Computational Analysis of Religious and Ideological Linguistic Behavior

Seth Green, Megan Stiles, Katherine Harton, Samantha Garofalo

Motivation and Objectives

Problem: In today's global environment, effective communication between groups of diverse ideological beliefs can mean the difference between peaceful negotiations and violent conflict.

At the root of communication is language, and researchers at the University of Virginia Center for Religion, Politics, and Conflict (RPC) hypothesize that the analysis of the performative character of a group's discourse (how words are used) provides valuable guidance for how to negotiate with groups of differing ideological beliefs. However, high-pressure situations leave little time for an exhaustive analysis of this nature. To help combat this need for immediate expertise in the field, the RPC team developed a manual method for measuring the performative character of discourse using a 1-9 linguistic rigidity scale, with a score of 1 indicating a rigid use of language and a score of 9 indicating a flexible use of language [1].

Objectives:

I. Expand on the signal processing approach of previous work in literature, which evaluates the efficacy of a computational approach to applying performative analysis to predict linguistic rigidity.

II. Evaluate the generalizability of computational performative analysis, considering text from non-religious groups. These include groups focused on political and social agendas rather than religion.

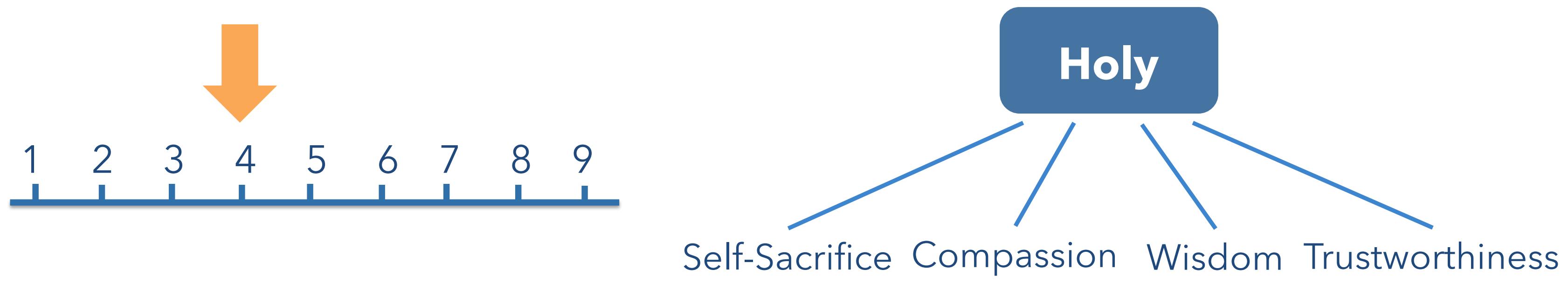

Figure 1. Depiction of Linguistic Rigidity scale. In this example, a group having four meanings for the word "Holy" (Self-Sacrifice, Compassion, Wisdom, and Trustworthiness) would be given a score of four on the scale.

Table 1. Data Sources

Group	# Documents	Group Rank
ACLU	40	3
American Ethical Union	15	8
Bahai	73	6
Dorothy Day	774	4
Integral Yoga/Yogaville	59	6
ISIS	48	1
John Piper	579	4
Liberal Judaism	166	6
Malcolm X	15	2
Meher Baba	265	8
Pastor Anderson	228	1
Rabbinic	58	4
Sea Shepherds	606	2
Steve Shepherd	728	4
Unitarian	301	7
Westboro Baptist Church	422	1

Methods and Signals

The Computational Linguistics (CL) team obtained data for this work from a number of different sources. The first was a repository that Venuti et al. collected [2]. This repository consists of 3,285 documents from nine different groups of various affiliations. The CL team collected additional sources of data using custom-built web scrapers using the Beautiful Soup package in Python [3]. This increased the total number of documents to 4,568, which consisted of 15 different groups of various affiliations. The CL team then pre-processed the documents for use in the models described below. The RPC team manually scored a random sample of single documents on their 1-9 scale, in addition to providing a group score.

Signals:

I. Keywords: Performative signals used in this analysis rely on the programmatic selection of keywords. The CL team ranked words using a custom TF-IDF algorithm with the IDF representing a word's frequency in a "general usage" Wikipedia corpus.

II. Performative Signals: Several signals measuring the diversity of contexts of the keywords.

III. Judgments: The judgment signal is defined as the percentage of sentences that contain a pronoun coupled with a keyword.

IV. Pronouns: Pronoun signals are defined as the percentage of words in the document that fall into the seven different pronoun categories.

The CL team approached this as an ordinal classification problem because, while the values on the scale correspond to a continuous estimate of contexts in which a group uses its deep-value words, there are a number of other performative traits of a group's speech that change as they move up or down the scale. The following machine learning algorithms were chosen based on their previous classification success [4] [5] [6] with noisy data: Support Vector Machines, Random Forest, and Gradient Boosted Trees.

