International Journal of Comparative Studies in International Relations and Development
p-ISSN: 2354-4198 | e-ISSN: 2354-4201
Volume 9, Number 1 November, 2023

IJCSIRD

How Language Gaps Constrain Generative AI Development

[1]Regina Ta & [2]Nicol Turner Lee
[1]The Brookings Institution
Washinton DC. United States
[2]Cerntre for Technology Innovation

Article DOI: 10.48028/iiprds/ijcsird.v9.i1.03

Abstract

Prompt-based generative artificial intelligence (AI) tools are quickly being deployed for a range of use cases, from writing emails and compiling legal cases to personalizing research essays in a wide range of educational, professional, and vocational disciplines. But language is not monolithic, and opportunities may be missed in developing generative AI tools for non-standard languages and dialects. Current applications often are not optimized for certain populations or communities and, in some instances, may exacerbate social and economic divisions. As noted by the Austrian linguist and philosopher Ludwig Wittgenstein, "The limits of my language mean the limits of my world." This is especially true today, when the language we speak can change how we engage with technology, and the limits of our online vernacular can constrain the full and fair use of existing and emerging technologies.

Keywords: *Language, Gaps, AI, Development*

Corresponding Author: Regina Ta

First Published: https://www.brookings.edu/articles/how-language-gaps-constrain-generative-ai-development/

Background to the Study

As it stands now, the majority of the world's speakers are being left behind if they are not part of one of the world's dominant languages, such as English, French, German, Spanish, Chinese, or Russian. There are over 7,000 languages spoken worldwide, yet a plurality of content on the internet is written in English, with the largest remaining online shares claimed by Asian and European languages like Mandarin or Spanish. Moreover, in the English language alone, there are over 150 dialects beyond "standard" U.S. English. Consequently, large language models (LLMs) that train AI tools, like generative AI, rely on binary internet data that serve to increase the gap between standard and non-standard speakers, widening the digital language divide.

Among sociologists, anthropologists, and linguists, language is a source of power and one that significantly influences the development and dissemination of new tools that are dependent upon learned, linguistic capabilities. Depending on where one sits within socio-ethnic contexts, native language can internally strengthen communities while also amplifying and replicating inequalities when coopted by incumbent power structures to restrict immigrant and historically marginalized communities. For example, during the transatlantic slave trade, literacy was a weapon used by white supremacists to reinforce the dependence of Blacks on slave masters, which resulted in many anti-literacy laws being passed in the 1800s in most Confederate states

Because of this historical artifact and other movements that have banned bilingual communications in preference for English-only rules and laws, it is important to consider the implications of constructing the same linguistic frameworks in the digital world, which exacerbate the digital divide in autonomous and generative systems.

Language Differences Starts with the Digital Divide

The resource disparities that exist across languages tend to perpetuate further disparities in technologies, such as generative AI systems and LLMs, due to their link to the digital divide. Most language-based systems are trained on internet data that researchers can scrape at scale. But only a few hundred languages are represented online with English taking up the largest proportion. As such, English has become one of the most data-rich languages, and the mass availability of English data has led to the creation of English-centric datasets and models.

Even before generative AI, most natural language processing (NLP) systems were designed and tested in "high resource" languages, like English. Of all the active languages worldwide, only 20 are considered to be "high-resource" languages, a categorization that refers to the amount of data available in a certain language to effectively train language-based systems. One reason for this extreme asymmetry is that speakers of under-resourced languages have limited access to digital services, which means they have a significantly smaller digital footprint and therefore are less likely to be included in web-scraped training data. Without enough data to train usable language-based systems, most of the world's AI applications will under-represent billions of people around the world. Not only are speakers of under-resourced languages at risk, but so are speakers of regional dialects of "high resource"

languages. A plurality online content including books, blogs, news articles, advertisements, and social media posts is written in "standard" U.S. English, which then becomes web-scraped training data for NLP systems and generative AI tools. In fact, ChatGPT was trained on 300 million words imagine how many of those words might have belonged to a non-standard English dialect.

Speakers of non-standard dialects, including AAVE (African American Vernacular English) or Chicano English (spoken primarily by Mexican American communities in the Southwest), are more likely to not be connected to the internet due to the lack of high-speed broadband, an internet-enabled device, or both, which makes them less likely to be productive online contributors. That is why the digital divide can be highly correlated with sparse and unequal representation in LLM training datasets, which results in generative AI and related resources being insufficiently built and representative to effectively serve more diverse communities.

The Digital Language Divide

We call the effects of these trends the "digital language divide," which will be explained further in the next section. English provides just one case study of how non-standard speakers of a high-resource language can be excluded. Mandarin, German, and other high-resource languages also have "standard" and non-standard varieties that may be under-represented online and in research, such as Kiezdeutsch (a German dialect used by first-generation immigrant youth in urban areas). While resource disparities among speakers derive from digital access and infrastructure, having technical leaders and developers who reflect linguistic diversity will also play a key role in building inclusive generative AI tools and beyond.

