EXPLORING LINGUISTIC JUSTICE AND DATA EQUITY IN AI

A Research Paper submitted to the Department of Engineering and Society
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Anusha Choudhary

March 30, 2023

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

In the Information Era, the world is connected not only by means of communicating over large distances but more importantly, by means of communicating despite differences in language. Natural Language Processing (NLP) is a wide sub-field of Machine Learning and Artificial Intelligence that aims to use machine learning to solve several tasks based on spoken and written natural language. Machine Translation is one such task within NLP, for which the current state-of-the-art technology uses a neural network to maximize translation performance (Bahdanau, Cho, & Bengio, 2016) and is thus aptly termed Neural Machine Translation (NMT). Current state-of-the-art translation models perform worse on language pairs for which there exists a smaller amount of data (Koehn and Knowles, 2017). Such language pairs are termed in the field as low-resource pairs. This means that in the world of Neural Networks, inequity in resource availability is synonymous with inequity in performance quality.

This research paper will explore not only why the problem of low-resource machine translation continues to exist from both a technical and a sociological perspective, but also why the problem is currently unexplored in the field of Science, Technology, and Society, as well as what frameworks can be used to overcome it.

To illustrate the relation between society, industry, and NMT and to explore different aspects of the imbalances in resource availability and translation quality across language varieties, we will look at three frameworks: Linguistic Justice, introduced by Nee et al. (2021), Social Construction, introduced by Pinch and Bijker (1984), and the Actor-Network Theory, introduced by Callon (1986).

The overall motivation for this research is two-fold: i) the state-of-the-art technical paper presents the most recent research in the field of low-resource language machine translation in a consolidated manner and analyzes any trends emerging from the presented information, and ii) the STS paper uses the frameworks of Linguistic Justice, Actor-Network theory, and Social

Construction to formulate a frame of reference that developers and researchers of NMT can use while improving and evaluating NMT models that includes all stakeholders of NMT tools and puts linguistic justice at the forefront. Thus, the technical and STS papers are tightly coupled and the results from one paper affect the other in a pivotal manner.

CURRENT STATE

The dichotomy of Artificial Intelligence simultaneously being far from a new concept within the field of Computer Science and being a fresh, new object of fascination in the public consciousness can be extended to the technical and social challenges faced by it as well: as Koehn & Knowles (2017) identified over half a decade ago, NMT worsens in quality on smaller datasets (p. 4), which is a challenge that has been faced by NMT models from the very beginning of their development. As it stands, English and languages with typological features similar to English such as Spanish, German, and French make up the majority of the available resources to train NMT models (Joshi et al., 2020). Consequently, populations dependent on translating between these low-resource pairs are at a disadvantage due to inequity in data. However, apart from sporadic research, little had been done to overcome the challenge of low-resource machine translation until the past two years, with the most notable achievement being the release of No Language Left Behing (NLLB) (Costa-Jussa et al, 2022).

It is productive to interpret the current state of the problem using a multi-layered approach, as shown in Figure 1 below, where an imbalance in the sociological layer cascades to problems in the technical sphere, which finally cascade to all users of machine translation tools.

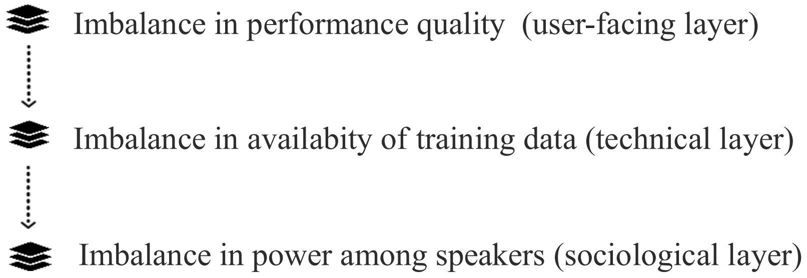
Imbalance in performance quality (user-facing layer)
Imbalance in availabity of training data (technical layer)
Imbalance in power among speakers (sociological layer)

Figure 1: Current State Diagram. This figure shows the current state of the three-layered problem of low-resource machine translation. (Choudhary, 2023)

In this paper, we focus on the sociological layer and attempt to answer why the problem has been unexplored on the sociological layer so far.

