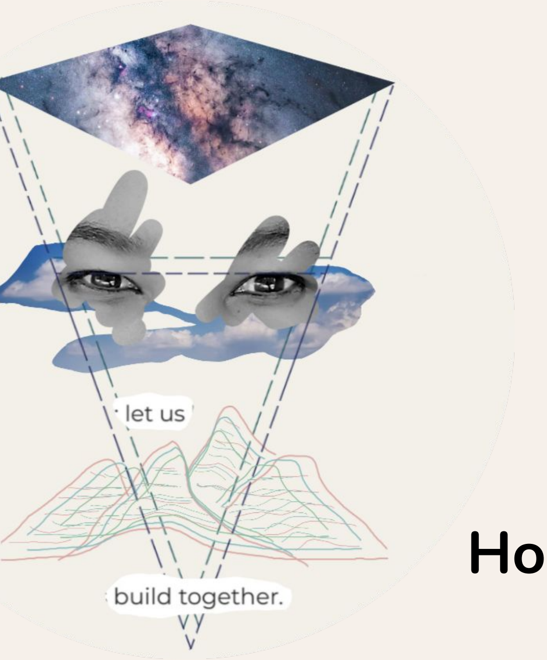
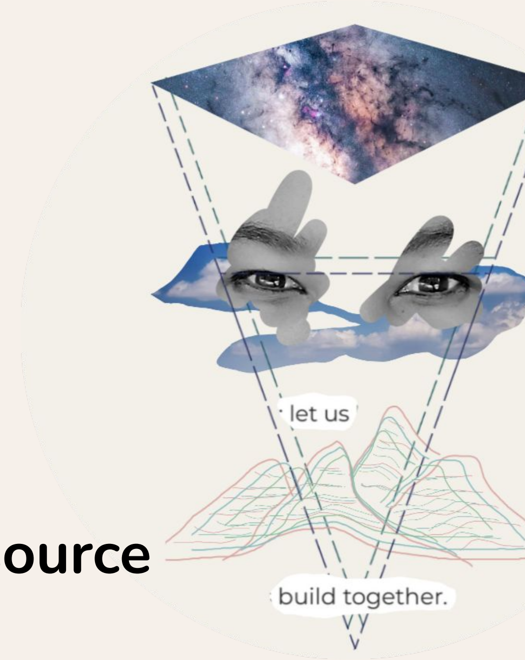


Transforming Africa

How we created a hub for low-resource languages



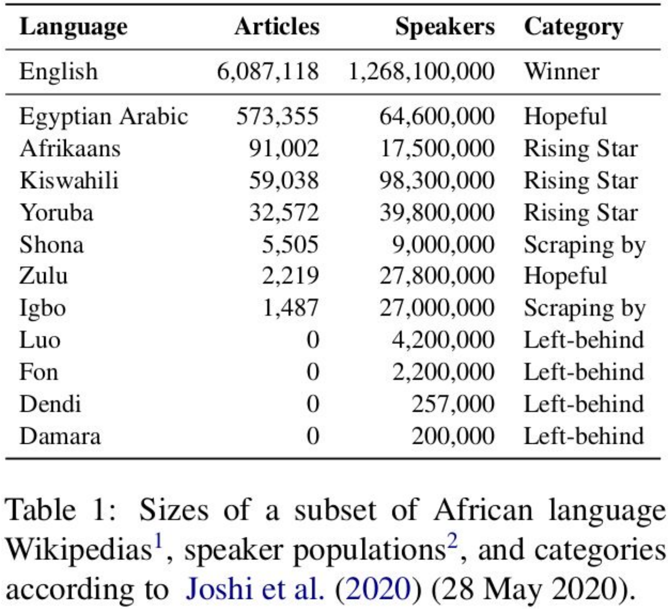
Presented By
Orevaoghene Ahia | Chris Emezue
For MASAKHANE

Open “NLP” Workshop 2021

Outline

- Introduction
- State of NLP in Africa before Masakhane
- How we are bridging the gap ?
- Success story
- Ongoing projects

