PhilologyBot Process Book

Parker Coady, Isabel Alvarado Blanco Uribe, Maya Sun, Yoni Weiner

Overview & Motivation

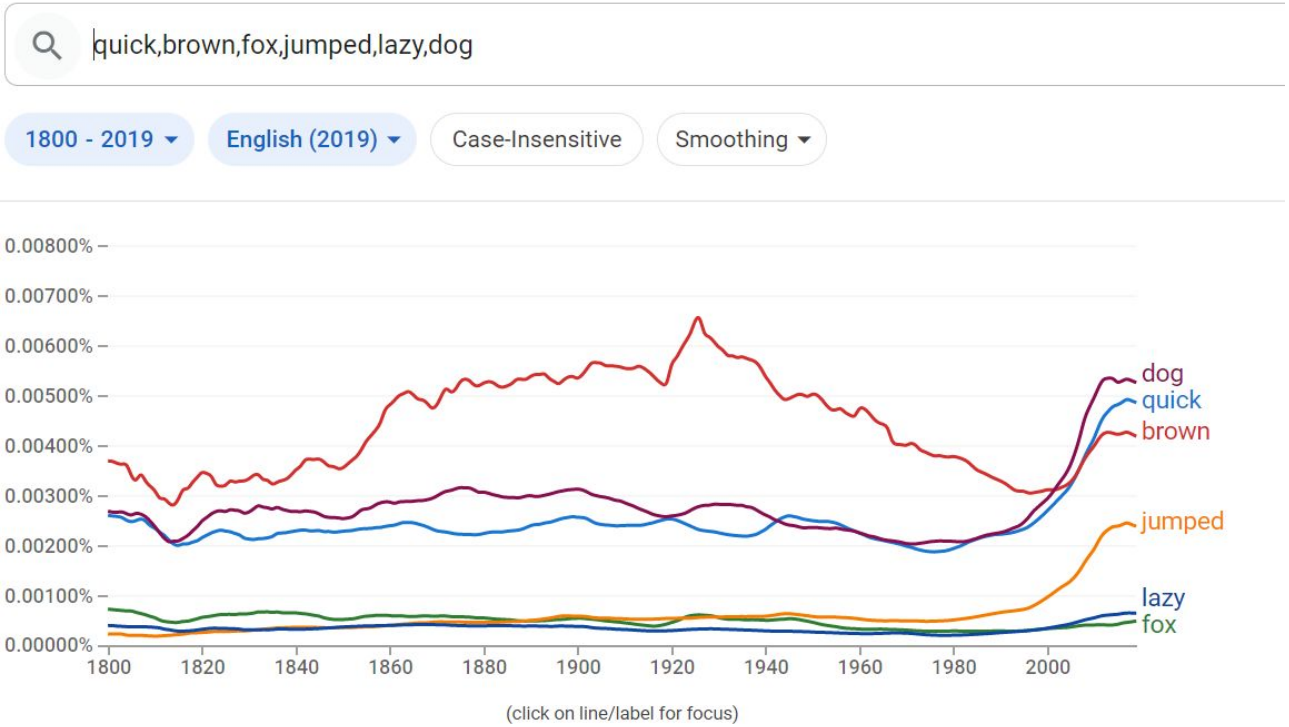
Project goal: A tool that allows people to explore how old the words in a sentence are.

Motivation:

Litographs - pictures composed of words (https://www.litographs.com/)

(https://verbosedavinci.github.io/Hyper-Text-DaVinci/)

Related work: Google Ngrams



Questions to Explore

- How old are the words that we usually use?
- How often do we use really old words?
- How do word frequencies change over time?

Data

To acquire the data, we created an AWS Lambda function which can be called through an API that gets data from the Google nGram viewer. To program our API, whenever a change is made to the `lambda_function.py` in the GitHub repository, GitHub Actions creates a zip of the `api/` folder and sends that to an S3 bucket for use.

The data is processed in the API, which returns a JSON object containing a 201xN dictionary, where N is the number of unique words passed into the API and 201 is the number of years of data that is present. The API is only able to get data for twelve words at a time, so we break up each API call into groups of 12 words and merge them into the main data table containing years and word use.

Though this process is not as fast as we would like, we believe it is far superior to loading in the raw dataset which is multiple gigabytes in size.

