A global AI community requires language-diverse publishing

Haley Lepp
Parth Sarin
Stanford University
Palo Alto, California, USA

Who produces a global AI? Who researches, designs, builds, trains, evaluates, and sells the models that are being deployed across the planet? An internationally diverse network of actors contributes to the technologies known as AI, including the "ghost" laborers who extract the material resources for energy and hardware[7], the barely-wage earners who label and label and label[5]; and the non-wage earning “users” whose labor has been classified by AI companies not as creation, but consumption[10]. In the spotlight are the AI researchers. Typically highly-paid corporate employees or scholars with high-earning potential, AI researchers have an out-sized voice in determining the trajectory of this technology, a voice which they project through publications and conference proceedings. Over and over again, this community has been interrogated for its lack of diversity in race, gender, country, and age. As a result, affinity groups have proliferated; conferences have adopted anti-bias policies, and volunteers have worked tirelessly to expand training opportunities for groups who have historically faced discrimination in computing communities. Yet a major component of diversity remains unaddressed, one which we argue helps uphold extractive and unbalanced nature of this powerful industry. To be an AI researcher in this global community, one must write and speak in English.

The top 100 ranked computer science proceedings and journals are published in English[1]. Despite their locations all over the world, in many countries in which English is not a primary language spoken, conference proceedings are officially held English. Even scholars who invest considerable time and expense into learning to produce academic English may face rejection in peer review. In fact, mining the historical reviews from the past six years of ICLR proceedings demonstrates that thousands of reviews over the years critique the language of authors either explicitly ("The paper is full of English mistakes.") or implicitly ("There are numerous grammatical errors and poorly-phrased sentences.")[12]. The ramifications of this monolingual research industry are wide: from heavily uneven linguistic output and poor attention to issues in so-called “low-resource” languages [4, 11], to limitations in global education and hiring[8], to a de facto tax for scholars who must pay for pre-submission copy-editing[2, 3].

What does a researcher do when they believe their work may be rejected based on reviewer perceptions of academic English? First, some will choose not submit work and simply not participate in the research community. Second, some will submit their work, but pay someone to translate or “professionalize” their writing. This process is expensive and requires extra time, which in such a fast-moving field puts such scholars at a competitive disadvantage.

As such, the increasing availability of automatic writing-assistance and translation software has been celebrated as a boon for inclusion, allowing scholars who might otherwise be excluded from monolingual publishing to instantly translate their writing prior to submission for peer review. Many scholars use ChatGPT to “fix" their writing for publication. However, we regard this phenomenon not as a solution for inclusion, but a symptom of linguistic exclusion. Authors should have the right to describe their findings in the language of their choice, without the mediation of translation. In preliminary interviews for a future study, multilingual ICLR scholars indicated that they feel that they have different personalities when they express themselves in one language over another. Translation is also never one-to-one, so by only publishing in one language, the community loses out on the vast diversity of other ways of knowing which might be represented. Furthermore, producing research in a single language also alienates readers in other languages. Widespread translation to English will drive our field away from, instead of toward, an inclusive global community.

Without intervention, two possibilities loom on the horizon. The AI publishing community will continue to exclude the majority of the world’s language groups and therefore people, and only the limited people in the world who receive long-term English training will participate. Alternatively, current publishing infrastructure will demand the increased assimilation to English-only education and publishing. Computing training will be paired with English-language training, pushing English around the globe as a requirement for high-paid employment and intellectual contribution to technology research. Neither of these are tolerable futures for a truly global research community. So, where can we go from here?

First, conferences should be administered in the language(s) of the country in which they are held. Organizers should hire translation services, and encourage scholars who speak English and other languages to present in their other languages, un-hiding the linguistic diversity of the existing computing community.

Second, we should include explicit instructions to peer reviewers to not adjudicate the language "appropriateness" of the papers they review[6]; instead, editors should explore other options for publishing multilingual scholarship, such as offering scholars the opportunity to share their work in multiple languages. English-speakers should share any cost burden of translation, instead of placing it entirely on those who do not speak English. Publications and conferences must set aside funds for translation services-not just into English, but from English. Publication infrastructure, such as OpenReview, should also include space for submissions of translations.

Finally, it is imperative that the burden of change not be upon people who are linguistically marginalized in publishing: everyone in this global community should be working toward learning to tolerate and embrace linguistic diversity. Graduate students should be encouraged, if not required, to take language courses[9]. If we truly aspire to have a global AI community, we must challenge the hegemony of English in computing.

REFERENCES

[1] 2022. Best Computer Science Conferences. https://research.com/conference-rankings/computer-science

[2] 2023. Scientific publishing has a language problem. *Nature Human Behaviour* 7, 7 (July 2023), 1019–1020. https://doi.org/10.1038/s41562-023-01679-6 Number: 7 Publisher: Nature Publishing Group.

[3] Tatsuya Amano, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White, et al. 2023. The manifold costs of being a non-native English speaker in science. *PLoS Biology* 21, 7 (2023), e3002184.

[4] Fran M Collyer. 2018. Global patterns in the publishing of academic knowledge: Global North, global South. *Current Sociology* 66, 1 (2018), 56–73.

[5] Kate Crawford. 2021. *Atlas of AI: power, politics, and the planetary costs of artificial intelligence.* Yale University Press, New Haven. OCLC: on1111967630.

[6] Nelson Flores and Jonathan Rosa. 2015. Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard educational review* 85, 2 (2015), 149–171.

[7] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass.* Eamon Dolan Books.

[8] Philip J. Guo. 2018. Non-Native English Speakers Learning Computer Programming: Barriers, Desires, and Design Opportunities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (, Montreal QC, Canada,) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173970

[9] Amy McDermott. 2023. English is the go-to language of science, but students often do better when taught in more tongues. *Proceedings of the National Academy of Sciences* 120, 40 (2023), e2315792120. https://doi.org/10.1073/pnas.2315792120 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2315792120

[10] R. Oldenziel. 2001. *Man the maker, woman the consumer : the consumption junction revisited.* University of Chicago Press, United States, 128–148.

[11] Surangika Ranathunga and Nisansa de Silva. 2022. Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 823–848. https://aclanthology.org/2022.aacl-main.62

[12] Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. 2022. Investigating Fairness Disparities in Peer Review: A Language Model Enhanced Approach. https://arxiv.org/abs/2211.06398