digital impact alliance

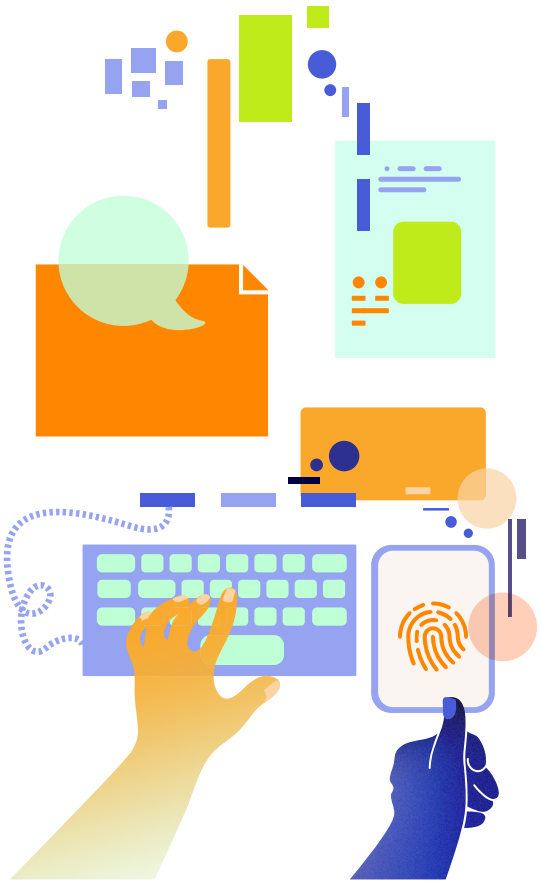
spotlight

Speaking in Tongues

Teaching Local Languages to Machines

Prasanna Lal Das

April 2023



ChatGPT[1], the charming, somewhat unreliable chatbot from OpenAI, went viral at the end of 2022, engaging people in conversations that ranged from the existential[2] to the mundane. Other generative AI tools such as DALL-E,[3] which can create images based on simple descriptions, and Make-a-Video,[4] which can generate short video clips based on simple text descriptions, have also sparked enthusiasm and have even won art competitions.[5] Similar tools are rapidly becoming fixtures in homes, where Alexa and Siri banter with and amuse people[6] when not following commands to turn off the lights or play the current No. 1 hit on a smart speaker.

But it's not all fun and games. If the evangelists are to be believed, the impact of these tools will soon show up in productivity data,[7] as chatbots begin to do things like write code, create public relations materials, and replace research assistants. The implications of such a productivity boost for international development are obvious.

As things stand now, however, people who don't speak or write English, or one of the other "major" languages that are generally spoken in advanced economies, are out of luck when trying to access

or use these tools and services. ChatGPT, for instance, largely understands the world through the eyes of English-speaking content creators. English comprises the bulk of the training data, while additional languages such as Spanish, French, German, Italian, Portuguese, Dutch, Russian, Arabic, Chinese, Japanese, Korean, and Hindi are also used to train the chatbot. Other tools like voice assistants also support only a small number of the world's languages. Currently, Google Home doesn't support Zulu,[8] which is widely spoken in South Africa, one of the more developed markets in Africa.

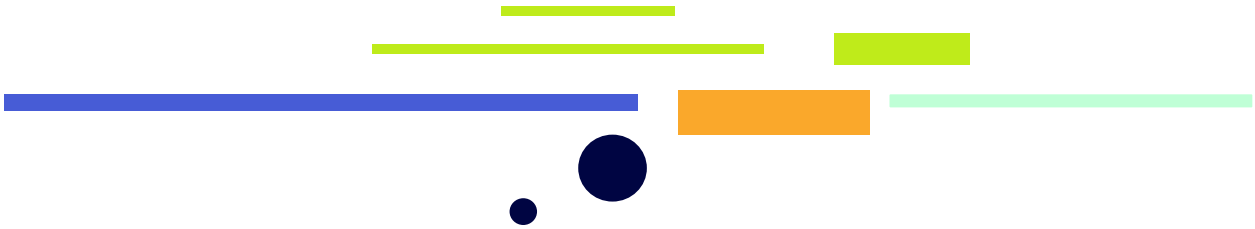
It's easy to see why this is the case. Machines and algorithms learn through exposure to a sufficiently large corpus of knowledge, which is typically available through written, video, and audio materials (e.g., books, articles, movies, and cartoons[9]), ideally online in digital format. This corpus powers natural language libraries (NLP) that provide the intelligence embedded in these machines.

