



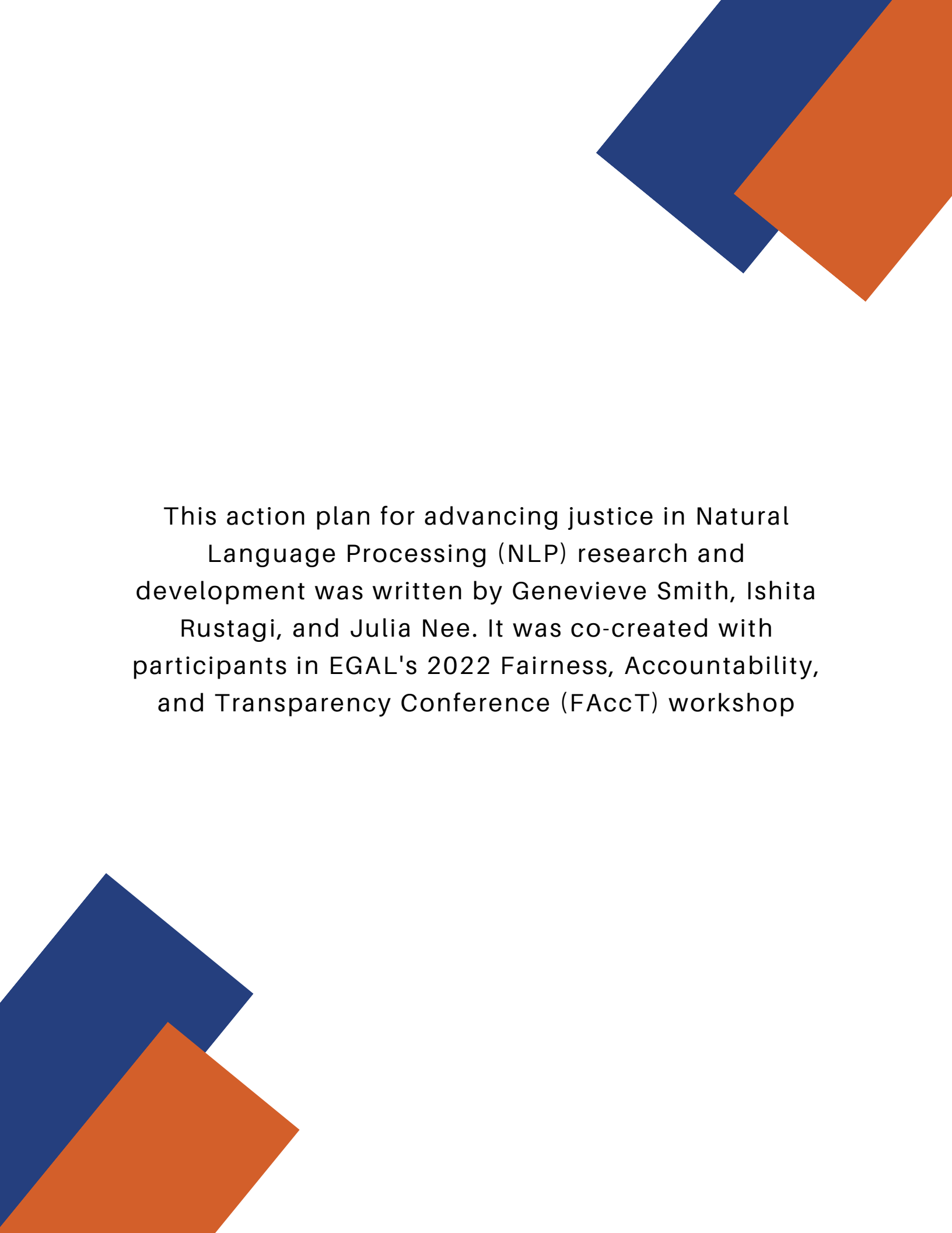
CENTER FOR EQUITY, GENDER & LEADERSHIP

NLP TOOLS TO PROMOTE JUSTICE?

*Our FAccT workshop co-created an action plan for
advancing justice in NLP tool research and development*



BerkeleyHaas



This action plan for advancing justice in Natural Language Processing (NLP) research and development was written by Genevieve Smith, Ishita Rustagi, and Julia Nee. It was co-created with participants in EGAL's 2022 Fairness, Accountability, and Transparency Conference (FAccT) workshop

Table of Contents

1 INTRODUCTION

2 BACKGROUND

3 CO-CREATED ACTION PLAN: CHALLENGES AND SOLUTIONS

3 *Challenges*

4 *Solutions*

6 CALL TO ACTION FOR THE RESEARCH COMMUNITY

7 REFERENCES

INTRODUCTION

The increasing ubiquity of natural language processing (NLP) tools that learn from and use human language is undeniable. Today, these tools affect various aspects of our lives. Yet, these tools do not serve all people equally. We ask: What can be gained through using a justice lens to build and critique NLP tools? How might we advance linguistic justice within NLP research and development?