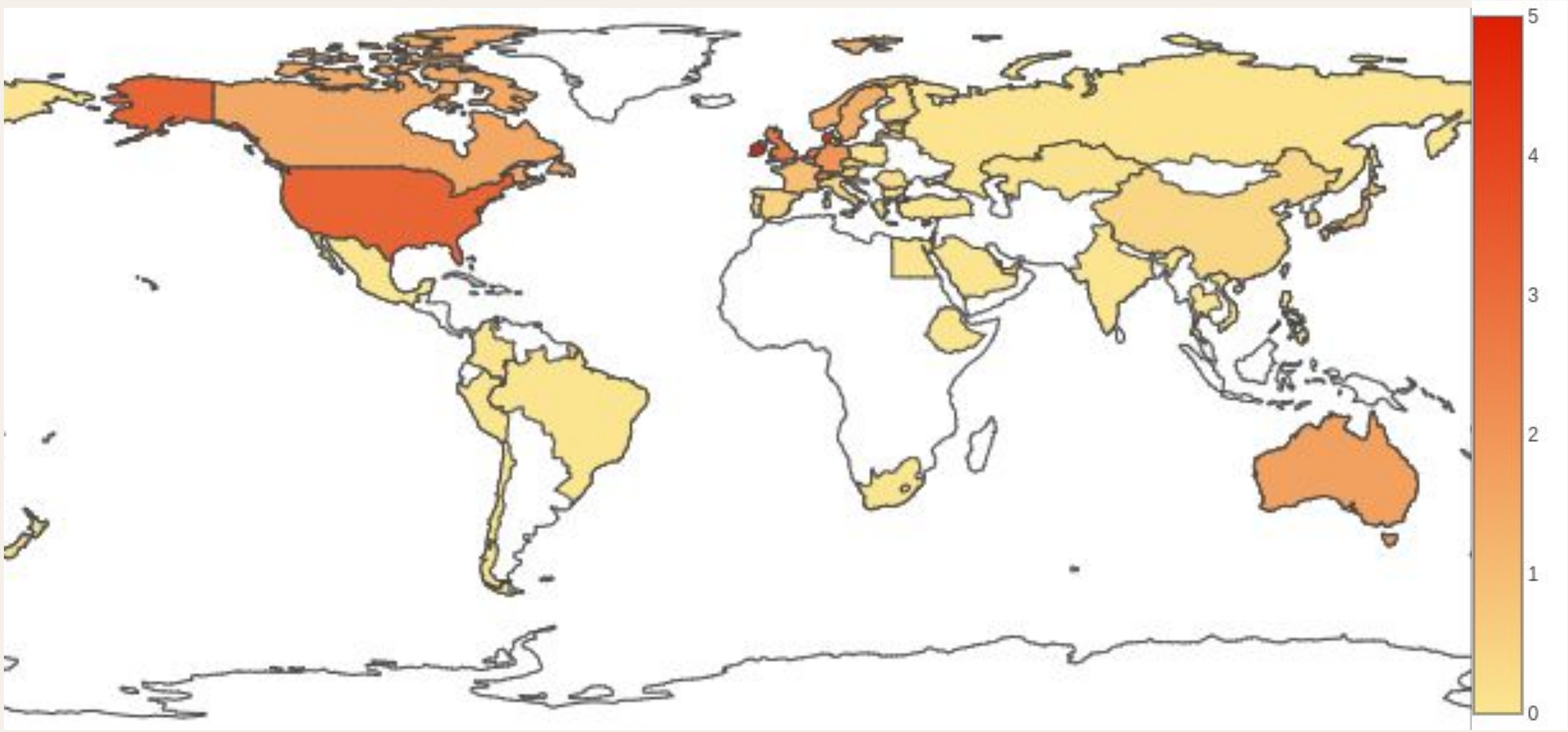
Low-resourcedness



Language	Articles	Speakers	Category
English	6,087,118	1,268,100,000	Winner
Egyptian Arabic	573,355	64,600,000	Hopeful
Afrikaans	91,002	17,500,000	Rising Star
Kiswahili	59,038	98,300,000	Rising Star
Yoruba	32,572	39,800,000	Rising Star
Shona	5,505	9,000,000	Scraping by
Zulu	2,219	27,800,000	Hopeful
Igbo	1,487	27,000,000	Scraping by
Luo	0	4,200,000	Left-behind
Fon	0	2,200,000	Left-behind
Dendi	0	257,000	Left-behind
Damara	0	200,000	Left-behind

Table 1: Sizes of a subset of African language Wikipedias[1], speaker populations[2], and categories according to Joshi et al. (2020) (28 May 2020).

Artwork by Keneilwe Mokoena



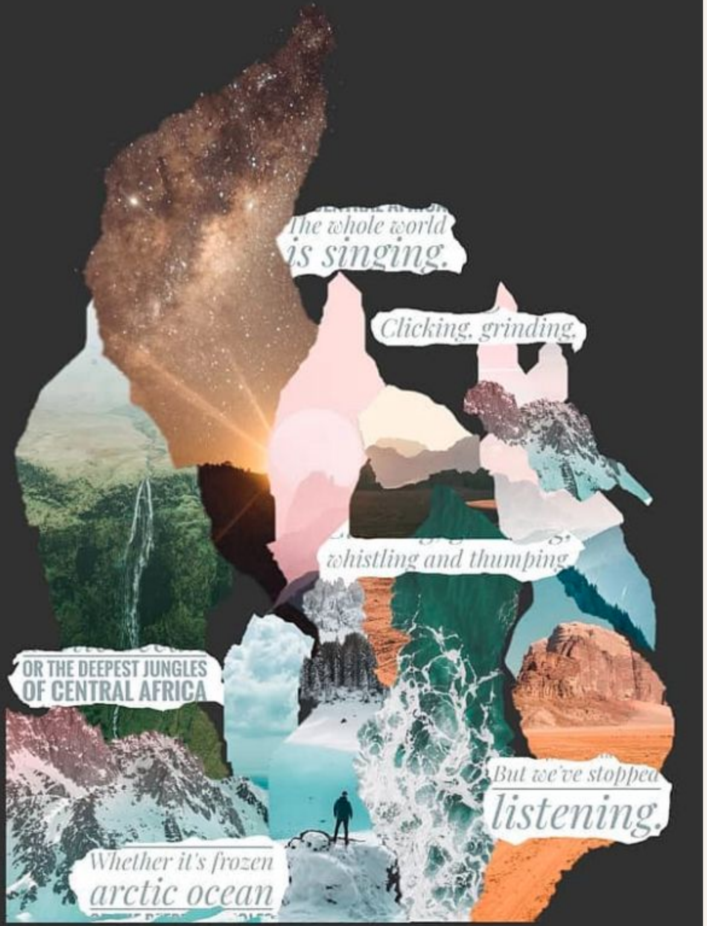
Normalized paper count by country at the 2018 NLP conferences (Caines, 2019)

30% of all living languages today are African languages

(Ethnologue)

I WANT TO DRINK IN ALL THE BEAUTY IN THE WORLD.
ESPECIALLY THE THINGS NOT SEEN.

Artwork by Keneilwe Mokoena

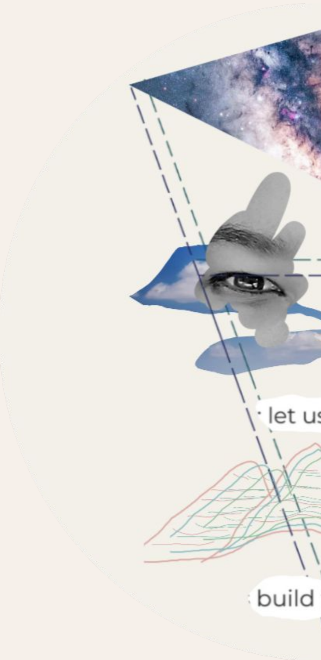


In all ACL conferences in 2018, only 0.1% of author affiliations were based in Africa

(Caines, 2019)

Artwork by Keneilwe Mokoena

Reflection



"The effect of a **cultural bomb** is to **annihilate a people's belief** in their names, in their languages, in their environment, in their heritage of struggle, in their unity, in their capacities and ultimately in themselves."