The unfortunate reality is that the quantity and quality of available explicit knowledge about developing countries is relatively low, and even lower in local languages. For example, 60% of the 10 million most popular websites on the internet are in English, according to one estimate.[10] Hindi, spoken by more than 600 million people worldwide, is the top South Asian language but accounts for only 0.1% of online content. Other languages like Bengali and Urdu, which are spoken by hundreds of millions of people, don't even appear on the list. Content in the African language of Igbo, spoken by at least 30 million people, makes up less than 0.1% of all online material. Only 85 of the world's 7,500 languages are represented in the major NLPs,[11] and the vast majority of NLPs support just seven languages, with English being the most advanced.[12]

In addition to excluding the more than 5 billion people who do not speak English, these tools reinforce the cultural and ethical norms of the English-speaking world or the English-speaking upper classes in countries such as India, thus perpetuating discrimination and other harmful effects. Not only do non-English speakers lack access to useful information, they might also be inundated with disinformation, since the tools developed to fight disinformation by platforms such as Twitter are much better at spotting and weeding out disinformation in English than in other languages.[13] The consequences are economic as well. It is expensive and time-consuming for local firms to develop products and solutions in local languages to meet the needs of unaddressed local markets.

Data Governance Initiatives on the Ground

While recognizing that the challenge is daunting, several data initiatives have emerged in the developing world and elsewhere that have the potential to bridge the language gap in digital tools and services. The following section highlights a few examples of data governance in action to ensure that countries and companies either harness existing data assets better or develop new ones to make human-machine interaction more inclusive.

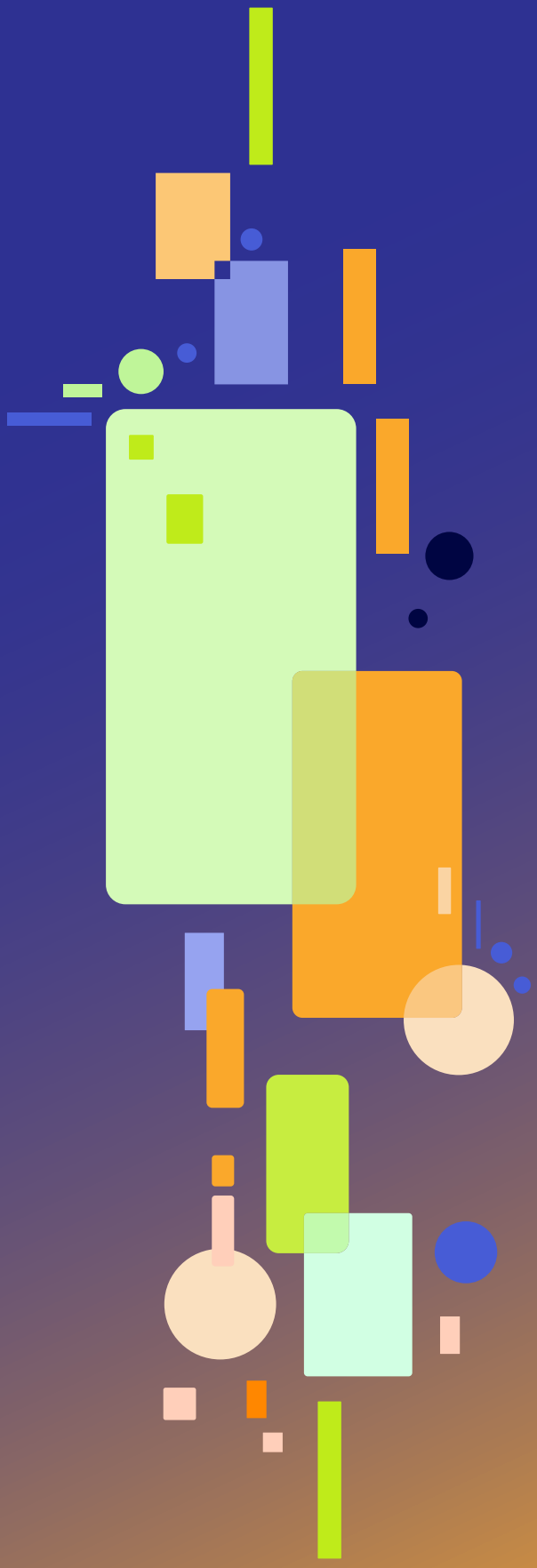


NLP of, by, and for the People: A Grassroots Approach to Bridging the NLP Divide in Africa