AN OBSCURED THREAT TO SOCIAL JUSTICE

The present situation of the problem of the poor performance of low-resource machine translation being previously unexplored in the field of Science, Technology, and Society is not entirely surprising when the risks associated with this problem are identified as a new risks. As defined by Martin (2023), some risks associated with technology can be classified as "new" risks, not because of their temporal newness but because the less obvious effects of technology are newly making their way to public consciousness. Martin (2023) describes such risks "are new only in the sense that (1) they are now identifiable—because of changes in the magnitude of the risks they present, because they have passed a certain threshold of accumulation in our environment, or because of a change in measuring techniques, or (2) the public's perception of them changed—because of education, experience, media attention, or a reduction in other hitherto dominant and masking risks." (pp. 108-109) In the case of the poor performance of low-resource machine translation, the risk is the threat to social justice. Inequity in the quality of machine translation across languages poses a limitation not only for the developers of neural machine translation tools, but it also stands as an obstacle to social justice. As Nee et al. (2021) argue, language and social reality are mutually enforcing (p. 2), and thus, linguistic injustice perpetuates social injustice. Both senses of newness of this risk presented by Martin (2023) apply here. To supplement the first sense of newness, Nee et al. (2021) provide recent cases that have made the risk of the threat to social justice identifiable (p. 2) along further instances of inequitable Natural Language Processing (NLP) tools such as Automated Speech Recognition (ASR) technology underperforming for dialects outside of Standard American English and inequity in algorithmic ranking of video search results for some language varieties.

To supplement the second sense of newness, a masking risk has been the inefficiency of high-resource machine translation so far. It is not unimaginable that researchers in the past decade have been so preoccupied on achieving good quality machine translation in the first place that the language pairs they were working with were not even an afterthought, but an inconsequential parameter hidden amongst a mound of more significant parameters. With the identification for its need and the relative convergence of the quality of high-resource machine translation, good quality machine translation across all languages has graduated from an inconsequential parameter to an afterthought, and finally to the central focus of many researchers in the field of Natural Language Processing. But what technical and social factors caused this imbalance in the first place?

The historical imbalance in the availability of resources for low-resource languages in machine learning models and the resulting inefficiency in the quality of machine translation for these languages is tightly coupled with the existing imbalances present in society and the technology industry. Accordingly, the motivation behind examining the role of society and the technology industry in the development of machine translation models is not only to improve the state of the existing machine translation tools for a wider population of language speakers, but also to advance social justice.

The technical cause is clear, and inherent to the nature of neural networks. Inefficiency in low-resource machine translation is caused by the inequity in availability of training corpuses for certain languages because of the prevalence of English, French, Spanish, and German data on popularly used text sources (Joshi et al., 2021). The inherent nature of large language models and neural networks is such that they perform well only on sufficiently large training datasets, which is the bottleneck constraint that all developers of machine translation models must work with.

The question of a social or industrial cause for this imbalance is one that has not been asked enough, and, as a consequence, the answers for it are elusive. This is what we attempt to answer next. We attempt to find out not only what social and industrial factors have influenced this inequity in big data, but also what can be done to make machine translation models more equitable and further the goal of linguistic and social justice.

REDISTRIBUTING POWER IN MACHINE TRANSLATION AND AI

So far, we have seen that when viewed from a technical standpoint, the imbalance in the performance quality in high and low-resource languages is a direct result of an imbalance in the *availability of training data* for high and low-resource languages. In the following sections, we will see that when viewed from a sociological standpoint, the imbalance in the performance quality in high and low-resource languages is a direct result of an imbalance in *power* between high-resource language speakers and low-resource language speakers.

