African Languages in the Technological Realm

Beata Wójtowicz
University of Warsaw

Martin Benjamin
Kamusi Project International

UniDive 2nd General Meeting - Naples, Italy – 8 February 2024

Presentation for Download:
kamu.si/unidive-africa-pdf
kamu.si/unidive-africa-ppt

UniDive

Universality, Diversity, and Idiosyncracy in Language Technology

EU Cost Action CA21167

University of Naples "L'Orientale"

Department of Asian, African and Mediterranean Studies

1969: Swahili Language and Literature established by Elena Bertoncini Zúbková

Vocabolario swahili italiano e italiano-swahili (1977)

1000s of usage examples for Kamusi Project (1995)

2004: Maddalena Toscano

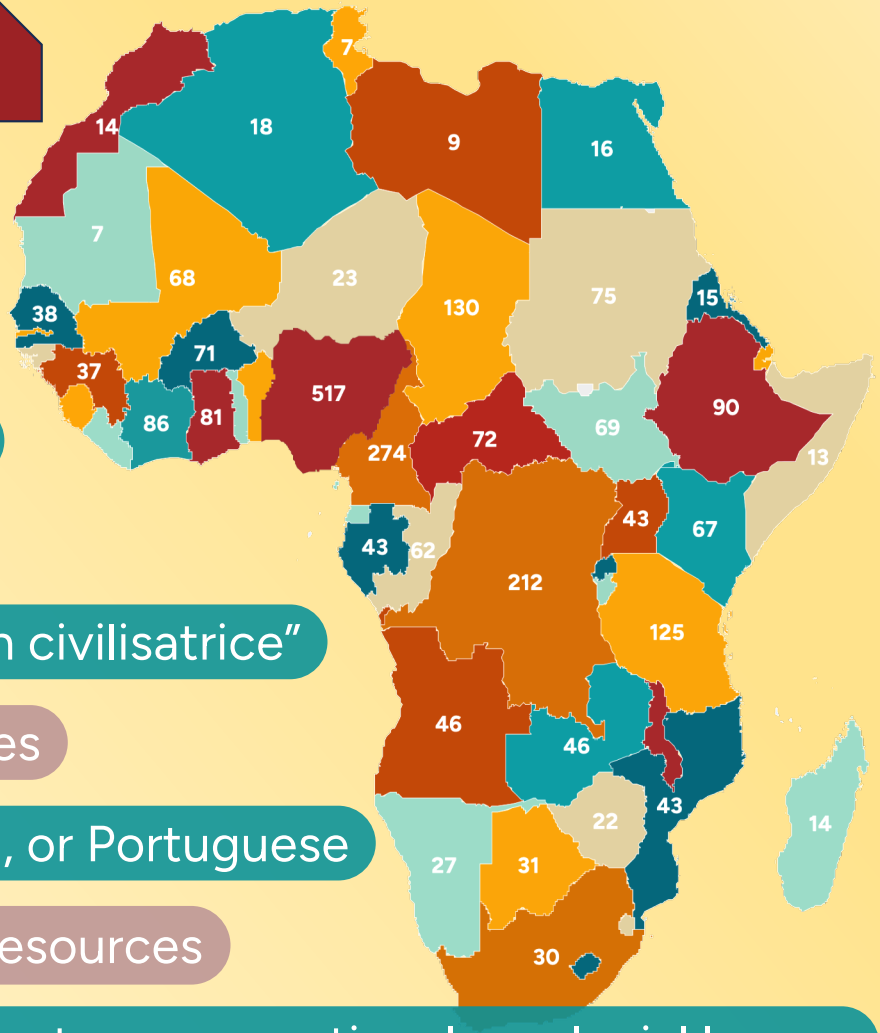
Dizionario swahili. Swahili-italiano, italiano-swahili

small Swahili untagged raw corpus

2018: Rosanna Tramutoli - Swahili-Italian linguistic glossary

African Languages Overview

- 2000-ish languages in 55 countries
- Mother tongues are African
- Early primary school is often in a local language
- Secondary school is in a colonial language
- European languages were central to the “mission civilisatrice”
- Governance usually occurs in European languages
- About 5% of Africans have good English, French, or Portuguese
- Dozens of “official” African languages, but few resources
- Rampant multilingualism (mother tongue + father tongue + national + colonial legacy



