Introducing the Transcripts of US Presidential Debates Dataset

James Martherus
Vanderbilt University
james.l.martherus@vanderbilt.edu
ORCID: 0000-0002-8285-3300
@JamesMartherus

June 23, 2020

Debates between candidates are an important part of political campaigns, particularly at the national level.[1] They help inform the electorate of candidate positions (Abramowitz, 1978), particularly for the challenger (Benoit and Hansen, 2001). They increase voter's confidence in their candidate selection (Benoit and Hansen, 2001) and their interest in the election (Lemert et al., 1983). Debates can also affect the bases on which voters evaluate candidates. Benoit et al. (2001) show that after viewing a presidential debate, voters cared less about issue positions and more about leadership qualities and specific policy proposals. When there is a large discrepancy between expected and actual debate performance, primary debates can even change vote choice (Yawn et al., 1998).

More recent work has begun to examine the *content* of presidential debates. Many of these studies are essentially case studies, using a single debate to learn about verbal styles (Hellweg and Phillips, 1981), the influence of question format on candidate's argument choices (Carlin et al., 2001), and how specific arguments affect voter attitudes (Fridkin et al., 2007). Only a few studies have used a large corpus of many debates, but these studies have answered interesting questions

[1] Even if they are not important, political scientists are clearly interested. In the last five years, three articles in the APSR, 31 articles in the AJPS, and 9 articles in the JOP use the phrase "presidential debates"

about power dynamics between candidates based on poll position (Prabhakaran et al., 2013), topic switching as a form of power (Prabhakaran et al., 2014), and identifying viral moments during debates (Lukito et al., 2019). My hope is that with ready access to high quality debate data, more scholars will explore these sorts of questions.

The Transcripts of US Presidential Debates Dataset

While transcripts of presidential debates are available from a variety of sources,[2][3] obtaining them usually requires web scraping skills, and scraped transcripts need to be extensively cleaned before they are useful. To aid researchers who wish to conduct these analyses but are unable to invest the time to scrape and clean the transcripts, I compiled a data set that includes transcripts for all US presidential and vice-presidential debates, along with transcripts for many Republican and Democratic primary debates.

I began by scraping all presidential and vice-presidential debate transcripts from debates.org, a site hosted by the Commission on Presidential Debates. To my knowledge, there is no official source for primary debate transcripts. To obtain transcripts for as many primary debates as possible, I manually searched a number of news sites including CNN, FOX, NBC, and ABC. I also found a number of transcripts made available by Rev, a company that offers transcription services.

After gathering these raw transcripts, I cleaned and standardized each transcript. This consisted of standardizing spacing between words, correcting misspelled names, standardizing file encoding, and removing unwanted artifacts. Next, I parsed each transcript into a tidy (Wickham, 2014) dataframe where each row represents a single debate statement, and each column represents some attribute of the statement. This was possible because all transcripts followed a similar structure, with speaker names in upper case letters followed by a colon, and the statement after the colon. Finally, I added additional fields to the dataframe and made existing fields easier to use. For example, I added full names in the speaker column. This makes it easier to run individual-level

[2] https://www.presidency.ucsb.edu/documents/democratic-presidential-candidates-debate-los-angeles-california-0
[3] https://www.debates.org/voter-education/debate-transcripts/

analyses when the dataset includes multiple candidates with the same last name (Bill and Hillary Clinton, for example). Following is a list of included columns:

Variable	Description
speaker	The first and last name of the speaker.
text	The text of the speaker's statement.
type	Debate type. Possible values include "Pres," "VP," "Rep," and "Dem."
election_year	The election year corresponding with the debate.
date	The date the debate actually took place.
candidate	A binary variable indicating whether or not the speaker is a candidate.

Table 1 – Variables in the TUPD Dataset

For ease of access, the Transcripts of US Presidential Debates dataset is available both in several data formats, and as an R package that allows access to the data with no download necessary. Installation instructions and examples for using the debates package are available here: https://github.com/jamesmartherus/debates

References

Abramowitz, A. I. (1978). The Impact of a Presidential Debate on Voter Rationality. *American Journal of Political Science*, 22(3):680–690. Publisher: [Midwest Political Science Association, Wiley].

Benoit, W. L. and Hansen, G. J. (2001). Presidential debate questions and the public agenda. *Communication Quarterly*, 49(2):130–141. Publisher: Routledge _eprint: https://doi.org/10.1080/01463370109385621.

Benoit, W. L., McKinney, M. S., and Lance Holbert, R. (2001). Beyond learning and persona: extending the scope of presidential debate effects. *Communication Monographs*, 68(3):259–273. Publisher: Routledge.

Carlin, D. B., Morris, E., and Smith, S. (2001). The Influence of Format and Questions on Candidates' Strategic Argument Choices in the 2000 Presidential Debates. *American Behavioral Scientist*, 44(12):2196–2218. Publisher: SAGE Publications Inc.

Fridkin, K. L., Kenney, P. J., Gershon, S. A., Shafer, K., and Woodall, G. S. (2007). Capturing the Power of a Campaign Event: The 2004 Presidential Debate in Tempe. *The Journal of Politics*, 69(3):770–785. Publisher: The University of Chicago Press.

Hellweg, S. A. and Phillips, S. L. (1981). A verbal and visual analysis of the 1980 Houston republican presidential primary debate. *Southern Speech Communication Journal*, 47(1):23–38. Publisher: Routledge _eprint: https://doi.org/10.1080/10417948109372512.

Lemert, J. B., Elliott, W. R., Nestvold, K. J., and Rarick, G. R. (1983). Effects Of Viewing A Presidential Primary Debate: An Experiment. *Communication Research*, 10(2):155–173. Publisher: SAGE Publications Inc.

Lukito, J., K Sarma, P., Foley, J., and Abhishek, A. (2019). Using time series and natural language processing to identify viral moments in the 2016 U.S. Presidential Debate. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 54–64, Minneapolis, Minnesota. Association for Computational Linguistics.

Prabhakaran, V., Arora, A., and Rambow, O. (2014). Staying on Topic: An Indicator of Power in Political Debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1486, Doha, Qatar. Association for Computational Linguistics.

Prabhakaran, V., John, A., and Seligmann, D. D. (2013). Power Dynamics in Spoken Interactions: A Case Study on 2012 Republican Primary Debates. page 2.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(1):1–23. Number: 1.

Yawn, M., Ellsworth, K., Beatty, B., and Kahn, K. F. (1998). How a Presidential Primary Debate Changed Attitudes of Audience Members. *Political Behavior*, 20(2):155–181.