CENTER FOR DEMOCRACY & TECHNOLOGY

July 27, 2023

To: National Science Foundation
2415 Eisenhower Avenue
Alexandria, Virginia 22314

Re: Request for Information (RFI) on Developing a Roadmap for the Directorate for Technology, Innovation, and Partnerships at the National Science Foundation | 88 FR 26345

Authors:
Gabriel Nicholas, Research Fellow, Center for Democracy & Technology
Aliya Bhatia, Policy Analyst, Free Expression Project, Center for Democracy & Technology

Introduction

The Center for Democracy & Technology (CDT) submits these comments in response to the National Science Foundation's (NSF) request for information regarding its newly-established Technology, Innovation, and Partnerships (TIP) Directorate. CDT is a nonprofit 501(c)(3) organization fighting to advance civil rights and civil liberties in the digital age. CDT's focus includes the impact of automated technologies on free expression, privacy and security, and other fundamental rights.

The rapid pace of development and adoption of large language models has led to exciting opportunities for research and scientific inquiry, economic opportunities, and social development across all sectors. The use of these technologies goes far beyond just generative applications. Recent research and commercial applications have demonstrated the utility of these language models to power systems used to provide healthcare and educational resources and moderate content on online services such as search engines and social media services (Kasneci et al., 2023; Nicholas & Bhatia, 2023b; Wiggers, 2023). As language models become embedded into more aspects of our social and technical systems, their limitations and biases will have larger ramifications on society at large.

One such limitation is how well language models work in languages other than English. A recent report from CDT entitled "Lost in Translation: Large Language Models in Non-English Languages" describes in detail the limitations of large language models' performance in languages other than English (Nicholas & Bhatia, 2023a). To help address this problem, we recommend that TIP help invest in use-inspired research to build training and test datasets in non-English languages, particularly those with limited data available, to make the development of language models more equitable across languages. Below, we explain why language models work better in English and a handful of other "high resource"

languages than in other languages, what effect that gap has, why others will not address the gap, and how TIP can help.

Language models do not work equally well in all languages

Today, state-of-the-art language models are able to convincingly analyze and generate text across dozens, if not hundreds of different languages. Nevertheless, these models work far better in some languages — particularly English — than others. The reasons are twofold. First, not all languages have the same amount of high-quality digitized text available to train language models. Second, research and development investment in the larger ecosystem further perpetuates the focus on the performance of language models in the English language and thus the availability and preponderance of English-language training data.

The asymmetry in available training datasets across languages is called the "resourcedness gap." A language's "resourcedness" refers to the quantity, quality, and diversity of data it has available. Languages vary widely in their resourcedness. English for instance is by far the most high-resource language, with magnitudes more high-quality data from a range of sources than in any other language (Joshi et al., 2020). For many language models, upward of 90% of their training data is in English (Touvron et al., 2023). Other high-resource languages include Spanish, German, and Mandarin.

Lower resource languages, such as Vietnamese, Bengali, Haitian Creole, and Farsi, have far fewer and often lower quality datasets available, despite having tens or hundreds of millions of speakers. In even lower resource languages, such as Tigrinya, Navajo, and Uyghur, longer-form examples of text that may exist on the web are either few and far between, such as a handful of Wikipedia articles or parliamentary proceedings, or are of low quality, such as posts replete with typos, profanity, and offensive language. Language models, having seen few examples of text in these languages, struggle in lower resource languages. As language models become more ubiquitous, becoming integrated into chatbots, search engines, and content moderation, non-English language speakers may find it more difficult to access accurate information (Murgia, 2023).

Some AI companies claim they can, or already have, overcome the resourcedness gap with only the data currently available. Companies tout their language models' ability to learn language-agnostic patterns across text by creating associations between words in lower resource languages through the intermediary of English. For instance, OpenAI claims that GPT-4 has state-of-the-art results in 26 languages, despite being tested and trained on text predominantly in "English and with a US-centric point of view" (OpenAI, 2023, p. 61). Google has also claimed that Bard has learned Bengali despite not seeing many examples of this text (Ngila, 2023). The text these models produce may sound like these languages but are often replete with errors because the examples of text these models have seen are of poor quality, are acontextual, or are machine-translated from English and contain errors or "translationese", words that language speakers don't actually use or understand. As a result, language

models more often produce hallucinations, errors, bias, and malicious outputs in languages other than English (Lin et al., 2021; Muller et al., 2021).

Language models' difficulty parsing non-English languages leads to disparate impacts for speakers and undermines safety measures

The resourcedness gap contributes directly to two specific harms: inequitable outcomes and disparate impacts on non-English speakers, and circumvention of English-language safety interventions. Despite struggling in languages for which they have little data, private entities are already using language models in many non-English contexts, and private and public entities will soon use them in many more. For instance, social media companies including Meta and Google already deploy language models to automatically identify and take action on content in dozens if not hundreds of languages (Lees et al., 2022; Meta AI, 2021; Nicholas & Bhatia, 2023a).