Official African Languages

Always: European Colonial Legacy Language

Tanzania: Swahili since independence in 1961

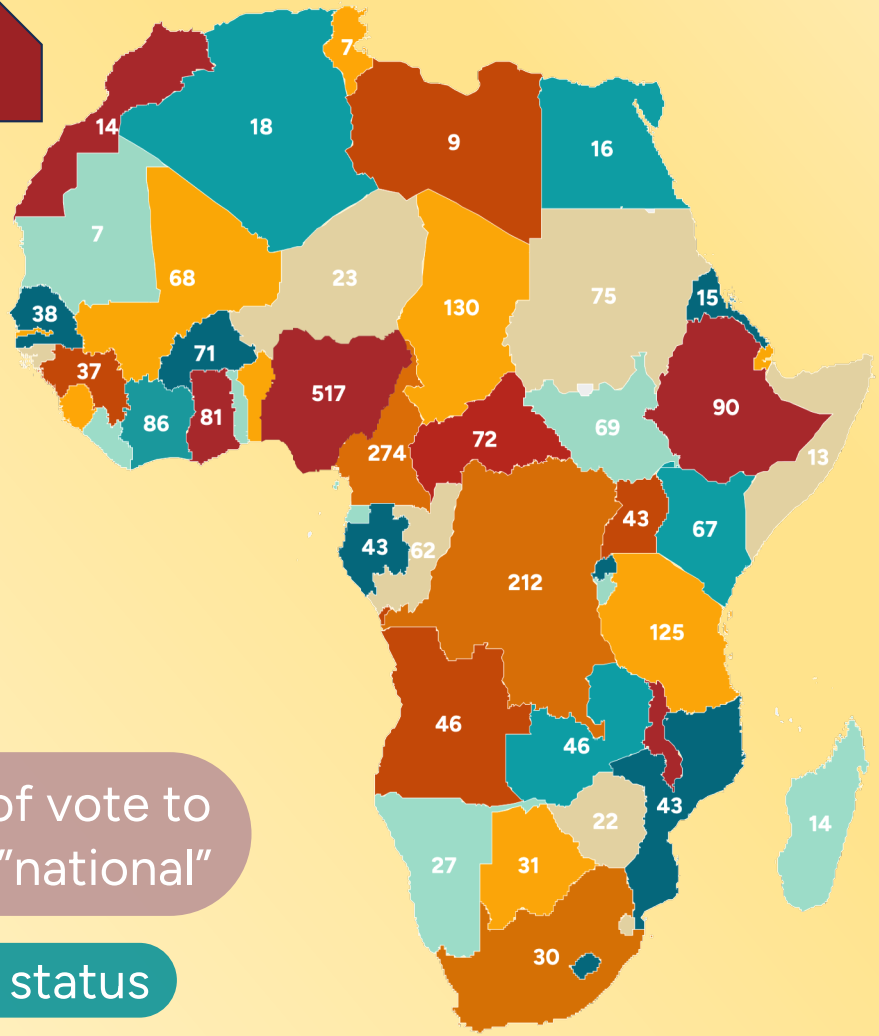
Kenya: Swahili official since 2010

Zimbabwe: 16 offiicial languages since 2016

South Africa: 9 African + Afrikaans + English

Mali: 2023 constitutional amendment with 97% of vote to make French "working" and 13 languages "national"