LINGUISTIC JUSTICE IN MACHINE TRANSLATION

In this section, we adapt the framework of Linguistic Justice (Nee et al., 2021) to the context of machine translation and explore ways to implement linguistic justice ideas into the practice of developing new models.

Linguistic justice as introduced by Nee et al. (2021) provides a four-layer approach to frame the development of NLP tools (pp. 3-6). The first layer focuses on equity and inclusion in the choices of words and phrases (p. 3), the second layer focuses on inclusive organization and labeling of words and phrases (pp. 3-4), the third layer emphasizes time, indexicality and context of words and phrases (pp. 4-5), and the fourth layer highlights power and accessibility inequities in NLP tools (pp. 5-6). Nee et al. (2021) refer to all Natural Language Processing technology when they present the framework of linguistic justice, placing no special emphasis on Neural Machine Translation. This leaves space for further exploration of linguistic justice in the specific context of NMT.

Viewing NMT from a linguistic justice lens, the second and fourth layers of linguistic justice emerge as the most relevant subjects for discussion, as inclusivity of language structure (implicated by the second layer in Nee et al. (2021, pp. 3-4)) and power and resource inequities experienced by speakers of low-resource languages (discussed in the fourth layer in Nee et al. (2021, pp. 5-6)) play the most pivotal roles in the inclusivity of machine translation. The second and fourth layers of linguistic justice will be further explored in the specific context of NMT in the following two sub-sections.

Linguistic Inclusivity in Model Development

This subsection focuses on the second layer of linguistic justice (Nee et al., 2021, pp. 3-4), inclusive organization and labeling of words and phrases. Nee et al. (2021, p 4) provide two questions for developers of NLP models to ensure inclusive organization and labeling of words and phrases. We will view both questions from a Machine Translation lens.

The first question that may be asked is, “How might an NLP tool be built to help individuals or organizations utilize patterns of organizing words and phrases for equity and inclusion? For example, can a tool flag for human review (e.g., for journalists and writers) potential uses of thepassive voice and personification of institutions?” From a Machine Translation perspective, this question can prompt developers of machine translation models to invest more time into researching the sentence structures specific to each individual low-resource language and adapting the model to that sentence structure rather than attempting to repurpose a model designed for languages of altogether different sentence structures.

The second question posed by Nee et al. (2021) is, “How might we ensure that datasets include accurate data that does not replicate deficit-based narratives?” (p. 4). When approaching this question from a Machine Translation perspective, one must think about the source corpuses from which training data is collected for low-resource machine translation. If the people collecting data are themselves speakers of high-resource languages, they may be searching in corpuses most familiar to them, which may be an unconscious manifestation of their deficit-based worldview that a limited pool of corpuses exists for training data. Low-resource language speakers across the world may prefer native databases, such as Yandex for Russian or Naver for Korean, which could be huge sources of text data if developers of NMT tools can expand their deficit-based mindsets of acknowledging Google, Wikipedia, and other popular corpuses as the only valid sources of training data.

Who Holds the Power?

This subsection focuses on the fourth layer of linguistic justice (Nee et al., pp. 5-6), power and accessibility inequities in NLP tools. Out of the seven questions posed under this layer of linguistic justice (Nee et al. 2021, p. 6), several apply directly to low-resource Machine Translation without the need for an intermediate rephrasing. The most important questions for developers of NMT tools to consider are: “Who is the target population for our tool? Why? Are our choices of target audience inclusive or do they reflect harmful stereotypes? Have we included members of the target audience in the development of the tool?”, “How might we be more transparent about the data our NLP tool is trained on and associated limitations of the tool? Have we audited our NLP systems to make sure that they work well for different language varieties, particularly target and potential user populations?”, “Are data labellers fluent in the language variety they are working with? Have data labellers been trained to counter their implicit biases?” and “What language varieties are represented in our training data and outputs? Do these varieties reflect the range of language used by the population of potential users? Is our target population maximally inclusive?” (Nee et al. 2021, p. 6).

