Statement of Research Interests

Venkata S Govindarajan

My research interests lie at the intersection of **Computational Linguistics** and **Computational Social Science** (CSS). Broadly, my goal is to use *linguistic data and features* and modern Natural Language Processing (NLP) and Machine Learning (ML) techniques to formulate hypotheses on intriguing questions in human communicative behavior. I accomplish my goals by carefully curating large datasets, and building interpretable models for prediction and analysis of my research questions. I am motivated by the firm belief that computational and statistical modeling can give us valuable insights into communication, and enrich our understanding of meaning in language in the process.

I have worked on a wide range of research topics — modeling the diversity of generalizations in the English language, advice-giving behavior in online communities, and the linguistic differences that characterize in-group from out-group speech. I have published primarily in Computational Linguistics venues (ACL, EACL, EMNLP); However, I plan to branch out into broader journals and conferences in Computational Social Science since I see Computer Scientists, Linguists and Social Scientists as the audience for my work. My work differs from most research in NLP studying social behavior by **encompassing the diversity in language use** through fine-grained schema and simple features, and building interpretable statistical models on top of the data **using modern ML and NLP techniques**. I believe that my data-driven modeling framework provides a unique perspective into human communication, while enabling nimble, iterative research.

Research Program & Trajectory

How can we use modern NLP tools and corpora to gain insights into human communicative behavior? This is the question that motivates my research endeavors, and I have pursued this question through research projects looking at the nature of advice-giving online [1], as well as how implicit intergroup bias manifests in language use online [2]. One of the hallmarks of human language is that we use it to do things with consequences in the real world [3]. Recent work in NLP has started to recognize the importance of this towards building better language technologies [4]. However, what about the *unintended* repercussions of our words — for instance, when we reinforce harmful stereotypes through implicit biases in our language use [5]?

My current research towards my dissertation tackles this question on the nature of biases in language through the lens of **social relationships** between the speaker and target of an utterance. Specifically, my focus is on **intergroup bias** in online communication — what differentiates in-group speech from out-group speech, and can we build

computational models to characterize and distinguish between the two? Bias (and its mitigation) in language technologies is an active area of study in NLP, and my work approaches the topic distinctly from existing work in NLP. As I lay out in Govindarajan et al. [2], I am motivated from the intuition that *all language is biased*, and we can gain insights into the nature of this bias by studying the social relationships between the speaker and the target of an utterance. I am inspired by work in social psychology on the Linguistic Intergroup Bias [LIB, 5] — but my aim is to go above and beyond existing work by incorporating insights from linguistics and NLP to look at *whole-sentence meaning in context*, combined with real-world language use online.

In Govindarajan et al. [2], we curate a dataset of tweets with transparent social relationships between the speaker and target of interpersonal utterances (in the domain of U.S. politics), and model two dimensions of the intergroup bias that we identified — the intergroup relationship (in-group vs. out-group), and interpersonal emotion. We find through human annotation that there are systematic relationships between the emotion expressed and the intergroup relationship, even with *the identities of the speaker and target masked*. This validates our idea that **social dynamics can have predictable effects on language behavior**. Furthermore, we found that trained annotators performed on-par at identifying intergroup relationships as a Support Vector Machine based on unigram and bigram features, but a transformer language model performed over 10 points better than both. This leads to a natural question — what are the latent, linguistic features that characterize the differences between in-group and out-group speech, and does a classifier based on a modern encoder based language model learn to use these latent features?

Generalization While there are a large number of possible language features at various levels of granularity that could distinguish in-group and out-group speech, the ability to **construct generalizations** through language is an important one, especially to communicate and reinforce stereotypes covertly [6]. In an earlier work [7], I studied *generalizations in the English language* through a decompositional lens. We found that there is rich diversity in how people can express generalizations, beyond existing, rigid linguistic schemas that could nonetheless by captured by simple, fine-grained, and crucially: easily computable and annotatable properties. Could subtle, distinctive forms of generalization underlie language differences in the intergroup bias as well?

