Jason Kushner, Jack McGreevy, Mehr Showkat

Kelsey Kraus

Ling 144

Politicians vs. The People: Who Uses Unique Words?

Introduction

The uniqueness of words used in a sample of speech can be an important indicator for its lexical richness. This uniqueness can be measured by dividing the number of unique words used by the total number of words used. In prepared speech, speakers have the opportunity to choose and review their word use to reduce or eliminate repetition. Spontaneous speech does not allow this, forcing speakers to insert pauses and unnecessary filler words to allow time to plan the next sentence. Spontaneous speech offers no opportunity for review, so if a speaker feels the need to elaborate on a subject, structures or topics may be repeated unintentionally. In contrast, political speeches are thoroughly revised and rehearsed to convey the speaker's intent. This opportunity for practice may eliminate unnecessary repetition and increase the uniqueness of words used. If this occurs, prepared speech will use more unique words than spontaneous speech. To test this theory, two English corpora, the Switchboard Corpus and the C-Span Inaugural Address Corpus, will be evaluated for their use of unique words. To measure this lexical richness, the Type/Token Ratio (TTR) and word frequency statistics will be computed for a sample from each.

Methods

Data gathered from the Switchboard Corpus and the C-Span Inaugural Address Corpus will be selected with criteria necessary to achieve comparable samples. Longer conversations from the Switchboard Corpus will be sampled in order to match the already lengthy samples from the C-Span Corpus. Both corpora exist solely in English, though the context of use is markedly different between the two. For the Switchboard Corpus, participants were given topics to discuss with an interlocutor over the phone. The C-Span Corpus contains inaugural speeches given by U.S. presidents. This incongruence in the context of speech does raise issues for this experiment because conversations on a prompted topic, as present in the Switchboard Corpus, may not be entirely representative of spontaneous speech. Similarly, inaugural speeches may not represent the entirety of prepared speech. Comparing the lexical richness found in these two corpora will explicitly compare only prompted speech with prepared political speech, but this data can be extrapolated to indicate trends amongst the larger question of spontaneous versus prepared speech. This dataset will contain long chains of sentences to analyze, split between two interlocutors, in the case of the Switchboard Corpus. The Python NLTK module will be used to compute the TTR of each. These values will then be saved into a string and manipulated with the “lower()” and “word_tokenize()” functions to tokenize the data. The token count will be the length of the new string. A Python regular expression can then be compiled to search for the number of words in the string, representing the type count. Type count will be divided by the number of tokens to obtain the final TTR value for each corpus. An additional

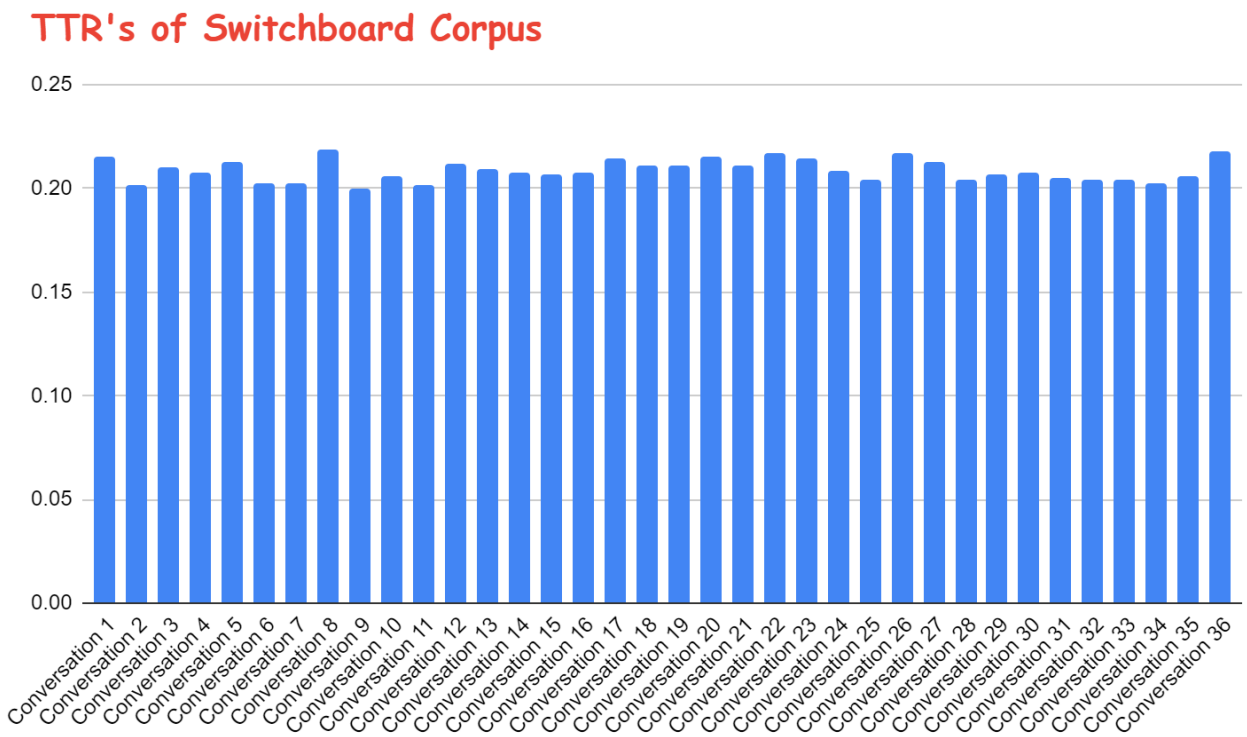
measure of lexical richness is the relative frequency of the words employed. Greater lexical richness should correspond to a higher number of low-frequency words. Final data will then be imported into Google Sheets to extract descriptive statistics and graphs showing the distribution of TTR values.