Ngũgĩ wa Thiong'o ~ Decolonising the Mind

Artwork by Keneilwe Mokoena

“Low-resourcedness”
is about more than
just data
- it’s societal

We Are

the effort for NLP for African languages that is

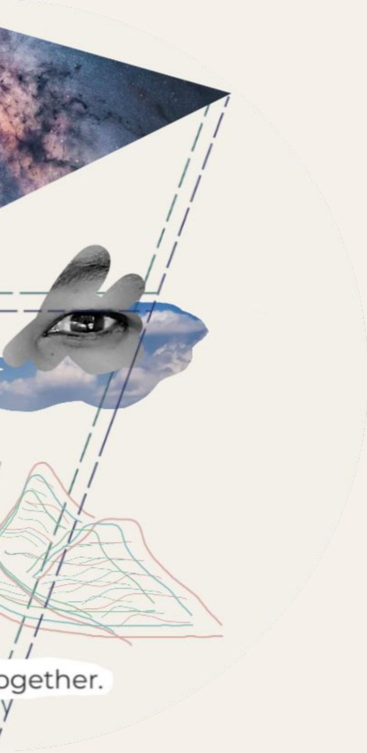
OPEN SOURCE

CONTINENT-WIDE

DISTRIBUTED

ONLINE

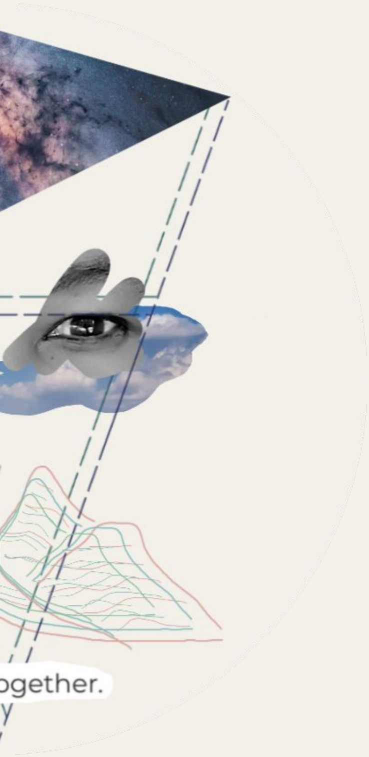
“Look Within” by Reggie Khumalo



Case Study:

Masakhane

#OurStory



Participatory Research
as a means to ensure
everyone who
should be in the room
is in the room

Goals

For Africa

To build a community of NLP researchers, connect and grow it, spurring and sharing further research, to enable language preservation, tool building and increase its global visibility and relevance

Artwork by Keneilwe Mokoena

Goals

For research communities

To discover best practices for distributed research, to be applied by other emerging research communities.

Artwork by Keneilwe Mokoena

#OurMethod

Artwork by Keneilwe Mokoena

1. How to find participants?

- *locally*: Indaba Deep Learning school, meetups, universities
- *remotely*: Twitter, conferences, press coverage, publications
- no academic prerequisites

2. How to assign roles? (agents)

- no fixed roles
- cross-functional collaboration
- adapt to interest and skills
- focus on knowledge transfer & sharing
  - from mentee to mentor

3. How to connect the distributed participants?

- *Open* digital communication channels
  - Active slack workspace & Google group
  - Weekly virtual meetings
  - Virtual reading group
- Making findings *accessible*
  - Public meeting notes
  - All results, models and code published for each submission
  - Data shared on GitHub
- Anyone can make calls for collaboration
  - *emerging* rather than imposed agendas

