Visualizing the English Lexicon with Google Ngram

| Andrew Kan  UW  Seattle  andrewk7@  uw.edu | Nathan Foster  UW  Seattle  njfoster@  uw.edu | Jon Shilling  UW  Seattle  jshill@  uw.edu | Kinsley Tran  UW  Seattle  kinstran@  uw.edu |
| --- | --- | --- | --- |

# ABSTRACT

Paste the appropriate copyright/license statement here. ACM now supports three different publication options:

* ACM copyright: ACM holds the copyright on the work. This is the historical approach.
* License: The author(s) retain copyright, but ACM receives an exclusive publication license.
* Open Access: The author(s) wish to pay for the work to be open access. The additional fee must be paid to ACM.

This text field is large enough to hold the appropriate release statement assuming it is single-spaced in Times New Roman 8-point font. Please do not change or modify the size of this text box.

Each submission will be assigned a DOI string to be included here.

Data from Google Books and Google Ngram has been analyzed many times over by various groups of researchers for many purposes. However, there is a lack of good visualizations regarding this enormous data set. We argue that there is great potential for Google Ngram data to be visualized, and laid out our method for attempting such a visualization in this paper. We took a brief look at the existing literature outlining insights gathered by researchers across fields, and used it as a starting point for our own observations. We explored the Google Ngram data set for noteworthy examples of changes in lexicon.

We took these examples and produced a visualization that allows users to interact with a small section of this enormous dataset. Due to performance concerns, inherent in visualizing hundreds of gigabytes of data, we decided to only use a small number of examples to enable a guided discussion of our findings through the visualization. Future work would be focused on expanding the size of usable data.

# INTRODUCTION

The origin of Google Book Ngram lies in a paper by Jean-Baptiste Michel et. al. titled *Quantitative Analysis of Culture Using Millions of Digitized Books.* In it, they describe their process of analyzing millions of books curated by Google books, and in the process generating a data set dubbed “Ngram.” This method was essentially to count the number of occurrences of phrases of n words in books, which produced a result set spanning millions of word combinations in millions of books, spanning hundreds of gigabytes of information. The separation by n number of words is what yields the term “Ngram.” From this data, Google released the Google Ngram Viewer, a publically available interface for accessing the data set.

The Google Ngram Viewer, while providing a basic level of utility, leaves much to be desired. It serves its purpose as a way to search for basic trends, but it provides no context or story for users to enjoy. This lessens the impact that such a large data set like this should have, and it was for the purpose of creating that story and context that we decided to create our own visualization drawing from prior research done using this data.

Such a visualization should make it easier to sample what the data set is, and how it can be used. It should contain several examples of trends present in the data, and should allow users to explore those trends for themselves. In making an algorithm to compute this visualization, great care should be taken in order to maximize the efficiency of producing charts, so as to make the experience as seamless and responsive as possible for the user.

The challenges we face while attempting this visualization will center on choosing only a few facets out of the possibly infinite stories we could choose to tell with the terabytes of information we have at our disposal. We chose to implement this visualization as a series of stories told through the lens of certain phenomenon present in the data. Using a select number of pre-defined stories is intended to ease the user into a vast data set, so they have enough context to not get lost.

Due to size and performance constraints, our algorithm will need to be able to iterate through the entire Ngram data set and make a list of Ngrams that are relevant to our visualization, and contain no other Ngrams. This will require a modest level of processing, the output of which we will further need to modify in order to use it most optimally in our final product.

# related work

The basis of Google Ngram comes from a paper that outlines how the data set was created in the first place [2]. In this paper, the authors describe their method for using OCR to scan millions of pages of books in order to aggregate their word counts. These word counts are expressed as the number of occurrences of that ngram across the entire corpus of works, for every year in the data set. In addition to single words (a “1-gram”) the researchers also performed the same aggregation on phrases of up to 5 words. During this aggregation process the team assigned meta data to each ngram, such as part of speech or most common replacements. These extra variables allowed the researchers to account for homonyms and other peculiarities within language, meaning they could make sure the words they were searching for were the correct versions of those words.

Other authors and teams have used the Google Ngram data set to highlight interesting trends, from whimsical pieces about profanity and other oddities [3], to more serious pieces focused on analysis of language and cultural trends present in the data. In all cases, these are examples of research that’s built off of the original paper above. However, most of these efforts were limited to what they could get out of Google’s provided visualization tool, the Google Ngram Viewer. This viewer provides basic search functionality on a limited set of inputs, and requires the user to think of words to search for themselves, making true exploration of the data very difficult.

Some even write about the limitations of the data set and its validity in accurately recording trends. Among the more interesting points related to these issues was an article in Wired written by Sarah Zhang [1]. Among the many potential pitfalls in using the data set, she pointed out OCR (optical Character Recognition) as one of the biggest, positing that errors in computer analysis of the text on a page might lead to significant errors. Other significant biases in the data can be attributed to a lack of diversity in the corpus of works used (a meaningful percentage of works used were either sermons or scientific articles) or that, since each book is only included once, the data represents “what is popular to publish rather than what is popular to read.”

# methods

We identified the following problems as key in addressing the lack of visualization of this data: the data set is too big, and the data exists in its current form with no context. To solve these problems, we applied our own analysis to prior research, combined the most interesting results, and combined them into a cohesive story to present to the user. We took those stories and used them to determine the data we needed to scrape from the main Google Ngram data set.