Background

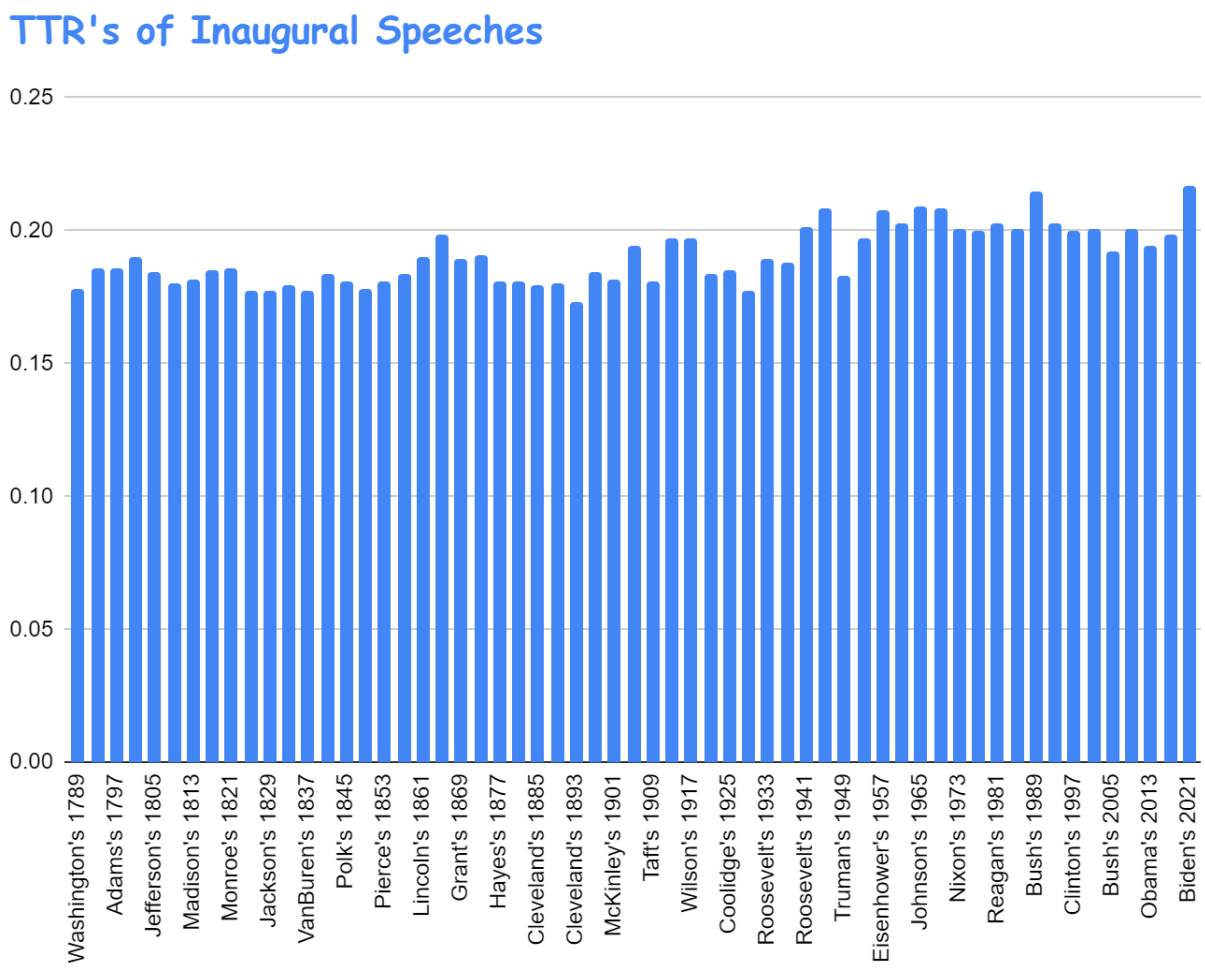
While much work has been done using the TTR as a measure of lexical richness, there has been no major work comparing TTRs of spontaneous vs prepared speech. TTR is used as a measure of lexical richness because it quantifies the number of unique words used in speech. Values range from 0-1, with values closer to 1 indicating greater lexical richness and values closer to 0 indicating lesser lexical richness. A sample with entirely unique words would have a TTR of 1, while a sample of the same size with only one repeated word would have a TTR closer to 0. Longer samples with a single repeated word would exponentially approach a TTR of 0 without ever actually reaching it. Preliminary research suggests that TTR will be higher in prepared speech, as the deliberate planning of speech enables speakers to exceed their normal vocabulary size. Spontaneous speech will likely use a smaller selection of more common words, and will include a greater value of connecting words and words used for pauses.