Artwork by Keneilwe Mokoena

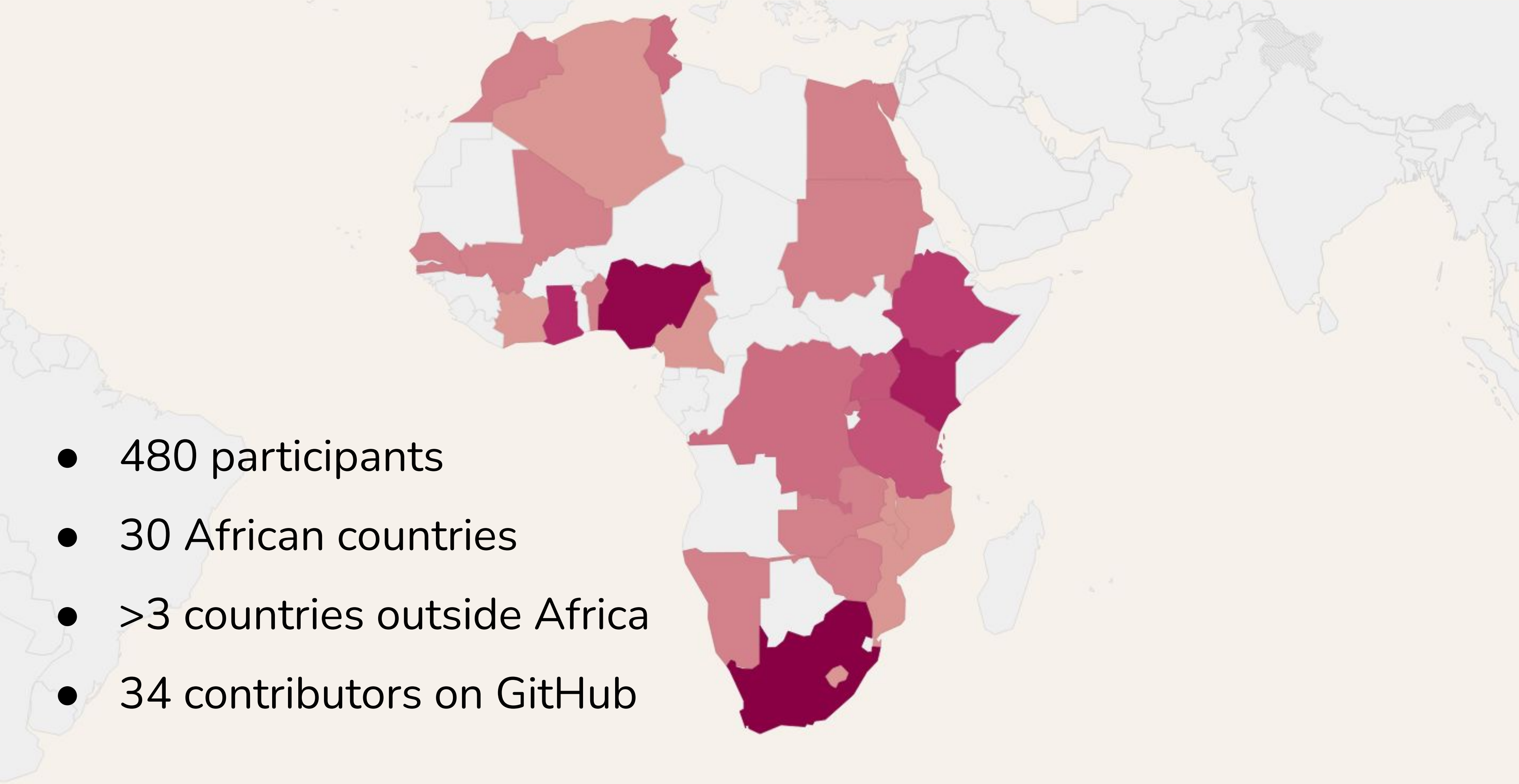
Artwork by Keneilwe Mokoena

4. How to lower participation barriers?

- a guided *hands-on* experience to get started
- making use of *freely* available resources
  - data: JW300 (Agić+, 2019) includes 101 African languages
  - code: JoeyNMT (Kreutzer+, 2019) 🐨
  - compute: Google Colab

  → beginner friendly Jupyter notebook to prepare data, train and evaluate MT models
- tutorials & guidelines

#OurCommunity



- 480 participants
- 30 African countries
- >3 countries outside Africa
- 34 contributors on GitHub

87% of participants found mentors of collaborators

Publications 2020 & 2021

∀, Masakhane:

- EMNLP Findings 2020
- AfricaNLP 2020
- AfricanNLP2021: MasakhaNER
- AFricanNLP2021: Quality at a Glance

13+ more papers:

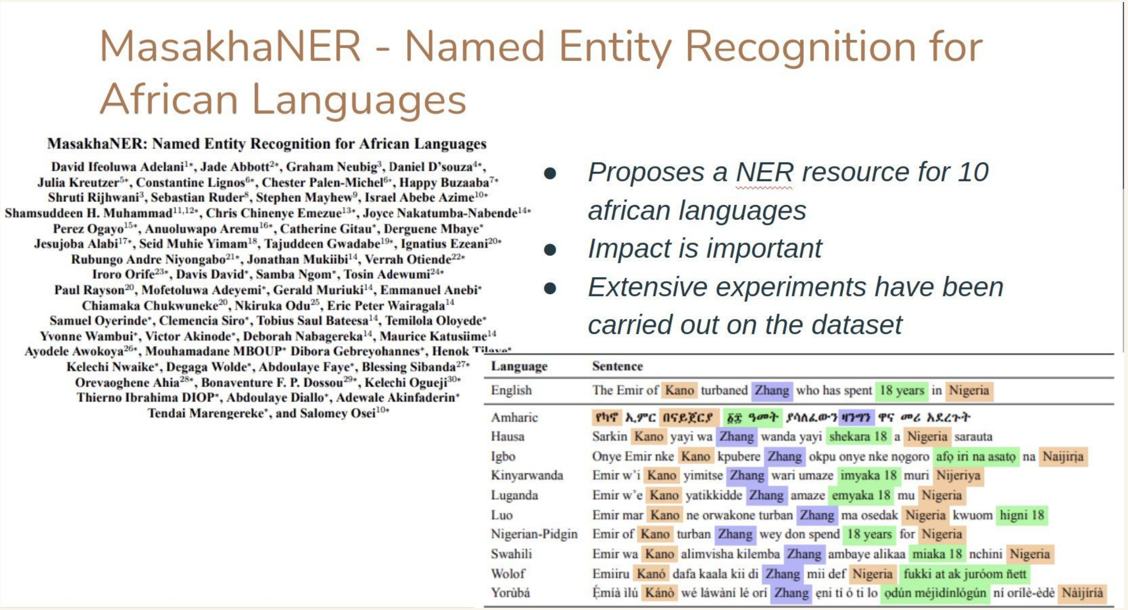
- NMT for Fon-French, Edoid languages, Nigerian Tigrinya, Hausa, South African languages
- Dataset creation for Setswana & Sepedi
- Transformer depth
- Law-Making in Kenya
- AI4D: Dataset challenges
- Text classification for Kinyarwanda & Kirundi
- Diacriticization for Yorùbá

