Leveraging Linguistic Content and Debater Traits to Predict Debate Outcomes

Alexandra Paxton (paxton.alexandra@gmail.com)
Rick Dale (rdale@ucmerced.edu)
Cognitive & Information Sciences, University of California, Merced
5200 N. Lake Road, Merced, CA 95343

Abstract

Since the earliest televised debates, cognitive and political sciences have been interested in how voters respond to political candidates and their messages, both verbal and nonverbal. The present work draws from this long tradition and combines it with work on persuasion and rhetoric to inform analyses of a new corpus of debate data: 48 transcripts from the Intelligence Squared U.S. series, televised Oxford-style debates on relevant sociopolitical issues (http://www.iq2us.org). As a first look at this corpus, we focus on how linguistic content (i.e., hedging and pronoun use) and debater traits (i.e., attractiveness and negativity) interact with arbitrary group identity (i.e., "for" vs. "against") to affect debate outcomes. Interestingly, we find that arbitrary group identity (i.e., "for" vs. "against" labels created by the framing of the debate rather than the actual opinions held) significantly affects the ways in which linguistic content and debater traits influence voters.

Keywords: communication; conflict; corpus analysis; debate; persuasion; politics; political psychology; political science

Introduction

Conflict is a regular part of the human experience. From legal battles to quarrels over chores, we regularly deal with conflict on personal, national, and international scales. While we may not necessarily enjoy these conflicts, we generally recognize that they are an essential part of our social experience. In fact, on a cultural level, it could be argued that we very highly value conflict in its proper place. Many democratic nations have adversarial judicial systems, requiring parties involved in legal action to argue their cases at the expense of the other, and hold debates as a key element of the electoral process.

In light of the importance of conflict, it is hardly surprising that so many have undertaken to try to explain it. Philosophers, political scientists, and cognitive scientists have attempted to answer questions of the origins of conflict, its purpose, and its essential characteristics using a variety of methods. The current project attempts to unite these perspectives to drive investigations of naturalistic debate using a newly compiled corpus of debate transcripts among experts on socially relevant topics.

We are specifically interested in investigating the ways in which debaters affect one another and their audience. With its interdisciplinary approach and computational focus, cognitive science is poised to uniquely and substantively add to our understanding of the topic. In the present work, we hope to spark such investigations by blending ideas from political science and pragmatics to shape linguistic analysis of a novel corpus of Oxford-style sociopolitical debates.

Persuasion, Political Science, and Pragmatics

Political science has been particularly enamored with debates since their first televised appearance (e.g., Baker & Norpoth, 1981). Related political and psychological research has investigated the effects of political advertising and campaigning (e.g., Geer, 2008). Both of these lines of research tend to incorporate an interest in the individual differences within the audience and in the manner in which the message is delivered. Given the nature of the corpus, we will focus more on the latter, although our hypotheses and analyses will also be shaped by ideas about the audience.

Voter Characteristics and Tendencies Integrating decades of reasoning studies, cognitive scientists Mercier and Sperber (2011) have recently suggested that human cognition appears to be geared toward argumentation and the defense of personal beliefs rather than reason. This view is consistent with a thread of findings in political science. While much of this work uncovers individual differences in the characteristics that drive political behavior, opinion, and opinion change (e.g., Brandt et al., 2014; Jost, Glaser, Kruglanski, & Sulloway, 2003; Koch, 1998), a recurring finding suggests that voters engage in motivated political reasoning and have a strong confirmation bias. Highly politically informed individuals tend to be swayed less than moderately or poorly informed individuals (Koch, 1998), and mass media and political advertising – while influential among undecided voters – are substantially less likely to change voters' opinions, once made (Forrest & Marks, 1999).

Political Figures and Media Research on characteristics of political figures and related media complements the aforementioned voter-based findings. Negativity in tone and message is perhaps one of the more well-studied of these concerns. While extreme negativity can cause voters to disengage from politics, a moderate amount of negativity surrounding legitimate concerns may prompt greater engagement (Kahn & Kenney, 1999). Such strategic negativity can be highly effective in winning over voters, especially undecided voters (Geer, 2008), but tangentially related negativity can reduce debater persuasiveness (Burgoon, Miller, Cohen, & Montgomery, 1978).