Results

Spontaneous speech from the Switchboard Corpus had a higher average TTR than prepared speech from the C-Span Inaugural Speech Corpus. This runs contrary to the initial prediction that prepared speech would have a higher TTR. Spontaneous Speech had an average TTR of 0.209 with a standard deviation of 0.005. Prepared Speech had an average TTR of 0.191 with a standard deviation of 0.011. Not only did spontaneous speech have a higher average, its standard deviation was lower, meaning the data was less volatile than that for prepared speech. Spontaneous speech also had a higher max TTR of 0.219 compared to a value of 0.217 for prepared speech, although these values are extremely close and may be within the same margin of error. The TTR values for spontaneous speech in the Switchboard Corpus are as follows:



Standard deviation was higher for Inaugural Speeches, so the data is much more volatile. Values were in a similar range to those of spontaneous speech. Data for prepared speech in the C-Span Inaugural Speech Corpus is as follows:



Overall statistics for the two datasets show that spontaneous speech clearly and consistently has a higher TTR value in these samples.

TTR Statistics for Corpus Data		
	Inaugural Speeches	**Switchboard Corpus**

Min	**0.173**	**0.200**
Max	**0.217**	**0.219**
Standard Deviation	**0.011**	**0.005**
Average	**0.191**	**0.209**

Discussion

Our original hypothesis, that the average TTR of planned speech would be higher than that of spontaneous speech, was proven wrong. We also see greater variation in TTR among the inaugural speech data. Spontaneous speech is more unique with a lower standard deviation and a higher average TTR.

This result indicates that an opportunity for revision and practice in speech does not increase the size of a speaker's utilized vocabulary. This may be influenced by the prepared speech corpus containing Inaugural Speeches, as these are targeted towards a "for the common man" audience. Inaugural Speeches are intended to be inspiring and motivational, which are appeals made more effective by keeping to a simple vocabulary of words with emotional impact. The intent of these speeches could affect the lexical richness of their content, especially if a higher lexical richness does not contribute to or even works against the speaker's goal.

Spontaneous speech is limited only by the individual speaker's vocabulary and the perceived vocabulary of the listener. The conversations in our corpora were extremely close in TTR values, which may be due to the prompts speakers were given. Speakers, with this prompt, may have stuck to similar lexicons when discussing their

topic. If the prompt was removed from this corpora, there may be a greater volatility in TTR values and lexical richness, as this can vary from topic to topic. If this proves true, it could be concluded that TTR variation is influenced more by topic than by the lexical knowledge of individual speakers. Further studies may be able to elaborate on this issue.

Conclusion

We find, based on TTR, that spontaneous speech has greater lexical richness than planned speech.

The defined intentions of inaugural addresses may have not served as the best representation of planned speech on their own. Similarly, the prompted speech samples taken from the Switchboard Corpus are imperfect representations of spontaneity. A greater diversity of data would be needed in future research. Our study also neglected to account for the great changes speeches undergo within a period of nearly two and a half centuries.

A conclusion is difficult to draw due to the apparent significance of collecting inaugural address data over such a wide range of years. It is worth pursuing the visible increase in TTR overtime within our planned speech samples. Possibly, a dataset controlling for speech patterns over a range of time may yield results more comparable to the higher TTRs of the Switchboard Corpus data.

References

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."

Godfrey, J., Holliman, E. & McDaniel, J. (1992). *SWITCHBOARD: telephone speech corpus for research and development*. Acoustics. IEEE International Conference on Speech and Signal Processing.

Ahrens, Kathleen. (2021). *The C-SPAN Inaugural Address Corpus.*

Djiwandono, P. I. (2016). Lexical richness in academic papers: a comparison between students' and lecturers' essays. Indonesian Journal of Applied Linguistics, 5(2), 209-216.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). Lexical diversity and language development (pp. 16-30). New York: Palgrave Macmillan.

Shi, Y., & Lei, L. (2021). Lexical use and social class: A study on lexical richness, word length, and word class in spoken English. Lingua, 262, 103155.

Wang, M., & Hu, F. (2021). The Application of NLTK Library for Python Natural Language Processing in Corpus Research. Theory and Practice in Language Studies, 11(9), 1041-1049.

Calhoun, S., Carletta, J., Brenier, J. M., Mayo, N., Jurafsky, D., Steedman, M., & Beaver, D. (2010). The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language resources and evaluation, 44(4), 387-419.

Neshkovska, S. (2019). Quitting with Style: Linguistic Analysis of Political Resignation Speeches. Thesis, 8(2), 3-30.