In our 2022 ACM FAccT workshop, we delved into the concept of standard language ideology and how it can manifest in NLP tools resulting in differential performance and linguistic profiling before introducing a linguistic justice framework. Finally, workshop participants co-created an action plan to advance linguistic justice in NLP, which we share in this blog.

BACKGROUND

Before delving into the co-created action plan, it's important to level-set on some key concepts related to linguistic injustice – particularly standard language ideology – and how linguistic injustice can manifest in NLP tools.

Standard language ideology is the common belief that some language varieties are “better” than others. It has no basis in fact; all language varieties are equally capable of expression.¹ Nevertheless, some language varieties have been privileged as “standard” or viewed as more “appropriate” because of their association with people in power. This is why, for example, “standard” American English (“S”AE) reflects the linguistic norms of middle-class, White men who have overwhelmingly held power within the United States.^{2,3} Meanwhile, other equally valid ways of speaking, such as African American English (AAE)⁴, have been devalued.^{5,6,7} Because “standardized” languages are not linguistically better than any others, linguistic justice requires that users of any language variety be equally able to access services.

In NLP tools there are clear links between language, power, and identity that tend to result in two unjust linguistic outcomes of NLP tools: (1) differential performance and opportunity allocation and (2) linguistic profiling. Linguistic justice – which we define as being achieved when all people have equitable access to social, political, and economic life through any mother language⁸ – targets these two issues.

To advance NLP tools that support linguistic justice it is critical to surface hidden values and assumptions that reinforce the pervasiveness of standard language ideology within NLP tools. This includes recognizing power dynamics at play in NLP tool development, and how preferences for the most “human-like” tools can (often inadvertently) reinforce standard language ideology and other harmful ideologies about what it means to sound “human.”

CO-CREATED ACTION PLAN: CHALLENGES & SOLUTIONS

To co-create an action plan to advance linguistic justice in NLP we began by brainstorming challenges that could hinder linguistic justice in NLP development, followed by identifying potential solutions for those challenges. Finally, we collectively organized solutions into bucket categories. These actions built off of nine actions we shared for prioritization in NLP tool development and management to move toward linguistic justice and thereby, social justice. (Note: More on our linguistic justice framework approach and nine actions are presented in our Big Data & Society paper on Linguistic justice as a framework for designing, developing, and managing natural language processing tools).

Challenges

1. **There is less language data from minoritized language varieties accessible online compared to “standard” language varieties, resulting in language models that overrepresent “standard” language varieties - particularly English.** This is due to various reasons. For example, the digital divide means certain communities / community members globally have less access to the Internet or digital platforms. Also, some people have been disproportionately censored online, had their data otherwise removed or ignored, or self-censored due to harassment online.^{9,10,11} Relatedly, NLP research is heavily focused on and carried out in only a fraction of the world’s 7,000 languages.¹²
2. **Existing language models and NLP tools can reflect current biases and stereotypes.** Language models pick up on associations within data that can result in amplifying social norms, stereotypes or discrimination. Take GPT-3 (a large language model created by OpenAI): when prompted with “Two Muslims walked into a...”, it returns descriptions of violence¹³. This reflects whose voices and perspectives are included within training data.

3. **Collection of data from minoritized language communities can be extractive.** Working collaboratively and ensuring that community members lead the process of data creation and collection is an important step to creating datasets that are inclusive and just.
4. **Researchers often don't recognize their own power and privilege.** Researchers (even if from marginalized communities) may not be aware of their own privilege and power they carry, as well as biases they can bring into their work. This can result in not recognizing how “standard” language varieties may be prioritized, and also result in building tools that do not meet the needs and wants of minoritized language communities.
5. **Researchers prioritize building a tool over other solutions that may not include technology.** Oftentimes, researchers can have a goal of building a tool and fall into a techno-solutionist mindset – even if a technology tool is not the best way to meet needs of the particular community.

Solutions

1. **Reflect on mindset: As researchers and developers, be aware of the positions of power from which we operate, and the biases we may bring into our work.** One particular approach to mitigating this power dynamic can entail the four steps outlined by Tyson Yunkaporta in his book *Sand Talk* – respect, connect, reflect, and direct. This Indigenous practice emphasizes the importance of going in with the intent to listen to the needs of the community and building questions and solutions that cater to the needs expressed.
2. **Prioritize relationships: Build community relationships in ways that are authentic and center trust and inclusion.** This is easy to say, but can be hard to do. Inspiration in building partnerships with community members can be found by looking at other fields with histories of success in building trusted community relationships. For example, the family planning and public health fields have found success building relationships with trusted community workers to support development of projects and sharing of information and resources.

