NEURAL MACHINE TRANSLATION ON LOW-RESOURCE LANGUAGE PAIRS

EXPLORING LINGUISTIC JUSTICE AND DATA EQUITY IN AI

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Anusha Choudhary

December 2, 2022

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.

ADVISORS

Catherine D. Baritaud, Department of Engineering and Society

Yangfeng Ji, Department of Computer Science

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. However, research has long established that the 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); these language pairs are termed in the field as low-resource pairs. Joshi et al. (2020) illustrates quantitatively how languages with typological features similar to English have an overall higher availability of resources across Wikipedia, the web, and popular language consortiums (p. 4). Consequently, populations dependent on translating between these low-resource pairs are at a disadvantage due to inequity in data. The efficiency and usefulness of popular, freely available translation tools such as Google Translate varies widely depending on the language pair one chooses for translation. Nee et al. (2022) points to more such 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.

Several efforts worth being summarized have been made in the field of Natural Language Processing and Neural Machine Translation to overcome this gap in translation quality. While it is imperative to consolidate the current research being done to achieve more equitable machine translation technology, it is all the more critical to ask what societal conditions led to the inequity in resource distribution across language varieties in the first place. Therefore, the overall motivation for this research is two-fold: i) the state-of-the-art technical paper attempts to present the most recent research in the field of low-resource language machine translation in a consolidated manner and to analyze any trends emerging from the presented information, and ii) the STS paper attempts to use the frameworks of Linguistic Justice, Actor-Network theory, and

Social Construction to illustrate the dominant societal and industrial conditions that continue to directly impact the current development of new machine-translation models. Thus, the technical and STS papers are tightly coupled and the results from one paper affect the other in a pivotal manner.

The personnel involved in this research include Anusha Choudhary, currently a fourth-year undergraduate student pursuing a major in Computer Science in the School of Engineering and Applied Sciences and a minor in Data Analytics at the University of Virginia. Yangfeng Ji, William Wulf Career Enhancement Assistant Professor at the Department of Computer Science and leader of the Information and Language Processing (ILP) Lab at the University of Virginia will serve as an advisor for the state-of-the-art technical paper and Catherine D. Baritaud, Senior Lecturer in the Science, Technology, and Society program at the University of Virginia will serve as the advisor for the STS paper. The following chart in Figure 1 outlines the timetable for these deliverables.