A pair of questions that deserve special emphasis are “Have we ensured that consent for use of language data has been given following culturally appropriate practices for the particular language community? Have we collaboratively and fairly engaged with marginalized language communities so that members of those groups can provide input and/or lead throughout the process from deciding whether or not to participate, to informing data collection, labeling and processing, to tool development and implementation? Does the tool address the needs and goals of the particular language community/ies?” and “Have we ensured appropriate privacy and ownership of language data?” (Nee et al. 2021, p. 6). Before aiming to achieve better quality machine translation across low-resource languages, few, if none, consider that the parameters that define better quality of performance could differ across languages. For example, accuracy of translation could not only mean preserving the meaning behind a sentence, but also preserving the honorific quality of a sentence, which may be a very important factor in languages such as Japanese and Korean, but relatively less important in a language like English. Asking speakers of low-resource languages which parameters they hold most important is imperative to improve the quality of translation in those languages. The second question that speakers of low-resource language speakers must be consulted about, especially in cases where they are the direct providers of the language data, is that of ownership and whether they consent to the use of their language data, and if so, what their terms of consent are.

ANALYZING PRESENT POWER IMBALANCES

In order to redistribute the power in the context of the development of machine translation models, we must first analyze the present power imbalances, which we do here through actor-network theory and social construction.

Actor-Network Theory

Pinch and Bijker (1984)'s Social Construction model places the engineer at the theoretical center of the discussion around technology, Actor-Network Theory provides an opportunity to explore the dynamics between the stakeholders of a piece of technology not only with the engineer but also with other stakeholders as well as the technology itself. Figure 4 uses ANT to contrast the dense network of speakers of high-resource languages**,** big tech corporations, research conferences, data labelers, NMT tools and developers of NLP tools with the sparsity of connections between speakers of low-resource languages and all other stakeholders. ANT also allows for the use of relative image size to highlight the amount of power any one actor has over other actors; consistent with arguments presented by Luitse and Denkena (2021), Nee et al. (2021), and Joshi et al. (2020), big tech corporations, research conferences, and the NMT models themselves occupy more power over users of NMT tools, developers of NMT tools, and data labelers and thus are portrayed as larger images in Figure 2 below.

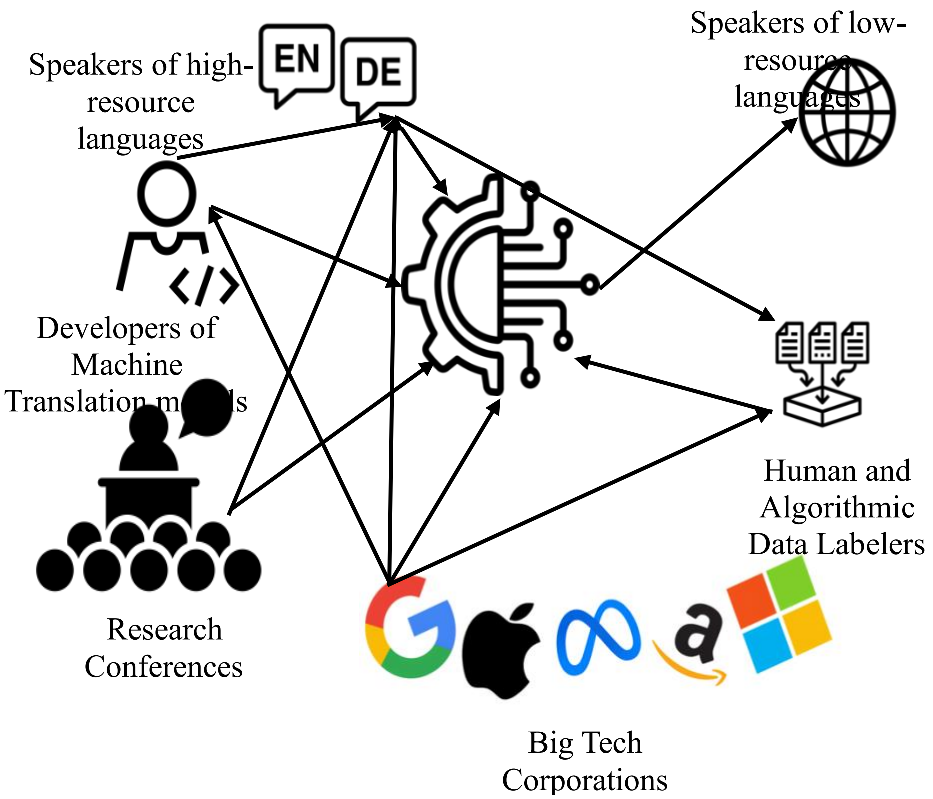
Speakers of high-resource languages
EN
DE
Speakers of low-resource languages
Developers of Machine Translation models
Human and Algorithmic Data Labelers
Research Conferences
Big Tech Corporations

