EVOLVING LINGUISTIC DIVERGENCE IN SOCIO-POLITICAL POLARITIES

Andres Karjus[*1] and Christine Cuskley[2]

[*]Corresponding Author: akarjus@tlu.ee
[1]ERA Chair for Cultural Data Analytics, Tallinn University, Tallinn, Estonia
[2]School of English Literature, Language and Linguistics, Newcastle University, Newcastle upon Tyne, United Kingdom

1. Introduction

All human languages change over time, as linguistic variants are discarded, innovated, and their meanings shifted. Most change stems from variation, be it geographical, cultural or social. Here we examine a division and source of variation intersecting these categories: political polarization. Particularly in the last decade, social and political scientists have been concerned with the causes and alarming social effects of increasing media polarization and partisan segregation particularly in the US (Brown & Enos, 2021). However, political polarization may also have an effect on the dynamics of language evolution and change, forming the basis for signals of in-group and out-group status (?) and potentially leading to more dramatic language speciation at deeper timescales (?; Altmann, Pierrehumbert, & Motter, 2011). Using existing political categorizations, we collect and use an extensive corpus of social media posts to quantify divergence in American English along the left-right axis, both in topics of conversation and lexical semantics.

Twitter data has been shown to be useful for mapping lexical innovation and variation (Grieve, Nini, & Guo, 2018; Bhat & Klein, 2020) and analyzing polarization effects (Chen, Salloum, Gronow, Ylä-Anttila, & Kivelä, 2021). Studies of linguistic divergence between political divides have often focused on politicians and activists (Adamic & Glance, 2005; Li, Schloss, & Follmer, 2017; Gentzkow, Shapiro, & Taddy, 2019). Those on the general population have grouped subjects based on self-reporting (Halpern & Rodrıguez, 2018) and social media activity (Sylwester & Purver, 2015; Demszky et al., 2019; KhudaBukhsh, Sarkar, Kamlet, & Mitchell, 2020). We focus on everyday interactions, not just political communication, and scale up the latter sampling approach, mining and grouping a total of half a billion follower listings across 72 large US news media accounts. We use the Allsides Media Bias Rankings (Allsides, 2021, v4) to delineate likely left and likely right biased news outlets (Allsides is not perfectly unbiased itself, but serves as a useful starting point). Having access to the entire follower bases of the news accounts allows us to carry out full set operations, and find users who only follow one side but not the other. Further limiting these to active and identifiably US-located users (a major bottleneck) leaves 6202 likely “left” and 4783 likely “right” aligned accounts. We mined their tweets between February-September 2021, yielding a corpus of 1.5 million tweets (750,814 and 732,521, respectively).

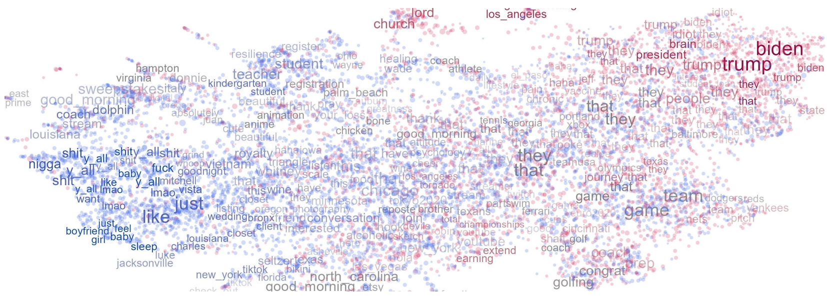

Figure 1. Divided they (micro)blog. UMAP dimension reduction of an LDA topic model of the tweet corpus, with illustrating keywords. Each dot is an account, accounts with similar content are arranged together (blue = “left”, red = “right”).

2. Results and Conclusions

We fit a topic model to the tweet corpus, visualized in Figure 1, showing how the two sides differ in their daily conversation topics. Word frequencies quantify this further, revealing magnitudes of difference in the usage of some terms. Phrases like *President Trump* and *communist* are used about 16x more by the “right” in 2021, who also prefer hand-shaped emoji in contrast to some face-shaped emoji used predominantly by the “left”. We also apply word embeddings to detect semantic change and identify potentially competing variants. While users on either side still of course speak the same language, a number of interesting divergences emerge, with differences in the meaning of some emoji and certain politically charged terms (e.g. *woke* referring to either “wokeness” or just waking up). Ongoing research aims to validate these findings with a crowd-sourced annotation task (Schlechtweg, Walde, & Eckmann, 2018), and compare with recent large-scale dialectal work (Grieve et al., 2018) to tease apart sources of variation.

By building on and scaling up previous methodologies of sampling utterances by speaker political alignment, our results show that existing political divisions are already being reflected in language use. Mapping this ongoing lexical and semantic evolution in American English provides a model applicable to studying similar phenomena in any language where sufficient data can be acquired. We emphasize the potential for studies of this kind to shed light on the interplay of interacting evolutionary dynamics at socio-cultural and linguistic levels.

References

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 36–43). Association for Computing Machinery.

Allsides. (2021). *Allsides Media Bias Ratings, version 4.* Available from: https://www.allsides.com/media-bias/media-bias-ratings.

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PLOS ONE*, *6*(5), 1–12.

Bhat, P., & Klein, O. (2020). Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter. In G. Bouvier & J. E. Rosenbaum (Eds.), *Twitter, the Public Sphere, and the Chaos of Online Deliberation* (pp. 151–172). Springer International Publishing.

Brown, J. R., & Enos, R. D. (2021). The measurement of partisan sorting for 180 million voters. *Nature Human Behaviour*, *5*(8), 998–1008.

Chen, T. H. Y., Salloum, A., Gronow, A., Ylä-Anttila, T., & Kivelä, M. (2021). Polarization of climate politics results from partisan sorting: Evidence from Finnish Twittersphere. *Global Environmental Change*, *71*, 102348.

Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., & Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv: 1904.01596*.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, *87*(4), 1307–1340.

Grieve, J., Nini, A., & Guo, D. (2018). Mapping lexical innovation on American social media. *Journal of English Linguistics*, *46*(4), 293–319.

Halpern, D. J., & Rodrıguez, P. L. (2018). Partisan Representations: Partisan Differences in Semantic Representations and their Role in Attitude Judgments. *Proceedings of CogSci 2018*, 6.

KhudaBukhsh, A. R., Sarkar, R., Kamlet, M. S., & Mitchell, T. M. (2020). We don't speak the same language: Interpreting polarization through machine translation. *arXiv: 2010.02339*.

Li, P., Schloss, B., & Follmer, D. J. (2017). Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behavior Research Methods*, *49*(5), 1668–1685.

Schlechtweg, D., Walde, S. Schulte im, & Eckmann, S. (2018). Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of NAACL 2018* (pp. 169–174).

Sylwester, K., & Purver, M. (2015). Twitter Language Use Reflects Psychological Differences between Democrats and Republicans. *PLOS ONE*, *10*(9), e0137422.