# Research Statement

**Vagrant Gautam**

The overarching goal of my research is to **bridge natural language processing (NLP) with linguistics**, to **build trustworthy NLP systems that work for everyone**. Anyone with an internet connection can now use NLP technology to reword their emails, write their university essays, and translate text between different languages. However, despite how convincingly these systems model human language, they also perpetuate biases, do not work equally well for different groups of language users, and are often just inconsistent with reality. My work aims to drive the evolution of trustworthy NLP, by *putting the "natural language" back into "natural language processing."* This is somewhat neglected in contemporary NLP research, but it is critical, as natural language forms the bulk of what NLP systems are trained on, it is what they are intended to process, and it is how most people interact with them. Therefore, I perform evaluations that are grounded in well-studied linguistic phenomena, and I build linguistically-motivated systems that are more reliable at doing our bidding. Under the umbrella of trustworthy NLP, my interests span three focus areas:

- **Fairness** (§1): NLP technology that works for everyone
- **Faithfulness** (§2): NLP technology that is faithful to facts and the input we give it
- **Meta-evaluation** (§3): NLP research practices that are valid and reliable

**The Role of Linguistics in My Work** My work in academia and industry draws heavily from several areas of linguistics: I have used principles of coherent discourse to evaluate errors in reference that cause misgendering or perpetuate stereotypical biases [TACL '24, EMNLP '24 WS], I have built on sociolinguistics research to disentangle identity bias from bias with terms that index identity [ACL '24 WS, EMNLP '24 WS], and I have created more reliable NLP systems that answer questions, through syntactic data substitutions [EMNLP '23 FI]. Prior to my PhD, I was a speech recognition engineer in industry, where I created an open-source tool for English syllabification [GitHub '22], using English phonotactic rules to reject implausible sequences of phones. My tool increased the quality of our pronunciation dictionary and improved our speech recognition accuracy, and is thus used in both industry and academic research [ACL '21 WS].

**Interdisciplinarity and Methodological Pluralism** Concepts like "fairness" and "interpretability" that are central to trustworthy NLP have been studied for much longer and with a variety of methods in other disciplines. To comprehensively understand and use these concepts in NLP, I embrace interdisciplinarity and methodological pluralism, using definitions from relevant disciplines and flexibly choosing the best methods (whether quantitative, qualitative, or a mixture) to answer my research questions. For example, I have used political science literature to reveal gaps in how "democratic" AI research is [EMNLP '24], critical theory texts to show that "intersectionality" is too narrowly construed in research on fairness [AIES '23], and bibliometrics and qualitative methods to analyze the importance of interpretability and explainability for NLP research today [EMNLP '24].

**Vision for the Future** I want to lead a research group that bridges NLP with linguistics, using what we know about natural language to build systems that are fair, faithful, and trustworthy by design. This requires infusing every step of the NLP pipeline with linguistics: conceptualizing tasks, creating data, building systems, and evaluating system components, internals and behaviours. Overall, I want my work to continue to serve as evidence that a linguistic lens vastly improves our practice of NLP. In §4, I outline some concrete future directions I want to explore with my group.

# 1 Fairness: NLP Technology that Works for Everyone

Although a person's gender is not often explicitly indicated in English text, names and pronouns can act as cues to it; many cultures assign names in gender-typed ways, and third person singular English pronouns mark grammatical gender, which tends to pattern with social gender. As a first step towards NLP systems that work for everyone, my work analyzes **gender fairness in systems and society, via names and pronouns**.

**Pronouns** In the context of fairness with pronouns, two popular tasks studied in NLP literature are coreference resolution (i.e., finding co-referring expressions in text), and correct pronoun use (i.e., the absence of misgendering), both of which I have made significant contributions to.

