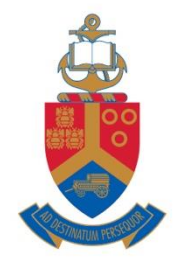




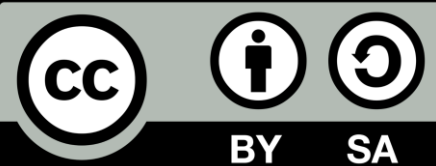
AI + Language+ Education *Opportunities for ALL*

Prof Vukosi Marivate (ABSA UP Chair of
Data Science)
vukosi.marivate@cs.up.ac.za
Feb 2025



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

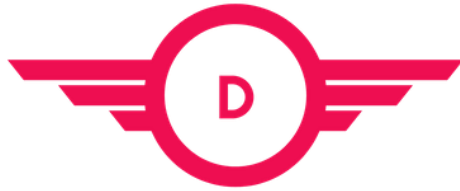
An AI driven Africa - Dalle 3



Make today matter

www.up.ac.za

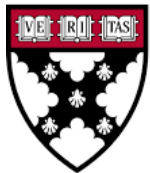
Research + Community Organiser



Data Science for Social Impact



DEEP LEARNING
INDABA



Harvard
Business
School

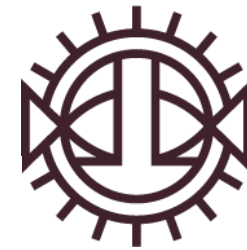


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

 **Lacuna Fund**
Our voice on data



THE FORUM OF
YOUNG GLOBAL LEADERS
SHAPING THE FUTURE



LELAPA AI



Masakhane

A grassroots NLP community for Africa, by Africans



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

All Members

Faculty

Post Docs

PhD

Masters

Honours

Research Assistants



Vukosi Marivate
Principal Investigator
and Leader



Herkulaas Combrink
Senior Member,
Director/Lecturer at
UFS



Abiodun Modupe
Senior Member



Seani Rananga
Senior Member, PhD
CS, Lecturer



Idris Abdulmumin
Postdoctoral Fellow



Kayode Olaleye
Postdoctoral Fellow



Dawit Bogale
PhD CS



Temitope Kekere
PhD IT



**Baphumelele
Masikisiki**
PhD



Mpho Mokoatle
PhD CS



Michelle Terblanche
PhD CS



Nombuyiselo Zondi
PhD education



Abiola Akinbowale
CS MIT Big Data
Science



Tsholofelo Gomba
CS MSc



Pitso Khoboko
CS MSc



Miehleketo Mathebula
CS MSc



Rozina Myoya
CS MIT Big Data
Science



Mohlatlego Nakeng
CS MIT Big Data
Science



Kwanele Radebe
CS MSc



Matimba Shingange
CS MIT Big Data
Science



Kathleen Siminyu
CS MSc



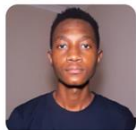
**Thapelo Andrew
Sindane**
CS MSc



David Walker
CS MSc



Masana Glad Baloyi
CS Honours



**Isheanesu Joseph
Dzingirai**
CS Honours



Arno Jooste
CS Honours



Reuben Jooste
CS Honours



Mokoena Lehlohonolo
CS Honours



Christiaan Lombard
CS Honours, Research
Assistant



Keabetswe Madumo
CS Honours, Research
Assistant



Manyako Manaswe
CS Honours



**Mulweli Patience
Mukwevho**
CS Honours



Makungu Ndlovu
CS Honours



Mosa Sethosa
CS Honours



Lesiba Setsiba
CS Honours

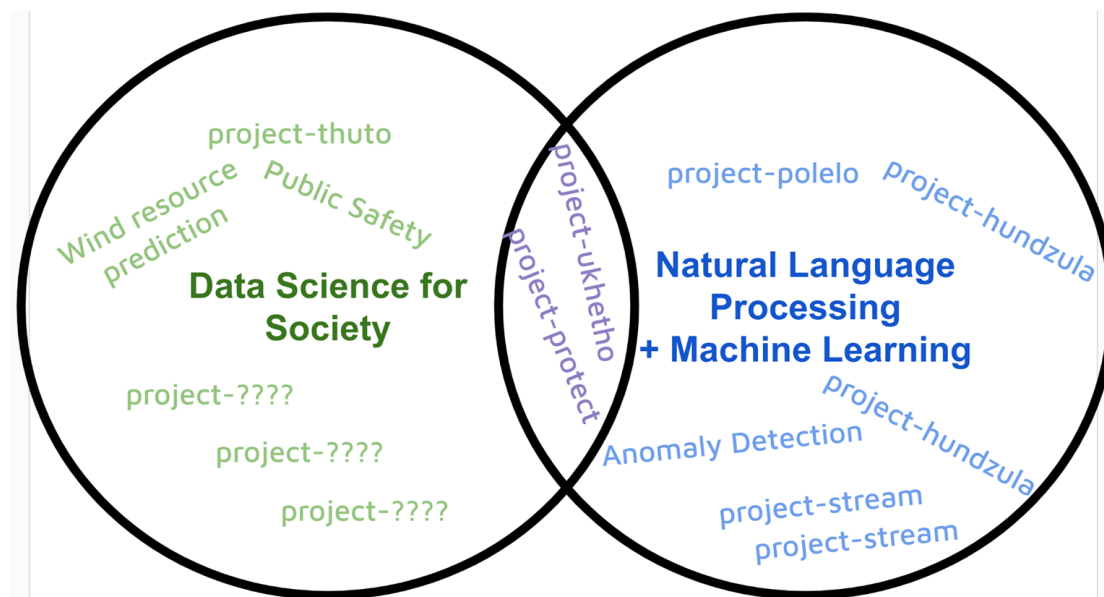


Richard Lastrucci
Research Assistant



Data Science for Social Impact

+ Others not pictured
+ Alumni



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Let's Get to the End



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

AI for Mother Tongue Education: Unlocking Potential

Personalized Learning in Local Languages

- AI-driven tutoring systems adapt content to individual student needs.
- Speech-to-text and text-to-speech tools enable accessibility.

Content Creation & Translation

- Machine translation enhances availability of learning materials.
- Large-scale corpus development supports textbook creation in African languages.

Teacher Support & Engagement

- AI-powered tools simplify content generation and assessment.
- Automated grading & feedback in mother tongue languages.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

AI for Mother Tongue Education: Unlocking Potential

Preserving Linguistic Diversity

- AI helps document, digitize, and revitalize endangered languages.
- NLP models can foster literacy & language learning from early education stages.

Bridging the Digital Divide with AI-driven Language Inclusion!



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Basic Education Opportunities for AI

Data Collection & Corpus Development

- **Open** Early-grade reading materials can help train language models for foundational understanding.
- Standardized tests and assignments in local languages can provide structured parallel text data.
- Schools can contribute to subject-specific terminology (math, science, etc.) in African languages.
- Collaboration with educators can help refine NLP-based spelling and grammar correction.

Government & Institutional Support

- Making public education materials openly available can assist research in general and NLP research.
- Partnerships between schools, universities, and tech firms can drive AI innovation.
- Empowering educators to integrate AI-driven tools in language teaching.

✍ By integrating NLP into basic education, we can not only improve AI models but also strengthen linguistic diversity and digital literacy in African languages.



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Why African Low Resource Languages Now?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matt8

www.up.ac.za

Low Resource Natural Language Processing - Languages

Many Speakers, Not Many Resources

Data

- Speech
- Text
- Etc.

Tools for language (digital dictionaries, grammar tools etc.)