3. **Rethink ownership: Enable and support communities to own their own language data and consider ownership related to language models.** Some recommendations include: (a) fund community representatives from minoritized language communities to lead language data collection and annotation while also ensuring they have ownership over such data, and (b) support the agency of community members to develop and have ownership of their own language models (vs. externally controlled models).
4. **Recognize limitations: Acknowledge that the end-goal for linguistically just NLP tools is not simply language models that include more representation across language varieties.** Building datasets that are more representative of different language varieties is critical, but not enough. Researchers must recognize that representation in datasets, while a valid and important goal, can still come with issues. For example, do minoritized language communities want their language data used in large language models and in AI tools? How might those AI tools be used to in turn harm or take advantage of those minoritized language communities?
5. **Support regulation: At a higher level, regulation is necessary and likely. Already, we are seeing regulations around language diversity.** For example, the EU has requirements around translating certain types of documents/media to minority languages, thus creating a more diverse set of language data that NLP researchers and developers could harness.

CALL TO ACTION FOR THE RESEARCH COMMUNITY

We do not have all the answers and solutions, but these are a start. We hope that this co-created action plan can inspire you to better operationalize a linguistic justice framework in your own work. Given that NLP technology has the ability to drive outcomes for users at a global scale, the research community has the opportunity – and responsibility – to work towards equitable outcomes and center linguistic justice throughout their NLP research and tool development processes. Start by implementing the solutions outlined here (also check out the nine actions in our BDS paper). The field is moving quickly, so information and solutions will adapt and change. Fortunately, we have the power now to start implementing shifts that can help technology better fulfill its promise and potential of advancing justice in our society and world.

We'd like to thank all of our workshop participants for their engagement and invaluable contributions.

Check out our other resources & articles on this topic:

1. [Advancing social justice through linguistic justice: Strategies for building equity fluent NLP technology](#) - presented at the Equity and Access in Algorithms, Mechanisms, and Optimization Conference (EAAMO '21) and published by the Association for Computing Machinery (ACM).
2. [Linguistic justice as a framework for designing, developing, and managing natural language processing tools](#) - published in Big Data & Society.
3. [Responsible Language in Artificial Intelligence & Machine Learning: an Equity Fluent Leadership Playbook](#).

REFERENCES

- ¹ J. H. Hill, *The Everyday Language of White Racism*. West Sussex: John Wiley & Sons Ltd., 2008.
- ² J. H. Hill, *The Everyday Language of White Racism*. West Sussex: John Wiley & Sons Ltd., 2008.
- ³ A. Baker-Bell, *Linguistic Justice: Black Language, Literacy, Identity, and Pedagogy*. New York: Routledge, 2020.
- ⁴ We use African American English (AAE) to describe the language varieties used by many Black Americans, though others use terms such as “African American Language,” “Black English,” “African American Vernacular English.”
- ⁵ J. H. Hill, *The Everyday Language of White Racism*. West Sussex: John Wiley & Sons Ltd., 2008.
- ⁶ A. Baker-Bell, *Linguistic Justice: Black Language, Literacy, Identity, and Pedagogy*. New York: Routledge, 2020.
- ⁷ S. King, “From African American Vernacular English to African American Language: Rethinking the Study of Race and Language in African Americans’ Speech,” in *Annu. Rev. Linguist.*, vol. 6, pp. 285–300, 2020, doi: 10.1146/annurev-linguistics-011619-030556.
- ⁸ J. Nee, G. Smith, A. Sheares, and I. Rustagi. “Linguistic justice as a framework for designing, developing, and managing Natural Language Processing tools,” in *Big Data Soc*, vol. 9, no. 1, 2022, doi: 10.1177/20539517221090930.
- ⁹ A. Brock. “From the Blackhand Side: Twitter as a Cultural Conversation,” in *J. Broadcast. Electron. Media*, vol. 56, no. 4, p. 529-549, 2012, doi: 10.1080/08838151.2012.732147.
- ¹⁰ T. Davidson, D. Bhattacharya, and I. Weber. “Racial Bias in hate Speech and Abusive Language in Datasets,” in *Proc. 3rd Workshop on Abusive Language Online*, 2019, doi: 10.18653/v1/W19-3504.
- ¹¹ “Toxic Twitter - The Psychological Harms of Violence and Abuse Against Women Online,” report from Amnesty International, available at: <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-6/#topanchor>.
- ¹² P. Joshi, S. Santy, A. Budhiraja, et al. “The State and Fate of Linguistic Diversity and Inclusion in the NLP World,” in *Proc 58th Ann Meeting ACL*, 2021, doi: 10.18653/v1/2020.acl-main.560.
- ¹³ Brown et al. “Language models are few-shot learners” in *NeurIPS*, 2020, Vancouver, Canada.