Computational methods for sociolinguistic analysis in online discussions

The language that people use to communicate both reflects and constructs the society that they live in, both in explicit statements and implicit word choices. A person's everyday language choices can reflect where they were raised, how they accommodate to their audience, and their attitudes toward other people. These social factors have an especially strong impact on the internet, where people from a variety of backgrounds meet to share information and build social connections primarily via text communication. My work investigates the relationship between language style and social factors on the internet, using computational methods to quantify structural linguistic factors. The internet provides researchers with a bird's eye view of social interactions that is difficult to replicate in offline settings.

I focus on intuitive social factors that are understood to affect linguistic choice in everyday conversation, which include social attitudes, community dynamics, and audience expectations. My work uses natural language processing (NLP) and statistical analysis to explain the variation in language structure using these concrete social factors. Whereas prior work often investigates word frequency, I use a variety of NLP methods to characterize **structural** patterns, such as syntax, that would otherwise be ignored by typical approaches such as word frequency. For example, will a word that is more **syntactically flexible** (occurring in diverse contexts) outcompete other words in an online community? My work extends sociolinguistic theory to the context of the internet, a domain with rich linguistic diversity.

As with other work in computational social science, studying language use on the internet can extend existing social science theories and provide new methods to draw insight from large-scale text data. In my dissertation work, I have explored linguistic variation in the domains of political attitudes, online community norms, and audience expectations in public discussions of crisis events.

How do social attitudes affect a multilingual person's choice between languages in public discussions?

Fig. 1: High support for Catalonian independence during the 2017 referendum paralleled the use of Catalan slogans (*per la republica* "for the republic") in protests.[1]

A person's attitude toward a particular topic can result in consistent patterns in their **language choice**: in political discussions, the use of a minority language is often connected to attitudes about the status of the language's culture (Shoemark et al. 2017). In 2017, the region of Catalonia in Spain voted for

[1] https://www.polgeonow.com/2017/10/catalonia-independence-vote-2017-results-map-graph.html, https://www.rt.com/news/189072-catalonia-independence-consultation-vote/

independence (see Fig. 1), which ignited a national debate over the cultural identity of Catalonia and whether it deserved to be a separate country. Through an analysis of Twitter discussion of the independence vote, we found that bilingual activists who were pro-independence more often wrote in Catalan than Spanish, even in posts unrelated to independence discussion. This effect was stronger than a similar study of another independence referendum, which supports the idea that political identity is particularly strong in the use of Catalan and more generally that minority language use can reflect social attitudes. In follow-up work, we are investigating the influence of cultural affiliation, such as active consumption of American media, on the grammatical integration of loanwords in social media discussions. This kind of work can reveal how language reflects cultural differences among multilingual people, which is especially relevant to internet discussions where cultures clash frequently.

How well does a word's diversity of linguistic contexts predict its adoption in a community?

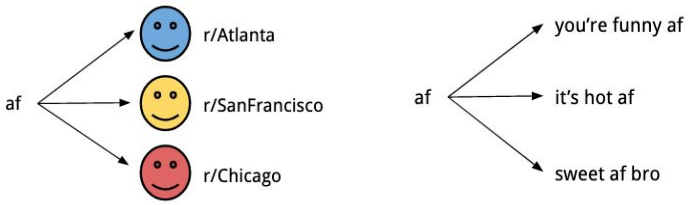

Fig. 2: A word may be socially disseminated (left) or linguistically disseminated (right), and we find that linguistic context dissemination predicts word growth more readily than social dissemination.

In online communities, new words emerge frequently via interactions between different sub-groups and the arrival of new internet users who bring their own unique vocabulary: *haha* today can be *lol* tomorrow. While the social factors leading to word adoption are well-studied (Altmann et al. 2011), the linguistic factors remain poorly understood, such as whether a word that can apply to many contexts will be adopted more readily than a competitor word. In a study on Reddit, I investigated the role of **linguistic context** as a factor in word adoption, to test the hypothesis that the spread of a word among diverse syntactic contexts can predict the adoption of nonstandard words (e.g. if the new intensifier "af" "as fuck" can occur with a wide variety of adjectives). I developed a new metric to measure linguistic *dissemination* among different contexts and found that this metric consistently predicted word growth and decline, even when compared to the standard metric of social dissemination. In contrast to prior work in innovation diffusion that focuses on social metrics (Altmann et al. 2011), I demonstrated that the linguistic variation of nonstandard words is an important factor in the eventual adoption of nonstandard words, which should inspire further study in word adoption that tests other linguistic metrics such as topical diversity. More broadly, the study shows how variation in linguistic structure can provide insight into large-scale social dynamics in online communities.

When do audience expectations in public discussions lead people to use more descriptive information?

Fig. 3: Newspaper headlines that mention city "San Juan" before Hurricane Maria (2013; with descriptive information) and after Hurricane Maria (2017; no descriptor information).