Figure 2: Actor-Network Theory for Machine Translation. This figure shows the links between the machine translation models (center) and the other actors in this network; notably, the fewest links exist between speakers of low-resource languages and the other actors. (Choudhary, 2022)

Social Construction

Consolidating the existing STS research related to Neural Machine Translation, five major societal and industrial stakeholders of NMT can be identified. First, a distinction must be drawn between users of NMT tools who speak high-resource languages and those who speak low-resource languages; while both groups of speakers are users of NMT tools, speakers of high-resource languages contribute to the development of NLP tools by way of providing training data for models but speakers of low-resource languages have restricted influence on the development of NLP tools since training data from low-resource language varities is rarely used. This one-way line of communication between low-resource language speakers and developers of NLP tools reinforces Nee et al. (2021)'s argument of language and power being intertwined (p. 2). Joshi et al. (2020) brings up research conferences, most notably the Association for Computational Linguistics (ACL), as another set of entities that influence development of NMT tools and are also influenced by emerging trends in new NLP tools. Nee et al. (2021) points to biases in data labelers as sources of bias in NLP tools, which points to how both human and algorithmic data labelers influence the development of NLP tools although they may not directly use the tools or be impacted by them. Lastly, Luitse and Denkena (2021, p.1) argue that the release of open-source models from big tech corporations such as Google's Bidirectional Encoder Representations from Transformers (BERT) model result in a monopolization of the market and a concentration of power in the hands of big tech corporations. This trend has been repeated with Meta AI releasing No Language Left Behind (NLLB) as open-source code on GitHub (Costa-jussà, 2022). Thus, big tech corporations play a big role in the development and accessibility of NMT tools. The interactions of the five major stakeholder groups mentioned in this section with the developers of NMT models are summarized using Pinch and Bijker (1984)'s Social Construction model in Figure 3 below.