In [EMNLP '24 WS] I showed that the conflation of pronouns with social gender is an unreliable way to measure gender bias, as different pronoun forms marking the same grammatical gender (e.g., *he*, *him* and *his*) are not associated with similar biases. For example, a system's tendency to associate *he* with *doctor* does not entail a tendency to associate *him* with *doctor*, even though prior work assumes that *he* and *him* are identical in their gendered associations.

I was the first to evaluate coherent pronoun use in a discourse-inspired multi-person context [TACL '24]. While prior work has focused on simplistic pronoun reuse, I designed a harder setup where *two* people are discussed with a combination of referring expressions and different pronouns. By creating a dataset of over 5 million instances, I showed that this setting is natural and easy for humans (who perform perfectly), while large language models do dramatically worse.

**Names** In [PLOS ONE '21], we computationally parsed news articles, extracted quotes from the parse trees, and then matched them up with the people quoted. We then used these names to infer the gender distribution of people quoted in Canadian news, and found significant disparities in the number of women quoted, compared to men. Mapping names to sociodemographic characteristics like gender and race in this manner is a common method in NLP fairness. In subsequent work [ACL '24 WS], I critiqued this practice, outlining issues of validity and ethics that lead to unreliable results and harms such as misgendering for real individuals. Our work also contributed concrete guidelines describing when to associate names with sociodemographic factors, as well as how to test the contextual validity of these associations and avoid harms.

# 2 Faithfulness: To Facts and Input

Beyond my work on faithful pronoun use [TACL '24], I also study **faithfulness in the context of facts and automatic systems for question answering**. If we humans are asked a question whose answer we don't know and can't find, we are capable of saying that we don't know, a trait that NLP systems of today are not typically endowed with. NLP systems instead make heuristic best guesses, which can have disastrous consequences if people believe them, e.g., for medical advice.

In [EMNLP '23 FI], I proposed a method to improve the reliability of NLP systems for question answering. The core of the approach is to give question answering systems both positive input (i.e., questions that can be answered) and negative input (i.e., questions that are unanswerable in context), to teach them to pay more attention to the question and available information, and to humbly abstain when there isn't enough information. My method involves creating unanswerable questions automatically, by replacing proper noun phrases in answerable questions with other relevant proper noun phrases of the same type. For example, although the question "Who is *Beyoncé*'s first child?" is answerable, the question "Who is *Blue Ivy*'s first child?" is not, as Blue Ivy (Beyoncé's daughter, a child herself) has no children. Compared to existing methods, mine is more lightweight and creates more human-interpretable questions, which in turn also creates question answering systems that are measurably more reliable.

# 3 Meta-Evaluation: Trustworthy NLP Research Practices

One of the challenges of research with NLP systems in 2024 is that people use these systems in a variety of creative ways that we may not be directly evaluating. We come up with "tasks" that are meant to formalize these human uses, we create datasets of examples, and metrics to help us automatically evaluate systems at scale. However, every step of this process can introduce gaps between what we want to measure and what we are actually measuring. This is why I am interested in meta-evaluation, and in reducing the gaps we introduce during conceptualization and operationalization. Much of the work I describe above makes important contributions to this area, e.g., [TACL '24] shows that system reasoning about pronouns can be overestimated without the more realistic multi-person setting, [EMNLP '24 WS] demonstrates empirical issues with a widespread conflation of pronoun forms with social gender, and [ACL '24 WS] provides concrete guidelines for the use of names in sociodemographic research.

