Should We Ban English NLP for a Year?

Anders Søgaard
Dpt. of Computer Science, Pioneer Centre for Artificial Intelligence, and Dpt. of Philosophy
University of Copenhagen
soegaard@di.ku.dk

Abstract

Around two thirds of NLP research at top venues is devoted exclusively to developing technology for speakers of English, most speech data comes from young urban speakers, and most texts used to train language models come from male writers. These biases feed into consumer technologies to widen existing inequality gaps, not only within, but also across, societies. Many have argued that it is almost impossible to mitigate inequality amplification. I argue that, on the contrary, it is quite simple to do so, and that counter-measures would have little-to-no negative impact, except for, perhaps, in the very short term.

1 Inequalities

If NLP makes people richer and happier, e.g., by allowing them more free time (Jin et al., 2021), it is unfortunate that NLP predominantly serves the needs of the richest and happiest among us. By and large, NLP supports languages spoken in the world's wealthiest regions (Blasi et al., 2022). Performance disparities within languages and across demographics are also well-documented (Amir et al., 2021; Zhang et al., 2021; Chalkidis et al., 2022) – and performance correlates strongly with income levels (Marrero, 2021; Blasi et al., 2022). Disparities may reflect data imbalances: If training data contains less data from a group, predictions for that group will tend to be worse. But disparities can also result from outlier behavior or higher degrees of variance in groups.

Disparities in resources, variation, priority, performance, and turn-around go hand-in-hand to create vicious circles that widen existing equality gaps. Take mobile assistants, for example. Mobile assistants help us organize our calendars, place calls, remind us of meetings, etc. Since they are typically speech-operated, their performance depends heavily on the performance of available speech recognizers. For most languages, speech recognizers were developed for young, urban subpopulations, seen as early adopters. As a consequence, while young urbans reap the benefits of mobile assistants, mobile assistants are often much less useful for other demographics (Feng et al., 2021; Markl, 2022).

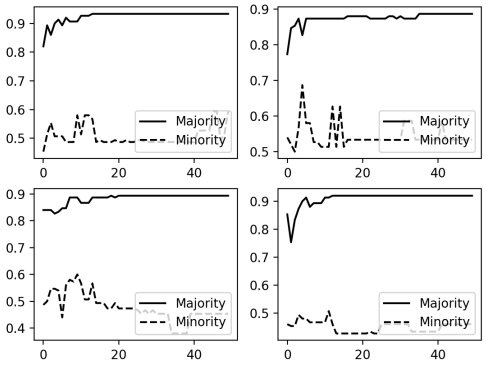

Figure 1: Validation performance over time when end user group growth is proportional to the performance on this group. Each time step corresponds to the inclusion of up to 20 new end users. Simulations on four circles datasets generated at random with https://scikit-learn.org/. The x-axis is time steps, y is classification accuracy.

Many have argued that it is almost impossible to mitigate inequality amplification (Fazelpour and Lipton, 2020; Lin and Chen, forthcoming). I argue that it is quite simple to do *something*, and that doing so would have little-to-no negative impact (except for, perhaps, in the very short term). I address two levels of inequality in NLP: inequality across languages and inequality across social groups. What languages and subgroups are favored is somewhat task-specific, but generally, NLP seems biased toward English and the tech-savvy:

The Dominance of English Existing estimates of how much of top venue NLP research is devoted to English vary a bit, but typically lie in the range

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 5254–5260
December 7-11, 2022

of 50-90%, averaging around two thirds.[1] Importantly, numbers do not seem to have changed for the better over the last 10 years. The vast majority of publicly available NLP datasets are limited to English. For this reason, it is much easier for start-ups and other companies to roll out products for the English-speaking markets. Naturally, this means that speakers of English have way more technologies at hand (Ananiadou et al., 2012).[2]

The Dominance of the Tech-Savvy The penetration rate of mobile assistants with young urbans leads to performance disparities in speech recognition. Similarly, chatbots are developed for an audience that more frequently interacts with dialogue systems. Training data is sampled from end users, and user feedback is leveraged as learning signals. As a result, performance disparities across demographic groups gradually widen until they become a matter of night and day. See Figure 1 for an illustration of this effect across four synthetic datasets.[3]

Outline I have argued how NLP favors English and the tech-savvy. This is not the only way in which NLP models are biased, but arguably the most important ones. §2 discusses how these biases have been justified in the past, arguing how few of these justifications are sound. §3 presents possible mitigation strategies, and §4 discusses their limitations.