Ngram Viewer Exports

https://storage.googleapis.com/books/ngrams/books/datasetsv3.html

English

Version 20200217

- 1-grams
- 2-grams
- 3-grams
- 4-grams
- 5-grams
- Dependencies

Version 20120701

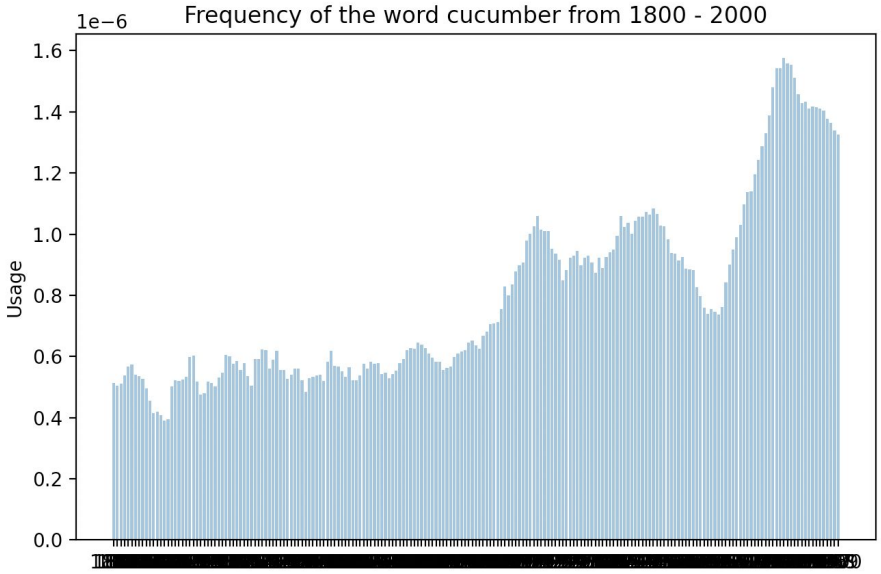
total_counts

1-grams 0 1 2 3 4 5 6 7 8 9 a b c d e f g h i j k l m n o other p pos punctuation q r s t u v w x y z

Exploratory Data Analysis

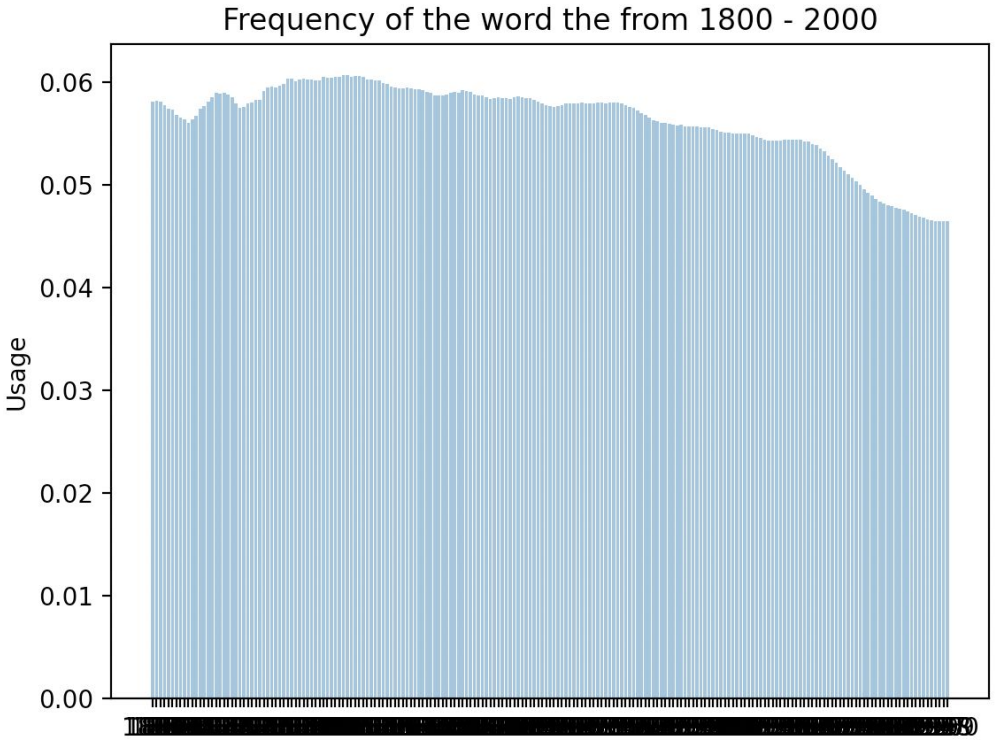
What visualizations did you use to initially look at your data? What insights did you gain? How did they inform your design?

