Online Appendix:

Let's Talk Politics

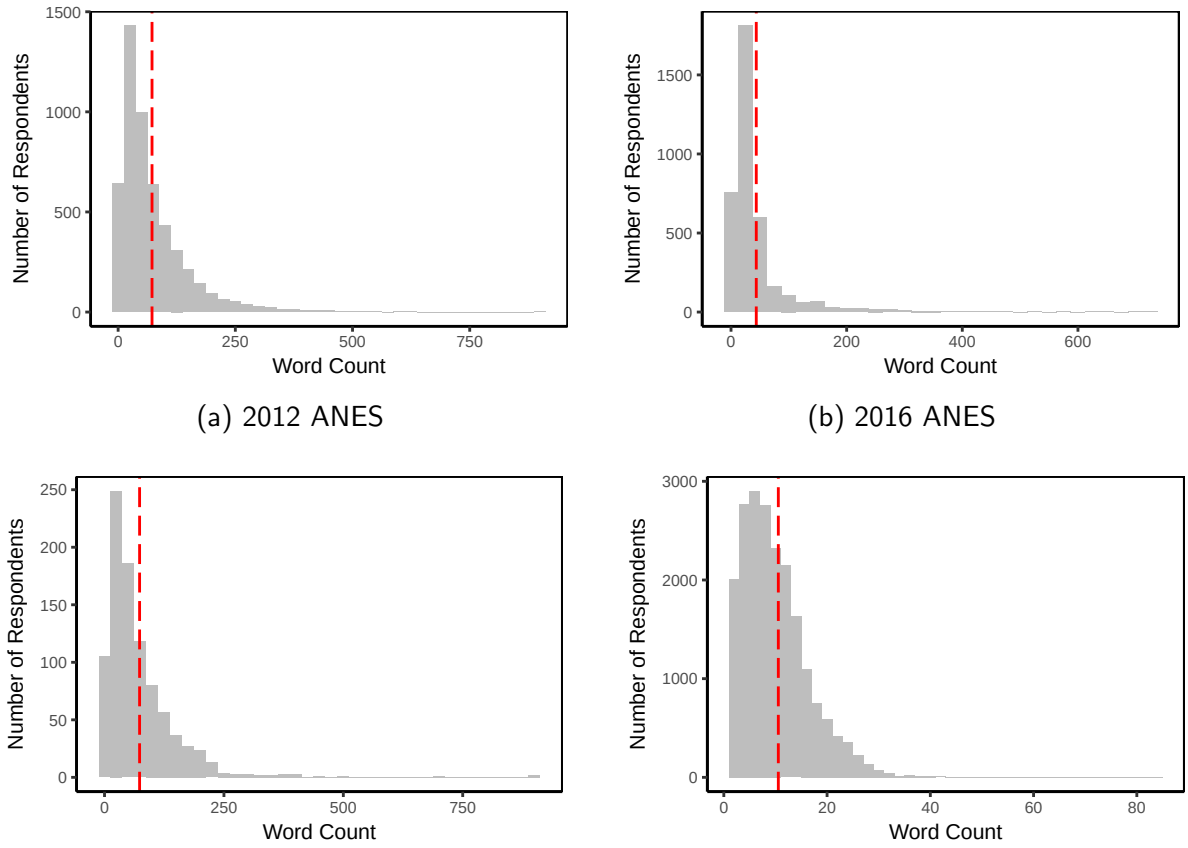
A Naive Approach for Measuring Political Sophistication

Patrick W. Kraft*

*Ph.D. Candidate, Stony Brook University, patrick.kraft@stonybrook.edu.

Appendix A Detailed Information on Open-Ended Responses and Discursive Sophistication Components

I Distribution of Word Counts in Open-Ended Responses



(a) 2012 ANES

(b) 2016 ANES

(c) 2015 YouGov

(d) Swiss Survey

Figure A.1: Histograms of total word count in the collection of open-ended responses for each individual. The dashed red lines indicate the average response lengths in each survey.