Africans speak approximately 200 languages,[21] but Apple's Siri, Google's Assistant, and Amazon's Alexa collectively speak zero African languages.[22] Masakhane,[23] a grassroots organization in Africa, has been spurring the growth of NLP research in Africa across several dimensions, including:

- Scientific corpora in African languages[14] to make scientific research more accessible to Africans and empower them to develop new scientific methods based on local traditions
- Named entity recognition (NER) in a wide variety of African languages[15] to make it easier to identify African people, locations, organizations, etc.
- Sentiment corpus and lexicon[16] in different African languages to make it easier to classify people's opinions and identify/address disinformation
- Open, accessible, and high-quality text and speech datasets that can be consumed by digital and data services and products[17]

Masakhane currently counts 60 active contributors in its community. A key component of its strategy is to promote African translations by Africans themselves and eventually bring African knowledge systems, which have been suppressed during colonial periods, to the fore. Community engagement techniques explored by Masakhane include participatory translation,[18] crowdsourced voice and speech contributions, community creation of datasets,[19] and community curation of speech.[20]



Information and Disinformation in Serbia – NLP in the Public and Private Sectors[28]

The World Health Organization (WHO) estimates that thousands of people lost their lives or were hospitalized during the

COVID-19 pandemic due to disinformation.[24] Governments and civil society organizations were keen to counter online disinformation but were stymied by the lack of tools in languages such as Serbian. The limited tools that do exist, often developed by large firms, are either inaccessible or have onerous or restrictive licensing requirements.

To fill the gap, the Artificial Intelligence Institute of Serbia, in partnership with WHO and the United States Agency for International Development (USAID) has implemented an NLP-based sentiment analysis tool to track information on social media platforms like Twitter. The tool, which is available to a small set of authorized users, highlights some of the challenges developers face when creating digital services that aren't backed by a sufficiently large language corpus:

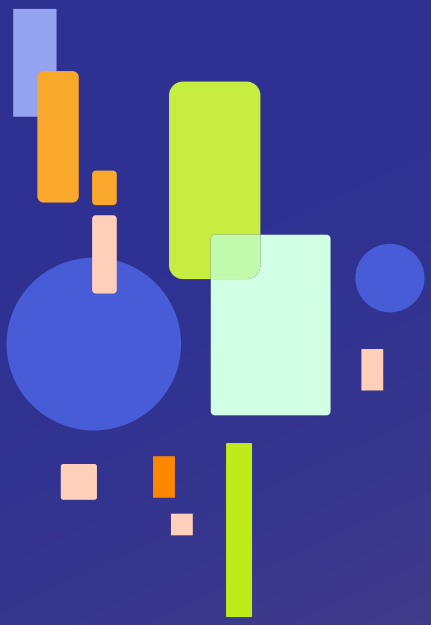
- The training data, sourced from Twitter, was only available for education purposes or use by international organizations as per Twitter's policies at the time. Such data can be cost-prohibitive for many entities.
- It was challenging to accurately identify tweets in Serbian due to similarities with neighboring languages like North Macedonian. The team did have the benefit of being able to use BERTic,[25] a language model recently developed for several Balkan languages. Many tweets did not contain geo-data that would have placed their origin in Serbia (or elsewhere).
- Significant human annotation was required to assemble the training dataset of 10,000 tweets.

One positive outcome of the initiative was that the language model developed by the team is now available to other users. The team recognizes the limitations of the Twitter-based training dataset and is keen to extend its work utilizing additional sources for Serbian language data, like Serbian language TV broadcasts and even emails.

Other Serbian language-based NLP projects[26] have emerged from the private and public sectors. For example, the Regional Linguistic Data Initiative[27] (ReLDI) is taking a community-based approach that will convene entrepreneurs, academia, CSOs, and government entities to develop NLP resources that are publicly accessible and easily reusable under permissive licenses. A core part of the group's plan is to invest in training and education so Serbia can strengthen its technical NLP capabilities.

Digitalizing Languages With a Little Help From the Government in Countries Large and Small

Many governments, especially those that represent people who speak and write in languages poorly served by the current set of NLP resources, have begun to see NLP as an important part of their data/AI infrastructure and, therefore, might need to make a basic investment in it. These countries might be homogenous and small like Estonia or large and heterogenous like India, where people speak many regional languages that are not readily available online.