Created with python



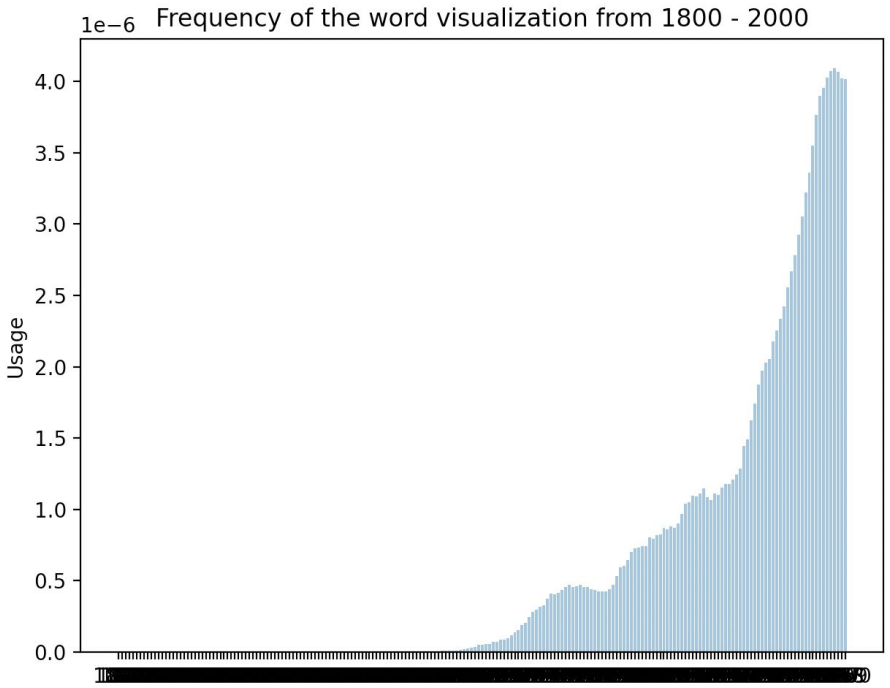
- Lots of years

Created with python



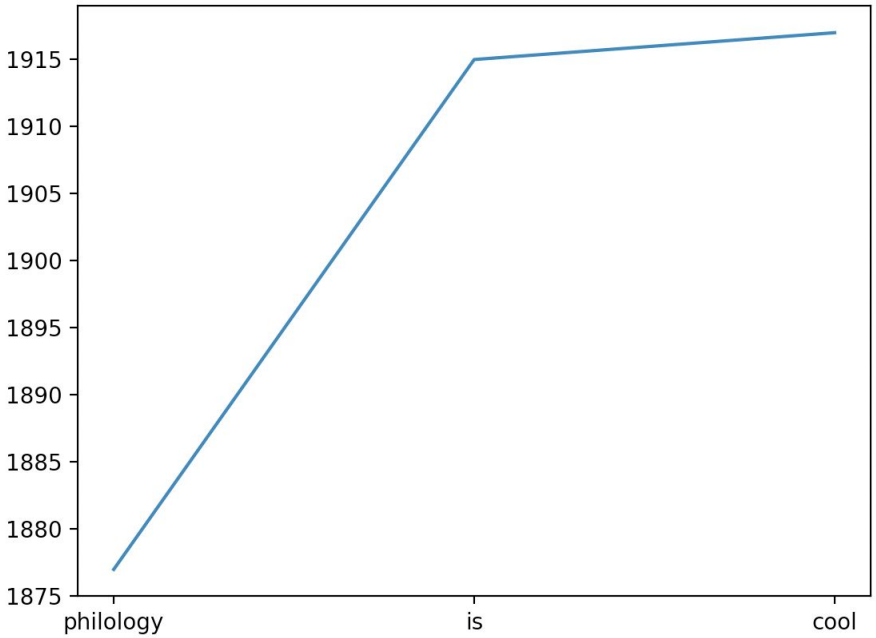
- Conjunctions are more uniform than other words
- Frequency range is very large