We investigated this question by **probing** the models we trained to distinguish between in-group and out-group utterances in Govindarajan et al. [8]. While previous studies into generalizations and LIB relied on word lists of more versus less 'abstract' (i.e. general) predicates, we used a sentence-level computational property: specificity [9], which measures the level of detail in a sentence and correlates highly with genericity. We formulated a hypothesis in line with the LIB — people are more likely to use general (less specific) language in utterances with positive and negative affect towards the in-group and out-group respectively. We calculated affect from robust annotations

for interpersonal emotion, which were collected by labelling sentences for simple, easy-to-identify emotions like anger, sadness, etc.

In our probing experiment, we modified sub-regions directly in *the embeddings of words* within the classification model (towards higher or lower specificity) rather than the surface level language that was input to the model. Thus, we can ask the **counterfactual** — would this tweet be considered in-group after intervening on the embeddings to make them more or less specific? While we found some evidence for specificity influencing intergroup relationship through these experiments, our overall takeaway was that there was no evidence for the asymmetry in abstractness hypothesized by the LIB. However, this negative result opens up more possibilities and informs my current work — one that studies how world and discourse context tie into the intergroup bias, and encompasses the extreme diversity of language use in online communities.

Immediate future work Building on top of the lessons learned from my probing experiments in [8], my goal is to study the intergroup bias on its own terms — incorporate more data across domains, model the surface nature of the bias using insights from common NLP techniques like masked language modeling and entity-tagging, and thus *discover* the underlying differences between in-group and out-group speech. I plan to do this by curating a dataset of comments by football fans on Reddit, and probing how their language changes when talking about their team (in-group) versus the opponent(out-group). Sports comments are an interesting testing ground for this analysis as they are grounded in clear external events (the final score), thus offering a clear **causal link from real-world event to the language used**. With a curated dataset in-hand, I plan to investigate different modeling approaches to automatically infer features of interest — specifically, predicting named entity expressions that refer to the in-group or out-group using various levels of contextual information, from sentence to discourse to the game result. Rather than relying on pre-defined hypotheses, this approach lets modern language models discover latent features and feature combinations from data, something they have proven to be very effective at [10]. By analyzing the behavior of the resultant model combined with probing, we can finally gain insights into our initial question — what characterizes in-group form out-group speech? Furthermore, how generalizable are the results from the sports domain to the politics domain in my initial dataset? These are just some of the questions that I am excited to be working on currently, and I hope to expound on them further in the coming months.

I also plan to investigate **how the intergroup bias manifests in synthesized text from Large Language Models** (LLMs) like LLaMA and ChatGPT. LLMs are trained on massive amounts of textual data, so their output is a marginal distribution over all authors in their training data, rather than the perspective of one person [11]. Do LLMs exhibit intergroup bias in their outputs, similar in nature to what I have shown in my work? Do they exhibit a consistent, hegemonic intergroup bias for individual mentions, or is this one that changes with the domain? Finally, is this bias steerable through interventions in the embeddings or through prompting? As LLMs are pushed

into real-world use without due consideration of their harms, my investigations will give a alternative insight into the **perspective that these models exhibit in their outputs** by offering a connecting thread across demographic dimensions.

Future Work

I envision my future research program to focus on the two themes which have played a central role in my research to date, and which I describe in what follows.

Generalization in Language While we found no evidence for our specific definition of generalization in language [specificity, 8] explaining intergroup language differences, I believe that generalizations over events and entities do still play a crucial role in intergroup language differences, and am currently working on characterizing these further by modeling intergroup entity references as described previously. In the long term, I wish to broaden my focus beyond generalizations in standard American English to non-standard English dialects and other languages, including languages underrepresented in NLP in general. Further, I want to investigate the transferability of computational features and models of generalizability across languages and domains. Do social pressures like the intergroup bias lead to similar patterns of generalization behavior across languages, and are there latent features that traditional NLP models can use to identify and predict these generalizations cross-linguistically? Linguistic generalization is an essential feature of language that will play a key role in building more intelligent language understanding and AI systems [12], and I am excited to delve into this in my future research.