According to the Federal Government's inventory of AI use cases, the federal government is also interested in using language models to power service delivery and provide access to information. Some examples from this inventory include the Department of Education's "Aidan Chat-bot" which provides answers to common financial aid questions and the Department of Labor's impending implementation of cloud-based language translation technologies to translate publicly-accessible policy guidance and other documents (US Department of Education Office of the Chief Information Officer, 2023; US Department of Labor Office of Data Governance, n.d.). These models are likely trained on predominantly English-language data and their inability to parse non-English languages is likely to create or reinforce the same barriers to information they seek to address. This could look like social media services failing to detect and take action on Spanish-language disinformation about vaccines or elections or a chatbot providing inaccurate information about financial services when prompted in a language other than English.

The second effect of the resourcedness gap is that language models' safety measures are likely to be easily circumvented in languages other than English due to the lack of available training resources in those languages. Language models may incorporate safety measures intended to prevent them from producing outputs containing dangerous, illegal, or personally identifiable material. To do so, these models are trained and tested using examples of this type of speech. Developers give models a description of a rule and then examples of text that violates the rule; they use the model to moderate itself, that is to determine whether the output adheres to or violates the rule. Red teaming, or efforts to conduct adversarial testing by domain experts, is one way companies test the strength of their tools and safety measures to reduce harmful outputs. Red teamers for Open AI's tools, for instance, found that when a model was prompted to create recruitment propaganda for terrorist groups in English, the model refused; however, when the model was given the same prompt in Farsi, the model fulfilled the request (Murgia, 2023). This kind of gap in the functionality of safety measures can be harmful to all users.

Others will not address the language data gap

Academia, industry, and society at large would benefit from shrinking the gap in available training and testing data in English versus other languages. However, neither computer science researchers nor private companies are currently incentivized to address this gap on their own, making support from TIP invaluable.

Private technology companies that build language AI are not incentivized to invest in lower-resource language development because it incurs high costs and offers only marginal commercial upside. To gather more training data in lower resource languages, companies cannot as easily employ web-scraping, the cheapest, most scalable, and industry-standard approach to gathering language data. Web data in low-resource languages is more often machine translated, misidentified by language detection software, and lower quality along many other dimensions (Kreutzer et al., 2022). Building new high-quality datasets therefore requires finding, creating, and scanning texts to build new language corpora, a far more expensive process that companies hesitate to invest in. Building datasets needed to test how well language models work in non-English, particularly lower resource languages, is even more difficult and expensive since it requires hiring many native language speakers as human data labelers.

Though it would improve and expand the audience for their products, technology companies are not financially incentivized to make this investment themselves. As popular wisdom goes, other languages offer smaller market opportunities relative to the cost, and many lucrative opportunities for language models, such as in scientific research and global commerce, already use English as a *lingua franca*. Although it may be financially worthwhile for companies to invest in training and testing their models in some languages spoken in larger, wealthier countries, it may not be commercially worthwhile for them to invest in lower-resource languages spoken predominantly in economically weaker countries, even if they have hundreds of millions of speakers.

Academics are also not incentivized to address the language gap. Although many datasets and benchmarks in the field of natural language processing do come from academia, academics focus far more on English than any other language. Between May 2022 and January 2023, there were likely 100 times more NLP publications about English than the next highest language (German).[1] Many of the lowest resource languages are overlooked by NLP researchers altogether.

This at least is in part because many languages do not have their own academic publications or conferences, and certainly not ones with the same reputational status as those focused on English. English and a handful of other high-resource languages experience a virtuous cycle of investment: researchers collect data, create benchmarks, and build models in these languages to publish their results

[1] Papers that do not explicitly mention any language in its abstract are almost always about English (Bender, 2019). Of the 5290 papers the Association for Computational Linguistics published between May 2022 and January 2023, 4720 mentioned no language in its abstract and 311 mentioned English. The next highest language mentioned was German, with 27 (ACL Rolling Review Dashboard, 2023)

in conferences and journals, burnishing the reputations of both themselves and their outlets and making it easier for other NLP researchers to do work in the future. Lower resource languages however experience this as a vicious cycle: research is not only difficult for all the reasons mentioned above, but it is difficult to get attention — and funding — for their work from the larger, more English-centric NLP community (Nicholas & Bhatia, 2023a).

How TIP can address the language gap

The gap in available testing and training data between different languages is a collective action problem. While companies and academics alike would benefit from having more datasets and benchmarks in non-English languages, and the US would benefit by safeguarding its globally dominant position in the AI landscape, no private entity or individual is incentivized to invest in these efforts themselves. With a little support from TIP though, many of these low-resource languages could shift from vicious to virtuous cycles of research.