Created with python



- Showing the years and comparing them is going to be difficult since the range is very large

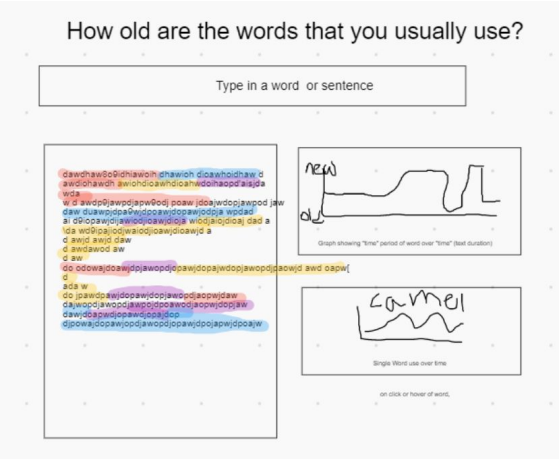
Max year of words in a sentence

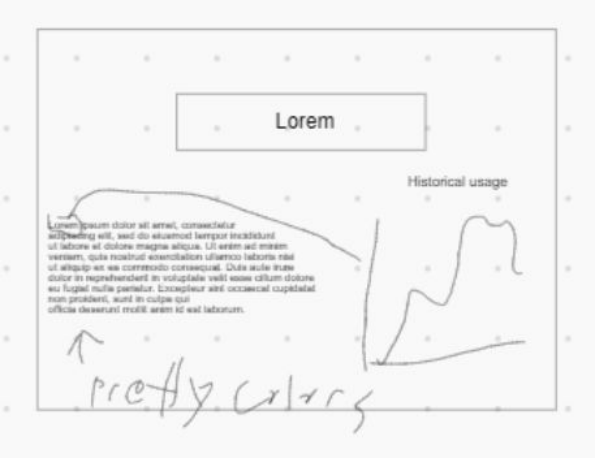


- Interesting to see when certain words in a sentence were the most common

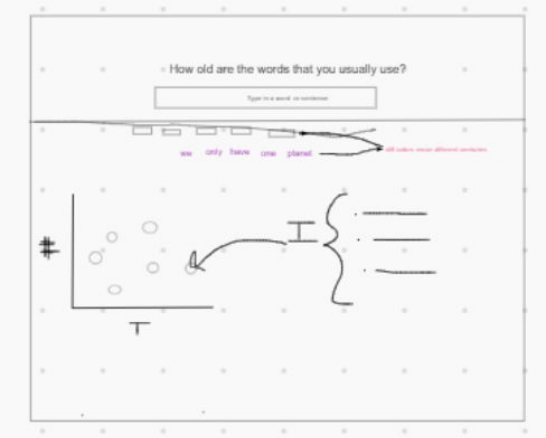
Design Evolution

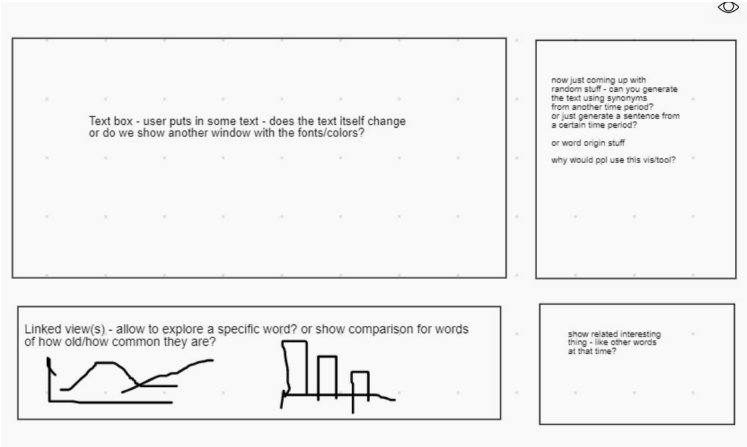
What are some diff visualizations you considered?