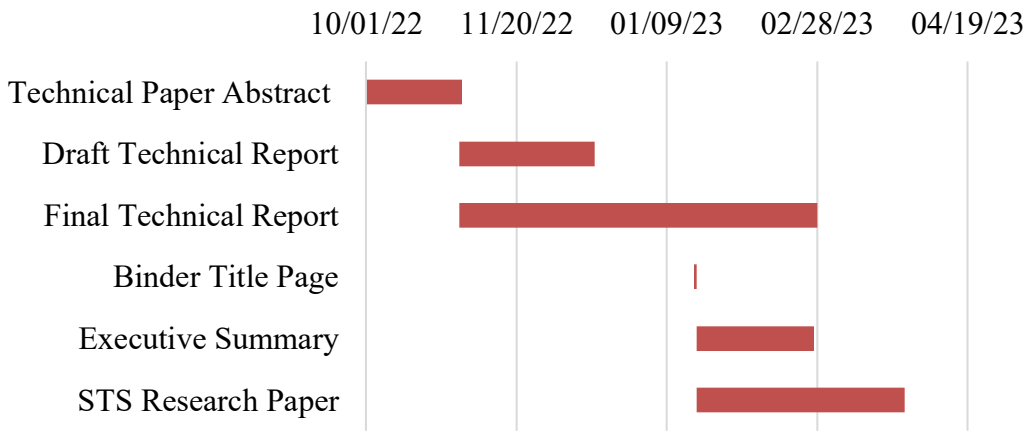


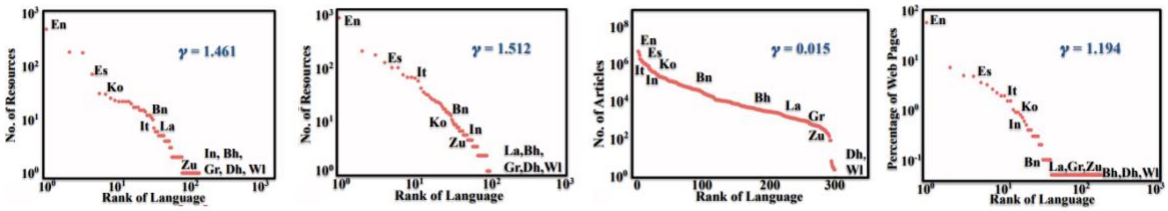
Figure 1: Gantt Chart for Tehnnical and STS deliverables: This chart displays the start and end dates for each deliverable required for the technical project and the STS portfolio. (Choudhary, 2022)

NEURAL MACHINE TRANSLATION ON LOW-RESOURCE LANGUAGE PAIRS

Natural Language Processing is a wide sub-field of Machine Learning and Artificial Intelligence that aims to use machine learning to solve several tasks such as analyzing the sentiment behind a piece of text or speech, recognizing speech, converting text to speech and vice-versa, generating text or speech, translating between texts in different languages, and many more. Of these tasks, the tasks involving translation between languages can be grouped under the task of Machine Translation. The current state-of-the-art technology for Machine Translation is Neural Machine Translation (NMT), which uses a neural network to maximize translation performance (Bahdanau, Cho, & Bengio, 2016). Neural Networks are a result of the Perceptron proposed by Frank Rosenblatt in 1958, and Rosenblatt (1962) describes the Perceptron as a brain model which attempts to explain the processes of the biological brain. As Koehn and Knowles (2017) explained, NMT worsens in quality on smaller datasets (p. 4, pp. 4). This means that in the world of Neural Networks, inequity in resource availability is synonymous with inequity in performance quality. 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) illustrate this resource inequity quantitatively in Figure 2 on page 5, by categorizing the world's languages into six classes based on shared typological features and plotting the quantity of resources available for each language class across the Linguistic Data Consortium (LDC), the Language Resources and Evaluation (LRE) Map, Wikipedia, and the Web.

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.0B	88.17%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	1.0B	8.93%
2	Zulu, Konkani, Lao, Maltese, Irish	19	300M	0.76%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.1B	1.13%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	1.6B	0.72%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.



(a) LDC (b) LRE (c) Wikipedia (d) Web

Figure 2: Plots of Different Available Resources for Different Languages. Languages to the far right do not have a representation in the resource category. Languages annotated are: Class 0- Dahalo (Dh), Wallisian(Wl); Class 1-Bhojpuri (Bh), Greenlandic (Gr); Class 2-Lao (La), Zulu (Zu); Class 3- Bengali (Bn), Indonesian (In); Class 4- Korean (Ko), Italian (It); Class 5- English (En), Spanish (Es). (Joshi et al., 2020, p.4)

Evidently, languages that make up the largest (88.17%) percentage of total languages have the smallest amount of available training data and languages that make up the smallest percentage (0.28%) of the world's languages have the largest amount of training data across all four resources in Figure 2. This poses a problem as it renders the state-of-the-art NMT models inefficient for 99.72% of total languages spoken by an aggregated 5B speakers. Consequently, it is also difficult to evaluate the performance of any NMT model as a language-agnostic model (Bender, 2011).

Several efforts have been made in the field of Natural Language Processing and Neural Machine Translation to overcome the gap in quality for low-resource language pairs such as Korean-English (Sennrich and Zhang, 2019), Mongolian-Chinese (Wenting et al., 2017), and Sinhala-English (Koshiya et al., 2021). Most recently, researchers at Meta AI developed a state-

of-the-art machine translation tool called No Language Left Behind (NLLB), which claims to offer human-centered machine translation for over 200 languages with an average improvement of 44% (in BLEU score) over the previous state-of-the-art models (Costa-jussà et al., 2022).

The goal of this state-of-the-art paper is to survey and summarize: (i) the recent models being developed for translating specifically between low-resource language pairs, (ii) the improvements made on Korean-English and the Indic languages, and (iii) the improvements in models built for translating between multiple low-resource language pairs such as NLLB. The aim is to understand the trends in the improvements made to NMT and analyze how generalizable they are to all low-resource language pairs.

EXPLORING LINGUISTIC JUSTICE AND DATA EQUITY IN AI

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. 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.