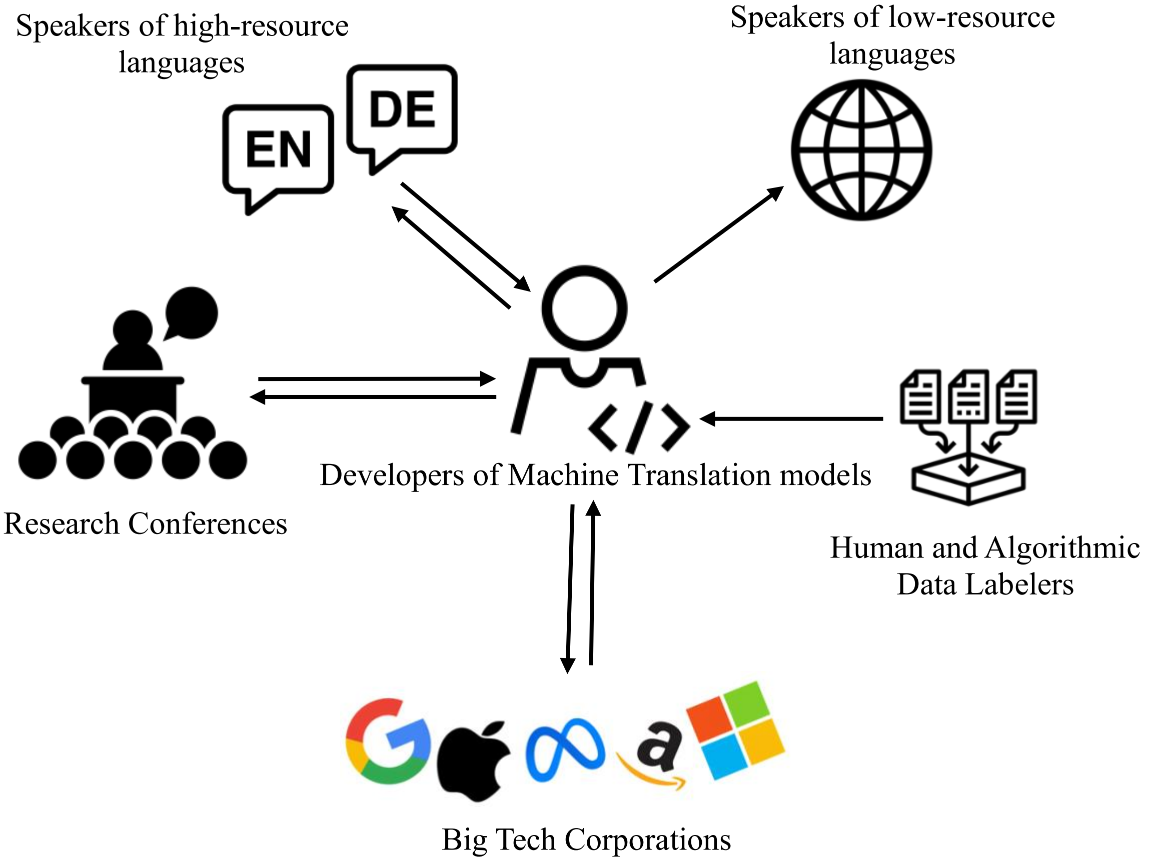
Speakers of high-resource languages
EN
DE
Speakers of low-resource languages
Research Conferences
Developers of Machine Translation models
Human and Algorithmic Data Labelers
Big Tech Corporations

Figure 3: The Social Construction Framework for Low-Resource Machine Translation. This figure shows the interactions of five major stakeholder groups of NMT tools with the developers of NMT models. (Adapted by Choudhary, 2022 from Bijker & Pinch, 1984)

FORMULATING A FRAME OF REFERENCE

Combining key concepts of Linguistic Justice and the results from Social Construction and Actor-Network Theory, we provide a frame of reference below that can act as a set of guiding principles for engineers and researchers in the field of NMT who wish to put linguistic justice, and hence, social justice, at the forefront of their work in this field.

1. Developers of NMT models must consider whether they are sufficiently inclusive of the diverse sentence structures, organization and labelling of words and phrases of each language they include in their machine translation models.
2. To build a bridge between the technology and all its user groups, language speakers must be included in all stages of model development.

a. Language speakers must be included in the data collection stage with appropriate discussions of ownership, privacy, consent and data labelling.

b. Language speakers of a variety of low-resource languages must be a part of the team of developers building NMT tools, not just third-party consultants. An example of this positively contributing the development of a NMT model is displayed in the video on Meta AI's website introducing NLLB, which features developers at Meta having deeply personal ties to the languages in the model.

c. Language speakers must hold significant power in the evaluation stage of the model, not only in deciding how well the model performs on certain metrics, but also in deciding the metrics themselves.