Other countries: African languages have various status

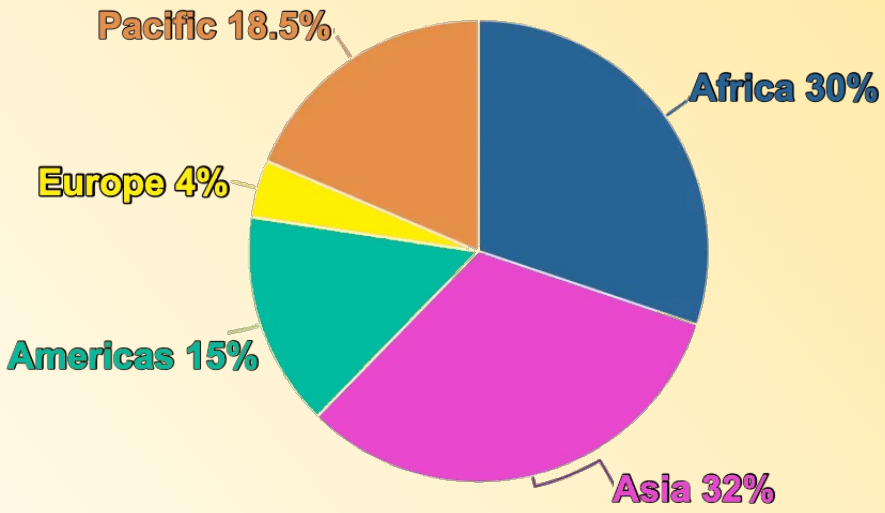


African Languages in Education

- African languages frequent for Early Primary
- Secondary is in colonial legacy language
- Poor performance on international exams in French or English
- STEM teachers often have weak language training
- Compare to & & Asia:
  - STEM is taught in local language
  - English and French are taught as course subjects
- Research: students learn better in their own language
- Ideology: success occurs in international language (many policy-makers, grant-makers, and parents)