II Overview of Topic Proportions

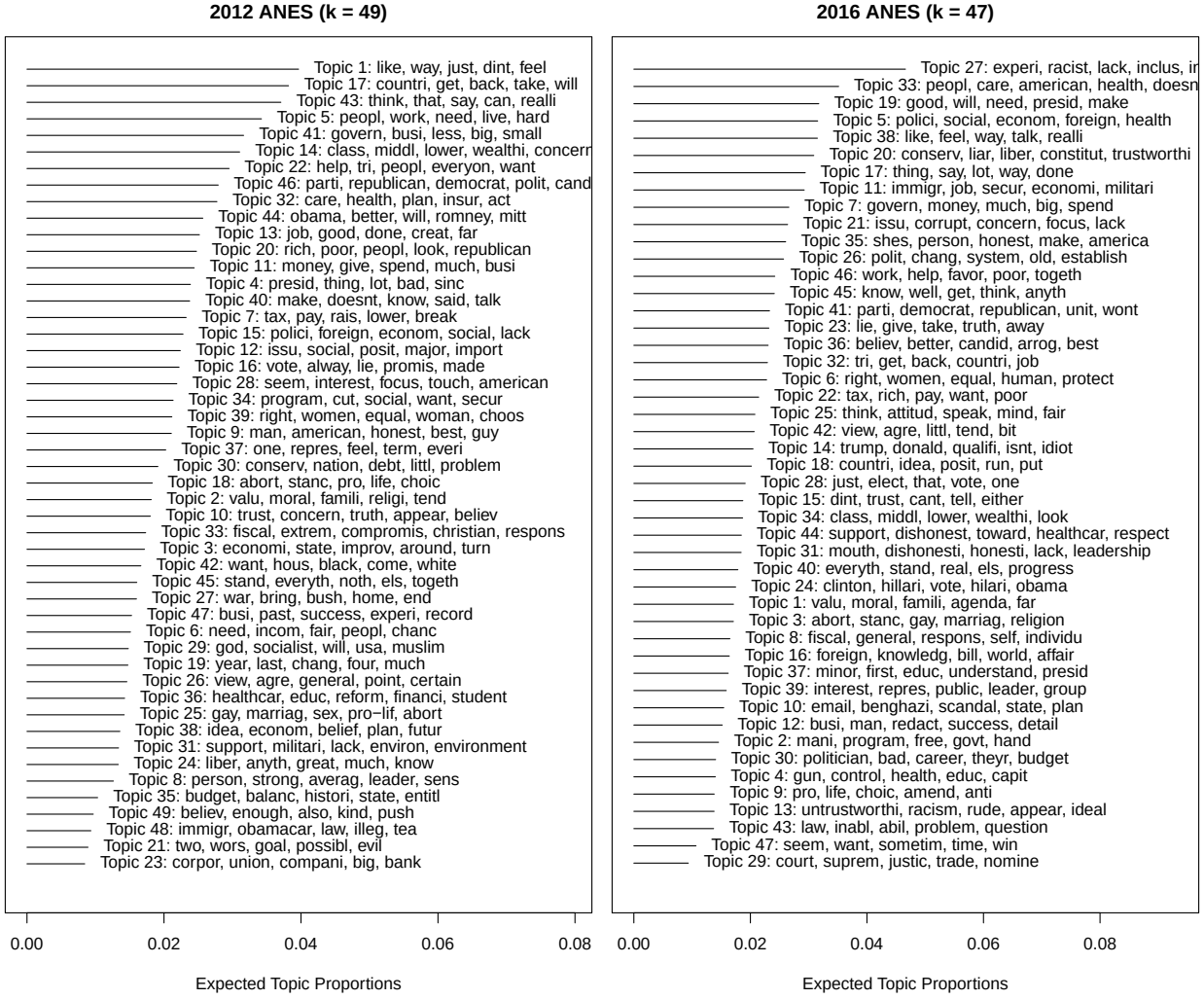


Figure A.2: Estimated topic proportions in the 2012 and 2016 ANES based on the structural topic model. See Appendix B for details on the model specification.