Why Does Digital Language Divide Matter

The language we speak determines how we engage with the world, as well as which worlds we can participate in. History has shown how language can be used as a tool of exclusion and oppression. From U.S. states that prohibited enslaved Black populations from learning how to read and write, to internment camps denying books and classroom resources to Japanese American children, the same pattern persists in the present, as far-right movements call for ending bilingual education for native Spanish speakers.

Repeatedly, language and who has access to it has been weaponized to disenfranchise vulnerable populations. The only difference today is that the stakes now involve language-based technologies, like generative AI, that can do the work of gatekeeping. As such, equitable distribution of generative AI's benefits and opportunities depends on equitable access to language data. Generative AI has the potential to close many existing equity gaps, from people with communication disabilities and low levels of literacy to learning for K-12 students across school districts. But when it fails to accurately capture the language registers of diverse speakers, it can also erase and contribute to the erasure of the historical contributions of people of color. For instance, when ChatGPT was asked to speak in the narrative voice of *The Hate U Give*, a young adult novel featuring an African American protagonist, its response was to simply insert "yo" at random intervals. Given the increased usage of generative AI, if its benefits are not accessible to or inclusive of all users, then we are only closing some equity gaps at the expense of widening others.

When "standard" varieties of a language are prioritized in training generative AI tools, those language users often get better performance from these tools, which further discriminates against other linguistic varieties and their speakers. For instance, AI detectors used to flag cheating, plagiarism, or misinformation have been found to be unreliable at distinguishing AI-generated text from human-written text, especially when the writer is a non-native English speaker. In one Stanford study, AI detectors erroneously flagged the majority of TOEFL (Test of English as a Foreign Language) essays as being AI-generated. Yet when tested with essays by students who were native English speakers, those same detectors performed with 100% accuracy.

This disparity reinforces social processes like prestige transfer, which establish "standard" U.S. English as the dominant mode of discourse, and any stylistic deviation in pronunciation or grammar is perceived as inferior or incorrect. Stark differences in performance with one language variety, as opposed to another, produce biased attitudes against non-standard speakers and burden non-standard speakers with the pressure of adapting to "standard" forms in order to reap the same benefits from generative AI. This is just part of the digital language divide.

Linguistic biases against non-standard speakers do not serve generative AI companies or developers well. If generative AI tools are striving for inclusiveness, representation, and scalability, then relying on language data that's not representative results in suboptimal performance that fails to fully capture the complexity of real-world contexts. Adhering to a "standard" language variety does not reflect reality, where many speakers code-switch or use different forms for different contexts. In fact, marginalized communities are often forced into code-switching to accommodate mainstream discourse. Accounting for linguistic varieties will create robust generative AI tools that are equipped to handle real-world conversations and encounters, as well as fulfill more nuanced use cases.

Further, as developers work to address many of these blind spots in the type of language data that is collected and aggregated, more open-source datasets, micro-data, and improved participatory involvement from non-standard English speakers could address inconsistencies in product accuracy. On the latter recommendation, more open-sourced language data may be far more inclusive than proprietary datasets that, again, may be constrained in diverse representation of language and contextual applications.

Conclusion/Recommendation

Too often, researchers reach for risk mitigation, which focuses on scaling back problematic models, instead of bias mitigation, which shifts the focus onto addressing issues head-on. To directly mitigate bias in generative AI tools, researchers can make region-specific or language-specific choices in model building and the creation of training datasets. This means involving a diverse set of "humans-in-the-loop" early on and inviting participation from local communities to bring their voices, dialects, and timing to LLMs. While there are numerous ways to engage underrepresented groups in existing and future training data, such collection must be done with transparency and some guardrails to ensure that cultural expertise is not an

exploitable asset. In addition to being a cultural attribute, language is also personal to individual speakers and households, and this should not be discounted. The transference of conversational and robust language tools are attributes of unique cultural efficacies that may not be coded in more homogeneous LLMs or AI more generally.

Many organizations and researchers have been paving the way toward emphasizing locality in training. To promote technical development in African languages, Masakhane is collecting linguistic data from African speakers with a variety of local dialects, operating at the grassroots level to involve the community they are trying to serve to capture culturally relevant data. Building more representative corpora collections of language and textual data will be essential. At the university level, a machine learning specialist from Stanford is addressing resource disparity by sharing open-source AAVE corpora featuring over 141,000 AAVE words. In addition, Universal Dependencies, a global research community for computational linguistics, has been sharing data for languages and dialects beyond "standard" U.S. English, including a corpusa of Hindi English representing code-switching from multilingual speakers.

Bridging the digital divide is essential, as increased usage of generative AI is only exacerbating the digital language divide, which, at its core, is a symptom of online disparities. Internet access varies by gender, geography, and socioeconomic status, all of which intersect with a user's regional dialect and linguistic variety. Communities with limited access to the internet will be underrepresented online, which then skews the textual data available for training generative AI tools. Ultimately, addressing what values and norms drive predominant language, acceptance, along with disparities in online access can help us build more inclusive online ecosystems that represent the full extent of our linguistic diversity.

Reference

https://www.brookings.edu/articles/how-language-gaps-constrain-generative-ai-development/