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Defining the challenges to Low Resource Languages

- Low availability of resources (Data, Tools, etc.)
- Discoverability 🔍
- Reproducibility
- Focus
- Benchmarks ⚙️
- Scale and Complexity

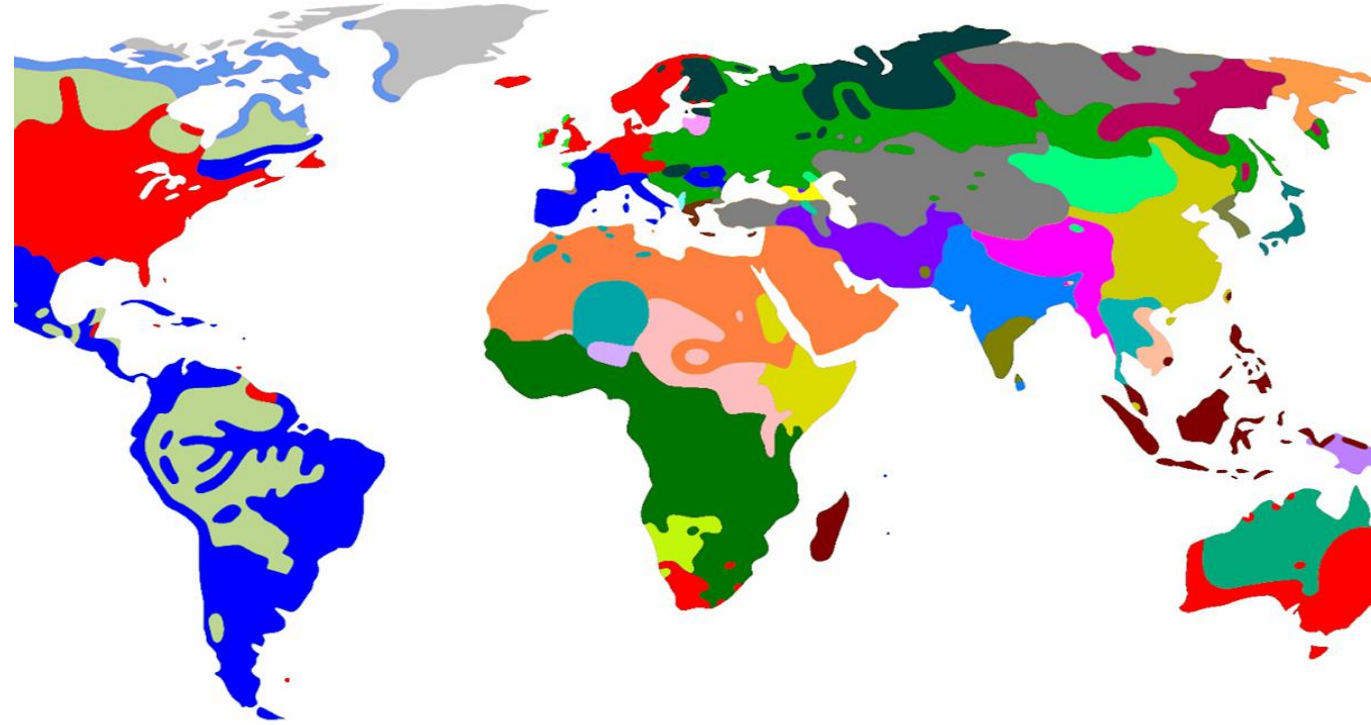
A Focus on Neural Machine Translation for African Languages

Laura Martinus

Explore / Johannesburg, South Africa Retro Rabbit / Johannesburg, South Africa
laura@explore-ai.net ja@rettorabbit.co.za

Jade Abbott

A Focus on Neural Machine Translation for African Languages



*Repartition map of the languages over the world (version blank of key)
[Wikimedia:User:Industrius]*



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Defining the challenges to Low Resource NLP

0. The Left-Behinds

1. The Scraping-Bys

2. The Hopefuls

3. The Rising Stars

4. The Underdogs

5. The Winners

Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

Table 1: Number of languages, number of speakers, and percentage of total languages for each language class.

The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Pratik Joshi* Sebastin Santy* Amar Budhiraja*

Kalika Bali Monojit Choudhury

Microsoft Research, India

{t-prjos, t-sesan, amar.budhiraja, kalikab, monojitc}@microsoft.com



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

The State and Fate of Linguistic Diversity and Inclusion in the NLP World

Make today matter

www.up.ac.za

AI current futures: Insatiable Appetite for Data and Compute



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

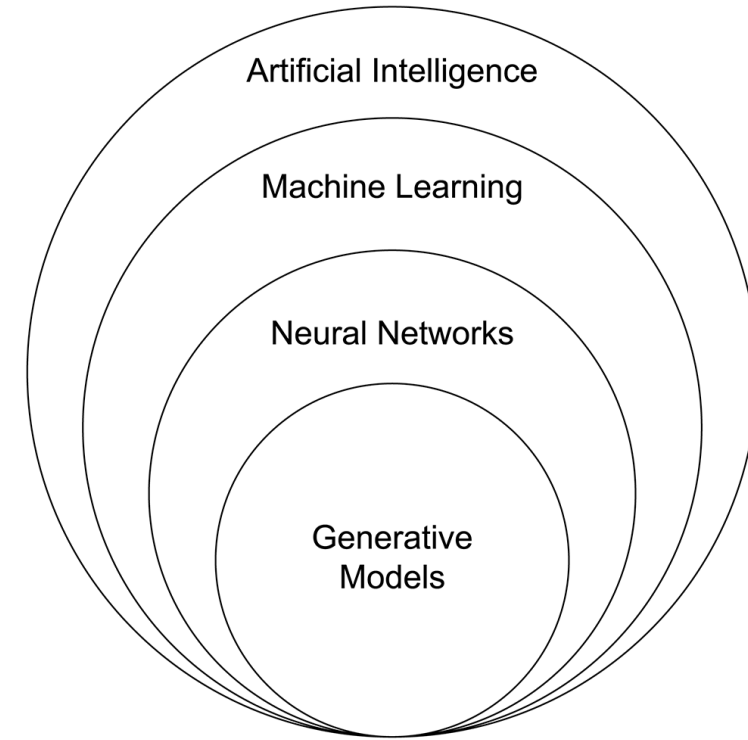
What is old is new again

This is not the first time we are going through an AI/ML hype cycle.

We are now in the age of Generative AI

It won't be the last.

How do we learn from the past to also get a grip on the future?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

There will be change

Many ideas that will be floating around.

Most will wither away, some will float to the top.

Predicting the Future of AI with AI: High-Quality link prediction in an exponentially growing knowledge network

Mario Krenn,^{1, *} Lorenzo Buffoni,² Bruno Coutinho,² Sagi Eppel,³ Jacob Gates Foster,⁴ Andrew Gritsevskiy,^{3, 5, 6} Harlin Lee,⁴ Yichao Lu,⁷ João P. Moutinho,² Nima Sanjabi,⁸ Rishi Sonthalia,⁴ Ngoc Mai Tran,⁹ Francisco Valente,¹⁰ Yangxinyu Xie,¹¹ Rose Yu,¹² and Michael Kopp⁶

¹Max Planck Institute for the Science of Light (MPL), Erlangen, Germany.

²Instituto de Telecomunicações, Lisbon, Portugal.

³University of Toronto, Canada.

⁴University of California Los Angeles, USA.

⁵Cavendish Laboratories, Cavendish, Vermont, USA.

⁶Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria.

⁷Layer 6 AI, Toronto, Canada.

⁸Independent Researcher, Barcelona, Spain.