To find our information, we researched what analysis had already been done on this information, and set out to representative examples of that analysis in the Ngram data. Our criteria for research analysis didn’t limit us to scholarly articles. Often, in fact, we relied on popular internet sites to supply us with interesting trends that have been explored through the Google Ngram Viewer. Our justification for this was that it would provide us with a more diverse set of stories to tell for our visualization. Once we had these we crafted narratives that would allow is to put each of those examples in context, allowing the user to understand the significance them.

In order to deal with the terabytes of data that make up the Google Ngram dataset, we decided to filter that data set to only use what we needed to present our story, rather than provide the entire dataset. This mainly narrowed down our data set to a list of words used in the analysis we found. By taking only a tiny set of words from the data set, we arrived at a file size that was reasonable to use in a JavaScript visualization.

To filter the information, we devised an algorithm that would search through all ngrams for the words we wanted to use, and construct a miniature list of Ngrams that corresponded directly to a dictionary of words we provided. In order to save processing time, we only selected original Ngram data that started with letters we knew our dictionary words started with. Hence, we never have to iterate through the list of Ngrams that started with letters like x or z, or that started with numbers. All the results were combined into a text file of the same format of the original Ngrams. From there we were able to convert that text file to a TSV for use in the visualization.

# results

As a result of our data optimization, all visualizations render nearly instantaneously, providing an optimal user experience. Coupled with a modern front-end framework, our final project is responsive, fast, and streamlined. The final result is a web site where users can interact with the analysis we’ve presented, and read about some of the research that’s been done in this area and what our own insights into that research are.

After constructing our data structure and determining the contents of our visualizations, we decided we would use line charts to communicate our ideas. Each of these line charts communicate on of the specific topics we wished to visualize. The scales are set up in order to highlight the curve of the line rather than highlight the specific frequency values, as we determined that the values themselves are less important than the change of those values over time.

The visualization that users will see will guide them through a series of charts, each a unique “story” about trends seen in the data set. Users will scroll through each story and learn more about each trend. We start with the most frequent words present English language literature in every decade since 1800. From there, users will see information about what topics have been popular throughout time, including technology.

From there, we will explore a series of trends that tell us about how people spoke and what they spoke about. We use the data set to depict how historical events enter the lexicon suddenly, then continue to see prevalence in later years. We also explore how profanity, the crux of informal language, transforms over time, or perhaps, even how it stays the same (some curse words have surprisingly long histories).

# Discussion

We hope our visualization work has made the research in this field more visible and accessible. By visualizing this data set and providing context through annotation, users should be better equipped to understand the basics of what you can do with Google Ngram, and be excited to make their own discoveries with this data set.

We began this project with many questions, and by going through the process of trying to answer them, we’ve only found more. The vastness of this data set is even bigger than the terabytes of information it represents digitally: it contains a decently representative record of an entire culture, and any use of this data set should reflect that fact. By interacting with a huge data set such as this and attempting to visualize that hugeness in a limited amount of space was as much a technical challenge as it was a conceptual one. While our goal was to enhance cognition and not introduce more vastness, We hope that these revelations are evident to the audience.

The audience should not conclude that what we have chosen to visualize is representative of all the work in the field of etymology or data science, but rather only that there is a vast amount of information that can be gleaned from this data, and that it is worthy of further study. We had to work to find the examples we displayed; with this visualization, we hope the audience takes advantage of that time and instead uses theirs to go beyond what we’ve done.

# future work

Throughout this process our biggest constraint was working with such an impossibly large data set. We would like to see an extension of this visualization that can utilize the entire Google Ngram data set. This would allow users much greater freedom in exploring our examples, rather than being constrained to a small amount of possible options. This would encourage a greater degree of interactivity, enabling users to make their own discoveries.

We would like to explore the possibility of making the entire data set query able as an API, hosted on an appropriate site that would be able to store all the necessary assets. We believe that if researchers had a way to access all of the Google Ngram data set instantaneously, rather than through downloads or even hard copies available directly from Google, that more work would be able to be done in this area. Such a solution would be very complex, but would yield many rewards. If such an API were open to the public as well, that would be even better.

We would also like to explore adding supplemental data sources to provide more context for our examples. As it stands, there is only so many conclusions you can come to with such a data set. It is one dimensional, and therefore is limited in the insights it can give. Several data sets exist that would enable us to tell more complicated stories, examples of which might be everything from other, similar word frequency analysis with different corpuses of work, or even sources of data that have nothing to do with word frequency, but that might paint a bigger picture when combined with what we already have. With these, we could even validate some of the research in these articles and prove or disprove the merits of using Ngram as an analytical tool, much as Sarah Zhang touched on in her piece in Wired [1].

# REFERENCES

1. Sarah Zhang. The Pitfalls of Using Google Ngram to Study Language. 2015. Retrieved May 31, 2017 from <https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/>
2. Jean-Baptiste Michel\*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden\*. *Quantitative Analysis of Culture Using Millions of Digitized Books*. **Science** (Published online ahead of print: 12/16/2010)
3. XKCD. Ngram Charts. Retrieved May 31, 2017 from <https://xkcd.com/ngram-charts/>