The debate context itself also affects perceptions of debaters. A debater who is perceived as winning not only improves his or her standing in the eyes of the audience but simultaneously causes opponents to be perceived less favorably as well (Schrott, 1990). Accordingly, increased audience engagement can improve the debaters' perception by individuals with less personal involvement in the issue (Axsom, Yates, & Chaiken, 1987).

Unsurprisingly, general traits of political candidates influence voters as well, with physical appearance and attractiveness being among the most influential. When evaluating political candidates, appearance can be more influential than personality traits and can even mitigate otherwise negative perceptions of the candidate stemming from differences of political opinions (Budesheim & DePaola, 1994). Analyses of Finnish voting records have linked attractiveness with vote increases of more than 20% over average-looking candidates (Berggren, Jordahl, & Poutvaara, 2010).

Contributions from Pragmatics and Rhetoric While political science may speak specifically to the context of debate, the study of pragmatics provides insight into contributors to opinion change at the discourse level. Though numerous other pragmatic influences exist, we find two types of metadiscourse particularly appropriate for the current study: hedges and personal pronouns. Metadiscourse markers reveal a speaker's relationship to the topic of discussion and the audience (Hyland, 1998). Hedges signal some level of uncertainty or tentativeness on the part of the speaker (e.g., "may," "might"); personal pronouns facilitate speaker-audience rapport (e.g., Dafouz-Milne, 2008). Previous research suggests that metadiscourse markers can trigger in-group sentiment within the audience (Hyland, 1998) and can increase persuasiveness and speaker-audience rapport (Dafouz-Milne, 2008).

The Present Study

Building from the contributions of political science and pragmatics, the current work integrates these areas with a "big data"-inspired approach to studying patterns of linguistics and persuasion in debate with established automated methods rather than traditional hand-coding methods. To do so, we gathered several dozen transcripts from the Intelligence Squared U.S. (IQ2) debate program. The resulting corpus provides fertile grounds for new insight into debate, thanks to its broad range of sociopolitical topics, its reliance on experts or professionals in the field, and its inherent measure of debater performance with its pre- and post-debate audience polls.

Given the effects of motivated reasoning (e.g., Brandt et al., 2014; Jost et al., 2003) and the relative difficulty of swaying partisan voters (e.g., Forrest & Marks, 1999), simply analyzing debates by outcomes should not be the sole focus of this research. We are interested not only in the characteristics of the winning side but also in those characteristics that contribute to more *effective* debating. We conceive of an effective debate team as one that is able to win over audience members to its side; therefore, we will include analyses of these change scores as well as absolute wins and losses.

Moreover, in the face of these effects (e.g., Brandt et al., 2014; Forrest & Marks, 1999; Jost et al., 2003), it should be difficult even for expert debaters to change the minds of an audience in a single program. We expect that the effects of these influences may be relatively be small in most cases: The majority of debate attendees will likely have made up their mind prior to the event. Debaters on both sides will be attempting to sway the same minority of undecided audience members during the program, leaving — in most cases — limited room for vote change.

The IQ2 corpus contains only Oxford-style debates, in which the topic to be debated is posed as a statement (e.g., "America doesn't need a strong dollar policy") and equal teams of debaters argue either for or against that statement. IQ2 strictly uses experts or professionals as debaters, and each debater argues a position that he or she truly believes. It is crucial to note that, although the opinions held by each debater are genuine, the direction in which the statement is posed is arbitrary. That is, any given argument could very well have been posed in the reverse direction (e.g., with the earlier example: "America *needs* a strong dollar policy"), with no change in the root issue under consideration or the participating debaters' actual stances. This debate structure affords us the opportunity to isolate the impact of arbitrary group identity, since the assignment of group title (i.e., "for"

or "against") is arbitrary rather than an essential part of the debaters' identities.

We approach our analyses of the IQ2 debates with hypotheses motivated by the existing work reviewed above. First, we hypothesize that debater traits will significantly impact effectiveness. Attractiveness should be positively associated with performance (e.g., Berggren et al., 2010). Moderate negativity should also be positively predictive of debater performance (e.g., Burgoon et al., 1978).

Second, guided by work from pragmatics, we anticipate that higher use of metadiscourse markers (here, hedges and personal pronouns) will predict team effectiveness. Based on previous research (Dafouz-Milne, 2008), debaters who effectively manage expert knowledge without appearing overconfident (i.e., through hedging) and who establish stronger personal ties to their audience (i.e., with personal pronouns) should win over additional votes.