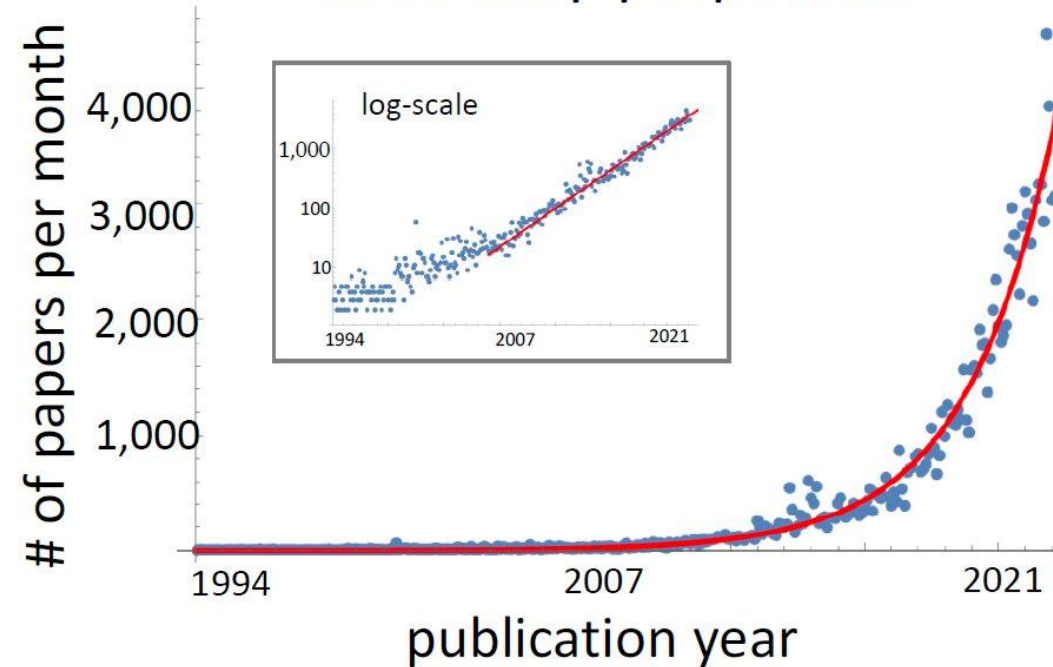
⁹University of Texas at Austin, USA.

¹⁰Independent Researcher, Leiria, Portugal.

¹¹University of Pennsylvania, USA.

¹²University of California, San Diego, USA.

ML+AI arXiv papers per month



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

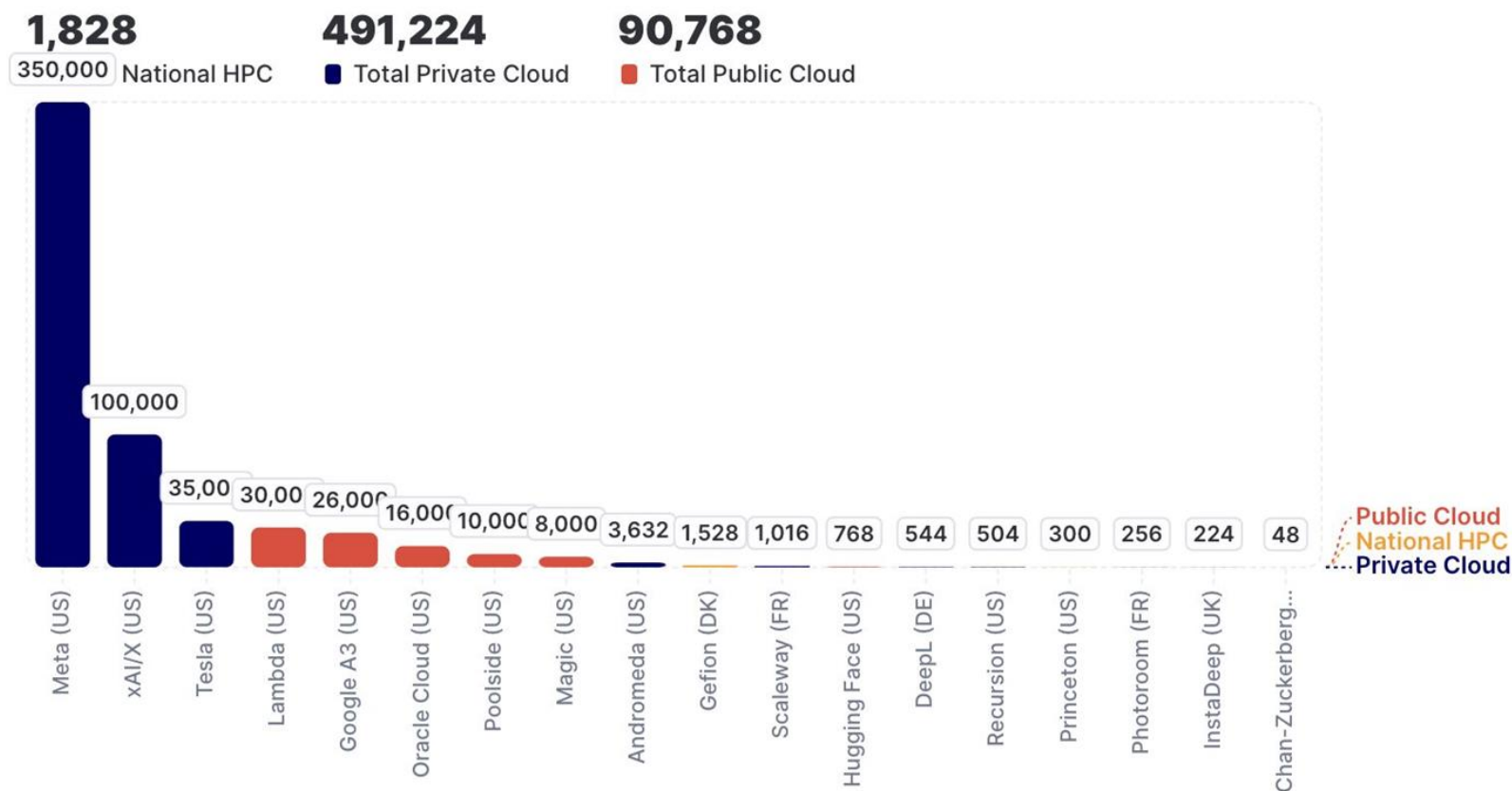
www.up.ac.za

Compute is all you need?

<https://www.stateof.ai/compute>

H100 GPU Count

Last update: November 2024



Notes: Public Cloud = capacity rented from hyperscalers; Private Cloud = owned and run by the company; National HPC = government owned and run. Meta's H100 figure represents its H100 purchase as per company earnings call 1 Feb 2024. Excludes a further 250,000 H100 equivalents worth of GPU.

Effectiveness of Data, At What Cost?



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

One of the largest gaps in the availability of data is in language

‘Brute force’ alone is unlikely to achieve parity in data generation in African languages compared to Anglophone, given the extent of the data imbalance.

	Language (Nov '23)	% Share of Internet content in local language
Global Benchmarks	English	52.60%
	Hindi	0.1%
Top African Languages	Afrikaans	0.003%
	Twi	0.00195%
	Swahili	0.00135%
	Malagasy	0.00022%
	Bambara	0.00025%
	Venda	0.000115%
	Hausa	0.00011%
Average with other African languages*		0.000999%
Sum of African Languages		0.01999%

English
53%

African
0.02%

2,650 x more content in English than African Languages

Lacuna Fund and Masakhane are key interventions, but only scratch the surface

* Afrikaans, Twi, Swahili, Bambara, Malagasy, Hausa, Venda, Haitian, Haitian Creole, Igbo, Luba-Katanga, Ndonga, Rundi, Tokelau, Tswana, Akan, Chichewa, Chewa, Nyanja, Fulah, Ganda, Masai

Data, Data, Data and More Data

The Unreasonable Effectiveness of Data

Publisher: IEEE

[Cite This](#)

[PDF](#)

Alon Halevy ; Peter Norvig ; Fernando Pereira [All Authors](#)

946

Cites in
Papers

24

Cites in
Patents

12111

Full
Text Views

Where is the data coming from to train all these models?