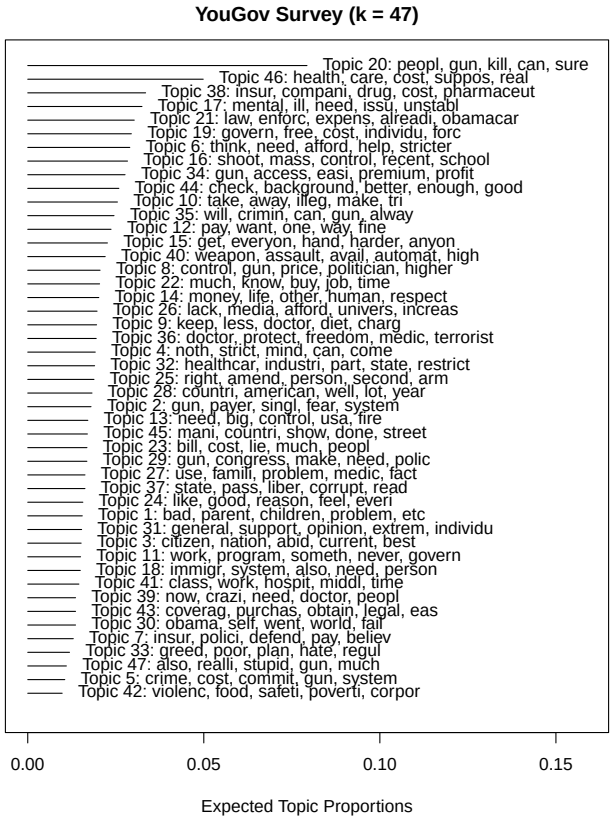


Figure A.3: Estimated topic proportions in the 2015 YouGov survey based on the structural topic model. See Appendix B for details on the model specification.

