute workshops or spit into smaller bite-sized portions over multiple sessions.

Dietary Guidelines

The overarching nutrient is one that supports critical thinking and analysis, particularly with respect to inherent biases and assumptions. Specifically, this recipe supports the following items from the ACRL Information Literacy Framework:

- Authority is Constructed and Contextual – We are using the Ngram tool as an indicator of authority and are learning what might temper this credibility.
- Information has Value -- Some individuals or groups may be underrepresented or systematically marginalized within the systems that produce and disseminate information.
- Research as Inquiry – We are encouraged to draw reasonable conclusions based on the analysis and interpretation of information.

Ingredients & Equipment

- Computer and projector for the facilitator
- Shareable online folder/dropbox where class can deposit their work (e.g. shared folder on Google Drive)
- Computers/laptops for students
- Supplemental document “Google Ngrams Illustrated” with suggested examples (give url)

Preparation

- Review the search tips for Google Ngram at https://books.google.com/ngrams/info.
- Create a Google Folder or similar place for the class to deposit their work. For simplicity, it is helpful to share the folder with “anyone with the link” and to create a short url so the link can be shared conveniently with the class.
- Prepare 2-3 general examples to demonstrate the tool. See Supplemental document for suggestions.

Cooking Method

Day One

1. Explain what Google Ngram is and its use in digital humanities as a text analysis tool. Text analysis identifies word patterns across a large number of digitized texts. The Ngram Viewer lets you type in a word or phrase and see how often these phrases have occurred over time in the books Google has digitized. From

this, researchers make conclusions about word use, culture, language, the rise and fall of various words, etc.

2. Demonstrate a basic Ngram from your prepared examples to show the class.
3. Ask students to generate an Ngram of their own using basic terms from their field or interests, especially one that might show changes over time in written works. Encourage experimentation. Subjects and graphs improve as students play with the tool.
4. Ask students to screenshot their Ngram to a Google Doc, write a brief paragraph explaining why they chose these terms and what the graph indicates, and save it to the Google folder you created for this purpose. Time permitting, ask for volunteers to share their Ngram with the class.

Day Two

1. Explain that Ngram can be useful but it also contains caveats that can affect results. Discuss and demonstrate how some of these can bias the Ngram:
    a. Manipulating scales – The Y-axis represents frequency and automatically adjusts to fit the space. This means the chart can be adjusted to make results seem less or more dramatic. **See search of thee,thy,thou adjusted by decades**. Notice how adjusting the time frame can emphasize or de-emphasize the amount of the increase after 2000. These are the same frequencies, but one looks much larger than the others.

b. Corpus errors -- Many of the 2009 and earlier corpora are not identified in the correct publication year (Younes; Zhang). Others incorrectly counted word proximity even when it crossed sentence boundaries but failed to count a valid proximity if it crossed page boundaries. Google also admits that optical character recognition (OCR) errors plague pre-1800 digitized texts (Google). **Search example: geek.**

c. Sampling bias – Books before 2000 were scanned from selected university libraries. Books after 2000 were from publishers who deposited their books with Google. This means the books behind the Ngram Viewer reflect the perspectives of scholarly agendas, of those with advanced education, and – later – of publishers' choices (Pechenick, et al. 2, Kestler; Younes). It is not a sample of all books nor of what the general population currently reads. It is a sample of what other groups chose according to their interests. **Example: berattle**. Examine apparent increase after 2000 by looking at the matching texts. Most of the matches occur in modern reprints of older works. This shows the influence of publishers on the overall sample.

d. Genre bias – The pre-2000 corpus has a preponderance of scientific publications. Later books are more heavily skewed toward fiction. There are tremendous differences in the writing styles of these genres, in their word choice, and in the interests of scientists versus fiction writers. These differences can influence results. The corpus also ignores serials and newspapers, meaning there is less representation of local terms and

everyday language. **Example: tissue,kleenex**. Much of the use of “tissue” is due to the science-heavy books before 2000, a genre that inflates the word frequency yet not in the intended context of “tissue” as a paper handkerchief.