Method

Corpus

For the current project, we compiled a corpus of 48 publicly available debate transcripts from Intelligence Squared U.S. (IQ2; http://www.iq2us.org), a series of Oxford-style debates initiated by the Rosenkranz Foundation (http://rosenkranzfdn.org). Debates spanned a wide variety of sociopolitical issues (see Table 1 for examples). Each debate (~105 minutes) featured equal groups of 2-3 experts on each side ("for" or "against") of the issue. Debates were structured to allow interactions among panelists, the moderator, and the live audience. Each panelist was first given 7 minutes for an opening statement, alternating between "for" and "against" groups. Panelists then challenged one another and answered audience questions.

Table 1. Example topics with pre- and post-debate votes.

Debate Name	Pre-Debate For	Pre-Debate Against	Post-Debate For	Post-Debate Against
Ban College Football	16%	53%	53%	39%
Obesity is the Government's Business	55%	19%	55%	35%
The Rich are Taxed Enough	28%	49%	30%	63%
California is the First Failed State	31%	25%	58%	37%

The final segment of the debate allowed each panelist 2 minutes for closing arguments.

Non-speech elements (e.g., audience laughter), moderator turns, and audience contributions have been discarded from the present analysis, and transcripts have been divided according to turn (i.e., all speech by one individual until another began talking). The subset of the data belonging to the debaters contains 229 unique debaters across 48 debates, and over 623,000 total words are included across more than 6,000 turns. Each group wins roughly the same proportion of the debates analyzed here ("for" wins = .51).

A unique feature of this corpus is its native measure of debater effectiveness through pre- and post-debate opinion polls. Before and after each debate, members of the live audience indicate their stance on the issue ("for," "against," or "undecided"). Assuming that effective debaters are those who best persuade the audience, debater effectiveness can be measured by overall winner and through change from pre- to post-debate opinion polls.

Debater Ratings

Physical appearance can heavily influence audience opinion (e.g., Berggren et al., 2010), and arguments can be weighed as much by perceptions of the debater as the content (e.g., Budesheim & DePaola, 1994). To capture the variance produced by this nonlinguistic factor, we collected participant ratings of attractiveness of debater headshots.

The rating procedure encompassed all debaters from debates available on the IQ2 website by December 2013. However, only ratings of the 229 debaters who participated in the 48 debates under consideration in the current study were analyzed in the present study. Headshots (90-100 pixels by 70-100 pixels) were downloaded from the IQ2 website and divided across 10 online surveys. Each survey presented 30-40 headshots (M = 39.3) to participants in random order. Participants rated each headshot on 7 personal dimensions, presented in random order, on a 1-5 Likert-style scale. Attractiveness ratings for each headshot were averaged across the individual ratings from 91-97 (M=95) undergraduate participants from the University of California, Merced.

Linguistic Analyses

Transcripts were first prepared for analysis using Linguistic Inquiry and Word Count (LIWC; Pennebaker, Francis, & Booth, 2007), a well-established linguistic analysis tool in social psychology research (see Tausczik & Pennebaker, 2010, for review). LIWC scanned transcript texts and generated the percentage of the text made up by each of a

number of categories. These categories ranged from affect words (e.g., "happy," "worried," "love") to pronouns to social processes (e.g., "talk," "child," "neighbor"). Although "bag of words" approaches are inherently less sensitive to contextual nuance than coder-based categorization, LIWC was chosen due to its relative level of acceptance within social psychology research and its high level of cost- and time-effectiveness.

Of these categories, the *negative emotion* category (e.g., "hate," "worried") became the measure of debater negativity. Pronoun use was measured with the *personal pronoun* category. A category for hedging was created by merging the *discrepancy* (e.g., "could," "would") and *tentative* (e.g., "maybe," "guess") categories.

The debater and linguistic information were then organized in a *by-word longform* or *B(eo)W(u)LF* matrix (Paxton & Dale, 2013). A text analysis tool created to facilitate multi scale analysis of language, B(eo)W(u)LF integrated the original transcript with the attractiveness ratings and LIWC output to create an expanded matrix annotating each word along each dimension. By creating the matrix at the word level, we were able to aggregate LIWC frequencies at the turn level for the logistic models and analyze changes in language use at the word level in our linear mixed-effects models (additional detail below).