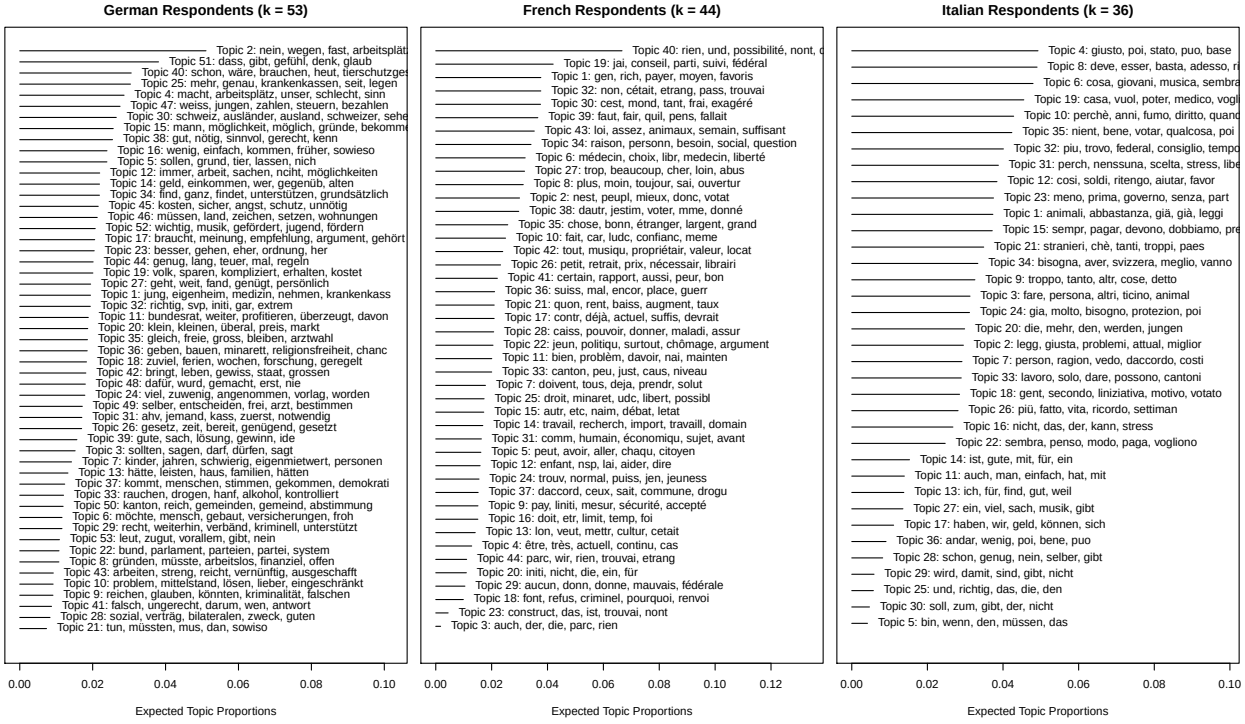
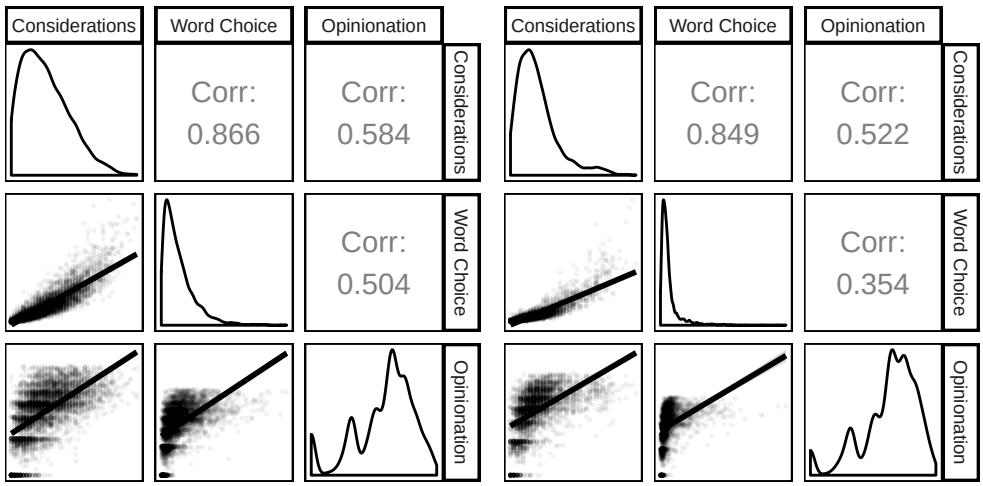


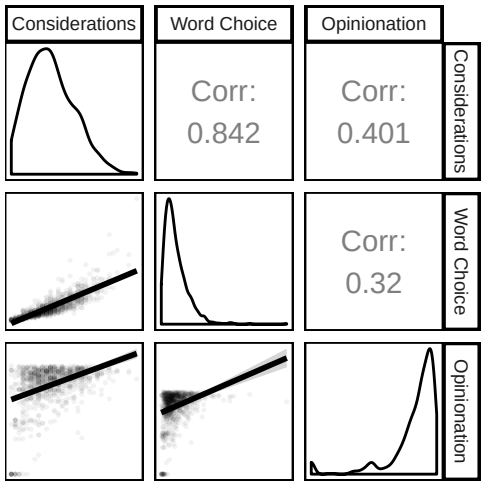
Figure A.4: Estimated topic proportions in the Swiss survey based on the structural topic model. See Appendix B for details on the model specification.

III Discursive Sophistication Components

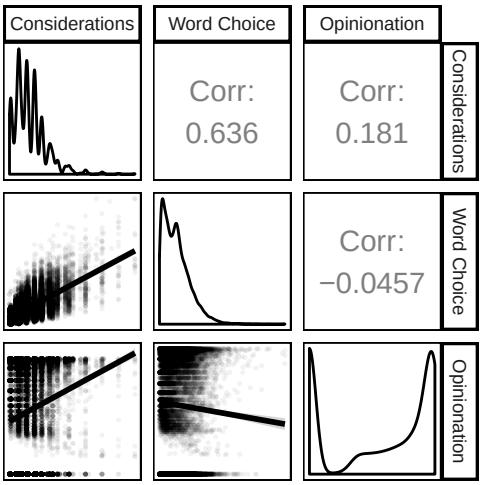


(a) 2012 ANES

(b) 2016 ANES



(c) 2015 YouGov



(d) Swiss Survey

Figure A.5: Correlation matrix of individual components of discursive sophistication. The plots on the diagonal display univariate densities for each component. The panels in the lower triangular display the scatter plot of two measures as well as a linear fit.