Full list

Artwork by Keneilwe Mokoena

2021 Wikimedia Foundation Research Award of the Year

Best Paper Award
@
AfricaNLP 2021



MasakhaNER - Named Entity Recognition for African Languages

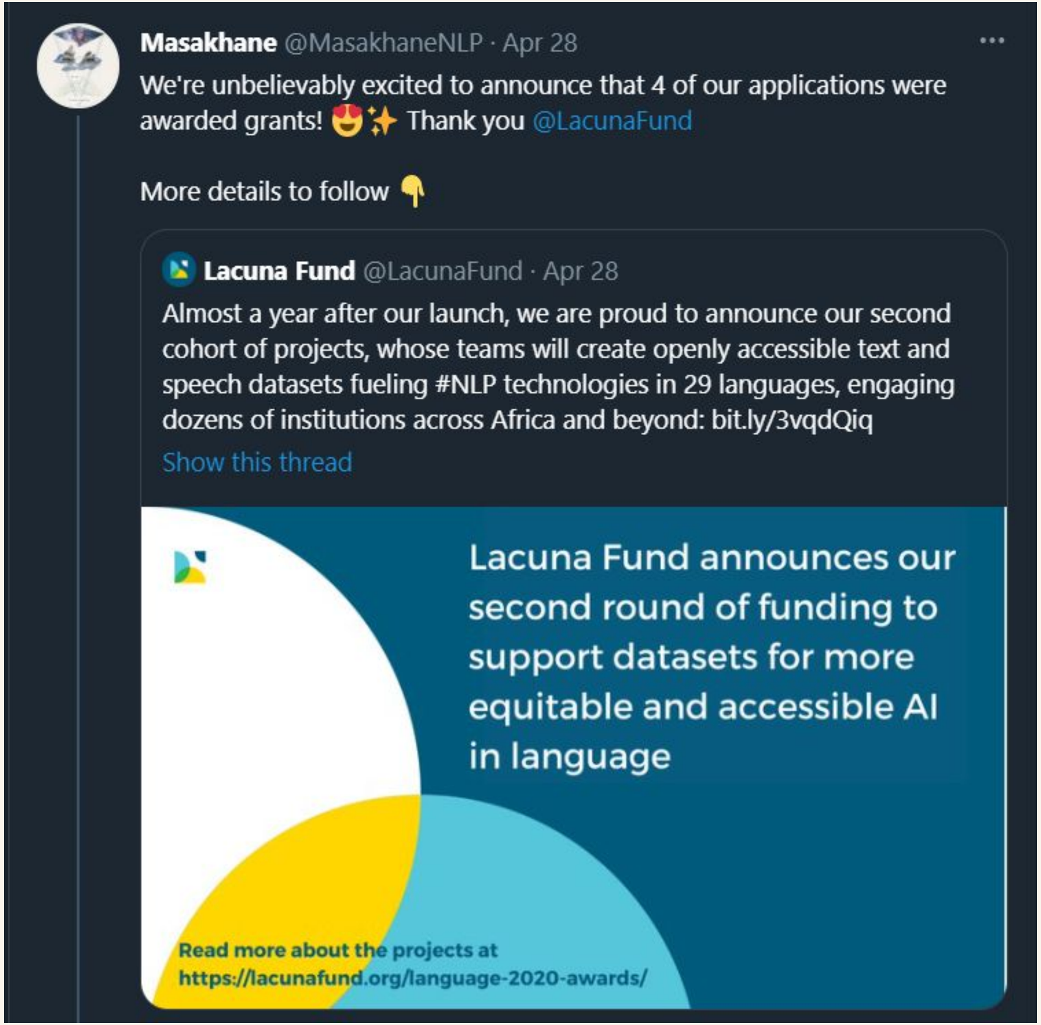
MasakhaNER: Named Entity Recognition for African Languages

**David Ifeoluwa Adelani[1*], Jade Abbott[2*], Graham Neubig[3], Daniel D'souza[4*],
Julia Kreutzer[5*], Constantine Lignos[6*], Chester Palen-Michel[6*], Happy Buzaaba[7*]
Shruti Rijhwani[3], Sebastian Ruder[8], Stephen Mayhew[9], Israel Abebe Azime[10*]
Shamsuddeen H. Muhammad[11,12*], Chris Chinenye Emezue[13*], Joyce Nakatumba-Nabende[14*]
Perez Ogayo[15*], Anuoluwapo Aremu[16*], Catherine Gitau[*], Derguene Mbaye[*]
Jesujoba Alabi[17*], Seid Muhie Yimam[18], Tajuddeen Gwadabe[19*], Ignatius Ezeani[20*]
Rubungo Andre Niyongabo[21*], Jonathan Mukiibi[14], Verrah Otiende[22*]
Iroro Orife[23*], Davis David[*], Samba Ngom[*], Tosin Adewumi[24*]
Paul Rayson[20], Mofetoluwa Adeyemi[*], Gerald Muriuki[14], Emmanuel Anebi[*]
Chiamaka Chukwuneke[20], Nkiruka Odu[25], Eric Peter Wairagala[14]
Samuel Oyerinde[*], Clemencia Siro[*], Tobius Saul Bateesa[14], Temilola Oloyede[*]
Yvonne Wambui[*], Victor Akinode[*], Deborah Nabagereka[14], Maurice Katusiime[14]
Ayodele Awokoya[26*], Mouhamadane MBOUP[*] Dibora Gebreyohannes[*], Henok Tilaye[*]
Kelechi Nwaike[*], Degaga Wolde[*], Abdoulaye Faye[*], Blessing Sibanda[27*]
Orevaoghene Ahia[28*], Bonaventure F. P. Dossou[29*], Kelechi Ogueji[30*]
Thierno Ibrahima DIOP[*], Abdoulaye Diallo[*], Adewale Akinfaderin[*]
Tendai Marengereke[*], and Salomey Osei[10*]**

- *Proposes a NER resource for 10 african languages*
- *Impact is important*
- *Extensive experiments have been carried out on the dataset*

Language	**Sentence**
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጀርያ ፩፰ ዓመት ያሳለፈውን ዛንግን ዋና መሪ አደረጉት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Nigeria sarauta
Igbo	Onye Emir nke Kano kpubere Zhang okpu onye nke ngoro afọ iri na asatọ na Naijiria
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Luganda	Emir w'e Kano yatikkidde Zhang amaze emyaka 18 mu Nigeria
Luo	Emir mar Kano ne orwakone turban Zhang ma osedak Nigeria kwuom higni 18
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Ẹ́míà ìlú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọdún méjìdínlógún ní orílè-èdè Nàìjíríà