Does it represent us? Bias!!!

How do make sure we protect citizens private information?

Africa cannot just be a [CHEAP] data market and AI consumer!!!!

Connecting back to our languages



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

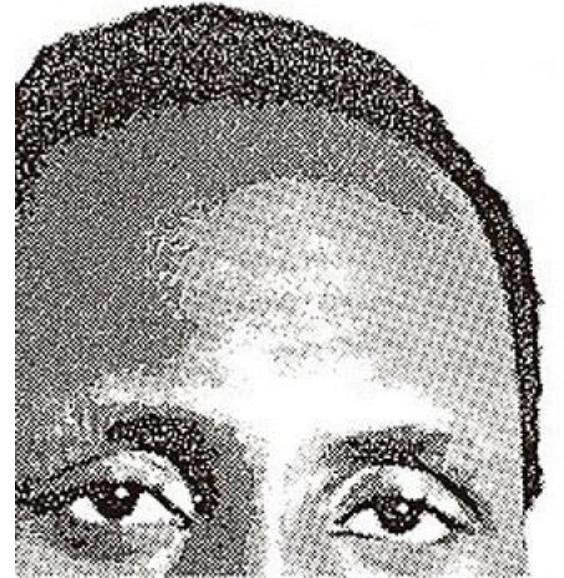
Make today matter

www.up.ac.za

Language is more than symbols

Language as communication and as culture are then products of each other. **Communication creates culture: culture is a means of communication. Language carries culture, and culture carries, particularly through orature and literature, the entire body of values by which we come to perceive ourselves and our place in the world.** How people perceive themselves and affects how they look at their culture, at their places politics and at the social production of wealth, at their entire relationship to nature and to other beings. Language is thus inseparable from ourselves as a community of human beings with a specific form and character, a specific history, a specific relationship to the world —
Decolonising the Mind (16)

NGŪGĨ
WA THIONG'O
Decolonising the Mind
THE POLITICS OF LANGUAGE
IN AFRICAN LITERATURE



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

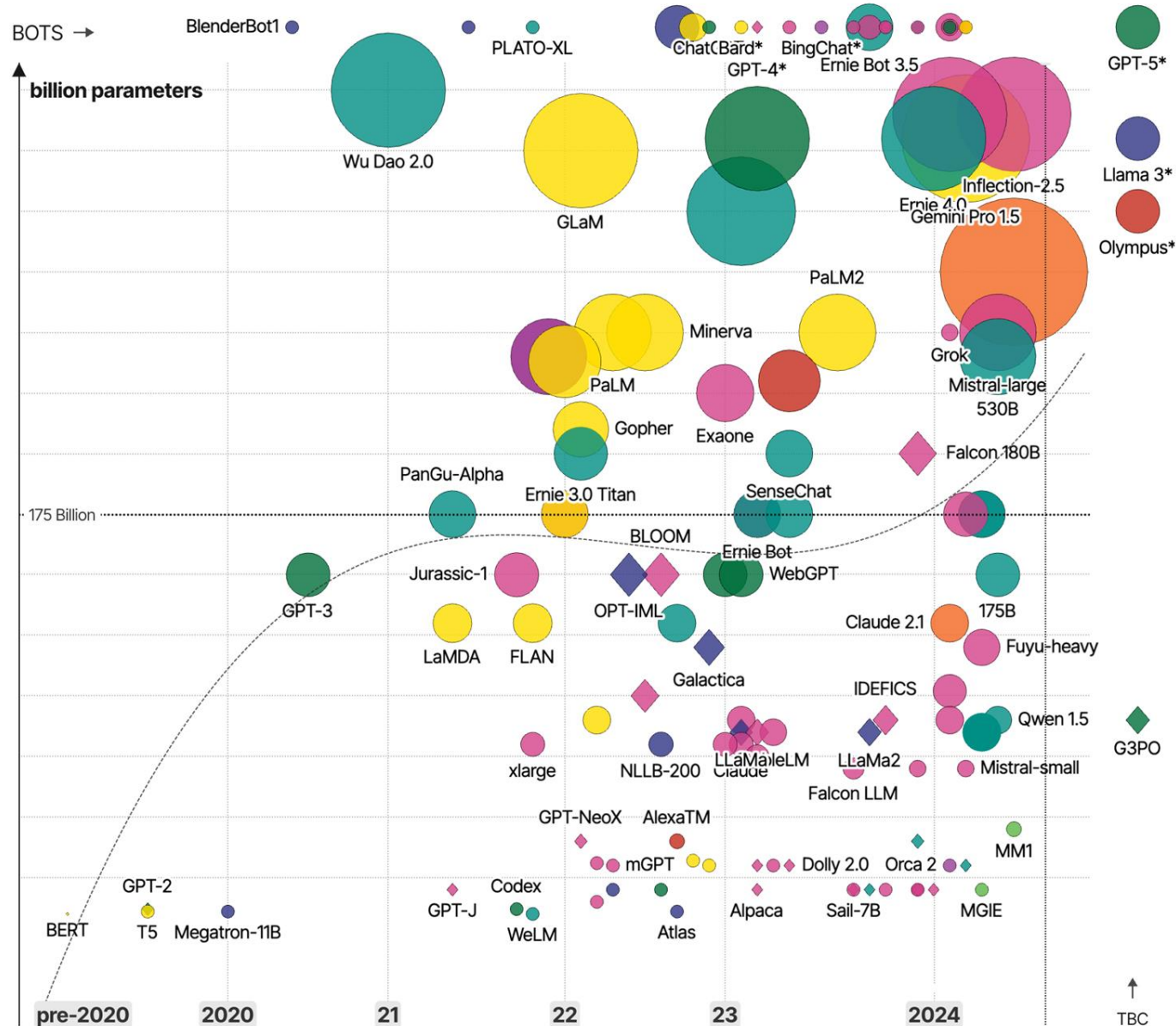
Make today matter

www.up.ac.za

The Rise and Rise of A.I. Large Language Models (LLMs) & their associated bots like ChatGPT

size = no. of parameters ◇ open-access

● Amazon-owned ● Anthropic ● Apple ● Chinese ● Google ● Meta / Facebook ● Microsoft ● OpenAI ● Other



Source <https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/>

All Wikipedias ordered by number of articles

The languages listed here are Wikipedias that have been created as also be sorted by other columns. It excludes [closed Wikipedias](#). Hist (or, after 15 October 2022, in the [history of the current source of the](#)

These statistics are updated four times a day. See [commons:Da](#)l doesn't update the numbers to what is seen at Commons, try [pu](#)

№	Language	Language (local)	Wiki	Articles
1	English	English	en	6,815,681
2	Cebuano	Cebuano	ceb	6,119,222
3	German	Deutsch	de	2,903,386
4	French	français	fr	2,607,022
5	Swedish	svenska	sv	2,583,112
6	Dutch	Nederlands	nl	2,156,687
7	Russian	русский	ru	1,975,846
8	Spanish	español	es	1,947,807
9	Italian	italiano	it	1,860,435
10	Egyptian Arabic	مصرى	arz	1,623,024
82	Swahili	Kiswahili	sw	80,020
110	Yoruba	Yorùbá	yo	33,971
162	Zulu	isiZulu	zu	11,311
300	Southern Sotho	Sesotho	st	949

Preparing the Vuk'uzenzele and ZA-gov- multilingual South African multilingual corpora



**Richard Lastrucci¹, Isheanesu Dzingirai¹, Jenalea Rajab², Andani Madodonga¹,
Matimba Shingange¹, Daniel Njini¹, Vukosi Marivate^{1,3}**

¹Department of Computer Science, University of Pretoria

²School of Computer Science and Applied Mathematics, University of the Witwatersrand

³Lelapa AI

richard.lastrucci@tuks.co.za, ishe.dzingirai@gmail.com, jenalea.rajab@gmail.com,
andanim412@gmail.com, mrosslyns@gmail.com, vukosi.marivate@cs.up.ac.za



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Setting the scene

Low Resource Language data is available, but may ***not be easily accessible*** for use in Natural Language Processing

Providing as much ***metadata*** as possible that identifies the source of the data, who might have written it and other translation information.