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 Carslon (2009), and the Actor-Network Theory.

LINGUISTIC JUSTICE

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 (pp. 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 STS paper.

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 direcetly 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 Carlson (2009)'s Social Construction model in Figure 3 below.

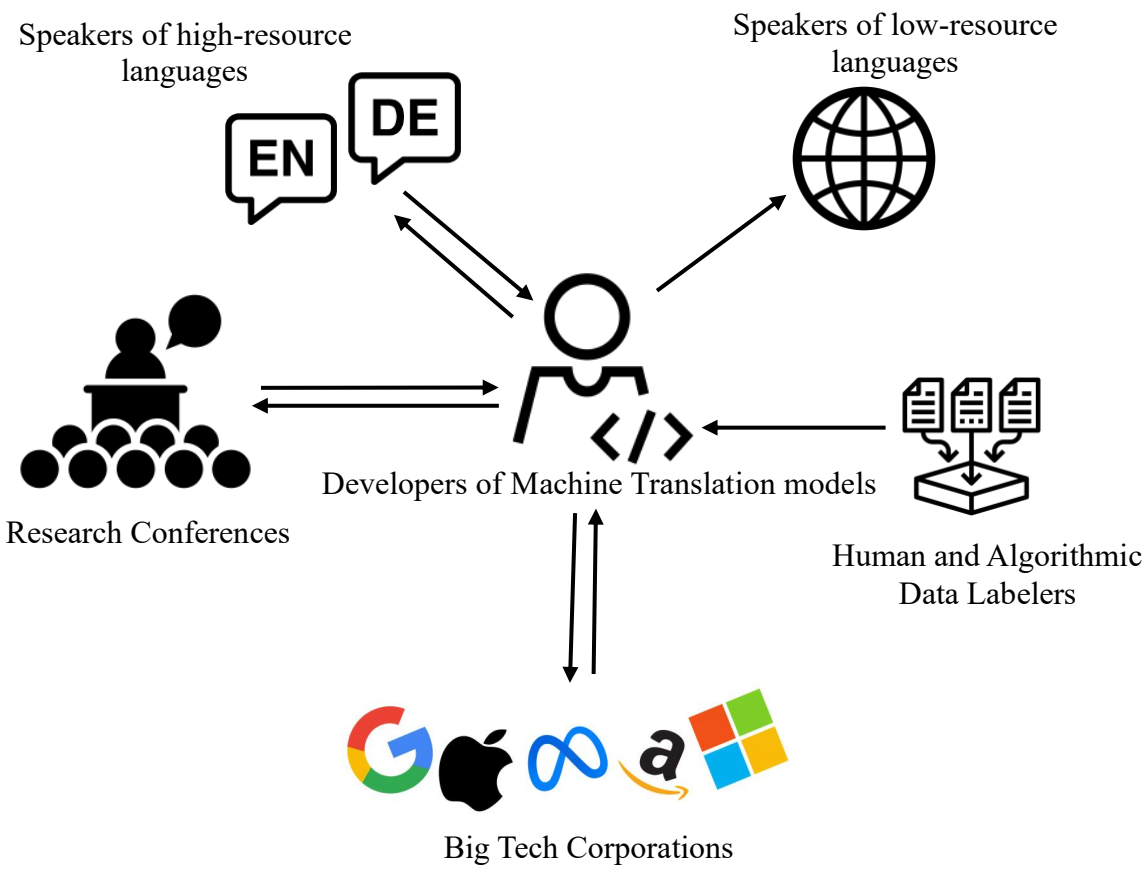


Figure 3: The Social Construction Framework for Low-Resource Machine Translation. This figure shows the interactions of five major stakeholder groups of NMT tools with the developers of NMT models. (Adapted by Choudhary, 2022 from Carlson, 2009)

ACTOR-NETWORK THEORY

While Carlson (2009)'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 4.