Estonia

The government of Estonia launched the Estonian Language Technology 2018-2027 program recognizing that "the world is increasingly interacting with machines in natural language, or by speech" and "it is not always profitable for the private sector to take on the risks associated with the development of technology for a language with a small number of speakers, as a small number of speakers also means a small market." The goals of the program are to both establish language standards and tools on par with the best international standards and ensure that they are "implemented in as many areas as possible, in the private, public and third sectors.[29]"

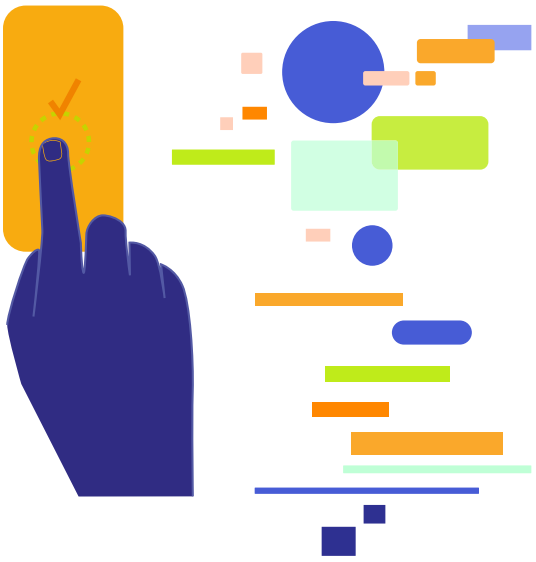
India

Indians speak 22 official languages, and a government estimate suggests that that there are 1,576 "rationalized" languages in India, apart from the 1,796 "mother tongues."[30] Hindi, spoken by more than 600 million people, appears on fewer that 0.1% of the websites worldwide.[31] Other languages such as Bengali, Urdu, Marathi, Tamil, and Telugu—each spoken by tens or hundreds of millions of people and among the 20 most spoken languages worldwide—account for even less online content. The Indian government, recognizing the importance of local languages for communication, outreach, knowledge, economic growth, and cultural preservation, recently launched the Bhasini program designed to transcend the language barrier and "build a National Public Digital Platform for languages to develop services and products for citizens by leveraging the power of artificial intelligence and other emerging technologies." An interesting feature of the platform is "Bhasha Daan," a crowdsourcing initiative inviting average citizens to create "an open repository of speech recognition data, parallel translations, and labeled images.[32]"

Big Tech: A Vital Resource If Not Always a Benevolent One

While ChatGPT may still speak only a few languages, Google Translate lets people converse with each other virtually everywhere in the world. Its companion commercial translation service is offered in 135 languages.[34] The No Language Left Behind[35] project from Meta claims to offer more than 200 languages for translation, giving "people the opportunity to access and share web content in their native language, and communicate with anyone, anywhere, regardless of their language preferences." One of its notable commitments is to translate Wikipedia into 200 languages, which will vastly increase the reach of the platform.

However, it is unclear if these large tech firms plan to offer NLP resources as a public good, and if they do, what kind of licensing arrangements may be available. Some researchers have also expressed concerns about dependence on big tech and the power imbalance that many developing country participants experience when interacting with global monopolies whose primary markets may be elsewhere.



Common Themes Across the Example Data Initiatives

The examples above are just a subset of the initiatives on the ground to make digital services accessible to people in their local languages around the world. Common themes across the initiatives include:

- Despite the excitement and enthusiasm, **most of the programs above are still at a very nascent stage**. Many may fail, and others will require investment and time to succeed. While countries such as India have initiated formal national NLP programs, others such as Serbia have taken a more ad hoc approach.

- Smaller countries like Estonia recognize the need for state intervention, as the local population isn't large enough to attract private-sector investment. **Countries will need to balance their local, cultural, and political interests against commercial realities** as languages become digital or are digitally excluded.

- **Community engagement is an important component of almost all initiatives.** India has set up a formal crowdsourcing program. Other programs in Africa are experimenting with elements of participatory design and crowd curation.

- While critics have accused ChatGPT and others of paying contributors from the Global South very poorly for their labeling and other content services,[33] it appears that **many initiatives in the south are beginning to dabble with payment models to incentivize crowdsourcing** and sustain contributions from the ground.

- **The engagement of local populations can ensure that NLP models learn appropriate cultural nuances** and better embody local social and ethical norms.