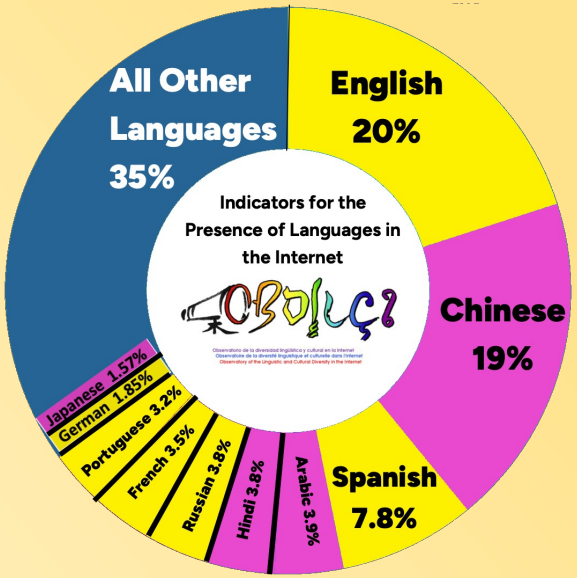
World Languages

% of Global Inventory

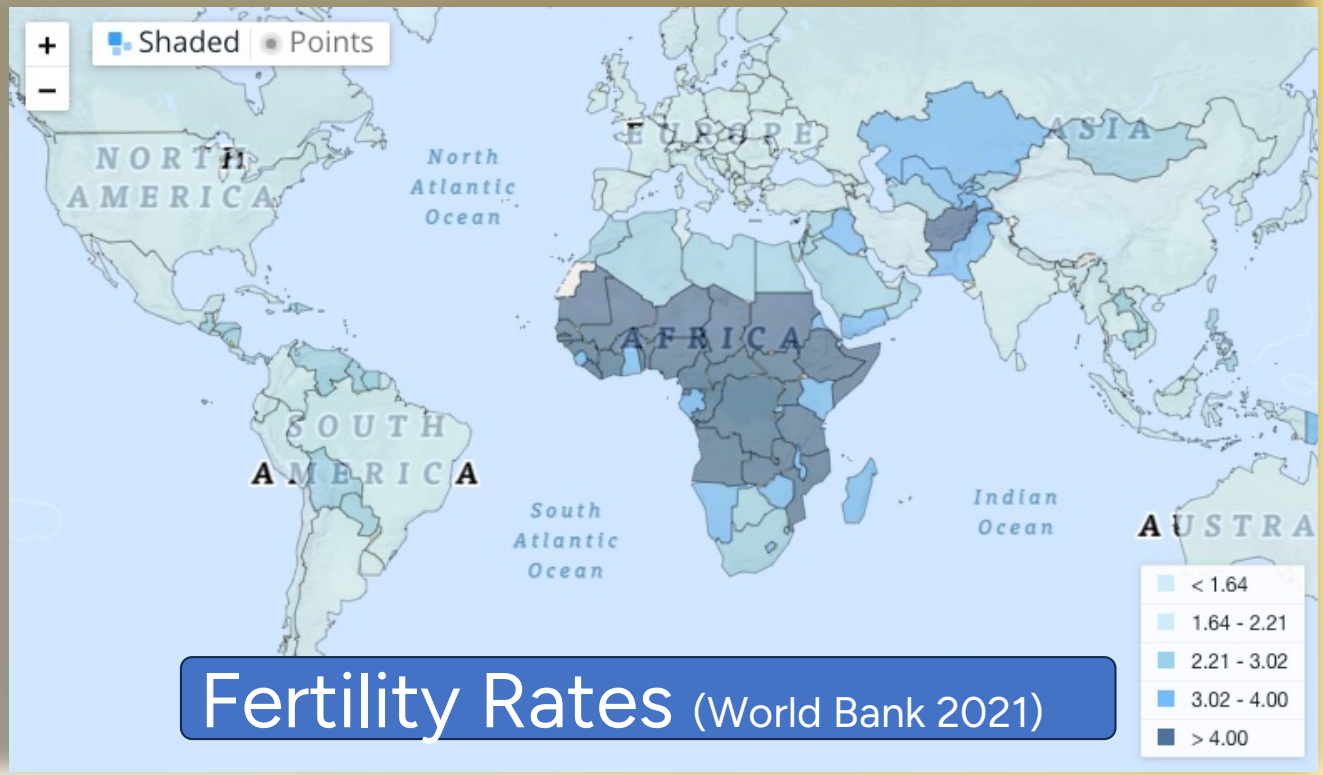


% of Technology

"Internet" is a rough proxy for all language tech

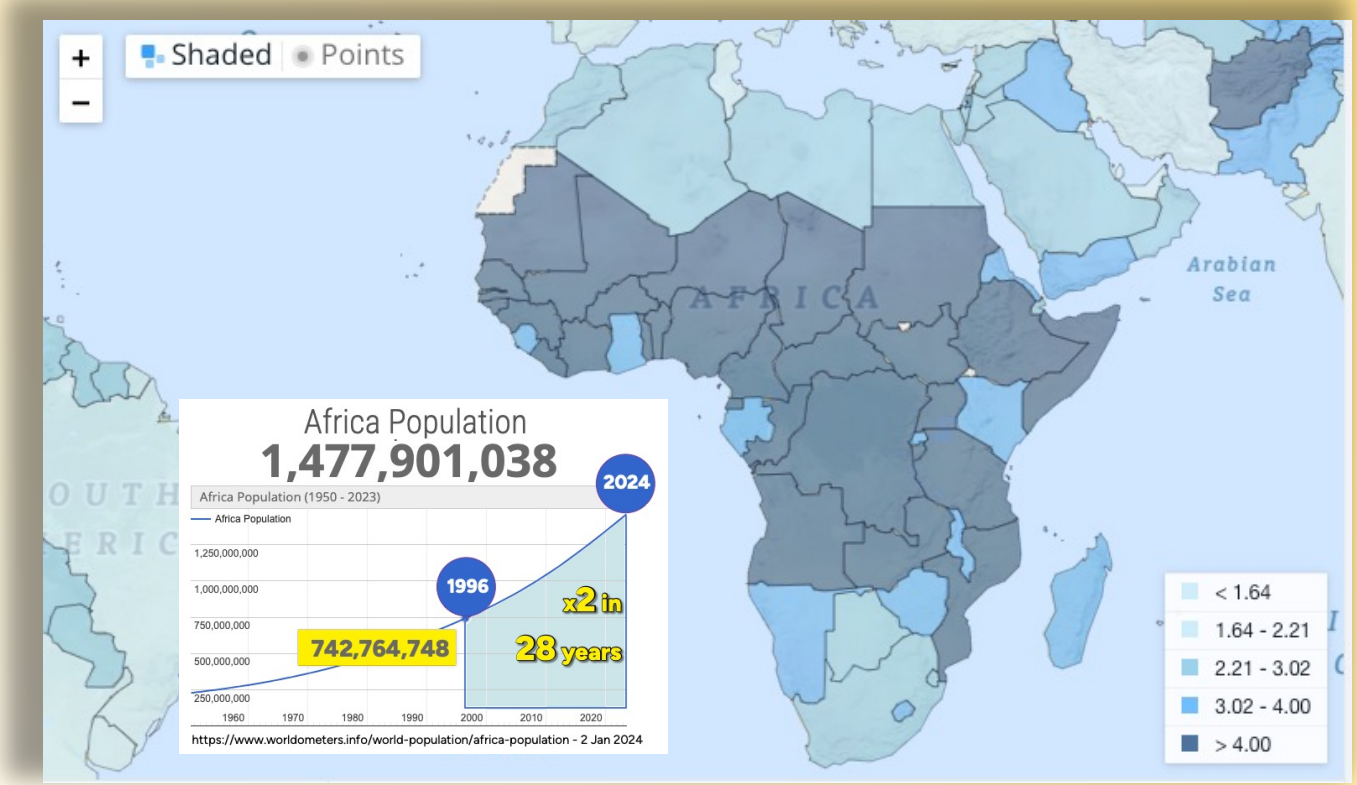


many African Languages are GROWING fast



Fertility Rates (World Bank 2021)