Some languages like ***isiNdebele*** has few resources and can be enhanced.

We build upon prior work such as that on ***Autshumato*** (Groenewald and Fourie, 2009; Groenewald and du Plooy, 2010)

- *Hendrik J Groenewald and Liza du Plooy. 2010. Processing parallel text corpora for three south african language pairs in the autshumato project.*
- *Hendrik Johannes Groenewald and Wildrich Fourie. 2009. Introducing the autshumato integrated translation environment.*



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Goals of the Project

Liberate and prepare textual multilingual data from Government Communication Information System [GCIS] of the South African government.

Extract and align sentence pairs for all South African languages within the data.

Automate as much of the process as possible.

Provide ***benchmarks for Machine Translation*** using this data.

Release the cleaned and machine readable data openly for other researchers to use.

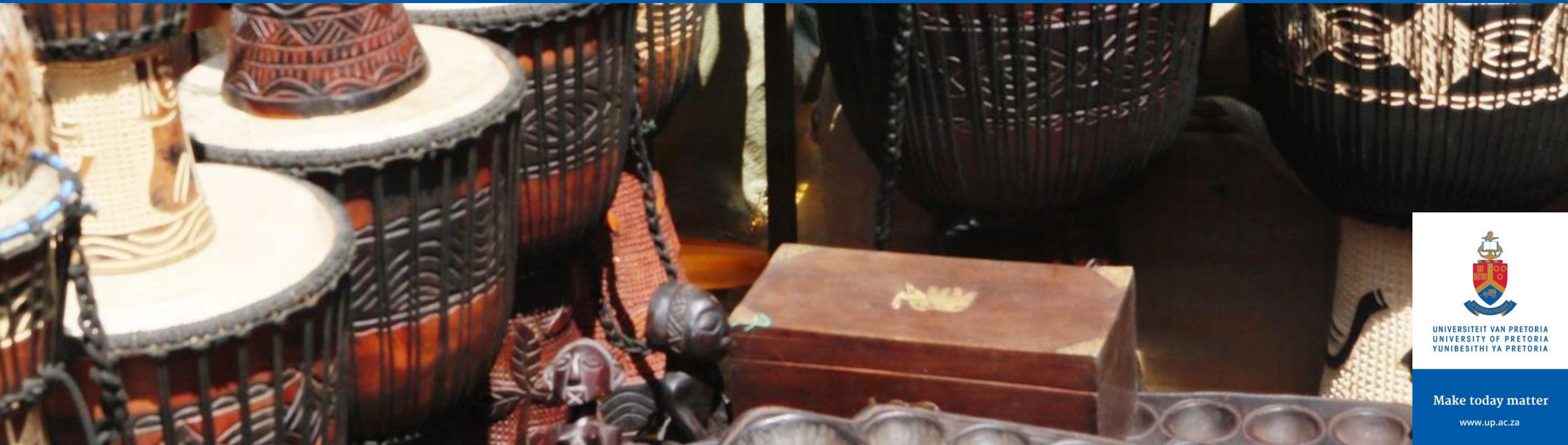


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Creating the datasets



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Cabinet Statements

What is this source?

- Statement of Cabinet Meetings of the South African government.
- A few weeks after the English statement is published, **translated statements are made available.**

Need

- **Extract statements** and make them easily accessible for NLP.
- Automate the process with **monitoring of website changes** and grow the dataset.



South African Government
www.gov.za

HOME ABOUT NEWSROOM SERVICES DOCUMENTS JOBS

Home » Newsroom » Cabinet Statements

Xitatimente hi mayelana na Nhlengeletano ya Khabinete ya le moyeni Ravunharhu, 31 Mhawuri 2022

1 Sep 2022

A. Xiyimo xa Khabinete eka Timhakakulu
1. Ikonomi

1.1. Khabinete yi amukele ku ehlanyana eka mpimo wa mpfumaleko wa mitirho etikweni, lowu wu ehleke hi 0.6%, kusuka 34.5% eka kotara yo sungula kufika ka 33.9% eka kotara ya vumbirhi ya 2022. Tinhlayo leti hangalasiweke eka Mbalango wa Nhlayo ya Vatirhi wa Kotara na Kotara hi Tinhlayonhlayo ta Afrika-Dzonga ti komba leswaku mivuyelo ya mitirho leyikulu kutlula hinkwayo yi rhekodiwile eka Vukorhokeri bya Miganga na Vanhu (276 000), Mabindzu (169 000), Timali (128 000) na Vuaki (104 000).

TRANSLATIONS

- English
- Afrikaans
- isiNdebele
- isiXhosa
- isiZulu
- Sesotho
- Sepedi
- Setswana
- Siswati
- Tshivenda
- Xitsonga

Cabinet Statements

You can use the filters to show only results that match your interests

Title	Speech Date
Statement on the Cabinet Meeting of 29 March 2023	30 Mar 2023
Statement on the Cabinet Meeting of 15 March 2023	16 Mar 2023
Statement on the Cabinet Meeting of 1 March 2023	2 Mar 2023
Statement on the Cabinet Meeting of 15 February 2023	16 Feb 2023
Statement on the Cabinet Meeting of 30 November 2022	1 Dec 2022
Statement on the Cabinet Meeting of 16 November 2022	18 Nov 2022



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Extracting Cabinet Statements

We extract the data from the HTML and prepare a JSON payload.

The JSON payload for each speech records:

- Date,
- Datetime,
- Title (in English),
- Url (top url for speech),
- Language payload for each language (eng, afr, nbl, xho, zul, nso, sep, tsn, ssw, ven, tso).
 - Title (in language),
 - Text (in language),
 - Url (for the translation).



The screenshot shows the South African Government website (www.gov.za) with the cabinet statement titled "Xitatimente hi mayelana na Nhlengeletano ya Khabinete ya le moyeni Ravunharhu, 31 Mhawuri 2022". The page includes a navigation bar with links like HOME, ABOUT, NEWSROOM, SERVICES, DOCUMENTS, and JOBS. A sidebar on the right lists translations in various South African languages. The main content area contains the text of the statement, starting with "1. Xiyimo xa Khabinete eka Timhakakulu" and "1. Ikonomi".



government
communications

Department:
Government Communication and Information System
REPUBLIC OF SOUTH AFRICA



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Prepared Dataset

We have 162 cabinet statements spanning 2 May 2013 to 1 December 2022.

The dataset will update automatically when new, translated, statements are available on the gov.za website.

The dataset, code and automated scrapers are available at <https://github.com/dsfsi/gov-za-multilingual> and Zenodo



The screenshot shows the South African Government website. At the top, the South African coat of arms and the text "South African Government" and "www.gov.za" are visible. Below this is a navigation bar with links: HOME, ABOUT, NEWSROOM, SERVICES, DOCUMENTS, JOBS. A search bar is on the right. The main content area displays a cabinet statement in Xhosa, dated 1 Sep 2022. The statement is titled "Xitatimente hi mayelana na Nhlengeletano ya Khabinete ya le moyeni Ravunharhu, 31 Mhawuri 2022". It discusses economic indicators and government performance. To the right of the statement is a "TRANSLATIONS" section with a list of languages: English, Afrikaans, isiNdebele, isiXhosa, isiZulu, Sesotho, Sepedi, Setswana, Siswati, Tshivenda, and Xitsonga.



government
communications

Department:
Government Communication and Information System
REPUBLIC OF SOUTH AFRICA



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

<https://doi.org/10.5281/zenodo.7635167>

Vuk'uzenzele Gov Newspaper

What is this source?

