FORUM ON INTERNET FREEDOM IN AFRICA 2021

Event Report

Linguistic and cultural diversity as an integral human right aspect of Internet freedom in the digital age

unesco
Centre
Under the auspices
of UNESCO

IRC AI
International Research Centre
of Artificial Inteligence
under the auspices of UNESCO

Internet Freedom

K4A
knowledge for all

REPUBLIKA SLOVENIJA
GOV.SI

Report: Linguistic and cultural diversity as an integral human right aspect of Internet freedom in the digital age

Dakar, Senegal / online

29 September at 16:30 EAT

(online)

As part of the Forum on Internet Freedom in Africa 2021 (September 28-30), this roundtable was supported by the International Research Centre on Artificial Intelligence (IRCAI) under the auspices of UNESCO, the Knowledge 4 All Foundation and the Ministry of Foreign Affairs of the Republic of Slovenia.

Abstract

Languages are a cornerstone of cultural identity. They are the means by which we bequeath our cultures and norms from generation to generation. However, the digital realm is an unfair ground where only a few global languages dominate, letting smaller African language groups behind. Therefore, in this panel discussion, a multi-disciplinary range of speakers will discuss how linguistic and cultural diversity can be preserved in the light of the digital age we are living in. Highlighting the efforts made by Masakhane, the focus of the panel discussion will lie on how openly accessible text and speech datasets can fuel *Natural Language Processing (NLP)* technologies. Thereby, a particular emphasis is put on how the African Union (AU), European Union (EU) and other players in the field can join forces in leveraging the presence of African languages in the information society.

Key Points

Moderated by Arthur Gwagwa (Utrecht University & UN Global Pulse), the roundtable is joined by a multi-disciplinary group of panelists, ranging from computer scientists to NLP researchers to ethicists:

- **Jade Abbott** (Co-Founder of Masakhane & Head of Data at Retro Rabbit)
- **Wilhelmina Onyothi Nekoto** (NLP Researcher at Masakhane)
- **Peter Nabende** (Lecturer at Makerere University)
- **Emre Kazim** (Co-Founder of Holistic AI & Research Fellow at UCL)

Why do dominating languages represent a danger for Africa?

Traditionally, when talking about internet freedom, we tend to refer to the civil and political rights we obtain from freedom of speech, or the rights people are deprived from under regimes that exerce internet shutdowns. However, linguistic diversity is just as much of a right to freedom of expression, as also stated in the *Universal Declaration of Human Rights*. Being able to express oneself in one's own language is an economic, social as well as cultural right. As Jade stated: "How can we participate in a digital democracy if we cannot even engage or express ourselves in the language we are most comfortable in?". Wilhelmina adds that Africans particularly struggle to tell their stories freely and openly in the digital realm, not merely because of the restrained press freedom under certain regimes, but again, simply because of the impossibility to engage in the languages they are most proficient in.

How did languages get to dominate the African digital realm?

The predominant languages that mediate communication in the digital age are languages that originate in Europe (English and to a lesser extent French or Spanish), creating an unfair ground in the digital language space. Arthur explains that, as a result of that, these languages and related cultures are then overproportionally reflected in our datasets. Meanwhile, algorithms are developed which not only rely on these datasets, but also train them. Jade adds that developers often just base their language technologies (e.g. *NLP*) on European languages and only after that they start replicating the same procedure with on African languages.

Arthur also brings up the role of economic interests in determining to which extent languages are represented in our data: In this digital era, *digital enclosures* allow companies to claim ownership over and operate with the information they obtain from their customers and potential target audience. Thereby, a firm that wants to enter a market in Africa can precisely tell which language group they want to harvest data on and which language they want to generate the data in, depending on which language their target audience is most likely to speak. However, according to Emre, the problem does not lie only in how the data is harvested but also how the data is categorized in first place: Categories often do not divide groups of people according to a meaningful common denominator but rather reflect how the data providers mistakenly perceive social phenomena.

How can language technologies be more inclusive towards underrepresented language groups?

According to Jade, language itself is not separate from the culture. It is the culture, in a way. But "anything else that we can do to augment that in a digital space is better". Peter adds that yet a lot of work needs to be done in order to account for the unaccounted languages in speech databases used in *NLP*. However, Jade asserts that since "we're often starting from scratch in Africa", we can at least be more intentional about how we represent culture online, instead of it being of a matter of circumstance. Peter thereby acknowledges that awareness about the underrepresentation of low-funded and endangered languages is rising, as various communities are taking action to boost the representation of these languages in design, datasets, and systems.

One of these communities is *Masakhane*, a grassroots organisation whose mission is to "strengthen and spur NLP research in African languages, for Africans, by Africans". Wilhelmina, for example, co-creates text datasets for language speech to power automated speech recognition for endangered languages. Thereby, she regularly engages in conversations with the elderly, seeking to understand their experiences and to co-create datasets capturing their speech to help them document their stories and preserve their cultural norms. Jade adds that *Masakhane*'s goal is to create tools which people, who have been unaccounted for in the digital environment up until now, can eventually engage in the digital space in the language of their choice. Open-source tools would thereby allow communities to actively contribute to the speech dataset that can then be used for translation or *Named Entity Recognition* (NER) tools.

How can different actors be engaged?

As indicated by Jade, Masakhane puts an emphasis on "collect[ing] data and translat[ing] texts into our languages, source texts from cultures who care about them". She considers Masakhane "a more sustainable initiative that is not torn by international priorities". Thus, Masakhane seeks to cooperate primarily with African institutions. By speaking to the African Union (AU), the organisation hopes to involve them more in this process. Emre accentuates that when looking into the ethical aspects of engaging different communities and accounting for them in the dataset, the political, psychological, and cultural context of Africa needs to be taken into consideration. The autonomy of communities must be respected, but polarities in contexts across the continent must also be born in mind.

Davor Orlić (COO at IRCAI, facilitator of this panel discussion) closes off the roundtable with an appeal is to the European Union and the African Union to create bridges, to release funding for developing AI technologies for connecting the common markets of the African Union (AU) and the European Union (EU), and to give researchers a chance to boost progress in the field of Language Technology (LT) for realising digital rights in Africa.

The full transcript of the panel discussion can be found here.

Quotes

"Linguistic and cultural diversity is also very important not only as a civil and political human right, but also as an economic, social and cultural right."

Arthur Gwagwa (Doctoral Researcher at Utrecht University & Expert on Governance of Data and AI at UN Global Pulse)

"If we can't even engage at all in our language on the internet, so much gets lost: How can we participate in this digital democracy? How can we participate in shaping the future and shaping these global discussions if we cannot even express it in the language which we're most comfortable with?"

Jade Abbott (Co-Founder of Masakhane & Head of Data at Retro Rabbit)

"We're storytellers as Africans, but we cannot tell our story freely and openly. This is for one because of internet freedom or press freedom issues, but many of us also still can't participate also because of our languages."

Wilhelmina Onyothi Nekoto (NLP Researcher at Masakhane)

"Instead of having Europeans coming to document, collect data and translate their texts into our languages, it's more about sourcing texts from the cultures who care about them."

Jade Abbott (Co-Founder of Masakhane & Head of Data at Retro Rabbit)

"These categories are not reflections of the data, but themselves imposed on the way in which we're thinking about science and how that creates a social phenomenon."

Emre Kazim (Co-Founder of Holistic AI & Research Fellow at UCL)

"Going into the future, we will have this phenomenon where you consider it in the designs, in the datasets you are using, in the systems and so on."

Peter Nabende (Lecturer at Makerere University)