Appendix B Pre-Processing and Topic Model Specification

I PreText Analysis

Two components of discursive sophistication (*considerations* and *word choice*) rely on quantities extracted from structural topic models (Roberts et al., 2014). As with any other text-as-data approach, a necessary first step before estimating the topic model is to pre-process the raw text and convert it into a document term matrix (DTM, see for example Manning et al., 2008). Common pre-processing procedures include stemming and lowercasing, as well as the removal of numbers, punctuation, stopwords, and infrequent terms. However, topic models and other unsupervised learning techniques can be sensitive to these pre-processing choices (c.f., Denny and Spirling, 2018). To address this issue, Denny and Spirling (2018) recommend that researchers compare DTMs under all possible pre-processing regimes. The authors propose *preText scores* as a measure to quantify the extent to which varying pre-processing regimes may yield unusual results compared to a baseline without any pre-processing.

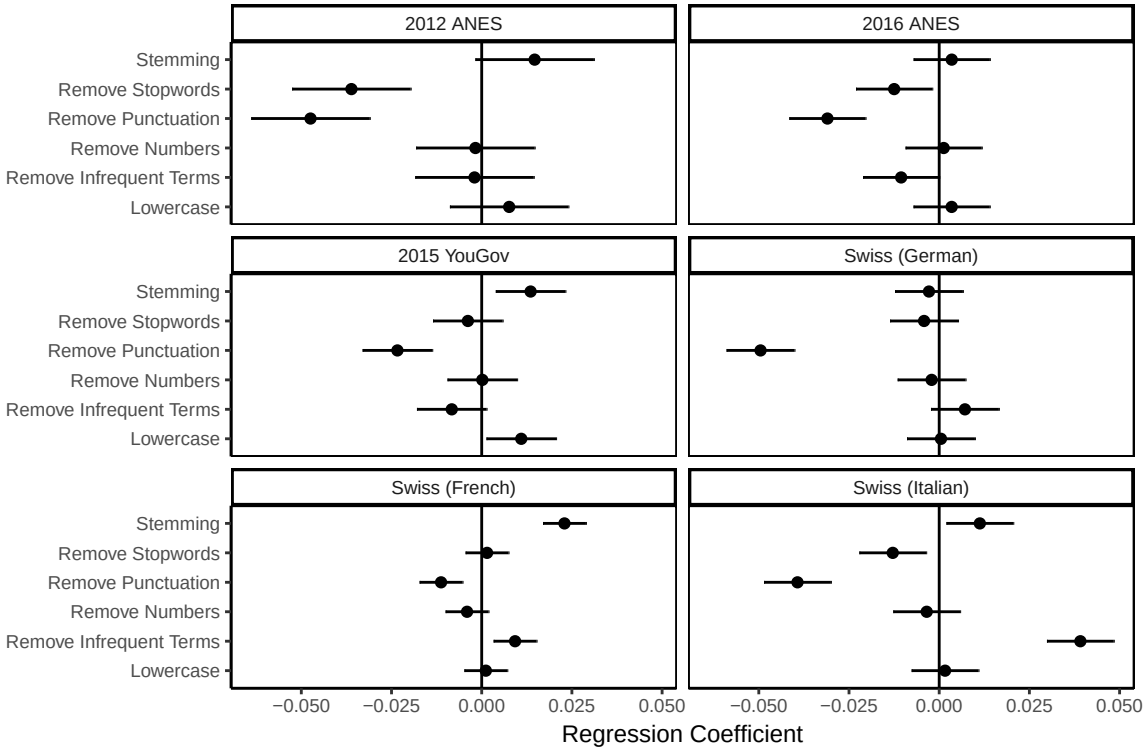


Figure B.1: PreText analysis of pre-processing decisions of open-ended responses across all datasets. Regression coefficients display the effects of each of the six pre-processing choices on the resulting preText score.