2 Justifications

Can the inequalities of NLP research be justified? One high-level justification of inequalities falls out of the way we have come to define fairness in NLP research. NLP researchers have almost uniformly adopted American philosopher John Rawls' definition of fairness (Larson, 2017; Vig et al., 2020; Ethayarajh and Jurafsky, 2020; Li et al., 2021; Chalkidis et al., 2022). Rawls' concern (Rawls, 1971) is with the *absolute* position of the least advantaged group rather than their relative position, even if we have to abandon strict equality of income and wealth. Rawls justifies inequality up to that point and thereby introduces a loophole. My concern with this form of justification is that the loophole turns out to be easily exploited because the effects of both policy and technology are indirect and must be evaluated in the long term (Mukhopadhyay and Mangal, 1997; Wörner and Reiss, 2001). Proponents of a free market argue that lowering high-income taxes may result in long-term benefits for the least advantaged, but what if they eventually do not? Similarly, can we be sure that our cutting edge research on English will have spill-over effects for other languages in the long run? It can–in econonomy, as well as in NLP–be difficult to predict what will get the wheels spinning for everyone, and what will create a downward spiral; see, again, Figure 1 for an illustration of this.

English resources are abundant.	*Opportunity*
Beta users invest time and energy.	*Desert*
It is up to industry and labs to decide.	*Procedure*
The tech-savvy have advanced needs.	*Need*
Other technologies are also for English markets first.	*Reference*

Table 1: Common justifications of NLP's excessive focus on English and the tech-savvy.

Bank (2018) considers five other common justifications of (economic) inequality; see Table 1:

(i) **Opportunity** arguments point to unique opportunities here and now. Some may, for example, justify beta testing technologies on English, because English is a big market with fast turn-around. Or justify developing prototype models tailored for narrow, but adoption-ready,

[1] See Bender (2009), Ruder et al. (2022), as well as https://sjmielke.com/acl-language-diversity.htm

[2] What explains the dominance of English in NLP research? Prestige seems to be an important factor. What is considered state of the art, is what achieves best performance on English. Consider, for example, the ACL Wiki's list of 'state-of-the-art' part-of-speech taggers

https://aclweb.org/aclwiki/POS_Tagging_(State_of_the_art)

This list is based on the performance of these models on the English Penn Treebank (Marcus et al., 1993). It is the same for ACL Wiki lists on NP chunking, question answering, and so on. Such prominent rankings make it more important for researchers to perform well on English benchmarks. Visibility and impact are determined by your doing so, not by your performance on other languages. Blasi et al. (2022) show that papers focusing on languages other than English are cited less often.

[3] For illustration of the dynamics of performance gaps, I created a simulation based on the following thought experiment. Consider two groups g and h, and a product p that relies on a predictive model trained on data from its end users. The predictive model in this case is a simple binary decision tree of depth 4. Since g is deemed more adoption ready than h, the company developing p reaches out to a group of beta users consisting mainly of members of g. The initial set of end users is 55, out of which 5 are from h. This initial inequality is justified by the adoption readiness of g, making fast feedback turn-around more likely. As the company says, performance on data from members of h will quickly improve, as more end users adopt the technology over time. The company, however, is wrong to assume that performance on the minority subgroup will 'catch up'. See Figure 1 for the simulation results. Under the plausible assumption that adoption is proportional to performance, it seems the opposite is true: The performance gap between g and h widens over time.

target audiences of tech-savvy end users in order to collect early feedback as fast as possible. *Rebuttal:* Such opportunism is unprincipled and ethically inferior (Smith, 1935). While desert and special needs can be saturated, opportunity seems forever self-reinforcing. Opportunity often relies on false premises: Take the claim, for example, that our excessive focus on English comes down to the limited availability of data for other languages. GPT-2, for example, was trained on 40GB of text. Such amounts of data are readily available for at least the top-100 biggest languages.[4] Even labeled data is available for many task-language pairs (Galeshchuk et al., 2019; Öhman et al., 2020; Nivre et al., 2020; Scialom et al., 2020; Hasan et al., 2021).

