

Google Books Ngrams and Political Science: Two Validity Tests for a Novel Data Source

Web Appendix¹

¹ Due to space limitations, the following tables are available as a Web Appendix. If accepted for publication, this Web Appendix would be available on our and the publisher's website.

Google Books Ngrams and the *Authors' Guild* Lawsuits

Beyond issues related to the sample of books in the corpus, the digitization process paused in late 2011 as a result of a lawsuit brought against Google by several major publishers (Cohen 2015; Somers 2017). In 2015, an agreement between Google and their opponents suggested they may be able to begin digitization again (*Authors Guild v. Google, Inc.* 2015), but the Department of Justice entered a decision on behalf of the publishers to throw the agreement into doubt. At present—despite the Supreme Court allowing the circuit court decision in Google’s favor stand (*The Authors Guild v. Google, Inc.* 2016)—Google’s digitization process is still in limbo, so any published works between 2012 and the present (2018) are not in the corpus. Furthermore, any books with copyright infringement at issue in *Author’s Guild* cases have yet to be added into the corpus. These issues, while obviously important, do not mean this data source is unusable provided scholars understand and accept the limits these facts place on their findings and conclusions. In fact, given these realities, the value of creating a validity test for Google Books Ngrams is all the more pressing.

References

- Authors Guild v. Google, Inc.* 2015, 804 F. 3d 202. Court of Appeals, 2nd Circuit.
- Cohen, Dan. 2015. “What the Google Books Victory Means for Readers.” *The Atlantic*, October 22, 2015. <https://www.theatlantic.com/technology/archive/2015/10/what-the-google-books-victory-means-for-readers-and-libraries/411910/>.
- Somers, James. 2017. “Torching the Modern-Day Library of Alexandria.” *The Atlantic*, April 20, 2017. <https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/>.
- The Authors Guild v. Google, Inc.* 2016, 136 S. Ct. 1658. Supreme Court.

Table A1: Correlation matrix

	Corruption Index	Political Corruption Ngrams
Corruption Index	1.00	
Political Corruption Ngrams	-0.87 (0.00)	1.00

Note: P-values in parentheses

Table A2: Correlation matrix

	Percent Asian American	Asian American Ngram
Percent Asian American	1.00	
Asian American Ngram	0.91 (0.00)	1.00

Note: P-values in parentheses

Table A3: Correlation matrix

	Percent Hispanic American	Hispanic Ngram	Latino Ngram
Percent Hispanic American	1.00		
Hispanic Ngram	0.96 (0.00)	1.00	
Latino Ngram	0.88 (0.00)	0.90 (0.00)	1.00

Note: P-values in parentheses