Results

Debates were analyzed with a series of linear mixed-effects models and mixed logistic models, predicting differences in pre- to post-debate votes (ΔV) and debate winner, respectively. Per Mirman's (2014) recommendations, *p*-values were obtained by assuming a *z*-distribution for the *t*-values. Models included turn as the sole random effect, as individual speaker and debate number strictly covaried with debate outcome, since very few speakers participated in multiple debates and each debate had only one outcome. The winner variable was dummy coded (0 = "against" group victory; 1 = "for" group victory).

As mentioned earlier, one goal of the current study was to examine differences in outcomes according to linguistic choices within arbitrarily assigned group membership (i.e., "for" group debaters, FD, versus "against" group debaters, AD). For the purposes of the current study, we were not interested in the differences in outcome based on general linguistic use but in differences based on linguistic use compared across groups. To do so, all models comprised *only* interaction terms between the target variables and group membership (i.e., $variable_{xGroup}$). However, main effects for the variables could be inferred if similar values are found for FD and AD.

Predicting Discrete Outcomes: Mixed Logistic Models

The first mixed logistic model combined debater traits ($attractiveness_{xGroup}$ and $negativity_{xGroup}$) and pragmatics variables ($hedging_{xGroup}$ and $pronoun_{xGroup}$) at the turn level to predict debate winner. As expected, the $attractiveness_{xGroup}$ (ß = .12, $p < .001$) and $negativity_{xGroup}$ (ß = -.002, $p < .05$) interactions significantly predicted differences in debate winners. $Pronoun_{xGroup}$ trended towards significance (ß=-.002, $p < .08$), but $hedging_{xGroup}$ did not (ß = -.002, $p > .25$).

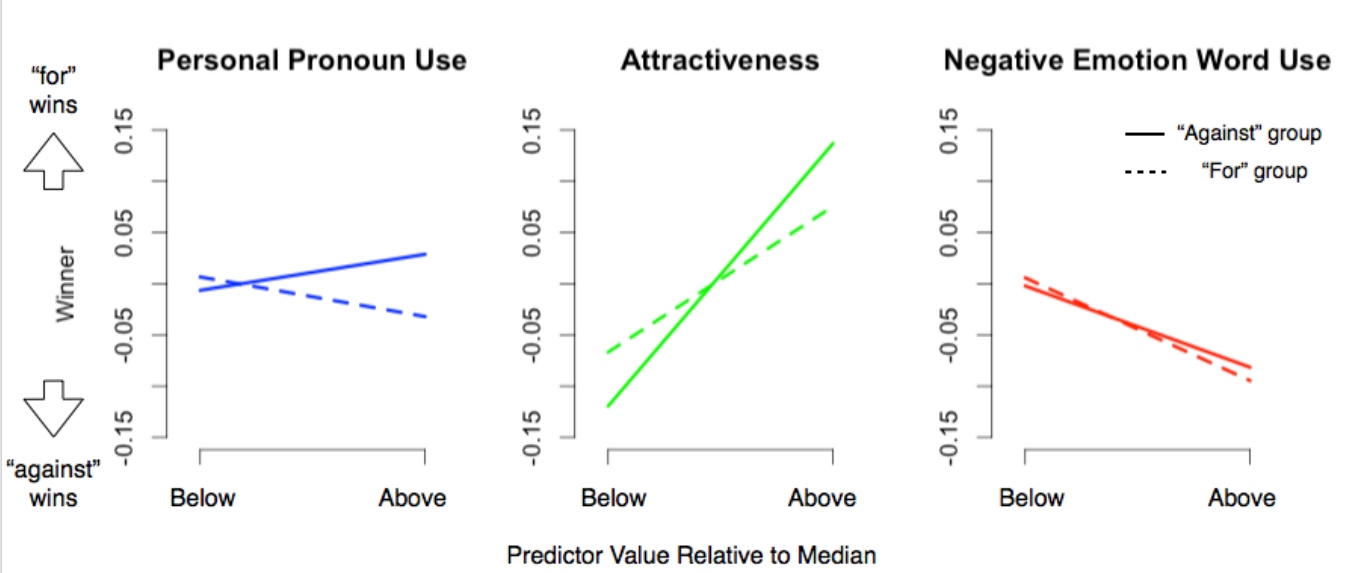

Figure 1. Results from the three significant predictors in the mixed logistic model predicting debate winner. All terms were interactions between the predictor (personal pronoun: right; attractiveness: center; negative emotion: left) and group membership (by line type). Predictor values were graphed as a median split.