(ii) **Desert** arguments justify inequality by pointing to history or characteristics that may lead us to think individuals deserve special treatment. Someone may have made sacrifices or worked hard for a common cause. Such arguments are based on moral merit. Consider again the tech-savvy beta users that volunteer to test-balloon a product. Beta users invest time and energy in products; should they not be given a certain priority? *Rebuttal:* Rewarding beta users creates a vicious circle, because beta users tend to be young, urban, well-educated and tech-savvy.

(iii) Procedural justifications of inequality say inequality can be excused if they are the result of accepted **Procedure**. Justifying CEO salary by saying it's up to the board to decide, is an example of a procedural justification of inequality. In NLP, such arguments are common: Someone may say, for example, that the focus on English is really just the organic result of local decisions by industry and research labs, and it is really not anyone's business to decide for them. Who are we to decide what research and industry labs focus on? some may ask. Or: Why should we favor some research over other research? *Rebuttal:* We already have Ethics Guidelines[5] to ensure that dual use research is rejected, thus already limiting the freedom of NLP researchers. Generally, there is widespread agreement that NLP and related technologies must be regulated for safety reasons and to avoid discrimination (Black and Murray, 2019).

(iv) Justifications for inequality may also refer to special **Needs**. In NLP, poly-synthetic languages may be in special need of extra annotation layers or hand-written finite-state transducers, whereas some target groups require specific technologies, e.g., text simplification for dyslexics. Some have argued Needs is the only good justification of inequality (Nielsen, 1979). I agree and will *not* provide a rebuttal for this type of justification.

(v) Finally, Bank (2018) lists a fifth type of justification for inequality. This type passes on blame by pointing to historical precedence or similar practices in other domains. You can argue, for example, that it is justified to develop NLP technologies for the English-speaking market first, because other technologies are also beta-released on this market first. This type of justification is called **Reference**. *Rebuttal:* Such reference arguments have the same problem as Opportunity in that they cannot be saturated.

[4]The web technology survey platform `w3techs.com` estimates that more than two million online web pages exist for these languages.

[5]`https://aclrollingreview.org/ethicsreviewertutorial`

I have, anecdotally, come across all of the five frames in discussions in the NLP community. The list is likely incomplete. Some frames are probably used more explicitly than others. Opportunity arguments (Utiyama and Isahara, 2007; Anastasopoulos et al., 2019), Desert arguments (Blasi et al., 2022; Lewis et al., 2020), and Need arguments (Paetzold and Specia, 2016; Yaneva et al., 2019) are abundant in the academic literature, whereas you rarely see explicit Procedure and Reference arguments. Either way, I have argued that only concerns for special needs (Need) seem to justify inequality.

3 Measures

What measures have NLP researchers proposed to mitigate inequalities? Way et al. (2022) argue that 'being able to build neural language models for other languages with the same quality as English is key for language equality', and that the stepping stone is collecting 'large amounts of publicly available corpora of good quality'. I think this is insufficient in my rebuttal of Opportunity in §2: Resources *are* often available. Our excessive focus on English and the tech-savvy is *not* primarily driven by data scarcity. Blasi et al. (2022) argue it is the economic prowess of the users of a language that drives the development of NLP technologies, but they do not present specific proposals for mitigating inequalities. They only refer to a need for global coordination.