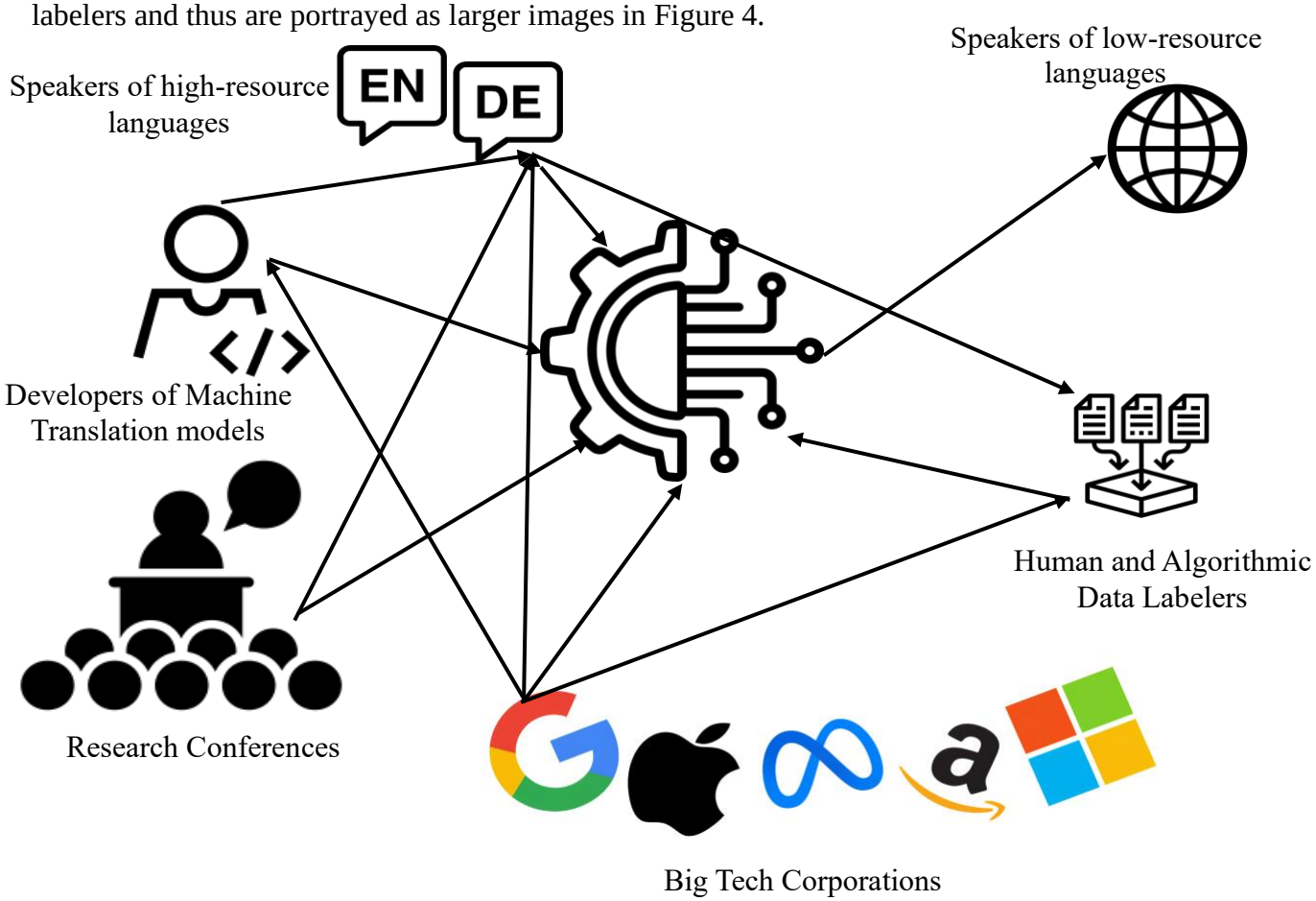


Figure 4: Actor-Network Theory for Machine Translation. This figure shows the links between the machine translation models (center) and the other actors in this network; notably, the fewest links exist between speakers of low-resource languages and the other actors. (Choudhary, 2022)

The ideal outcome of this STS paper is 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.

TRANSLATING BETWEEN TECHNOLOGY AND SOCIETY

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.

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

Carlson, B. (2009). *Social Construction*. [Figure 4]. Class handout (Unpublished). 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

Choudhary, A. (2022). *Gantt Chart for Tehnnical and STS deliverables.* [Figure 1]. *Prospectus* (Unpublished undergraduate thesis). 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.

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

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

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