To determine the best model for the data, additional mixed logistic models then were calculated across all possible combinations of the four predictors. A model predicting debate winner with attractiveness$_{xGroup}$ (ß = .12, p < .001), negativity$_{xGroup}$ (ß = -.002, p < .05), and pronoun$_{xGroup}$ (ß = -.003, p < .05) best described the data, as measured by lowest AIC (see Figure 1). Interestingly, these were almost all of the same variables from the original model, suggesting that most of the hypothesized relations effectively captured audience voting behavior under consideration here.

The significant interaction terms revealed an effect of nominal group membership on debate performance across the three significant variables. Lower pronoun use was not associated with a win by either team. We do, however, see a positive main effect of high pronoun use.

Additionally, higher attractiveness ratings were unilaterally predictive of the FD wins, regardless of the attractive debater's own group membership. Less attractive AD were *more* likely to win than more attractive AD, against the hypothesized direction. Data for the FD, however, behaved as anticipated: Attractive FD were much more likely to win than their less attractive counterparts.

An unexpected relation was found between debate winner and negativity, as well. Higher negativity in both groups was associated with an AD win. However, lower negativity in both groups was not strongly associated with a win by either group.

Predicting Continuous Outcomes: Linear Mixed-Effects Models To increase model sensitivity to initial vote, a linear mixed-effects model with the same variables (attractiveness$_{xGroup}$, negativity$_{xGroup}$, hedging$_{xGroup}$, and pronoun$_{xGroup}$) predicted the difference in pre- to post-debate vote change between AD and FD (ΔV). Higher ΔV would signal a greater pre- to post-debate vote change for FD relative to AD. Lower ΔV would indicate a greater pre- to post-debate change for AD relative to FD.

Surprisingly, the results of this model were very similar to those of the logistic model reported above. Attractiveness$_{xGroup}$ (ß = -.06, p < .001), negativity$_{xGroup}$ (ß = -.002, p < .05), and pronoun$_{xGroup}$ (ß = -.009, p < .05) significantly predicted ΔV, while hedging$_{xGroup}$ did not (ß = -.0008, p > .5). The patterns generally adhered to those reported above, although the effect of pronoun$_{xGroup}$ was larger. This was unexpected due to the fact that we anticipated that this would be a more sensitive measure of factors related to debater performance, as it which washes out effects of wins resulting from heavily skewed starting votes.

Additional models were tested to find the best fit for the data. As with the logistic model, a model predicting ΔV with attractiveness$_{xGroup}$ (ß = -.06, p<.001), negativity$_{xGroup}$ (ß = -.002, p < .05), and pronoun$_{xGroup}$ (ß = -.009, p < .001) was found to best capture the data, having the lowest AIC of the permutations tested. Examination of the data revealed mostly similar results to those outlined in the description of the logistic model, with several exceptions. Lower and higher negativity were more strongly associated with FD and AD wins, respectively, than in the win-based model. Attractiveness overall shifted more slightly towards predicting AD wins, and more extreme differences in low and high attractiveness were seen for AD rather than FD. Finally, lower pronouns use in both of groups was more predictive of AD wins, and higher use of pronouns were inversely predictive of wins for each group (i.e., higher FD pronoun use predicted an AD win and vice-versa).

Discussion

The present study blended ideas from political science and rhetoric to answer a basic question: Does the language we use change based on how we frame group membership? To answer this, we used corpus analysis techniques from cognitive science to explore debate effectiveness in a new corpus of debate transcripts from the Intelligence Squared U.S. program. One of the most striking findings of the current study is the consistent difference in winning behaviors according to arbitrary group membership. While additional research must be done to account for other potential explanatory effects not examined here, we find strong evidence that the framing effects imposed by the naming of the debate significantly affect the audience's perception of debaters.

