MASAKHANE - MACHINE TRANSLATION FOR AFRICA

∀, **Iroro Fred Ọ̀nọ̀mẹ̀ Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin,*** **Alp ktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, Abdallah Bashir**
Masakhane, Africa
masakhane.io
masakhane-mt@googlegroups.com

1 THE STATE OF AFRICAN NLP

2144 of all 7111 (30.15%) living languages today are African languages (Eberhard et al., 2019). But only a small portion of linguistic resources for NLP research are built for African languages. As a result, there are only few NLP publications: In all ACL conferences in 2019, only 5 out of 2695 (0.19%) author affiliations were based in Africa (Caines, 2019). This stark contrast of *linguistic richness versus poor representation* of African languages in NLP is caused by multiple factors.

First of all, African societies do not see hope for African languages being accepted as primary means of communication (Alexander, 2009). As a result, few efforts to fund NLP or translation for African languages exist, despite the potential impact. This *lack of focus* has had a ripple effect.

The few existing resources are not easily discoverable, published in closed journals, non-indexed local conferences, or remain undigitized, surviving only in private collections (Mesthrie, 1995). This *opaqueness* impedes researchers' ability to reproduce and build upon existing results, and to develop, compete on and progress public benchmarks (Martinus & Abbott, 2019).

African researchers are disproportionately affected by *socio-economic factors*, and are often hindered by visa issues (Johnson, 2019) and costs of flights from and within Africa (Hattem, 2017). They are distributed and disconnected on the continent, and rarely have the opportunity to commune, collaborate and share.

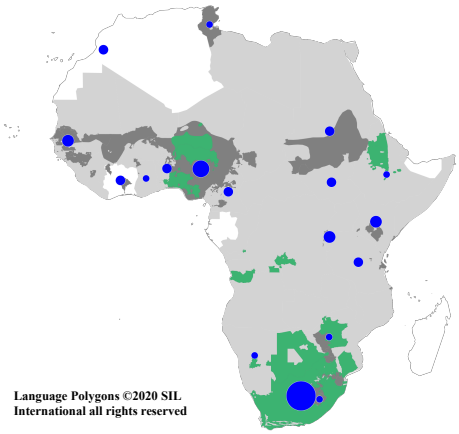
Furthermore, African languages are of *high linguistic complexity and variety*, with diverse morphologies and phonologies, including lexical and grammatical tonal patterns, and many are practiced within multilingual societies with frequent code switching (Ndubuisi-Obi et al., 2019; Bird, 1999; Gibbon et al., 2006). Because of this complexity, cross-lingual generalization from success in languages like English are not guaranteed.

2 CONTRIBUTION

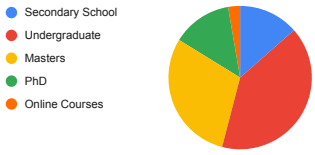
Founded at the *Deep Learning Indaba 2019*, MASAKHANE constitutes an open-source, continent-wide, distributed, online research effort for machine translation for African languages. Its goals are threefold:

1. **For Africa**: To build a community of NLP researchers, connect and grow it, spurring and sharing further research, to enable language preservation and increase its global visibility and relevance.

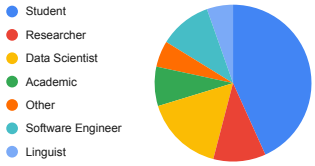
*In rebellion against the status attributed to author order, the order of authors has been randomised (Strange, 2008). The symbol ∀ furthermore takes the place as first author to represent the whole community.



(a) Origin of Participants & Focus Language.



(b) Highest Level of Education.



(c) Occupation.

Figure 1: Participants with African origin are represented by blue markers in (a), indigenous areas that are covered by the languages of current benchmarks in green, benchmarks in progress in dark grey, and countries where those languages are spoken in light grey. Education (b) and occupation (c) of a subset of 37 participants as indicated in a voluntary survey in February 2019.

2. **For NLP researchers**: To build data sets and tools to facilitate NLP research on African languages, and to pose new research problems to enrich the NLP research landscape.

3. **For the global researchers community**: To discover best practices for distributed research, to be applied by other emerging research communities.