e. Confirmation bias -- Usually the words researchers enter into Ngram are words they suspect will have an interesting graph. They suspect the graph will show something, so it is easier for them to see what they expect. **Example: environmentalism**. Results increase 1950-2000 after the publication of the book *Silent Spring*, but does this mean *Silent Spring* accounts for that increase?

f. Print vs. Culture – **Example: Frodo,Captain Kirk**. The graph shows declining use but ignores the popularity of that text among subcultures of fantasy and sci-fi literature. It is easy to make false assumptions about culture based on printed works. But print culture does not equal culture as a whole. To assume it does is to assume that culture is one homogenous, undifferentiated body having a common set of shared meanings with little variation. The reality, however, is that there are many diverse groups with unique language and word use that may not be reflected in the larger print corpus. What about magazines and newspapers produced by and for ethnic, racial, indigenous, or immigrant populations? What about the pulp fiction novels read by many in the 1930s, or manga popular with many readers today? Culture consists of diversity and if not represented, can

easily be subsumed by the larger voices that make it into mainstream print.

g. Print vs. popularity – There is a tendency to think that frequent appearances of a phrase in print means that term is more popular in society. This may or may not be true. **Example: Hercule Poirot,Rhett Butler**. Does this mean one character is more popular than another, or does it simply mean that one was written about more than the other? Consider, too, that Agatha Christie wrote many, many books with the character of Poirot, whereas *Gone with the Wind* featuring Rhett Butler was the only book that author wrote. This is also a good example of how a prolific author can dominate results over another with fewer books but no less influence.

2. After demonstrating Google Ngram and exploring the validity of inferences based on its results, ask students to return to their initial Ngram and write an additional paragraph about the possible limitations of the chart. They may also choose to refine their Ngram after learning more about how the tool works.
3. Summarize Lessons Learned

   Discussion prompt: Google Ngram is powerful. It lets you instantly examine millions of texts, more than you could possibly read on your own. Consider the implications of that last statement. You are making inferences about books that you have never actually read. What are the advantages of that? The disadvantages? How can the tool be used to minimize the negatives? Encourage the class the reflect on these questions and discuss them.

Key points:

- Google Ngram is a convenient tool, but it is not perfect. It is important to acknowledge its limitations, to recognize what conclusions are supported, and what may not be.
- Be cautious about accepting a time series where frequency is attributed to an external event (like *Silent Spring*). There are way too many other variables occurring at the same time that could account for a change.
- Be very wary of Ngrams that claim a term is popular in culture or everyday language based on its appearance in books. Word frequency is not a measure of pop culture, nor is the formal academic writing that much of the Google corpus is based on an indicator of how people speak everyday.

Main takeaway: Above all, understand the corpus. What is included? What is being left out? Who is overrepresented? Who is underrepresented? Whatever biases are in the sample will bias your results. Whoever or whatever centralizes information and controls the presentation of it can affect what ideas and names mean to society.

Chef's Notes

An interesting variation – and one that adds more spice – is to have students find examples of Google Ngrams in published research or social media (Twitter, Substack, Reddit, blogs, etc.) and critique it.

Additional Resources

Google, “Google Books Ngram Viewer.” 2024. https://books.google.com/ngrams/info. Accessed 7 Oct. 2024.

Kestler, Thomas. “Big Data and Ideational Institutionalism. Reconsidering the Possibilities and Limitations of Google Ngram in the Study of Ideas.” *Politische Vierteljahresschrift*, Jan. 2024. *EBSCOhost*, https://doi.org/10.1007/s11615-024-00569-4.

Pechenick, Eitan Adam, Danforth, Christopher M. Danforth, and Peter Sheridan Dodds. “Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution.” *PLoS ONE* 10, no. 10 (2015): e0137041. https://doi.org/10.1371/journal.pone.0137041.

Younes, Nadja. “Guideline for Improving the reliability of Google Ngram Studies: Evidence from Religious Terms.” PLoS ONE 14, no. 3 (2019): e0213554. https://doi.org/10.1371/journal.pone.0213554.

Zhang, Sarah. “The Pitfalls of Using Google Ngram to Study Language.” *Wired*, October 12, 2015. https://web.archive.org/web/20230318034709/https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-Ngram/.