How old are the words that you usually use?
Type in a word or sentence
Graph showing "time" period of word over "time" (text duration)
Single Word use over time
on click or hover of word,


Lorem
Historical usage
pretty colors

merge


How old are the words that you usually use?
Type in a word or sentence
we
only
have
one
planet
#
T


Text box - user puts in some text - does the text itself change
or do we show another window with the fonts/colors?
now just coming up with
random stuff - can you generate
the text using synonyms
from another time period?
or just generate a sentence from
a certain time period?
or word origin stuff
why would ppl use this vis/tool?
Linked view(s) - allow to explore a specific word? or show comparison for words
of how old/how common they are?
show related interesting
thing - like other words
at that time?

How old are the words you usually use?

Enter text
(multiple words)

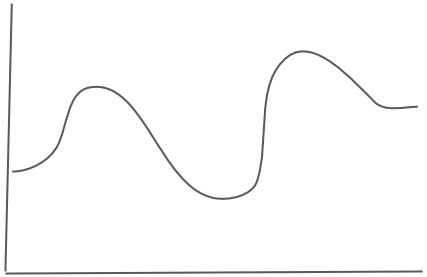
Show colors
& fonts prettily

1800s

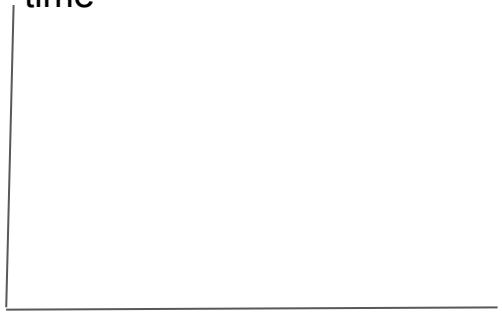
1900s

Possible settings / description / legend

Age vs sentence linked view (from yoni's)

