



# 4IR IN AFRICA CANNOT TAKE PLACE IN ENGLISH ONLY

**Of the roughly 7 000 languages in the world, 2 000 of them are African – yet African languages are barely represented in technology. The Masakhane Natural Language Processing (NLP) research project aims to give these languages a voice and spur NLP research in African languages, for Africans, by Africans.**

Researcher: Dr Vukosi Marivate, ABSA Chair of Data Science

**I**n times of crisis, language can fail us. What does “sanitise” mean in most African languages? What does “social distancing” mean? We don’t know, because there are not many systems to consider how such concepts are presented in these languages.

The Data Science for Social Impact (DSFSI) research group in the Department of Computer Science at the University of Pretoria has created tools to make better use of the small amount of data for these “low-resourced languages”.

The Masakhane project, an African natural language processing research initiative, was named a joint winner of the inaugural 2021 Wikimedia Foundation Research Award of the Year. Dr Vukosi Marivate, ABSA Chair of Data Science in the Faculty of Engineering, Built Environment and Information Technology (EBIT), is one of the leaders of the project. The UP team has gone on to create an online translation interface to the

Masakhane project. Collaborating with Dr Marivate on this project are Abiodun Modupe, his colleague in UP’s Department of Computer Science, and Catherine Gitau and Salomon Kabenamualu of the African Master’s in Machine Intelligence programme, which is part of the African Institute for Mathematical Sciences in Ghana. Dr Marivate considers this Wikimedia Foundation award a great honour. “It also shows that we can innovate, even in the ways we conduct research,” he says.

The DSFSI research group received funding from Mozilla Open Source Support (MOSS) to build a web tool to show off the translation tools developed with Masakhane. Funding provided by the MOSS Foundation was also used to build the machine translation tool. The beta version of the tool made provision for feedback from test users, as well as from members of the Masakhane community.



The final tool has since been launched, and is available as an open-source resource.

## Machine-learning tool

“The idea for a machine-learning tool to assist in the translation of 50 of the regional languages on the continent was developed at the #SautiYetu African NLP Unconference 2020,” Dr Marivate says. “This event is linked to the Deep Learning Indaba, an organisation focused on strengthening African machine learning (ML) and supporting Africans to be owners of technology advances and artificial intelligence (AI). An objective of the organisation is to create leadership and recognise excellence in the development of ML and AI across Africa.”

The research that formed part of the development of the Masakhane machine translation tool describes a novel approach to machine translation for African languages, illustrating how the challenges these languages face to join the web can be overcome, and some of the technologies from which other languages benefit today. The Wikimedia Foundation calls this research an inspiring example of work towards knowledge equity, which is one of the two main pillars of the 2030 Wikimedia Movement Strategy.

The project also gave rise to the establishment of the Masakhane grassroots community, which aims to develop NLP systems for Africa by Africans. Masakhane, which roughly translates as “we build together” in isiZulu, has as its goal for Africans to shape and own technological advances towards human dignity, well-being and equity through inclusive community building, open participatory research and multi-disciplinarity. Members of the Masakhane community provide data for the research project and assist in building models and testing sample translations to improve the accuracy of the machine translation tool.

They represent the African countries and languages that form part of the research, and comprise individuals from a range of relevant professions, including data scientists, researchers, language practitioners, translators and software developers.

According to Dr Marivate, several related developments are in the pipeline to support African language tasks, including new machine-learning models and a speech-to-text translation tool.

“An inspiring example of work towards knowledge equity.”

The Masakhane translation tool is available as an open-source resource and can be accessed at <http://translate.masakhane.io>