The main way TIP can help bolster non-English AI is by sponsoring non-English research agendas both in the US and abroad. The most direct way for TIP to do this is to sponsor the creation of new publications, conferences, academic and industry research collaborations, and competitions in specific low-resource languages. One model for how this can work is exemplified by EVALITA, an event hosted by the Italian Association for Computational Linguistics. In it, researchers submit datasets for new language tasks and benchmarks, such as dating documents or identifying misogyny. Then, researchers compete to train models to maximize those benchmarks and publish the best results in conference proceedings, thereby driving interest and attention toward Italian NLP and creating resources companies and external stakeholders can use to evaluate Italian-language hate speech detection systems (Basile et al., 2020).

TIP can also improve non-English and low-resource language research efforts by supporting local collectives of language-specific research networks. Collectives such as Masakhane (African languages) (Orife et al., 2020), IndoNLP (Indonesian languages) (Aji et al., 2022), AmericasNLP (Indigenous languages) (Mager et al., 2021), and ARBML (Arabic dialects) (Alyafeai & Al-Shaibani, 2020) often have deep knowledge of where the largest gaps are in their language's specific research but are sorely lacking the funds necessary to address them.

Finally, TIP can support parallel use-inspired social science research to help better understand the larger effects of AI on different non-English language communities and how to mitigate the harms it may cause. Improving how language models work in lower resource languages can have positive effects, such as economic inclusion and protection from linguistic erasure, but also negative effects, such as exposing speakers to disinformation and labor displacement. TIP should fund research in areas such as auditing that can help maximize the benefits while minimizing the harms of LLMs to language communities.

References

ACL Rolling Review Dashboard. (2023). Papers mentioning >0 languages. *Papers Mentioning >0 Languages*. http://stats.aclrollingreview.org/submissions/linguistic-diversity/

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500

Alyafeai, Z., & Al-Shaibani, M. (2020). ARBML: Democratizing Arabic Natural Language Processing Tools. *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, 8–13. https://doi.org/10.18653/v1/2020.nlposs-1.2

Basile, V., Maro, M. D., Croce, D., & Passaro, L. (2020, December 17). *EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Online.

Bender, E. (2019, September 15). *The #BenderRule: On Naming the Languages We Study and Why It Matters*. The Gradient. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters

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 for Computational Linguistics*, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P. O., … Adeyemi, M. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, *10*, 50–72. https://doi.org/10.1162/tacl_a_00447

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022). A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3197–3207. https://doi.org/10.1145/3534678.3539147

Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., … Li, X. (2021). Few-shot Learning with Multilingual Language Models. *ArXiv:2112.10668 [Cs]*. http://arxiv.org/abs/2112.10668

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L.,

Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., & Kann, K. (2021). Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, 202–217. https://doi.org/10.18653/v1/2021.americasnlp-1.23

Meta AI. (2021, December 8). Harmful content can evolve quickly. Our new AI system adapts to tackle it. *Meta AI*. https://ai.meta.com/blog/harmful-content-can-evolve-quickly-our-new-ai-system-adapts-to-tackle-it/

Muller, B., Anastasopoulos, A., Sagot, B., & Seddah, D. (2021). When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 448–462. https://doi.org/10.18653/v1/2021.naacl-main.38

Murgia, M. (2023, April 14). OpenAI's red team: The experts hired to 'break' ChatGPT. *Financial Times*. https://www.ft.com/content/0876687a-f8b7-4b39-b513-5fee942831e8

Ngila, F. (2023, April 17). Google says its AI developed unexpected skills. *Quartz*. https://qz.com/google-ai-skills-sundar-pichai-bard-hallucinations-1850342984

Nicholas, G., & Bhatia, A. (2023a, May 23). Lost in Translation: Large Language Models in Non-English Content Analysis. *Center for Democracy & Technology*. https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/

Nicholas, G., & Bhatia, A. (2023b, May 23). The Dire Defect of 'Multilingual' AI Content Moderation. *WIRED*. https://www.wired.com/story/content-moderation-language-artificial-intelligence/

OpenAI. (2023, March 27). *GPT-4 Technical Report*. https://cdn.openai.com/papers/gpt-4.pdf

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., Kabongo, S., Meressa, M., Murhabazi, E., Ahia, O., van Biljon, E., Ramkilowan, A., Akinfaderin, A., Öktem, A., Akin, W., Kioko, G., … Bashir, A. (2020). *Masakhane—Machine Translation For Africa* (arXiv:2003.11529). arXiv. https://doi.org/10.48550/arXiv.2003.11529

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023, July 18). Llama 2: Open Foundation and Fine-Tuned Chat Models. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/

US Department of Education Office of the Chief Information Officer. (2023, February 3). *Inventory of Department of Education AI Use-Cases*. US Department of Education (ED). https://www2.ed.gov/about/offices/list/ocio/technology/ai-inventory/index.html

US Department of Labor Office of Data Governance. (n.d.). *Artificial Intelligence Use Case Inventory*. Retrieved July 27, 2023, from http://www.dol.gov/agencies/odg/ai-inventory

Wiggers, K. (2023, May 16). Hippocratic is building a large language model for healthcare. *TechCrunch*. https://techcrunch.com/2023/05/16/hippocratic-is-building-a-large-language-model-for-healthcare/