Semantic invariance One of the reasons I was drawn towards studying the intergroup bias was the variation in language posited by hypotheses like the LIB — the same event being described in different ways by different people due to contextual social pressures. Current work in NLP tries to understand different ways of expressing semantically similar content through tasks like style transfer and paraphrase generation. I wish to build on this work and my research through the lens of semantic invariance. What are the pressures and features *modulating* the need to communicate in a different surface form, while retaining as much of the core meaning as possible? As I have observed with my work studying intergroup bias, the unconscious linguistic changes that follow from intergroup pressures lead to reinforcing social stereotypes — thus, while the lexical and sentence-semantics might have remained the same, the social meaning of an utterance [13] is radically different. I wish to study tease apart this distinction in meaning further — Can ostensibly surface-level transformations like style transfer truly retain *all meaning*? Characterizing various surface-level language alterations in terms of the social and non-social pressures that lead to them is a rich vein of research that I believe will have huge ramifications on how we analyze bias in communication and language technologies.

Mentorship As a faculty member at the University of X, I intend to mentor students in the use of computational methods to study human communicative behavior. I have been fortunate to have had numerous mentors who have supported and guided me through 7 years of graduate school. As I begin my journey into an academic career, 2 principles that I've learned from my mentors will govern interactions with students that I mentor. First is a **proud ownership of their research**, a quality that has pushed me to refine my work constantly as I strive for excellence. Second, is the ability to **zoom out from the details of the work to take a sweeping view** of where I am in my research journey towards my goals — a valuable skill in research projects which may take months or years. Both of these have been crucial lessons, that I will hope to inculcate among students embarking on research under me.

To conclude, I am excited at the prospect of continuing my academic journey at the University of X, where in addition to contributing to the broader research community, I will be able to guide students through rewarding research projects and unique classroom experiences.

References

[1] Venkata S Govindarajan, Benjamin Chen, Rebecca Warholic, Katrin Erk, and Junyi Jessy Li. "Help! Need Advice on Identifying Advice". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 5295–5306.

[2] Venkata S Govindarajan, Katherine Atwell, Barea Sinno, Malihe Alikhani, David Beaver, and Junyi Jessy Li. "How people talk about each other: Modeling Generalized Intergroup Bias and Emotion". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2488–2498.

[3] John Langshaw Austin. *How to do things with words*. Vol. 88. Oxford university press, 1975.

[4] Rowan Zellers, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. "TuringAdvice: A Generative and Dynamic Evaluation of Language Use". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, June 2021, pp. 4856–4880.

[5] Anne Maass. "Linguistic Intergroup Bias: Stereotype Perpetuation Through Language". In: *Advances in Experimental Social Psychology*. Ed. by Mark P. Zanna. Vol. 31. Academic Press, Jan. 1, 1999, pp. 79–121.

[6] D. Wigboldus, G. Semin, and R. Spears. "How do we communicate stereotypes? Linguistic bases and inferential consequences." In: *Journal of personality and social psychology* 78 1 (2000), pp. 5–18.

[7] Venkata Govindarajan, Benjamin Van Durme, and Aaron Steven White. "Decomposing Generalization: Models of Generic , Habitual, and Episodic Statements". In: *Transactions of the Association for Computational Linguistics (TACL)* 7 (2019), pp. 501–517.

[8] Venkata S Govindarajan, David Beaver, Kyle Mahowald, and Junyi Jessy Li. "Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 12853–12862.

[9] Yifan Gao, Yang Zhong, Daniel Preoțiuc-Pietro, and Junyi Jessy Li. "Predicting and Analyzing Language Specificity in Social Media Posts". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.1 (July 17, 2019). Number: 01, pp. 6415–6422. ISSN: 2374-3468.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[11] Jacob Andreas. "Language Models as Agent Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5769–5779.

[12] Lisa Bauer, Yicheng Wang, and Mohit Bansal. "Commonsense for Generative Multi-Hop Question Answering Tasks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4220–4230.

[13] Penelope Eckert. "Variation, convention, and social meaning". In: *Annual meeting of the Linguistic Society of America. Oakland CA*. Vol. 7. 2005.