LACUNA FUND
Grant Recipients



Masakhane @MasakhaneNLP · Apr 28

We're unbelievably excited to announce that 4 of our applications were awarded grants! 😍✨ Thank you @LacunaFund

More details to follow 👇

Lacuna Fund @LacunaFund · Apr 28

Almost a year after our launch, we are proud to announce our second cohort of projects, whose teams will create openly accessible text and speech datasets fueling #NLP technologies in 29 languages, engaging dozens of institutions across Africa and beyond: bit.ly/3vqdQiq

Show this thread

Lacuna Fund announces our second round of funding to support datasets for more equitable and accessible AI in language

Read more about the projects at https://lacunafund.org/language-2020-awards/

#OurData

❤️

When data curators originate from the
places the languages are spoken

Coffee & Translate Sessions for Namibia’s KhoekhoeGowab Languages

https://github.com/masakhane-io/masakhane-khoekhoegowab

Community Driven Creation of Nigerian Language Datasets

https://github.com/masakhane-io/masakhane-wazobia-dataset

Data Curation for African NER Datasets

https://github.com/masakhane-io/masakhane-ner

Opening up existing but unpublished public data from South African Parliament for 11 official languages

https://bit.ly/raw-parliamentary-translations

Data Collection for Fon Language

https://github.com/bonaventuredossou/ffr-v1

#OurExperiments

Our NMT results

- **47+ translation models for 35 African languages** have been published
- Evaluation: post-edits for 9 languages
  - Training on JW300
  - Post-editing for COVID Surveys and TED talks
- **MasakhaneWEB**

Full list

Artwork by Keneilwe Mokoena

Artwork by Keneilwe Mokoena

Ongoing Initiatives

- Machine Translation models and evaluation for more and more languages
- Data Gathering
- Named Entity Recognition
- Building language models for African languages
- Building tools for people to try out our trained models
- Weekly NLP reading group

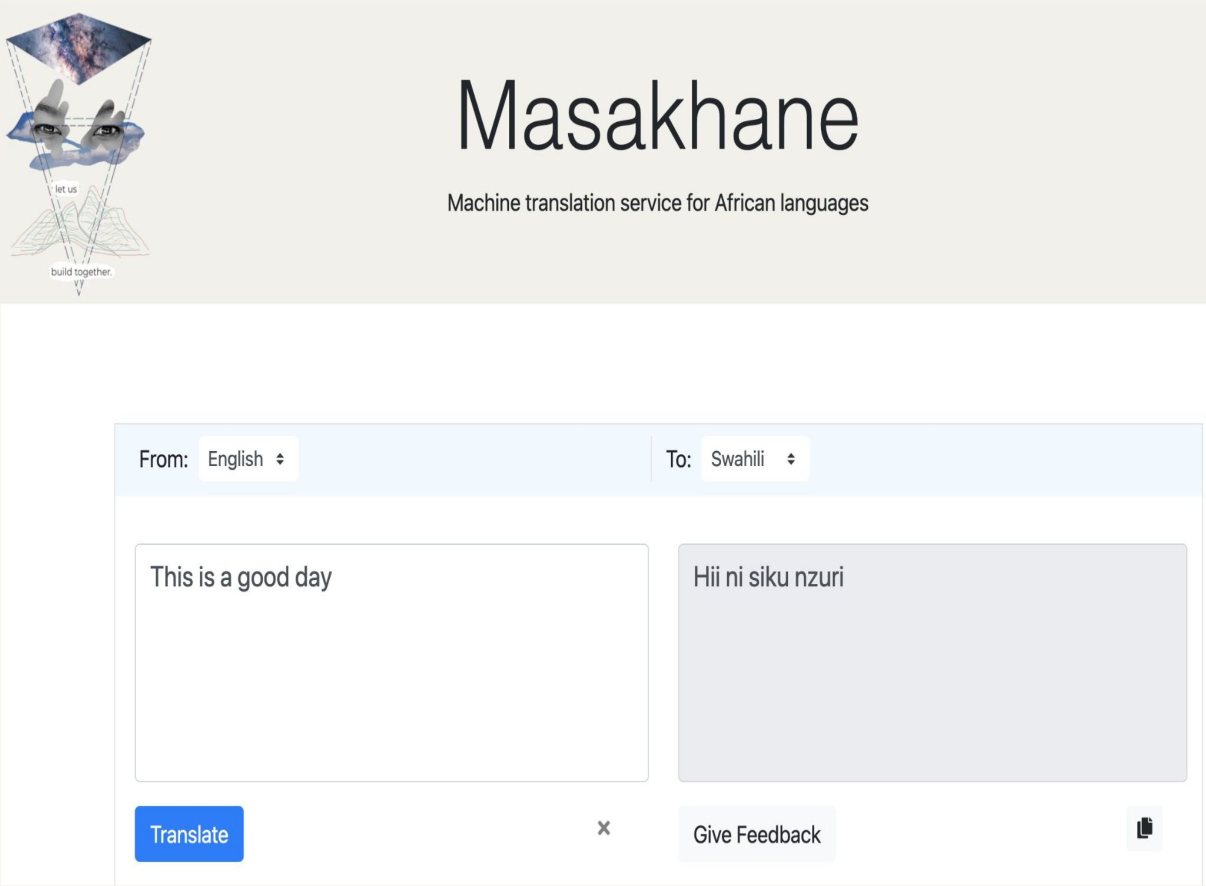
Challenges

- Domain Gap
- **Dialectical** differences impact performance
- Diacritics
- Evaluation
- Ineffective **keyboards**