- “Vuk'uzenzele is a free Government Newspaper, committed to making a difference in the lives of South Africans”
- Published in 2 editions per month. Available physically and digitally [PDF] in all 11 South African languages.

Need

- **Extract text from PDFs.** Clean and make them easily accessible for NLP.
- Make the **extraction process repeatable** with humans in the loop.

<https://www.vukuzenzele.gov.za/>



Vuk'uzenzele

Yi humelerisiwile hi: Vufambisi bya Mfumo bya Vuhlanganisi na Mahungu (GCIS)

English/Xitsonga

Dzivamisoko 2

Ku endla swimakiwa swa mi na swo tisasekisa hi mpepo



Sihle Manda

Loko a karhatiwile pfaleleka eka fulele ni ra Kapa hi madyankarhi wa ntungumhla Dotwana a ri na xin ra glulupu leri ri nga yisa masungulo ya bindzu lero x 31 wa malembe, musungurition, a hlamusela riendo ra nzele.

Namhla Collection i nongolohlayisa nhlonge swa ntumbuluk tikosikaya ra Egotyibeni eTsolo, e Riendo ra bindzu leri ri sung n'okisa nakambe khandhele ra risiwa hi sinamono leri a ri gov matimba ya mona fasilitereni ra y Loko kondiraka ya yena na Dor xa nhlurukiso wa vaaki ya ku vuhangalasamahungunyinyi yil Dotwana u pakile tibege ta yena ra yena egirachini ya manana w "Ndzi endla khosimetiki na sw ntsungulo wa miri ku nga ri tihisa xikhwati xa hina xa nd

Swiya ema



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Extracting Vuk'uzenzele PDFs

To clean it, a team member goes through each extracted text file and formats it as follows:

- *Line 1:* Title of article (in language)
- *Line 2:* empty line
- *Line 3:* Author of article (if available. If not, defaults to Vukuzenzele Unnamed)
- *Line 4:* empty line
- *Line 5-end:* body of article



government
communications

Department:
Government Communication and Information System
REPUBLIC OF SOUTH AFRICA



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Prepared Dataset

We have 53 editions of the newspaper spanning January 2020 to July 2022. (More are being added constantly)

Automations have been built to download and archive the PDFs, however manual effort is still required to extract and identify translated articles.

The dataset, code and automated scrapers are available at at <https://github.com/dsfsi/vukuzenzele-nlp> and Zenodo

<https://doi.org/10.5281/zenodo.7598539>



government
communications

Department:
Government Communication and Information System
REPUBLIC OF SOUTH AFRICA

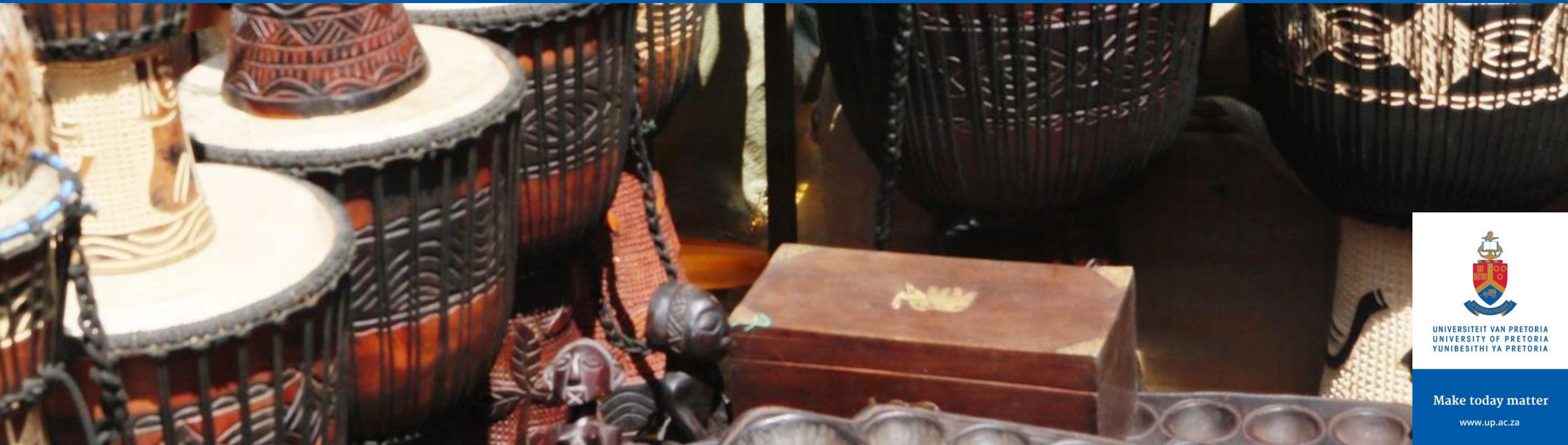


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Sentence Alignment and MT Benchmarks



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

Sentence Alignment: *Output*

The output of the sentence alignment task is 55 distinct parallel corpora for both the Vukuzenzele & Cabinet Statements.

Table 1: Language List with ISO 639-2 codes

Name	Code
isiZulu	zul
isiXhosa	xho
Afrikaans	afr
English	eng
Sepedi	nso
Setswana	tsn
Xitsonga	tso
Sesotho	sot
siSwati	ssw
Tshivenda	ven
isiNdebele	nbl

Table 2: Top ten datasets with the most observations with a cosine score greater than or equal to 0.65 in Vuk'uzenzele.

Language pair	No. of observations in Vuk'uzenzele
ssw-xho	2,202
ssw-zul	2,183
xho-zul	2,102
nso-xho	2,081
nso-tso	2,071
ssw-tso	2,034
nso-ssw	2,021
tsn-tso	2,020
tsn-xho	2,009
tso-xho	2,009

Table 3: Top ten datasets with the most observations with a cosine score greater than or equal to 0.65

Language pair	No. of observations in ZAgov Multilingual
nbl-ven	18,984
nso-ssw	18,697
zul-ssw	18,563
xho-ssw	18,387
xho-zul	18,145
xho-nso	18,110
xho-tso	17,954
ssw-tso	17,880
zul-tso	17,789
zul-nso	17,630



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA



government
communications

Department:
Government Communication and Information System
REPUBLIC OF SOUTH AFRICA

Make today matter

www.up.ac.za

Data Characteristics for Aligned Translations

Table 5: Characteristics of the translation data for the Vuk'uzenzele (Vuk.) datasets

Translation Direction	Size #sents (#src / #trg tokens)
nbl→eng	136 (3.4k / 3.9k)
nso→eng	1715 (53.7k / 41.6k)
ssw→eng	1588 (29.9k / 37.6k)
sot→eng	260 (9.9k / 7.5k)
tsn→eng	1366 (49.8k / 31.7k)
tso→eng	1998 (58.9k / 46.6k)
ven→eng	230 (9.1k / 7k)
xho→eng	1338 (25.8k / 31.5k)
zul→eng	1874 (34.1k / 43k)

Table 6: Characteristics of the translation data for the ZA-gov-multilingual (Gov.) datasets

Translation Direction	Size #sents (#src / #trg tokens)
nbl→eng	3513 (63.9k / 107k)
nso→eng	14742 (460.9k / 375k)
ssw→eng	15139 (291k / 377.8k)
sot→eng	4995 (145.9k / 153.5k)
tsn→eng	14068 (493.1k / 362.2k)
tso→eng	15393 (466.4k / 381.2k)
ven→eng	3404 (68.2k / 96.6k)
xho→eng	15853 (318.2k / 389.5k)
zul→eng	15503 (327.5k / 384.1k)



NMT Benchmarks

Training translation models on a small amount of data can hinder the quality

Therefore we leverage a Massively Multilingual Model (M2M100)(Fan et al , 2021) and fine-tune it on our relatively small datasets (Adelani et al, 2022)

To provide our results in context and for comparison we also fine-tune the M2M100 model on subsets of the existing **Autshumato** corpora (Which exist only in the 'eng-xxx')

We focus our efforts on providing NMT benchmarks for the low resource African languages

- *David Adelani et al. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation*
- *Angela Fan et al. 2021. Beyond english-centric multilingual machine translation.*



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

MT Results

The highest BLEU scores are distributed across the ZA-gov-multilingual and Autshumato NMT models

ZA-gov-multilingual models achieving a higher score for **Setswana, Xitsonga and isiZulu**

Highest benchmark result for **Xitsonga** across all datasets, demonstrating the effectiveness of transfer learning for new low-resource language datasets.