3 Methodology and Results

MASAKHANE's strategy is to offer *barrier-free open access* to first hands-on NLP experiences with African languages, fighting the above-mentioned opaqueness. With an easy-to-use open source platform, it allows individuals to train neural machine translation (NMT) models on a parallel corpus for a language of their choice, and share the results with an online community. The *online community* is based on weekly meetings, an active Slack workspace, and a GitHub repository (`github.com/masakhane-io`), so that members can support each other and connect despite geographical distances. *No academic prerequisites* are required for participation, since tertiary education enrolments are minimal in sub-saharan Africa (Jowi et al., 2018).

A *Jupyter Notebook* features documented data preparation, model configuration, training and evaluation. It runs on Google Colab with a single (free) GPU for a small limited number of hours, such that participants do not require expensive hardware. The NMT models are built using Joey NMT (Kreutzer et al., 2019), which comes with a beginner-friendly documentation. Participants submit and publish their data, code and results for training on their language to improve reproducibility and discoverability. To lower the barrier of data collection, the JW300 multilingual dataset (Agić & Vulić, 2019) with parallel corpora for English to 101 African languages is integrated into the notebook. With the goal of improving translation quality by transfer learning across languages in the future, *global test sets* with English sources are extracted from JW300, and excluded from training data for any language pair to avoid potential data leakage for cross-lingual transfer.

As of February 14, 2020, the MASAKHANE community consists of 144 participants from 17 African countries with diverse educations and occupations (Figure 1), and 2 countries outside Africa (USA and Germany). So far, 30 translation results for 28 African languages have been published by 25 contributors on GitHub.

4 FUTURE ROADMAP

MASAKHANE aims to continue to grow and facilitate engagement within the community, especially helping inactive users contribute benchmarks and fostering mentoring relations. In the next year, the project will expand to different NLP tasks beyond NMT in order to reach a broader audience. Qualitative analysis on model performance as well as investigations of automatic evaluation metrics will spur healthy competition on results. MASAKHANE will also provide notebooks for transfer and un/self-supervised learning to push translation quality. In terms of data collection, the size and domain of global test sets will be expanded.

REFERENCES

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. doi: 10.18653/v1/P19-1310. URL https://www.aclweb.org/anthology/P19-1310.

Neville Alexander. Evolving african approaches to the management of linguistic diversity: The acalan project. *Language Matters*, 40(2):117–132, 2009.

Steven Bird. Strategies for representing tone in african writing systems. *Written Language & Literacy*, 2(1): 1–44, 1999.

Andrew Caines. The geographic diversity of nlp conferences, Oct 2019. URL http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/.

David M Eberhard, Gary F. Simons, and Charles D. Fenning. Ethnologue: Languages of the worlds. twenty-second edition, 2019. URL https://www.ethnologue.com/region/Africa.

Dafydd Gibbon, Eno-Abasi Urua, and Moses Ekpenyong. Morphotonology for tts in niger-congo languages. In *Speech Prosody*, 2006.

Julian Hattem. African air travel is awful. why?, Nov 2017. URL https://www.citylab.com/transportation/2017/11/why-is-african-air-travel-so-terrible/546422/.

Khari Johnson. Canada is denying travel visas to ai researchers headed to neurips - again, Nov 2019. URL https://venturebeat.com/2019/11/09/canada-is-denying-travel-visas-to-ai-researchers-headed-to-neurips-again/.

James Jowi, Charles Ochieng Ongondo, and Mulu Nega. Building phd capacity in sub-saharan africa, 2018. URL https://www.britishcouncil.org/education/ihe/knowledge-centre/developing-talent-employability/phd-capacities-sub-saharan-afric.

Julia Kreutzer, Joost Bastings, and Stefan Riezler. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, Hong Kong, China, 2019.

Laura Martinus and Jade Z Abbott. A focus on neural machine translation for african languages. *arXiv preprint arXiv:1906.05685*, 2019.

Rajend Mesthrie. *Language and social history: Studies in South African sociolinguistics*. New Africa Books, 1995.

Innocent Ndubuisi-Obi, Sayan Ghosh, and David Jurgens. Wetin dey with these comments? modeling sociolinguistic factors affecting code-switching behavior in nigerian online discussions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6204–6214, 2019.

Kevin Strange. Authorship: why not just toss a coin? *American Journal of Physiology-Cell Physiology*, 295(3):C567–C575, 2008. doi: 10.1152/ajpcell.00208.2008. URL https://doi.org/10.1152/ajpcell.00208.2008. PMID: 18776156.