Artwork by Keneilwe Mokoena

#Tools

MasakhaneWEB



Masakhane

Machine translation service for African languages

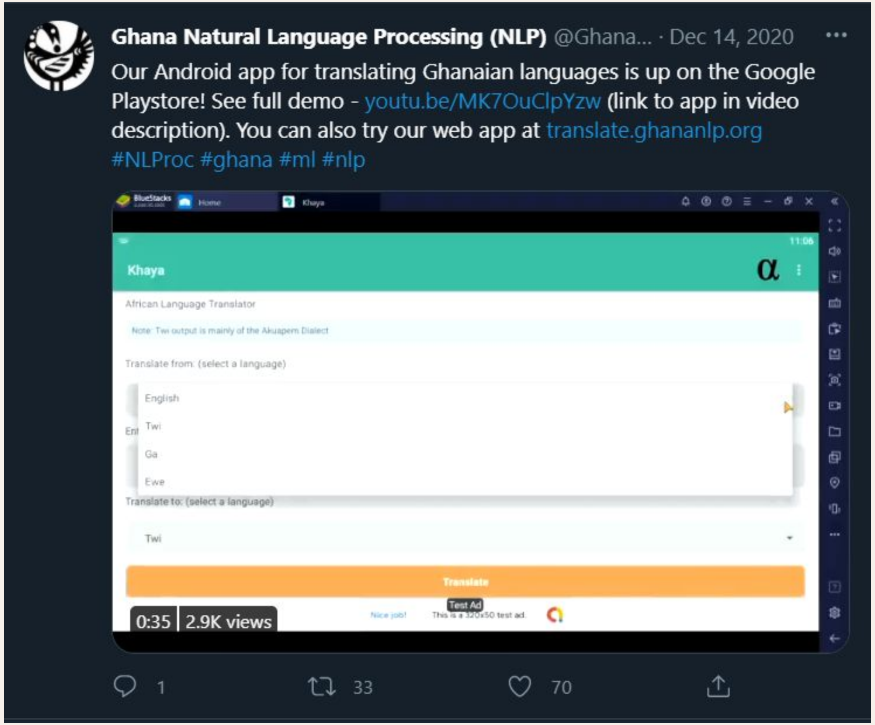
From: English

To: Swahili

This is a good day

Hii ni siku nzuri

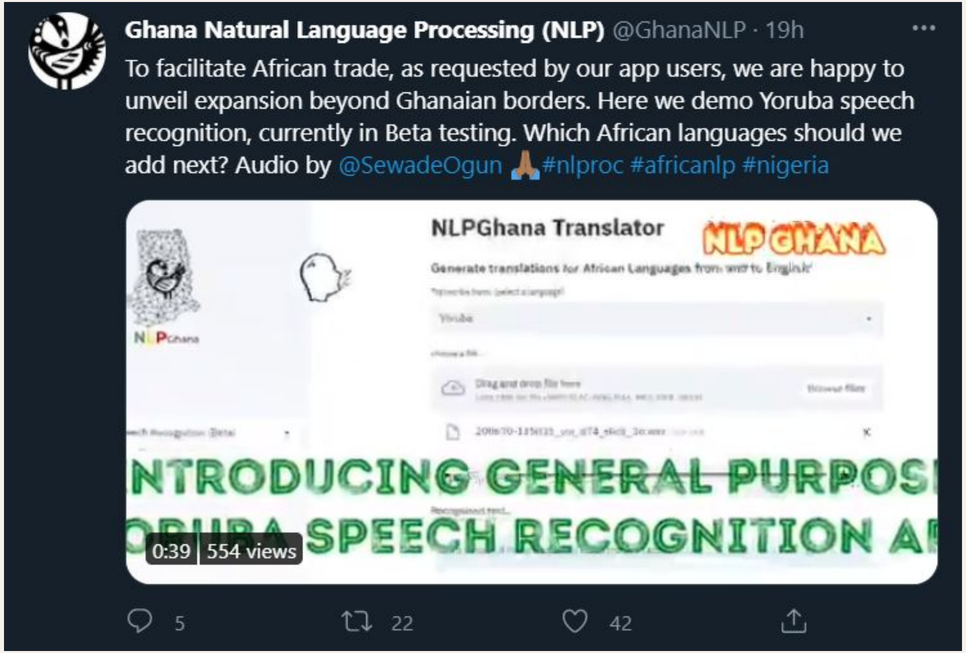
GhanaNLP



Ghana Natural Language Processing (NLP) @Ghana... · Dec 14, 2020

Our Android app for translating Ghanaian languages is up on the Google Playstore! See full demo - youtu.be/MK7OuClpYzw (link to app in video description). You can also try our web app at translate.ghananlp.org #NLProc #ghana #ml #nlp

BlueStacks
Home
Khaya
11:06
Khaya
α
African Language Translator
Note: Twi output is mainly of the Akuapem Dialect
Translate from: (select a language)
English
Twi
Ga
Ewe
Translate to: (select a language)
Twi
Translate
Test Ad
Nice job!
This is a 320x50 test ad.
0:35 | 2.9K views
1
33
70