Endnotes

1 https://openai.com/blog/chatgpt/

2 “ChatGPT and 20 questions,” https://rodfaulkner.medium.com/chatgpt-and-20-questions-7ac5fc9c4aea

3 https://openai.com/blog/dall-e-introducing-outpainting/

4 https://ai.facebook.com/blog/generative-ai-text-to-video/

5 “Artwork made by AI wins competition,” https://impakter.com/art-made-by-ai-wins-ine-arts-competition/

6 “Funny conversations with Siri and Alexa,” https://www.facebook.com/youarepeachy/videos/funny-conversations-with-siri-and-alexa/2199455656838392/

7 “How generative AI will supercharge productivity,” https://www.fastcompany.com/90836481/how-generative-ai-will-supercharge-productivity

8 “Conversational AI: Africans disproportionately disadvantaged,” https://www.context.news/ai/opinion/conversational-ai-africans-disproportionally-disadvantaged.

9 “CIDEr: Consensus-based image description evaluation,” https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vedantam_CIDEr_Consensus-Based_Image_2015_CVPR_paper.pdf.

10 “Usage statistics of content languages for websites,” https://w3techs.com/technologies/overview/content_language.

11 “New languages for NLP: Building linguistic diversity in the digital humanities,” https://cdh.princeton.edu/projects/new-languages-nlp-building-linguistic-diversity-digital-humanities/

12 “The importance of natural language processing for non-English languages,” https://towardsdatascience.com/the-importance-of-natural-language-processing-for-non-english-languages-ada463697b9d

13 “Election disinformation in different languages is still a big problem in the US,” https://cdt.org/insights/election-disinformation-in-different-languages-is-a-big-problem-in-the-u-s/

14 “Ethologue: Languages of the world,” https://www.ethnologue.com/region/Africa

15 “Why AI needs to be able to understand the world’s languages,” https://www.scientificamerican.com/article/why-ai-needs-to-be-able-to-understand-all-the-worlds-languages/

16 https://www.masakhane.io/

17 “Masakhane MT: Decolonise science,” https://www.masakhane.io/ongoing-projects/masakhane-mt-decolonise-science

18 “Masakhane NER: Know our names,” https://www.masakhane.io/ongoing-projects/masakhaner-know-our-names

19 “Naija NLP: Sentiment lexicon and hate speech,” https://www.masakhane.io/ongoing-projects/naijanlp-sentiment-lexicon-hate-speech

20 “Makerere NLP: Text and speech for East Africa,” https://www.masakhane.io/ongoing-projects/makererenlp-text-speech-for-east-africa

21 “Participatory translation of Oshiwambo: Towards cultural preservation through language technology,” https://openreview.net/forum?id=BFbg59zVUZc

22 “Machine translation for African languages: Community creation of datasets and models in Uganda,” https://openreview.net/forum?id=BK-z5qzEU-9

23 TCNspeech: A community-curated speech corpus for sermons,” https://openreview.net/pdf?id=r_-PYcf4LZc

24 Based on inputs provided by the Government of the Republic of Serbia.

25 “Fighting misinformation in the time of Covid-19: One click at a time,” https://www.who.int/news-room/feature-stories/detail/ғghting-misinformation-in-the-time-of-covid-19-one-click-at-a-time

26 “BERTic: The transformer language model for Bosnian, Croatian, Montenegrin, and Serbian.

27 Based on input from Slobodan Markovic, Digital Advisor, UNDP Serbia

28 https://reldi.spur.uzh.ch/

29 The Language and Technology Research Development Program ‘Estonian Language Technology 2018-2027’ of the Ministry of Education and Research, https://www.keeletehnoloogia.ee/en/the-language-technology-research-and-development-program-201cestonian-language-technology-2018-2027201d-of-the-ministry-of-education-and-research/view

30 The Indian linguistic space, https://www.education.gov.in/hi/sites/upload_ғles/mhrd/ғles/upload_document/languagebr.pdf

31 Usage statistics of Hindi for websites, https://w3techs.com/technologies/details/cl-hi-

32 https://bhashini.gov.in/en/about

33 Translation Hub, https://cloud.google.com/translation-hub#section-9

34 https://ai.facebook.com/research/no-language-left-behind/

35 “The future of the work we seek,” https://itforchange.net/sites/default/ғles/2261/ITfC_Future_of_Work_Report_2022_.pdf?mkt_tok=Njg1LUtCTC03NjUAAAGJz977qF_qldFk-m2ElQ9Y_XVA2cHNZxIV3UmiLKJY0clzP08S3cUfUH-mH12U8itmUwFXs_YBHKuHmiGDcHho4jYTYVoxb2WdlpI7dHjlsHOU