My research also includes meta-evaluation work of a second kind: **highly interdisciplinary, multimethod work interrogating the use and impact of trustworthy NLP concepts** such as democracy, intersectionality and interpretability. "Democratization" has become a popular word to use in the NLP and machine learning literature, and indeed, truly democratic AI would be more trustworthy. However, in practice, our conceptual analysis of democracy in NLP and machine learning research [EMNLP '24] shows that it is mostly used as a buzzword instead of drawing from the 3000+ years of thought on democracy and democratization. "Intersectionality" is a framework from critical theory that is used in a similarly superficial way in fairness research. We find that the framework is flattened to a one-dimensional view of merely *intersecting* biases [AIES '23]. Finally, in [EMNLP '24], we perform a mixed-methods analysis to investigate the impact of "interpretability and analysis" work on the field of NLP. This kind of work builds trust in NLP systems by examining their internals and explaining their predictions, rather than treating them just as black boxes. All three projects used a mixture of quantitative and qualitative methods, studied varied sources of data, and involved reading and citing interdisciplinary work from several fields, including political science, critical theory, legal studies, and bibliometrics. I intend to continue this style of research in future work as my practice of both linguistics and NLP is enhanced by different disciplinary and epistemological perspectives.

# 4 Future Directions

## 4.1 Speech Technology for Everyone

While most of my work studies language in the modality of text, I also have an industry background and long-standing interest in speech. I want to **use sociolinguistics, phonetics and phonology to study and improve how speech systems handle variation in speech**.

Current research into building systems that process speech view dimensions of variation such as accents, dialects, and code-switching in an overly simplistic way. For example, researchers sometimes use country names to label English accents, which erases the existence of immigrants in Canada, and glosses over interesting dialectal variation in the US. These problematic labels are part of datasets that are used to create speech systems and even evaluate them for accent biases. In collaboration with sociolinguists, I want to develop higher-quality datasets of accent variation, to more accurately measure how well current systems handle variation, and to build systems that are better.

Another angle that I am interested in investigating is how these systems encode notions of variation: Systems that transcribe speech might cluster accents differently in their internals, while systems that synthesize speech might absorb and replicate reductive stereotypes about what people sound like. With each of these directions, I want to ensure that speech variation is handled fairly, represented faithfully, and conceptualized and operationalized in a linguistically-informed way.

### 4.2 Pragmatics in Interaction and Context

The latest and most popular incarnation of NLP technology is generative language models such as ChatGPT. Systems like this are increasingly used with chat interfaces and trained to mimic human conversational partners. This opens up several interesting avenues of study within human-computer interaction. Thus far, a neglected dimension of this has been pragmatics, which is why I want to **study pragmatic aspects of interaction in the context of fairness and factuality**.

The last year has seen an explosion of work on culture in NLP, analyzing how systems reflect cultural values and avoid or replicate stereotypes about different cultures. I am interested in applying the lens of cross-cultural pragmatics to NLP technology, to investigate whether aspects such as politeness, sarcasm and humour are represented in culturally-appropriate ways. This is critical for language-based assistants that are deployed in different cultural contexts. From a more technical perspective, I want to investigate which adaptation strategies lead to systems that follow cross-cultural pragmatic norms more reliably.

In the context of factuality, a body of work in NLP studies factual "knowledge conflicts," i.e., what happens when facts that a system has memorized conflict with a user's input. While existing work takes a limited, binary view of conflicts, I am more interested in the nuanced pragmatic strategies with which humans resolve conflicting information in interaction. For example, while reading a dystopian sci-fi novel, we might accept that the capital of Canada is now Vancouver. Similarly, Calgary is a more plausible fake capital for Canada than Cairo is. I am interested in investigating knowledge conflicts from these dimensions of situational and commonsense plausibility in NLP systems, as well as linguistic strategies for resolving knowledge conflicts, such as asking clarification questions.

### 4.3 Linguistically-Informed Non-English NLP

Current approaches to non-English NLP frequently involve simply applying methods that work for English to other languages, regardless of differences in their linguistic features, typological relationships, and available data. Perhaps unsurprisingly, this doesn't always work, as we showed in [ACL '24 FI]. We studied in-context learning, a popular technique to adapt NLP systems to specific tasks, and showed that even though it works well for English, it does not work as well with other languages and can sometimes even *worsen* their performance. In line with the high-level goals of my research, I want to depart from current Anglocentric trends in NLP and instead **use language-specific linguistic features to build non-English NLP systems**.