II Robustness Checks for Varying Model Specifications

Following the procedure outlined in Denny and Spirling (2018), Figure B.1 displays the results of a linear model regressing preText scores resulting from all possible pre-processing regimes on each individual step for a random subset of 500 open-ended responses in each of the studies included in the analyses. Significant coefficients indicate that the topic model results my be sensitive to the respective pre-processing step. As such, removing stopwords and punctuation, as well as removing infrequent terms (at least in the 2016 ANES) might be problematic. Denny and Spirling (2018), however, emphasize that the most important consideration in choosing pre-processing steps are theoretical. Given that the purpose of the topic model is to extract considerations related to political preferences, there are strong theoretical reasons to remove stopwords and punctuation from open-ended responses as they do not contain any relevant content. Furthermore, I apply lowercasing and stemming of terms to reduce resulting document term matrix to a computationally more manageable size and since these pre-processing steps should not be influential according to the preText analysis.

It is less obvious from a theoretical perspective whether to remove infrequent terms from open-ended responses, although it is preferred in order to make the estimation of the discursive sophistication components computationally efficient. Since the preText analysis for the 2016 ANES suggests that this pre-processing step might be influential, I compare discursive sophistication for both alternative regimes below (c.f., Denny and Spirling, 2018). Before turning to this sensitivity check, however, I consider another crucial modeling choice when working with topic models: determining the total number of topics k to be estimated. For all analyses reported below, the number of topics was selected using the algorithm proposed by Lee and Mimno (2014) and implemented in the `stm` package in **R** (Roberts, Stewart, and Tingley, 2014).[1]

Figure B.2 examines whether the proposed measure of discursive sophistication is sensitive to the removal of infrequent terms as well as the chosen number of topics k. The y-axis depicts the preferred pre-processing regime including all steps discussed above while the x-axis plots results for alternative specifications. The panels on the left compare the preferred specification to discursive sophistication based on a reduced number of topics ($k = 20$). The middle panels additionally include infrequent terms instead of removing them.[2] The panels on the right omit do not perform stemming as part of the pre-processing step. Across all panels, discursive sophistication scores are highly correlated and therefore insensitive to pre-processing choices and varying numbers of topics.

In summary, open-ended responses in the analyses reported below are pre-processed by stemming and lowercasing, as well as the removing numbers, punctuation, stopwords, and infrequent terms (i.e., terms that appear in fewer than 10 responses).[3] While the results discussed in the manuscript are based on this preferred specification, the substantive results are robust for alternative pre-processing regimes or varying numbers of topics.

[1] I used measures for age, gender, education, party identification, as well as an interaction between education and party identification as covariates for topic prevalence. This variable selection—with the exception of including gender—is equivalent to the procedure model specification described in Roberts et al. (2014).

[2] Calculating discursive sophistication with large numbers of topics while including infrequent terms is computationally prohibitive.

[3] Prior to applying these pre-processing steps, open-ended responses in the 2012 & 2016 ANES as well as the 2015 YouGov survey are cleaned by correcting spelling errors using an implementation of the Aspell spell-checking algorithm (`www.aspell.net`).