One common strategy for mitigating performance disparities *across languages* is to make models language-independent (Bender, 2009). Multilingual models often still exhibit cross-language disparities (Singh et al., 2019), but can be augmented with an objective minimizing the loss of the worst-off language (Ponti et al., 2021; de Lhoneux et al., 2022). Similarly, many learning algorithms have been developed to maximize performance on the groups with the worst performance. Examples include square root sampling (Stickland and Murray, 2019), adaptive scheduling (Jean et al., 2019), loss-balanced task weighting, (Liu et al., 2019), group-distributional robust optimization (Sagawa et al., 2020), and worst-case-aware automated curriculum learning (Zhang et al., 2020). This does not and will not bridge existing gaps: The algorithms are commonly thought to ensure equal performance but in fact, because they implement Rawlsian fairness, they only prescribe inequality up to a point.

Lin and Chen (forthcoming) highlight the chal-

lenges of achieving fairness in the context of structurally unjust societies; see also Fazelpour et al. (2022). Such considerations, as well as the urgency of the matter, has made me wonder what holds us back in adopting more radical measures. Inspired by policies proposed to mitigate climate change, I briefly discuss the pros and cons of three possible pathways forward:

NLP Cap and Trade Under cap and trade (Peace and Stavins, 2010), lawmakers establish a limit (or "cap") on the overall cost or risk, say, the amount of greenhouse gases. Such caps can be negotiated from year to year, and are ideally supported by commonly agreed-upon objectives and scientific evidence. In NLP, this could be a cap on monolingual language models, a cap on technologies for or research publications on English, a cap on male annotators, etc. Just like governments could initially auction off emission allowances to the highest bidder or allocate them evenly or in light of special needs, ACL could distribute quota for English models, biased end user groups, or biased annotator pools. Subsequent to the initial allocation, research labs could reduce their 'emissions' and sell excess allowances to other research labs for quota.

NLP Carbon Tax A carbon tax (Martin et al., 2014) is the obverse of cap and trade: rather than fixing the amount of allowable emissions, it specifies their price. In the same way, NLP researchers could incur a cost–by paying higher conference fees, subtracting from their reviewer scores, or by disqualifying them from paper awards–if they 'emit greenhouse gases' by, say, working on English or with a biased set of end users or annotators.

NLP Car-Free Sundays An alternative to the above is the equivalent of car-free sundays. Car-free Sundays produce significant mean carbon emission reductions and reduce overall traffic activity. While these effects are variable (Glazener et al., 2022), car-free Sundays also help to promote the cause of mitigating climate change. Car-free Sundays are also less intrusive and less bureaucratic than cap and trade or the equivalent of a carbon tax. Examples of regulatory steps in NLP that would be comparable to car-free Sundays, would include a one-year ban on English models, biased end user groups, biased annotator pools, etc. In practice, bans could, for example, be implemented by automatic desk rejection of all such papers submitted to ACL 2023 or to all the main conferences of that year. It is easy to see the positive effects of such an initiative: ACL 2022 accepted 702 papers. 702 papers on other languages than English and/or with annotator pools would be a big step toward course-correcting and mitigating existing biases. [6]

4 Discussion

NLP Cap and Trade, NLP Carbon Tax, and NLP Car-Free Sundays are all possible ways of reducing the widening digital language divide and to reduce performance disparities across groups. If these ideas seem radical, it is worth remembering that the public perception of carbon tax has changed much over time (Jagers et al., 2021). The regulation discussed in the above nevertheless goes well beyond the regulation previously proposed. The European Union, for example, recently presented a legal framework for artificial intelligence.[7] In the framework, NLP and related technologies are classified as high-risk to low-to-no-risk, and low-to-no-risk technologies, e.g., spam filters or syntactic parsers, are left unregulated. A one-year ban on English NLP would also mean a one-year ban on English spam filters. My motivation for extending regulation to low-to-no-risk technologies is about inequality, not safety. A one-year ban on English would also be more radical than earlier attempts by the ACL to promote linguistic diversity and bias mitigation, by thematic research tracks and best paper awards. So why go further now?