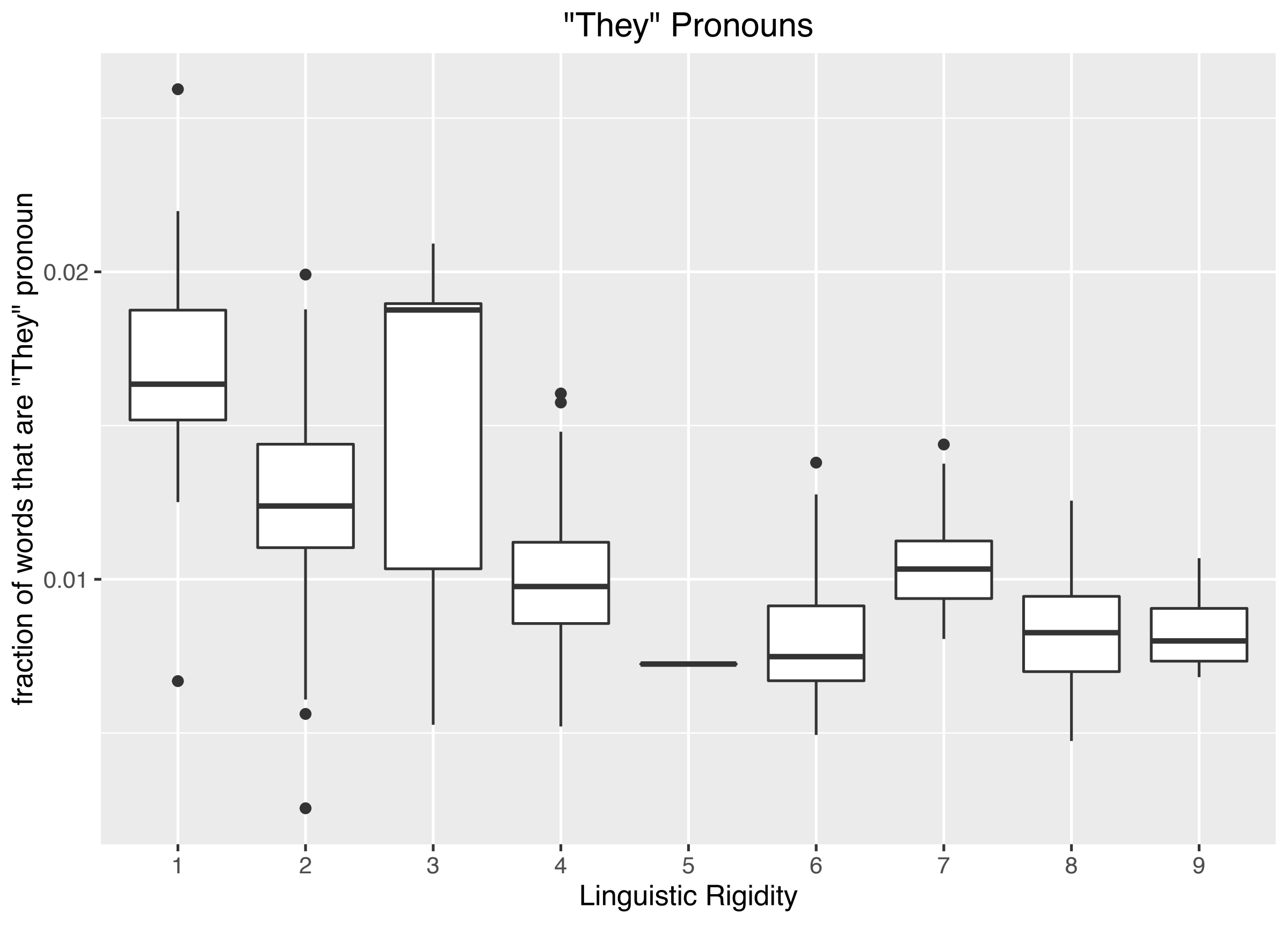

Figure 2. "They" Pronoun Usage

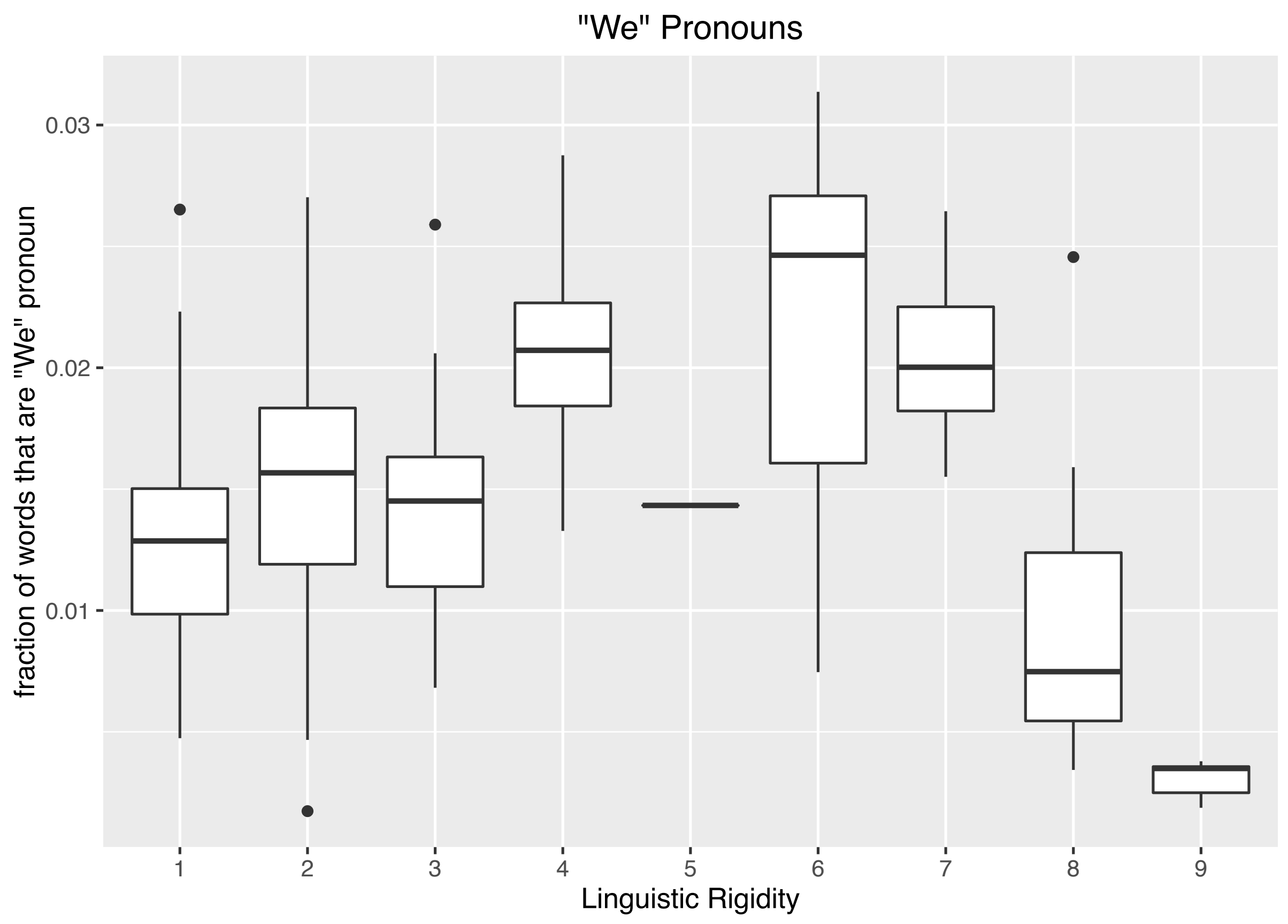

Figure 3. "We" Pronoun Usage

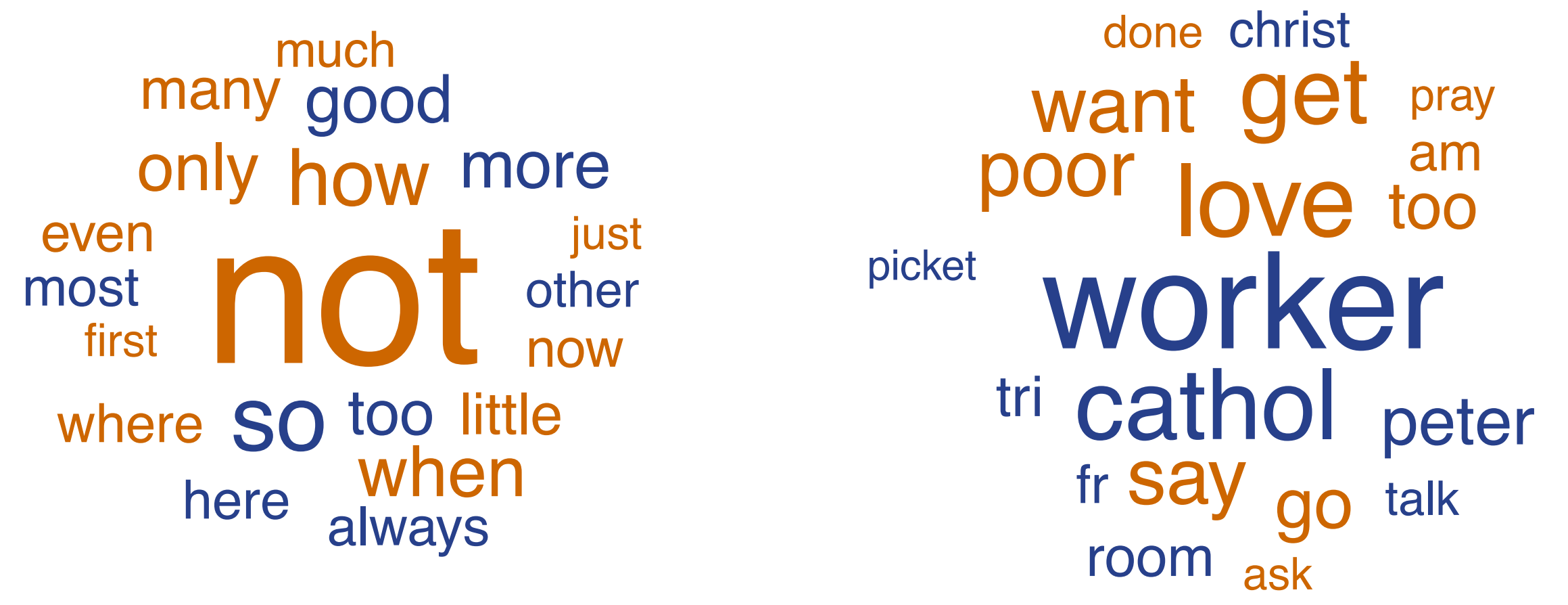

Figure 4. Comparison of keywords selection process from previous work [2] (on left) and the CL team's method (on right). Word cloud is from an example Dorothy Day speech.

Results

The models using the improved signals increased the predictive accuracy of [2]'s models by over 13%. For the binned document runs, the three models, random forest, support vector machine, and boosted trees all had accuracy rates within 1%, with an average accuracy of 96.6%. In addition, the models retained their high accuracy with the addition of non-political and more diverse religious groups, with an average accuracy of 96.0%.

Single document runs had a maximum accuracy of 77.5%. These initial results suggest the potential to perform real-time on-the-ground analysis on a single document level, increasing the practicality of this method.

Table 2. Binned Document Results.

Model	Accuracy (%)
Random Forest	96.1
Support Vector Machine	96.3
Gradient Boosted Tree	95.4
Average	*96.0*

Table 3. Single Document Results.

Model	Accuracy (%)
Random Forest	77.5
Support Vector Machine	70.5