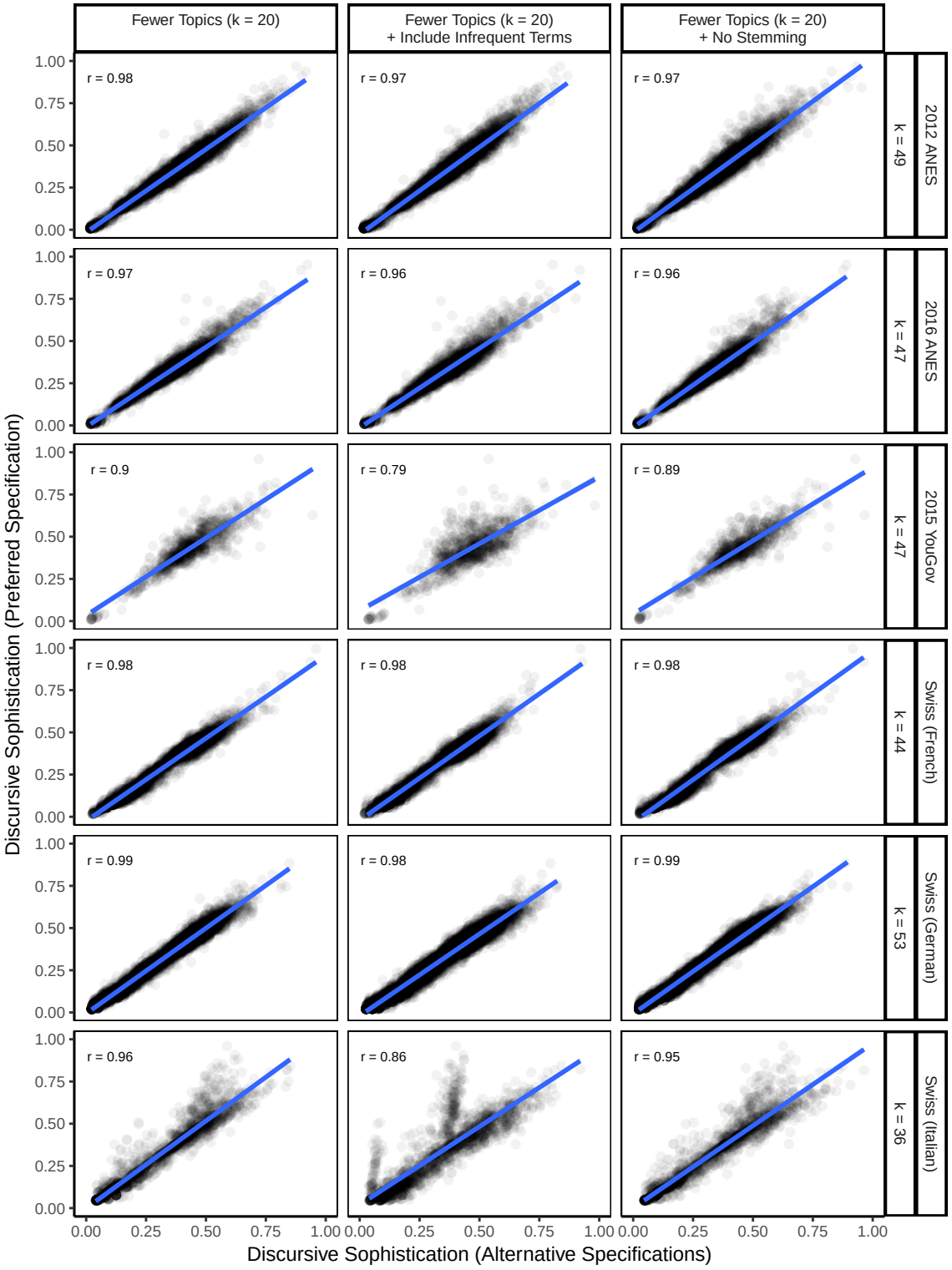


Figure B.2: Robustness of discursive sophistication measure for different pre-processing choices and topic model specifications.

References

Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *Political Analysis* forthcoming.

Lee, Moontae, and David Mimno. 2014. "Low-dimensional embeddings for interpretable anchor-based topic inference." In *Proceedings of Empirical Methods in Natural Language Processing*. Citeseer.

Manning, Christopher D., Prabhakar Raghavan, Hinrich Schütze et al. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2014. "stm: R package for structural topic models." *Journal of Statistical Software* 1: 1–49.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–1082.