Our contributions:

- Extend the benchmark translation resources (in the government data domain) to **isiNdebele, isiXhosa, siSwati and Tshivenda**
- Broaden the translation direction **beyond English** as the source language.

It is noted that variations in the Autsumato subset selections will yield different results and an in-depth analysis is left for future work.

Table 4: BLEU scores for Massively Multilingual Transfer on xxx-eng translations using the Vuk'uzenzele (Vuk.), ZA-gov-multilingual (Gov.) and subsets of the available Autshumato datasets (Aut.)

Translation Direction	BLEU		
	Vuk.	Gov.	Aut.
nbl →eng	7.33	8.04	12.24
nso→eng	9.29	26.50	-
ssw→eng	4.80	28.72	-
sot →eng	4.55	10.21	14.83
tsn→eng	2.80	29.68	28.04
tso →eng	13.86	35.40	32.10
ven →eng	2.32	9.68	17.24
xho→eng	6.05	26.81	-
zul→eng	9.97	30.03	25.90

• Source languages which were not including in the original M2M100 pre-training are highlighted



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za



SABC Tindzaba
@SABCTindzaba

Badlali beBafana Bafana Ronwen Williams na Percy Tau banelitsemba lekutsi lelicembu liyowucudzelana ku AFCON e Ivory Coast hhayi nje kutsi liyowungeta sibalo.

#SABCNews
#SABCTindzaba

⚡ **Hosted inference API** ⓘ

📄 Text2Text Generation

Badlali beBafana Bafana Ronwen Williams na Percy Tau banelitsemba lekutsi lelicembu liyowucudzelana ku AFCON e Ivory Coast hhayi nje kutsi liyowungeta sibalo.

Compute +Enter

0.0

Computation time on Intel Xeon 3rd Gen Scalable cpu: 3.522 s

Bafana Bafana Ronwen Williams and Percy Tau are confident that the team is competing to AFCON on Ivory Coast and not only raise awards.

Breaking News



Mo, 11 sep. ☁ 30°

Tu, 12 sep. ☀ 24°

We, 13 sep. ☀ 23°

Th, 14 sep. ☀ 27°

Fr, 15 sep. ☀ 27°



Our **NEW** facebook
page name & Look



[Home](#) / [News](#) / [Dikgang](#) / Ba lebisitswe molato wa petelelo le go kgothosa

BA LEBISITSWE MOLATO WA PETELELO LE GO KGOTHOSA

📅 10 Sep 2023

Borre ba le babedi ba ba tlholegang kwa lefatsheng la Zimbabwe ba itshupile fa pele ga lekgotla ka Labone mabapi le molato wa go tsaya ka dikgoka le go thubetsa mme yo o dingwaga tse di masome a matlhano le bobedi (52) wa kgotla ya Kanamo kwa Mahalapye.

Borre bao, wa dingwaga tse di masome a mararo le bobedi (32) le wa tse di masome a mararo le borataro (36) mo bekeng e e fetileng ba ne ba tsenelela mme yo mo ntlong ba mo amoga megala ya letheke e le mebedi ba bo ba mmeteleta.

Mookamela Mapodisi a Mahalapye Central Supt. Boitshepho Mudongo o kaile fa ka lesego ba kgonne go ba tshwara mme megala yotlhe ka bobedi e bonwe.

A re bobedi jo bo amanngwa le dikgetsi tse dingwe tsa go kgothosa di le pedi ka ba fitlhetswe ka dilwana tse di neng di begilwe fa di utswilwe.



⚡ Hosted inference API ⓘ

🔍 Token Classification

Borre ba le babedi ba ba tlholegang kwa lefatsheng la Zimbabwe ba itshupile fa pele ga lekgotla ka Labone mabapi le molato wa go tsaya ka dikgoka le go thubetsa mme yo o dingwaga tse di masome a matlhano le bobedi (52) wa kgotla ya Kanamo kwa Mahalapye.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

Borre ba le babedi ba ba tlholegang kwa lefatsheng la Zimbabwe LOC ba itshupile fa pele ga lekgotla ka Labone DATE mabapi le molato wa go tsaya ka dikgoka le go thubetsa mme yo o dingwaga tse di masome a matlhano DATE le bobedi DATE (52) wa kgotla ya LOC Kana PER mo kwa LOC Mahalapye LOC .

</> JSON Output

🖥 Maximize

⚡ Hosted inference API ⓘ

Token Classification

Borre ba le babedi ba ba tlholegang kwa lefatsheng la Zimbabwe ba itshupile fa pele ga lekgotla ka Labone mabapi le molato wa go tsaya ka dikgoka le go thubetsa mme yo o dingwaga tse di masome a matlhano le bobedi (52) wa kgotla ya Kanamo kwa Mahalapye.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.042 s

Borre NOUN ba le DET babedi NUM ba DET ba PRON tlholegang VERB kwa DET lefatsheng PROPN la DET Zimbabwe PROPN ba PRON
itshupile VERB fa PUNCT pele ADV ga ADP lekgotla NOUN ka ADP Labone mabapi NOUN le AUX molato NOUN wa DET go tsaya VERB ka ADP
dikgoka NOUN le CCONJ go thubetsa VERB mme ADV yo DET o dingwaga NUM tse DET di masome NUM a PROPN matlhano NUM le CCONJ
bobedi NUM (PUNCT 52 NUM) PUNCT wa DET kgotla NOUN ya DET Kana PROPN mo kwa DET Mahalapye PROPN . PUNCT

</> JSON Output

Maximize

Mavito: South African Terminology, Lexicon and Glossary Project

Vukosi Marivate (PI)* Fiskani Banda Richard Lastrucci Mohlatlego Nakeng Kayode Olaleye Thapelo Sindane
Data Science for Social Impact
University of Pretoria

<https://dsfsi.github.io/za-mavito/>

oertb-termbank-IATE

5,744 rows

- column - =

Apply

[View and edit SQL](#)

This data as [json](#), [CSV \(advanced\)](#)

Link	rowid ▼	eng	afr	zul	xho	ssw	nbl	nso	tsn
1	1	acculturation	akkulturasie	ukunikezelana ngamasiko	utshintshiselwano ngenkcubeko	kutselelana kwemasiko	ukuthatha amasiko wakosobantu okubangelwa kuqongelana ngokwamasiko	tlwaetšo ya setšo	tlwaetso ya setso
2	2	acephalous society	leierlose samelewing, asefale samelewing	umphakathi ongenanhloko ngokwezombusazwe	uluntu olungenayo inkokeli	ummango lote inhlokho	umphakathi onganabadosiphambili	setšhaba seo se se nago moetapele wa sepolotiki	setšhaba tihoka-moeteledipele
3	3	achieved status	verworwe status	isikhundla esisetshenzelwe	iwonga elifikelelweyo	simobunjalo lesizuziwe setenhlalo	ubujamo obuzuziweko	phihlelelo ya maemo	seemo se se fitheletsweng
4	4	adult socialisation	volwasse sosialisering	ukulungiselelwa impilo yobudala	uqhelaniso lobuntu lwabantu abadala	kukhulisa umuntfu lomdzala	ikambiso yokulungiselela ubudala	thuto ya go tlwaetša go phela le batho	thuto ya bogolo
5	5	affinal kin	aangetroude familie	izihlobo ngokomshado	umzalwane ngokomtshato	tihlobo ngekwemendvo, umshado	ukuhlobana ngokuthathana	balapa ka lenyalo	balosika ka nyalo
6	6	affluent society	welvarende gemeenskap	umphakathi onothile waseMelika	uluntu olunobutyebi	ummango lonjingile	umphakathi onjingileko	setšhaba se se atlegilego	setšhaba se se humileng



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

We need to get tools in people hands!