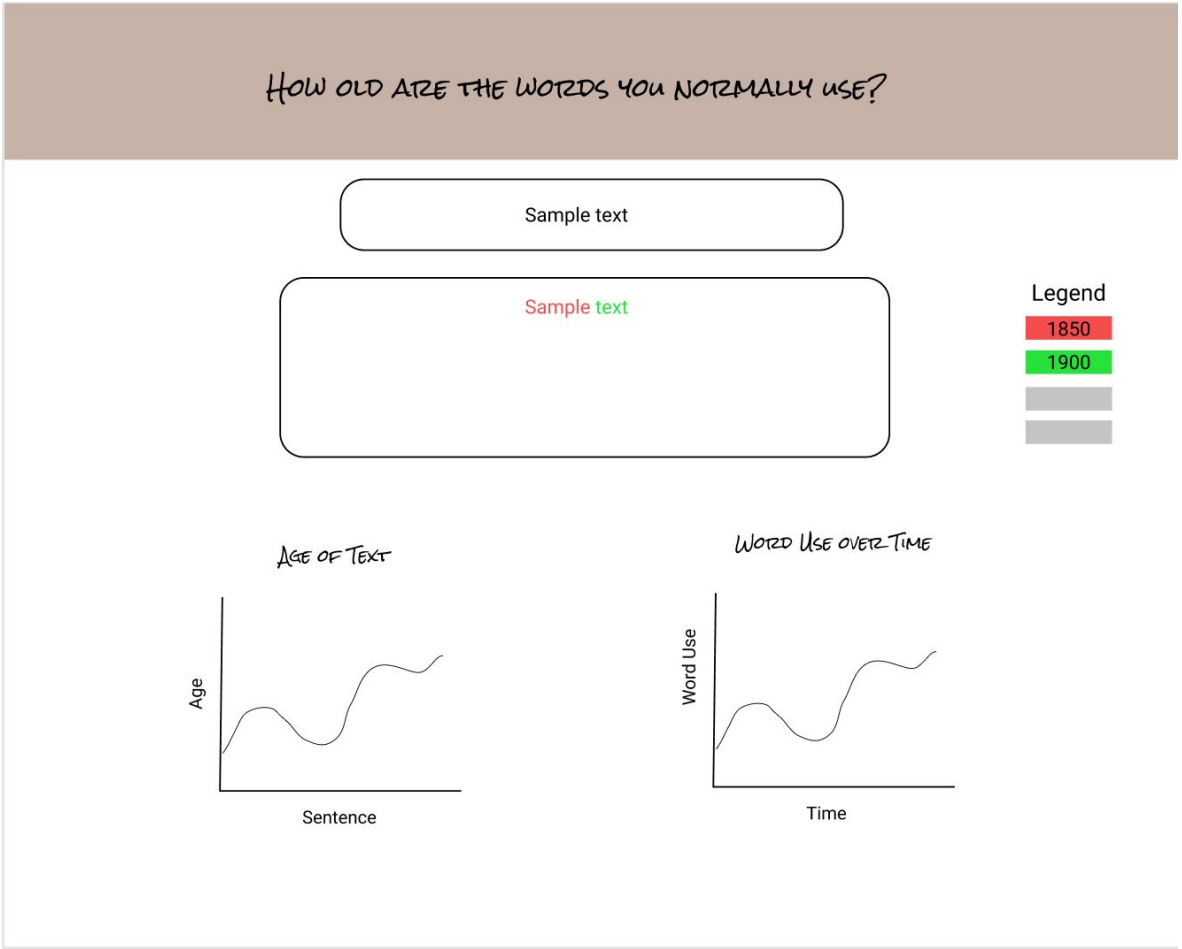


Specific word - word use over time



? possible related words thing if time

Design created with Figma



Implementation

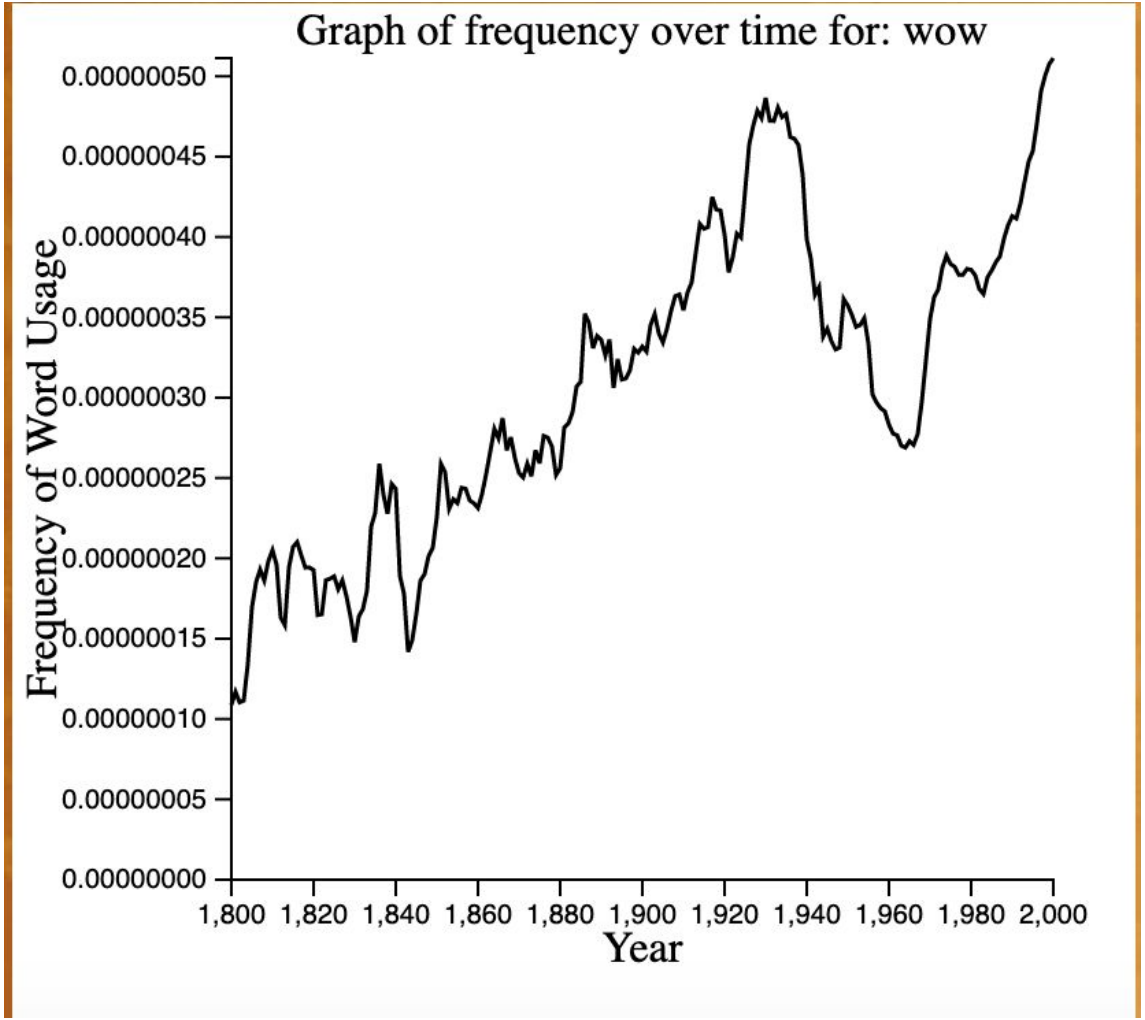
Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