As a result, although we found a positive link between each group's pronoun use and respective wins, two of our hypothesized effects exhibit interesting interactions with group membership, behaving as expected for one group but not the other. First, as anticipated, attractiveness in the "for" group positively predicts "for" group wins; conversely, attractiveness in the "against" group is negatively associated with "against" group wins. This could point to a conflict within audience members between the "against" debater's negative group identity and a positive personal attribute (i.e., attractiveness), harming the coherence of the debater's argument.

Similarly, we found that all high negativity was associated with "against" group wins, suggesting that the "for" group may be punished for negativity in a way that the "against" group is not. It may also be possible that increased mention

of the "against" group label may be associated with increased wins, given LIWC's contextual blindness. These interactions highlight the powerful role of arbitrary group identity as a salient framing effect for participants.

Because debate is a higher-level discourse activity, it is unsurprising that even significant linguistic and debater effects have relatively low effect sizes. While we believe that the effects reported do exist, we cannot expect the audience to react to the debate without the weight of their preexisting opinions affecting their votes. More recent Intelligence Squared U.S. debates have begun tracking votes in more detail, providing breakdowns of post-debate votes by pre-debate votes. This additional data will allow us to isolate even further the effects contributing to persuasion beyond what we have found here.

Acknowledgments

The authors' thanks go to Stephanie Huette (University of Memphis) for manuscript feedback and to undergraduate research assistants Chelsea Coe, Alex Lau, Nicholas Rodriguez, and Amanda Varela for help in data preparation.

References

Axsom, D., Yates, S., & Chaiken, S. (1987). Audience response as a heuristic cue in persuasion. *Journal of Personality and Social Psychology*, *53*(1), 30.

Baker, K. L., & Norpoth, H. (1981). Candidates on television: The 1972 electoral debates in West Germany. *Public Opinion Quarterly*, 45(3), 329–345.

Berggren, N., Jordahl, H., & Poutvaara, P. (2010). The looks of a winner: Beauty and electoral success. *Journal of Public Economics*, *94*(1), 8–15.

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science*, *23*(1), 27–34.

Brooks, D. J., & Geer, J. G. (2007). Beyond negativity: The effects of incivility on the electorate. *American Journal of Political Science*, *51*(1), 1–16.

Budesheim, T. L., & DePaola, S. J. (1994). Beauty or the beast? The effects of appearance, personality, and issue information on evaluations of political candidates. *Personality and Social Psychology Bulletin*, *20*(4), 339-348.

Burgoon, M., Miller, M. D., Cohen, M., & Montgomery, C. L. (1978). An empirical test of a model of resistance to persuasion. *Human Communication Research*,*5*(1), 27-39.

Dafouz-Milne, E. (2008). The pragmatic role of textual and interpersonal metadiscourse markers in the construction and attainment of persuasion: A cross-linguistic study of newspaper discourse. *Journal of Pragmatics*, *40*(1), 95–113.

Forrest, J., & Marks, G. N. (1999). The mass media, election campaigning and voter response: The Australian experience. *Party Politics*, *5*(1), 99–114.

Geer, J. G. (2008). *In defense of negativity: Attack ads in presidential campaigns*. Chicago, IL: University of Chicago Press.

Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics*, *30*(4), 437–455.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375.

Kahn, K. F., & Kenney, P. J. (1999). Do negative campaigns mobilize or suppress turnout? Clarifying the relationship between negativity and participation. *American Political Science Review*, *93*(4), 877–889.

Koch, J. W. (1998). Political rhetoric and political persuasion: The changing structure of citizens' preferences on health insurance during policy debate. *Public Opinion Quarterly*, *62*(2), 209–229.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Chapman and Hall / CRC.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.

Onraet, E., & Van Hiel, A. (2014). Are right-wing adherents mentally troubled? Recent insights on the relationship of right-wing attitudes with threat and psychological ill-being. *Current Directions in Psychological Science*, *23*(1), 35–40.

Paxton, A., & Dale, R. (2013). B(eo)W(u)LF: Facilitating recurrence analysis on multi-level language. arXiv: 1308.2696 [cs.CL].

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2007). Linguistic inquiry and word count: LIWC [Computer software]. Austin, TX: liwc.net.

Schrott, P. R. (1990). Electoral consequences of "winning" televised campaign debates. *Public Opinion Quarterly*, *54*(4), 567–585.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.