My argument for considering temporary regulation is a) that we urgently need to act, and b) that NLP turn-around is fast, and the field has proven incredibly adaptive. In other words, it would not have

[6] Bans are easily gamed, and one strategy in the face of, say, a ban on English NLP, would be to machine translate your favorite English dataset into one or more languages. Often this will not be necessary because native data sets exist already, and there will be a natural pressure to invest in annotations where possible, because the community is well aware of the biases that haunt machine translated datasets (Hershcovich et al., 2022). Even if some adopted this strategy, however, I still think a ban or some form of regulation would move the needle in the right direction. Another important concern is whether students would suffer more from a temporary ban than their professors, as they are operating at a different time scale. This is hard to foresee. Most PhD students will be more adaptive than industry labs and could actually benefit from a ban. A few others would explore other venues, which would be great for inter-disciplinary cross-fertilization. If the community expects there to be a sizeable portion of students left, who would incur a loss, a possible strategy could be to combine a one-year ban with a taxation model, giving each institution limited 'emission allowances'.

[7] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

many negative side-effects to impose such regulation. If tomorrow researchers were told that papers on English NLP would be desk-rejected from ACL 2023 or ACL 2024 (as a form of NLP Car-Free Sundays), many of us would have to course-correct a bit. Many papers would have to report results on different data, and new data would have to be annotated. But that, arguably, would serve NLP well. Course-correcting would, by the end of the day, require limited effort.

5 Conclusion

This paper is a position paper, arguing that NLP is contributing to global inequalities through a digital language divide, and by implicitly favoring the tech-savvy. While many of us have promoted linguistic diversity in recent years, numbers suggest our field is still massively biased toward developing English NLP for tech-savvy demographics. Maybe time is ripe to consider more concrete measures? To get the discussion off ground, I briefly considered three types of measures inspired by regulation advanced to mitigate climate change: cap-and-trade, carbon tax, and car-free Sundays. I argue that the NLP community would quickly adapt to most such initiatives, course-correcting in little or no time. Perhaps initially, the equivalent of car-free Sundays is the least intrusive and bureaucratic form of regulation, but exactly what our next steps should look like, I leave up for community-wide discussion.

6 Limitations

Generating the above opinions, required no GPUs and no manual annotations. One major limitation of this work, though, is that it is all words, no action. Empty vessels make the loudest sound, Plato said. On the other hand, I would add, they change course more easily.

References

Silvio Amir, Jan-Willem van de Meent, and Byron Wallace. 2021. On the impact of random seeds on the fairness of clinical classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3808–3823, Online. Association for Computational Linguistics.

S. Ananiadou, John McNaught, and P. Thompson. 2012. *The English Language in the Digital Age*. META-NET White Paper Series. Springer Nature, United States. EC FP7 PSP Project METANET4U.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. 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)*, pages 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Julian Bank. 2018. Mr. Winterkorn's pay: A typology of justification patterns of income inequality. *Social Justice Research*, 29:228–252.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Julia Black and Andrew Douglas Murray. 2019. Regulating ai and machine learning: Setting the regulatory agenda. *Eur. J. Law Technol.*, 10.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland. Association for Computational Linguistics.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Sina Fazelpour and Zachary C. Lipton. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 57–63, New York, NY, USA. Association for Computing Machinery.

Sina Fazelpour, Zachary C. Lipton, and David Danks. 2022. Algorithmic fairness and the situated dynamics of justice. *Canadian Journal of Philosophy*, 52(1):44–60.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *CoRR*, abs/2103.15122.

Svitlana Galeshchuk, Ju Qiu, and Julien Jourdan. 2019. Sentiment analysis for multilingual corpora. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 120–125, Florence, Italy. Association for Computational Linguistics.

Andrew Glazener, James Wylie, Willem van Waas, and Haneen Khreis. 2022. The impacts of car-free days and events on the environment and human health. *Current Environmental Health Reports*, 9(2).

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Sverker C. Jagers, Erick Lachapelle, Johan Martinsson, and Simon Matti. 2021. Bridging the ideological gap? how fairness perceptions mediate the effect of revenue recycling on public support for carbon taxes in the united states, canada and germany. *Review of Policy Research*, 38(5):529–554.