Enter text here

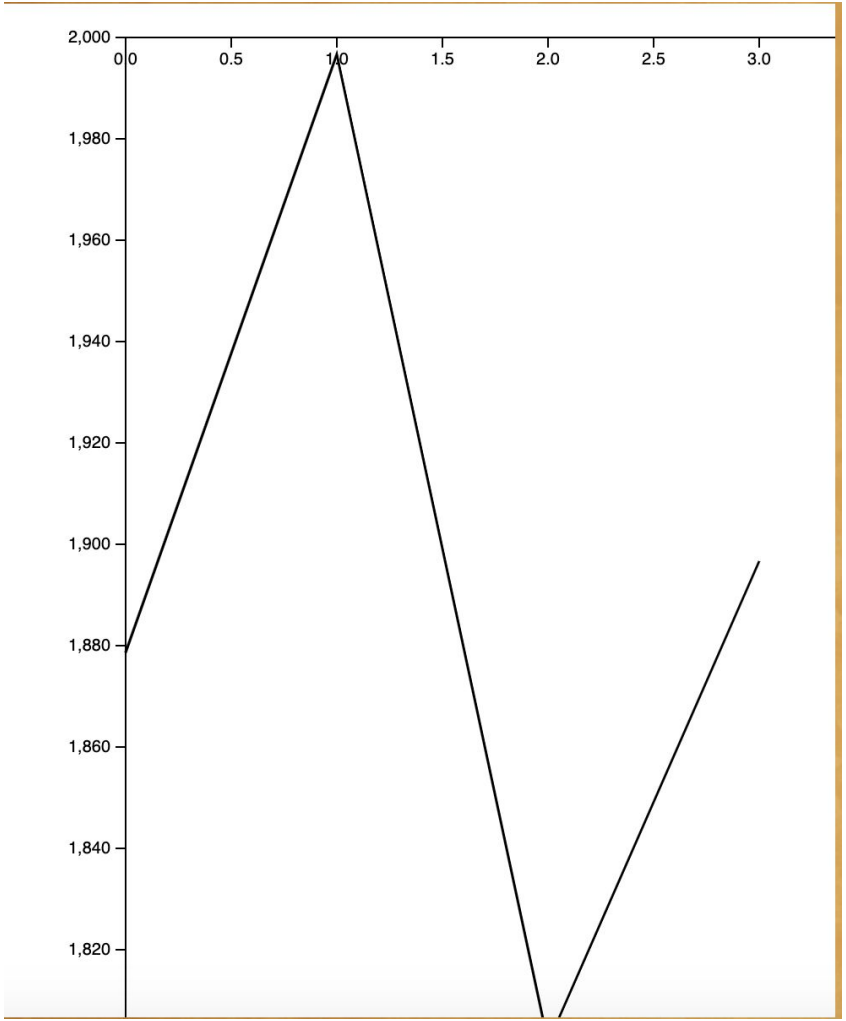
Processed passage shows up here

You can click on the words to select the graphs

Legend of the text is here.
The font correlates to the font most frequently used in that decade and the colors are separated by 50 year spans. Color gradients are used within the 50 year buckets.



Graph of frequency over time of the selected word wow



Average year of sentences in the passage

- Sentence order is based on index

Evaluation

Through this vis we were able to view the use of words over time and see what kinds of words (and fonts!) are used in speeches today versus in the past. It’s a really fun vis but could be made aesthetically more pleasing. We wanted to add more graphs and linked views but ran out of time.