many African Languages are GROWING fast



100+ big languages are getting bigger

1000+ smaller languages are under threat

Mega languages and **Endangered** languages have **different** needs and challenges

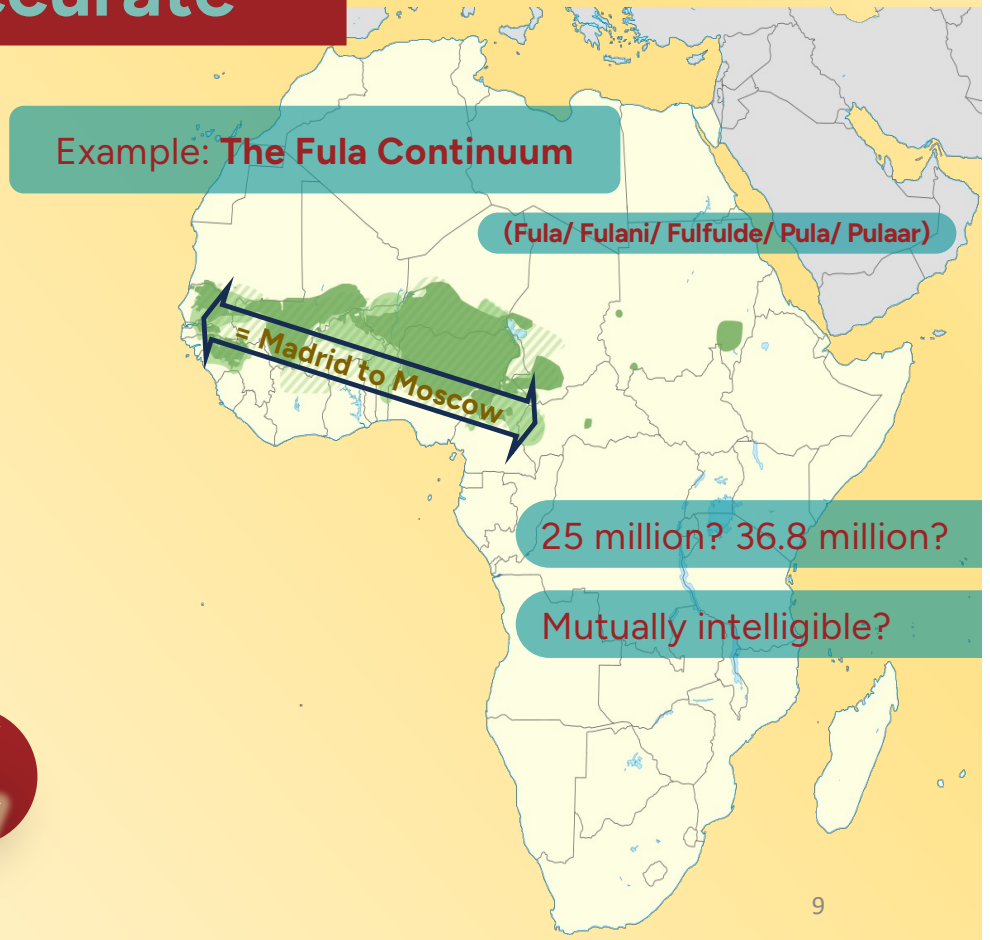
Caution: Language # Estimates are Guaranteed to be Inaccurate