The problem of Neural Machine Translation on low-resource languages is inherently both a technological and a social problem. When a problem is both technological and social, applying either an exclusively technological fix or an exclusively sociological fix may leave gaps in the solution and thus in the equitability of machine translation. It is imperative that technologists and researchers keep questions of linguistic justice at the forefront when improving and evaluating NMT models. The hope is that the technical and the STS papers be treated as complementary entities that successfully provide both an account of the technological improvements in the current state-of-the-art Neural Machine Translation models as well as an exposition on the sociological areas for improvement in the context of Neural Machine Translation.

CONCLUSION: FINE-TUNING TOWARDS LINGUISTIC JUSTICE

Fine-tuning in the context of machine learning is the process of improving the performance of a model by modifying quantitative parameters in the model. To achieve linguistic justice in machine translation models, the idea of fine-tuning must be extended to the qualitative aspects surrounding the development of such models, resulting in the fine-tuning of the societal and industrial actors, their relationships, and their power balances in the network of machine translation models. As we saw, redistributing the power when it comes to model development and actively involving speakers of low-resource languages in all stages of the model development process is the key to more efficient machine translation across all languages.

REFERENCES

Bahdanau, D., Cho, K. & Bengio, Y. (2015, May 7-9). *Neural machine translation by jointly learning to align and translate* [Conference presentation]. ICLR 2015, San Diego, CA, United States.

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Interaction of Linguistics and Computational Linguistics*, *6*(1). https://doi.org/10.33011/lilt.v6i.1239

Callon, M. (1986). The sociology of an actor-network: The case of the electric vehicle. *Mapping the dynamics of science and technology*.

Carlson, B. (2009). *Social Construction*. [Figure 4]. Class handout (Unpublished). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Choudhary, A. (2022). *Actor-Network Theory for Machine Translation.* [Figure 4]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Choudhary, A. (2022). *Current State Diagram.* [Figure 1]. *Exploring Linguistic Justice and Data Equity in AI.* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & Wang, J. (2022). No language left behind: Scaling human-centered machine translation (Unpublished). *arXiv:2207.04672*. https://doi.org/10.48550/arXiv.2207.04672

Epaliyana K., Ranathunga S. & Jayasena S. (2021). Improving back-translation with iterative filtering and data selection for Sinhala-English NMT. *Proceedings for the Moratuwa engineering research conference*. https://doi.org/10.1109/MERCon52712.2021.9525800

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). *Plots of Different Available Resources for Different Languages.* [Figure 3]. The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th annual meeting of the association of computational linguistics.* https://doi.org/ 10.18653/v1/2020.acl-main.560

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th annual meeting of the association of computational linguistics*. https://doi.org/ 10.18653/v1/2020.acl-main.560

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. *Proceedings for the first workshop on neural machine translation*. https://doi.org/10.48550/arXiv.1706.03872

Luitse, D., & Denkena, W. (2021). The great Transformer: Examining the role of large language models in the political economy of AI. Big Data & Society, 8(2), 191–205. https://doi.org/10.1177/20539517211047734

Martin, M. & Schinzinger, R. (2009). *Introduction to Engineering Ethics* (2nd ed). McGraw-Hill Education.

Nee, J., Smith, G.M., Sheares, A. & Rustagi, I. (2021). Advancing social justice through linguistic justice: strategies for building equity fluent NLP technology. *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*. https://doi.org/10.1145/3465416.3483301

Nee, J., Smith, G.M., Sheares, A. & Rustagi, I. (2022). Linguistic justice as a framework for designing, developing, and managing natural language processing tools. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517221090930

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, 14(3), 399–441. http://www.jstor.org/stable/285355

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books, Washington DC. http://catalog.hathitrust.org/Record/000203591

Sennrich R. & Zhang B. (2019). Revisiting low-resource neural machine translation: a case study. *Association for Computational Linguistics*. https://doi.org/10.18653/v1/P19-1021