Sébastien Jean, Orhan Firat, and Melvin Johnson. 2019. Adaptive scheduling for multi-task learning. In *ArXiv 1909.06434*.

Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. How good is NLP? a sober look at NLP tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Miryam de Lhoneux, Sheng Zhang, and Anders Søgaard. 2022. Zero-shot dependency parsing with worst-case aware automated curriculum learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 578–587, Dublin, Ireland. Association for Computational Linguistics.

Mike Li, Hongseok Namkoong, and Shangzhou Xia. 2021. Evaluating model performance under worst-case subpopulations. In *Advances in Neural Information Processing Systems*, volume 34, pages 17325–17334, Vancouver, CA. Curran Associates, Inc.

Ting-An Lin and Po-Hsuan Cameron Chen. forthcoming. Artificial intelligence in a structurally unjust society. *Feminist Philosophy Quarterly*.

Shengchao Liu, Yingyu Liang, and Anthony Gitter. 2019. Loss-balanced task weighting to reduce negative transfer in multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9977–9978.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Nina Markl. 2022. Language variation and algorithmic bias: Understanding algorithmic bias in british english automatic speech recognition. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 521–534, New York, NY, USA. Association for Computing Machinery.

Gustavo Marrero. 2021. *Does Race and Gender Inequality Impact Income Growth?* The World Bank, Washington, D.C.

Ralf Martin, Laure B. de Preux, and Ulrich J. Wagner. 2014. The impact of a carbon tax on manufacturing: Evidence from microdata. *Journal of Public Economics*, 117:1–14.

Tridas Mukhopadhyay and Vandana Mangal. 1997. Direct and indirect impacts of information technology applications on productivity: A field study. *International Journal of Electronic Commerce*, 1(3):83–102.

Kai Nielsen. 1979. Radical egalitarian justice: Justice as equality. *Social Theory and Practice*, 5(2):209–226.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference*

on Computational Linguistics, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. Understanding the lexical simplification needs of non-native speakers of English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 717–727, Osaka, Japan. The COLING 2016 Organizing Committee.

Janet Peace and Robert N. Stavins. 2010. *Meaningful and Cost Effective Climate Policy: The Case for Cap and Trade*. Pew Center on Global Climate Change, Arlington, VA. F-28.

Edoardo Maria Ponti, Rahul Aralikatte, Disha Shrivastava, Siva Reddy, and Anders Søgaard. 2021. Minimax and neyman–Pearson meta-learning for outlier languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1245–1260, Online. Association for Computational Linguistics.

John Rawls. 1971. *A Theory of Justice*, 1 edition. Belknap Press of Harvard University Press, Cambridge, Massachussets.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

T. V. Smith. 1935. Opportunism. *The International Journal of Ethics*, 45(2):235–239.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *ICML*.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, Vancouver, CA. Curran Associates, Inc.

Andy Way, Georg Rehm, Jane Dunne, Maria Giagkou, José Manuel Gomez-Perez, Jan Hajič, Stefanie Hegele, Martin Kaltenböck, Teresa Lynn, Katrin Marheinecke, Natalia Resende, Inguna Skadiņa amd Marcin Skowron, Tea Vojtěchová, and Annika Grützner-Zahn. 2022. Report on the state of language technology in 2030. Available at `https://european-language-equality.eu/wp-content/uploads/2022/05/ELE___Deliverable_D2_18__Report_on_State_of_LT_in_2030_.pdf`.

Stefan Wörner and Thomas Reiss. 2001. The direct and indirect impacts of new technologies on employment: The example of the German biotechnology sector. *Science and Public Policy*, 28(5):371–380.

Victoria Yaneva, Constantin Orasan, Le An Ha, and Natalia Ponomareva. 2019. A survey of the perceived text adaptation needs of adults with autism. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1356–1363, Varna, Bulgaria. INCOMA Ltd.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2021. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2020. Worst-case-aware curriculum learning for zero and few shot transfer.