When sharing information in discussion, people must determine how much *context* they need to provide for their audience. One type of linguistic context is **descriptive information** for location names, which may or may not be known to an audience: many Americans know about Puerto Rico, but they may not recognize its capital without additional description (see Fig. 3). Prior work supports a regular decrease over time in descriptive information for names in news coverage (Staliunaite et al. 2018), but it is unclear how much of that decrease is due to audience needs. To address this gap, I investigated how Twitter and Facebook users changed their use of descriptive information for location names during their discussion of the crisis events. I leveraged named entity recognition and dependency parsing to detect descriptive information, which captured the notion of descriptive information with high precision without sacrificing data diversity. I found that discussion participants decreased their use of descriptive information after the peak in collective attention, and locations that were local to a particular audience (e.g. mentioning "San Juan" to locals from the area) had fewer descriptions. This suggests that the discussion participants accommodate to their audience's lower perceived need for information during the event, i.e. more collective attention paid toward affected locations leads to increased expectations of shared knowledge.

Future work: Detecting and evaluating linguistic polarization in online discussions

My previous work has focused on a variety of social factors that influence language style variation, including social attitudes, online community norms and audience expectations. In my future work, I will focus on developing linguistically-motivated metrics for social cohesion and division, and testing the ability of such metrics to generalize across domains. Developing such metrics will provide more accurate estimates of political division, which will guide interventions to address division in online discussion such as exposing online commenters to opposing opinions. Studying social cohesion and division will also lend itself to interdisciplinary collaborations with political and sociology researchers, who can benefit from expanding their typical methods toolkit from limited surveys to more open-domain text data.

How well can linguistically-motivated metrics for differences in opinions capture political polarization?

Political polarization is typically quantified as the split between social groups based on divergent beliefs, such as disagreements between Democrats and Republicans about policy. Online discussions about political issues often result in polarized opinions between different social groups, which can inhibit information sharing between groups. While prior computational work has quantified polarization using word count differences between predefined social groups (e.g. Demszky et al. (2019)), I am interested in developing linguistically-motivated metrics to quantify the level of polarization in online discourse. For

example, knowing that Republicans tend to talk about guns more often than Democrats does not imply that their opinions are polarized, but observing a difference in the relative *valence* of gun-related discussion (positive vs. negative) can reveal polarization. Developing high-precision metrics will be useful in situations with relatively low word counts (i.e. high-variance), such as comparing individual speakers rather than large-scale political parties. I will leverage distributed representations of word and sentence semantics to measure differences in valence across discussion posts on news articles related to politics. I will experiment with several structured linguistic representations of expressed opinions using sentence structure, including restricting the scope to information connected to named entities ("Trump is wrong") and to concrete nouns ("abortion is wrong"). Ideally, such a metric will be able to determine the degree of difference between two texts even without being given *a priori* knowledge of what topics should be considered, which is an assumption made by stance detection.

How well do polarization metrics generalize across domains?

I plan to evaluate the utility of linguistically-motivated polarization metrics with both intrinsic and extrinsic comparisons to guarantee the generalizability of such metrics. Typical studies of large-scale polarization evaluate such metrics against expert judgment, which is often sparse and limited to well-understood domains such as American politics. In intrinsic evaluation, I plan to compare the estimated aggregate polarization against ground-truth data from voting: for a given newspaper in a state, the state's level of Republican versus Democrat support is available from voting records. Extrinsic evaluations will include predictive tasks, such as inferring whether a user will agree with another user's comment based on their prior computed degree of polarization, as well as descriptive tasks, such as comparing relative aggregate rates of polarization across more or less divisive topics (e.g. political elections versus sports games). The predictive tasks can also include community-level predictions such as whether a community will split into multiple sub-groups with differing opinions (e.g. when the subreddit r/News generated r/WorldNews).

As more diverse text corpora become available, quantifying polarization will become more important as a lens for understanding broad trends and changes in society. The metrics that my research develops can be extended to other domains related to social groups: the integration of immigrants into society can be better understood by comparing the relative alignment of immigrant-written texts with non-immigrant texts. Evaluating more linguistically-motivated metrics in collaboration with domain experts will encourage computational social scientists to leverage text data with a more critical lens, without relying on word frequency alone.

References

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2011). Niche as a determinant of word fate in online groups. *PloS one*, *6*(5).

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

Shoemark, P., Sur, D., Shrimpton, L., Murray, I., & Goldwater, S. (2017). Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *EACL*.

Stewart, I., & Eisenstein, J. (2018). Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. In *EMNLP*.

Stewart, I., Pinter, Y., & Eisenstein, J. (2018). Sí o no, ¿què penses? Catalonian independence and linguistic identity on social media. In *NAACL*.

Stewart, I., Yang, Y., & Eisenstein J. (2019). Characterizing collective attention via descriptor context in public discussions of crisis events. In submission.