Gradient Boosted Tree	69.7
Average	*72.6*

Conclusions

The performative signals, judgments, and pronoun usage of different groups have shown to predict a group's linguistic rigidity score with a high degree of accuracy, when using the RCP team's manual process of document scoring as ground truth.

In addition, the ability to track a group's use of language over time, identifying potential spikes or deviations from the group's baseline is a more practical application of this work. Running the model on a single-document level rather than on binned documents presented numerous challenges in that each input contained fewer words. However, the CL team's successful analysis of single documents has significant implications for real-time, on-the-ground analysis of ideological and religious groups.

Contact

Seth Green, Megan Stiles, Katherine Harton, Samantha Garofalo
University of Virginia, Data Science Institute
Email: smg5b, mes5ac, klh8mr, smg7un@virginia.edu
Website: https://github.com/seth127/DSI-Religion-2017
Phone: (434) 924-4262
Address: P.O. Box 400249, Charlottesville, VA 22904

References

1. Consultation with researchers at the University of Virginia Center for Religion, Politics, and Conflict (RPC). 1540 Jefferson Park Avenue, University of Virginia, Charlottesville, VA 22904-4126.
2. Venuti, Nicholas, Sachtjen, Brian, McIntyre, Hope, et al. "Predicting the Tolerance Level of Religious Discourse Through Computational Linguistics," presented at the 2016 IEEE Systems and Information Engineering Design Conference (SIEDS '16), Charlottesville, VA, 2016.
3. Richardson, Leonard. Beautiful soup documentation, [Online] Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.
4. Palomino-Garibay, Alonso, Camacho-González, Adolfo T., Fierro-Villaneda, Ricardo A., et al. 2015. "A random forest approach for authorship profiling." *Cappellato et al. [8].*
5. Treeratpituk, Pucktada and Giles, C. Lee. June 2009. "Disambiguating authors in academic publications using random forests." In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries,* pp. 39-48.
6. Joachims, Thorsten, 1998. "Text categorization with support vector machines: Learning with many relevant features." *Machine learning: ECML-98*, pp.137-142.

Acknowledgments

The team would like to thank Professor Peter Ochs, Jonathan Teubner, Essam Fahim, and Syed Moulvi of the University of Virginia's Religious Studies department for their expertise and critical insights. The work for this paper is partially supported by a grant from the U.S. Army Research Laboratory.