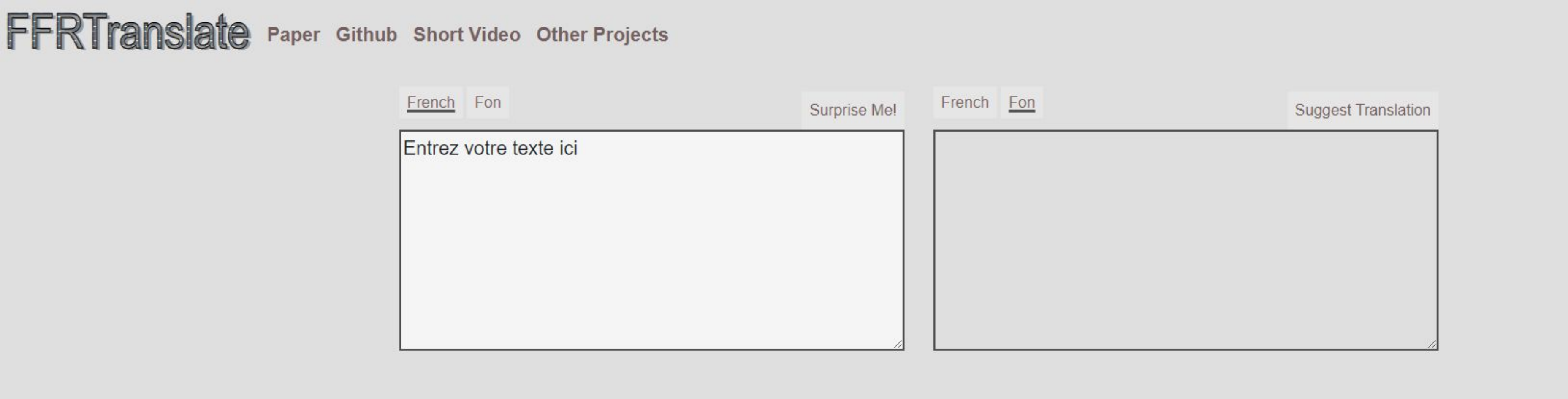


Ghana Natural Language Processing (NLP) @GhanaNLP · 19h

To facilitate African trade, as requested by our app users, we are happy to unveil expansion beyond Ghanaian borders. Here we demo Yoruba speech recognition, currently in Beta testing. Which African languages should we add next? Audio by @SewadeOgun 🙏🏾#nlproc #africanlp #nigeria

NLPGhana Translator
NLP GHANA
Generate translations for African Languages from and to English
Yoruba
Drag and drop file here
NLPGhana
NTRODUCING GENERAL PURPOSI
ORUBA SPEECH RECOGNITION AI
0:39 | 554 views
5
22
42

FFRTranslate



Entrez votre texte ici

#JoinUs

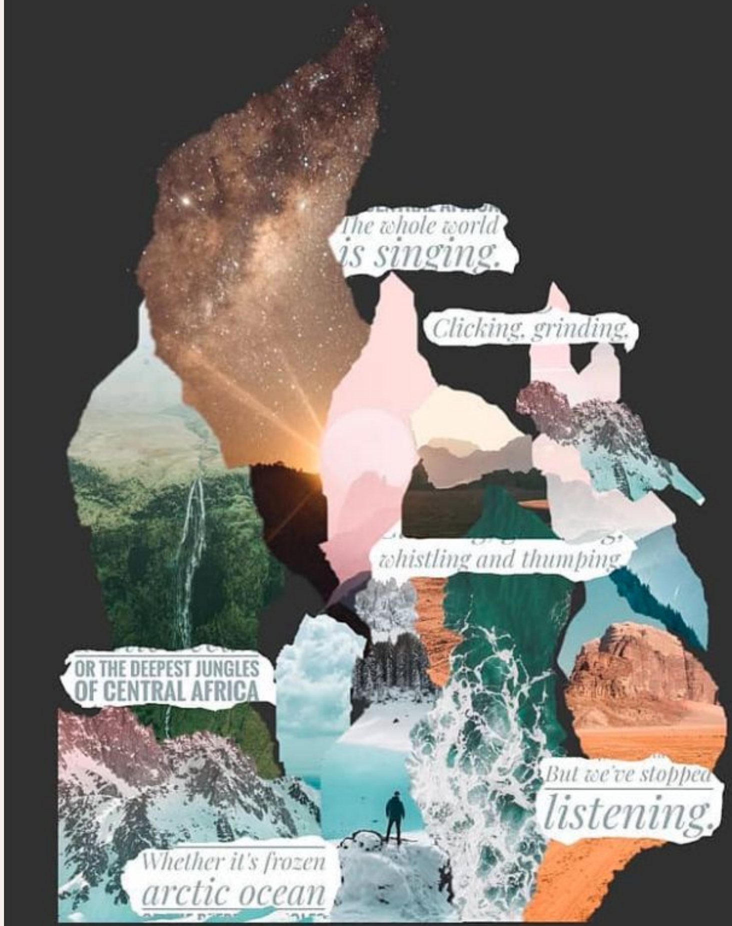
It's on us (including you) to color this map for 202X!

- **TRAIN A MODEL** - Contribute a trained model and related code for your language
- **ANALYSIS** - Contribute analysis of data/models for any African language(s).
- **DATA** - Help build or find datasets for your language
- **DOCUMENTATION** - Help document our discussions, progress.
- **MENTORSHIP** - Provide advice or help tune models for their languages and datasets, or help people get started
- **COLLABORATION** - On papers, workshops, data collection, ...

- **ADMIN-** Help out with administrative tasks and coordinate research outputs
- **COMPUTE** - Help with infrastructure and compute! Do you have spare compute to donate?
- **BRAINSTORM** - Join our weekly meetings, provide advice or ideas
- **STORY-TELLING** - Tell our stories to the world by doing talks about the community, contributing to our Medium publication, or engaging with media outlets
- **MLOps & ML Engineering-** If you are a software developer, ML engineer or support reproducibility, data gathering, and model sharing

Road Ahead | #OurFuture

- Community growth
- Expand to more NLP tasks
- Qualitative analysis of models
- Evaluation of metrics
- Notebooks for transfer, self-supervised learning and data augmentation
- Expand size & domain of global test sets
- Reproducibility & comparability



Artwork by Keneilwe Mokoena

Artwork by Keneilwe Mokoena
Her email: keneilwe.bk@gmail.com
Instagram: @ke_neil_we

Masakhane
“We build together”

masakhane.io

Twitter: @MasakhaneNLP

https://github.com/masakhane-io/community-guidelines

https://bit.ly/masakhane-starting-together

https://bit.ly/masakhane-calendar

https://bit.ly/masakhane-mt-how-to-start