I am particularly interested in targeting aspects of fairness and faithfulness that I have already studied in English (e.g., queer languaging) in other languages. For instance, while singular they and English neopronouns have recently begun to be studied in NLP, there is limited work on queer languaging practices in morphologically more complex languages like German. Here, pronouns are not the only parts of speech that mark grammatical gender, and using gender-neutral language beyond the binary requires correctly inflecting nouns and adjectives as well.

Beyond language-specific systems, I also want to explore linguistic approaches to making multilingual systems more efficient. One approach might be to bake linguistic universals into the design of systems. I also want to test the feasibility of *automatically* building systems that separate facts from linguistic capabilities in different languages (similar to a separation between syntax and semantics), in a way that doesn't compromise overall system quality. Since different languages are represented on the internet in different quantities, NLP systems learn more facts in the languages that have more data, and cannot always reproduce these facts in a different language. Taking steps towards separating memory and languaging could potentially solve this open problem in NLP.

## References


[PLOS ONE '21] Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, **Vagrant Gautam**, Junette Gonzales, Prashanth Rao, and Maite Taboada. "The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media". *PLOS ONE* (2021). URL: https://doi.org/10.1371/journal.pone.0245533.

[GitHub '22] **Vagrant Gautam**. *Arpabet Syllabifier*. 2022. URL: https://github.com/dippedrusk/arpabet-syllabifier.

[TACL '24] **Vagrant Gautam**, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. "Robust Pronoun Fidelity with English LLMs: Are they Reasoning, Repeating, or Just Biased?" *Transactions of the Association for Computational Linguistics* (2024). URL: https://arxiv.org/abs/2404.03134.

[ACL '21 WS] **Vagrant Gautam**, Wang Yau Li, Zafarullah Mahmood, Fred Mailhot, Shreekantha Nadig, Riqiang Wang, and Nathan Zhang. "Avengers, Ensemble! Benefits of ensembling in grapheme-to-phoneme prediction". *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. 2021. URL: https://aclanthology.org/2021.sigmorphon-1.16.

[EMNLP '24 WS] **Vagrant Gautam**, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow. "WinoPron: Revisiting English Winogender Schemas for Consistency, Coverage, and Grammatical Case". *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*. 2024. URL: https://aclanthology.org/2024.crac-1.6.

[ACL '24 WS] **Vagrant Gautam**, Arjun Subramonian, Anne Lauscher, and Os Keyes. "Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP". *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. 2024. URL: https://aclanthology.org/2024.gebnlp-1.20.

[EMNLP '23 FI] **Vagrant Gautam**, Miaoran Zhang, and Dietrich Klakow. "A Lightweight Method to Generate Unanswerable Questions in English". *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023. URL: https://aclanthology.org/2023.findings-emnlp.491.

[EMNLP '24] Marius Mosbach, **Vagrant Gautam***, Tomás Vergara Browne*, Dietrich Klakow, and Mor Geva. "From Insights to Actions: The Impact of Interpretability and Analysis Research on NLP". *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024. URL: https://aclanthology.org/2024.emnlp-main.181.

[AIES '23] Anaelia Ovalle, Arjun Subramonian, **Vagrant Gautam**, Gilbert Gee, and Kai-Wei Chang. "Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness". *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. 2023.

[EMNLP '24] Arjun Subramonian*, **Vagrant Gautam***, Dietrich Klakow, and Zeerak Talat. "Understanding "Democratization" in NLP and ML Research". *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024. URL: https://aclanthology.org/2024.emnlp-main.184.

[ACL '24 FI] Miaoran Zhang, **Vagrant Gautam**, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. "The Impact of Demonstrations on Multilingual In-Context Learning: A Multidimensional Analysis". *Findings of the Association for Computational Linguistics: ACL 2024*. 2024. URL: https://aclanthology.org/2024.findings-acl.438.