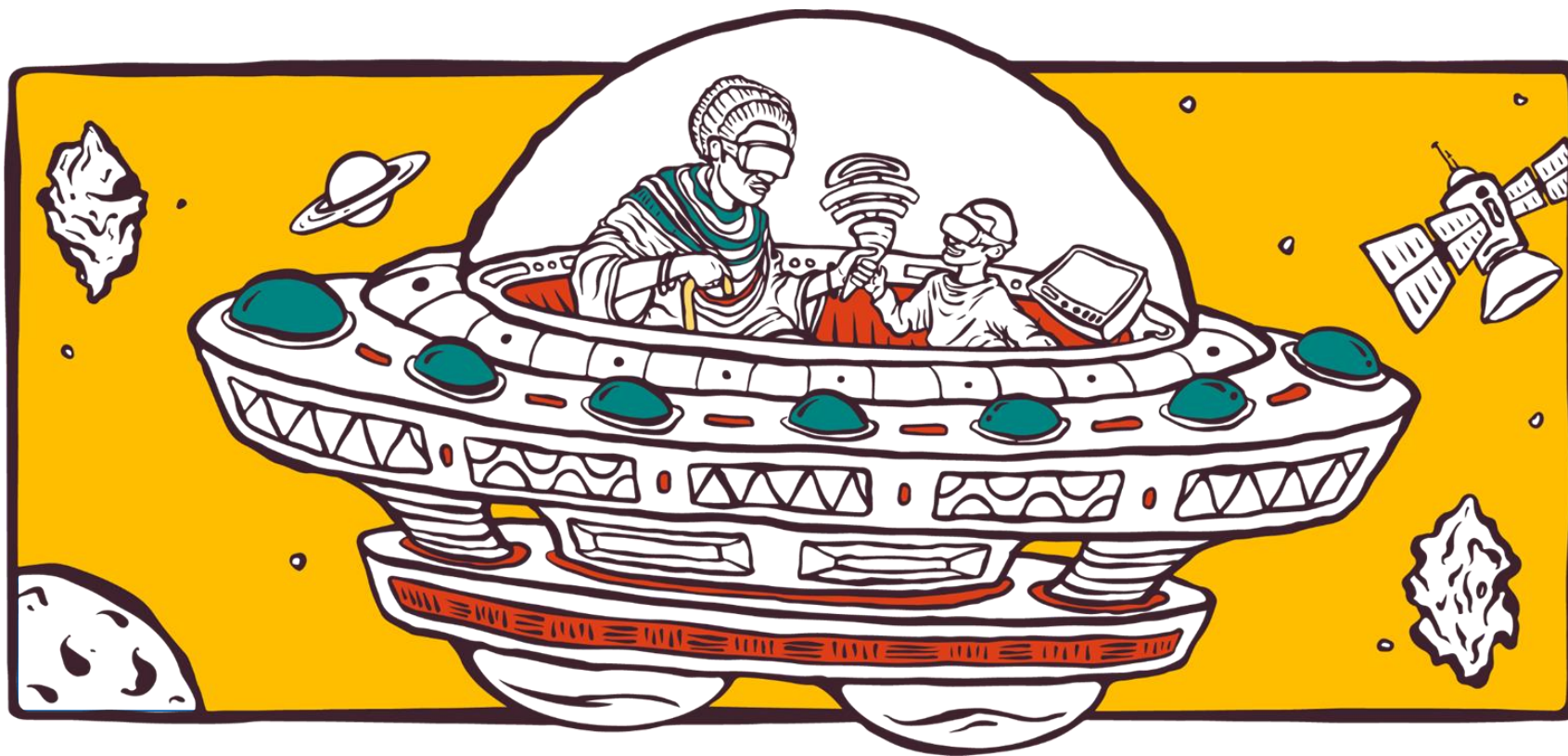
UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za



LELAPA AI



**Grow your market, know your market, foster loyal
connections**



At Lelapa AI, we are targeting digital products and services, to serve **520 million Africans** in their **own languages**



VULAVULA FEATURES



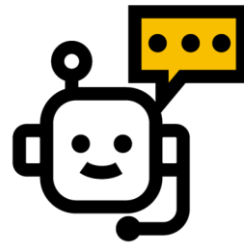
TRANSCRIBE

Voice to text
conversion



ANALYSE

Text Analysis
functions



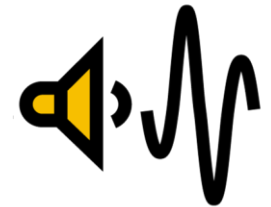
CONVERSE

NLU capability
to power chat



TRANSLATE

Text to text
translation



SPEAK

Text to speech
conversion

Solutions

EGRA

Support the enhancement of EGRA-AI for home languages

Teacher support

Virtual Teaching Assistant:
Support in creating multilingual lesson plans and chat for live teacher support.

Education language model

Train an education (literacy & numeracy) specific model with robust data for use across multiple literary programmes.

Partner org AI enablement

Enable AI capabilities of partners & their current programmes- to prioritise based on most feasible over a period



BRIDGE THE
DIVIDE OF
COMMUNICATION

Collaborations



government
communications

Department:
Government Communication and Information System
REPUBLIC OF SOUTH AFRICA



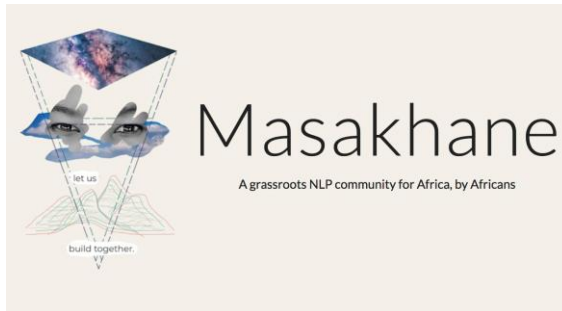
DEEP
LEARNING
INDABA



**MEDIA
MONITORING
AFRICA**



Titimuleni
Nthavela



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Make today matter

www.up.ac.za

DSFSI Support - Thank You

Grants/Gifts/Fellowships



Canada



Data Science for Social Impact



FUTURE PROFESSORS PROGRAMME (PHASE 01)

CREATING THE NEXT GENERATION OF SOUTH AFRICAN PROFESSORS

The Future Professors Programme (Phase 01) is a flagship programme of the [Department of Higher Education and Training](#). It is a national, collaborative initiative, aimed at developing the qualities of academic excellence and leadership in university scholarship in order to contribute to the development of a future South African professoriate.



science & innovation

Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA



ARTIFICIAL
INTELLIGENCE
FOR
DEVELOPMENT
AFRICA



Make today matter
www.up.ac.za

Thank you Questions



Prof. Vukosi Marivate

vukosi.marivate@cs.up.ac.za

<https://dsfsi.github.io>

@vukosi

@DSFSI_Research

Keep in touch

Join our research group newsletter

<https://dsup.substack.com/>

Made with ❤️ in Tshwane



**Faculty of Engineering,
Built Environment and
Information Technology**

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknološi ya Tshedimošo

Make today matter

www.up.ac.za