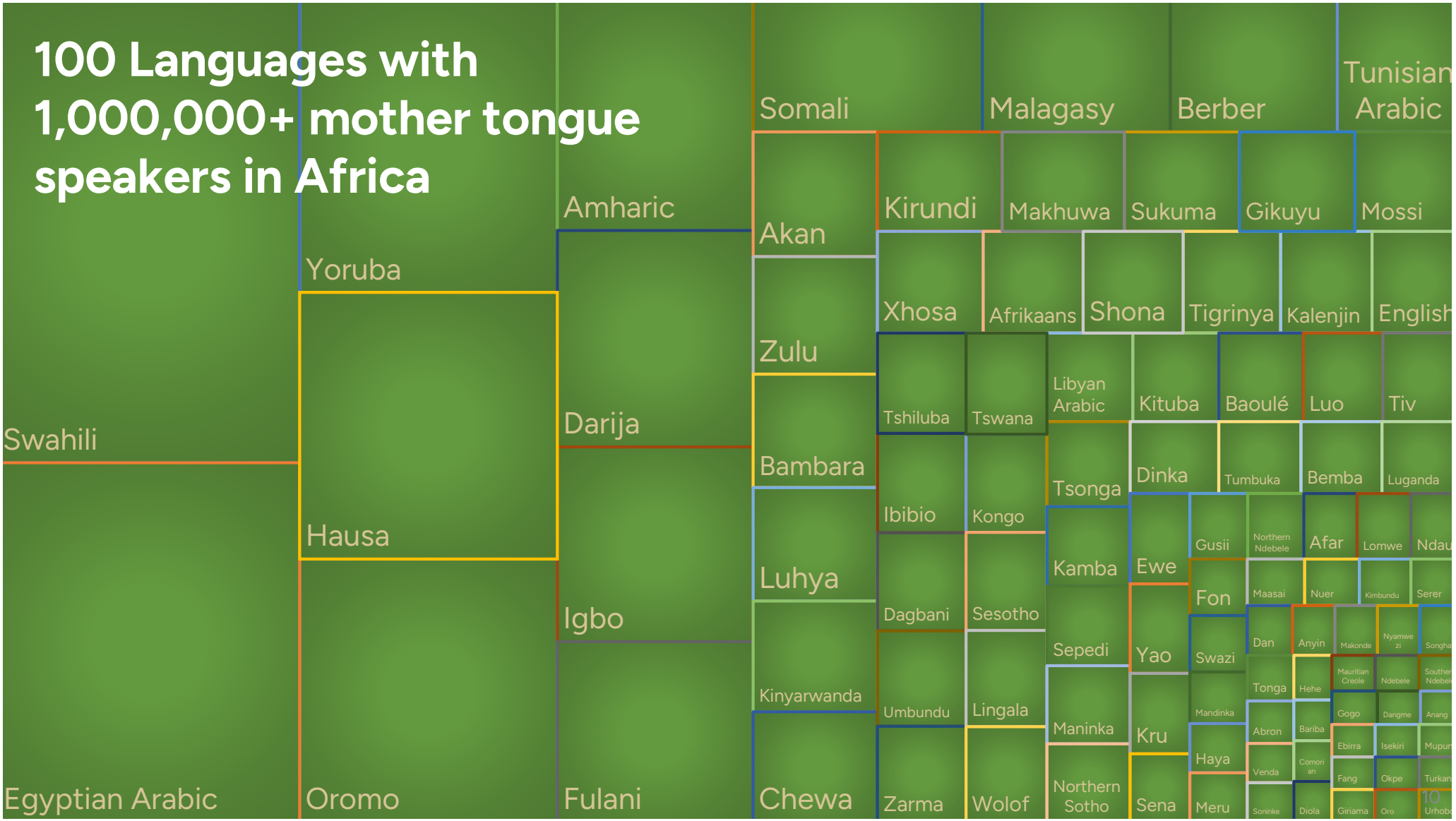
Language is not usually a census question

Estimates for different languages might be from decades apart

Multilingualism: a speaker counted for X might also speak Y and Z

The boundaries between languages/ dialects/ macro-languages are very blurry

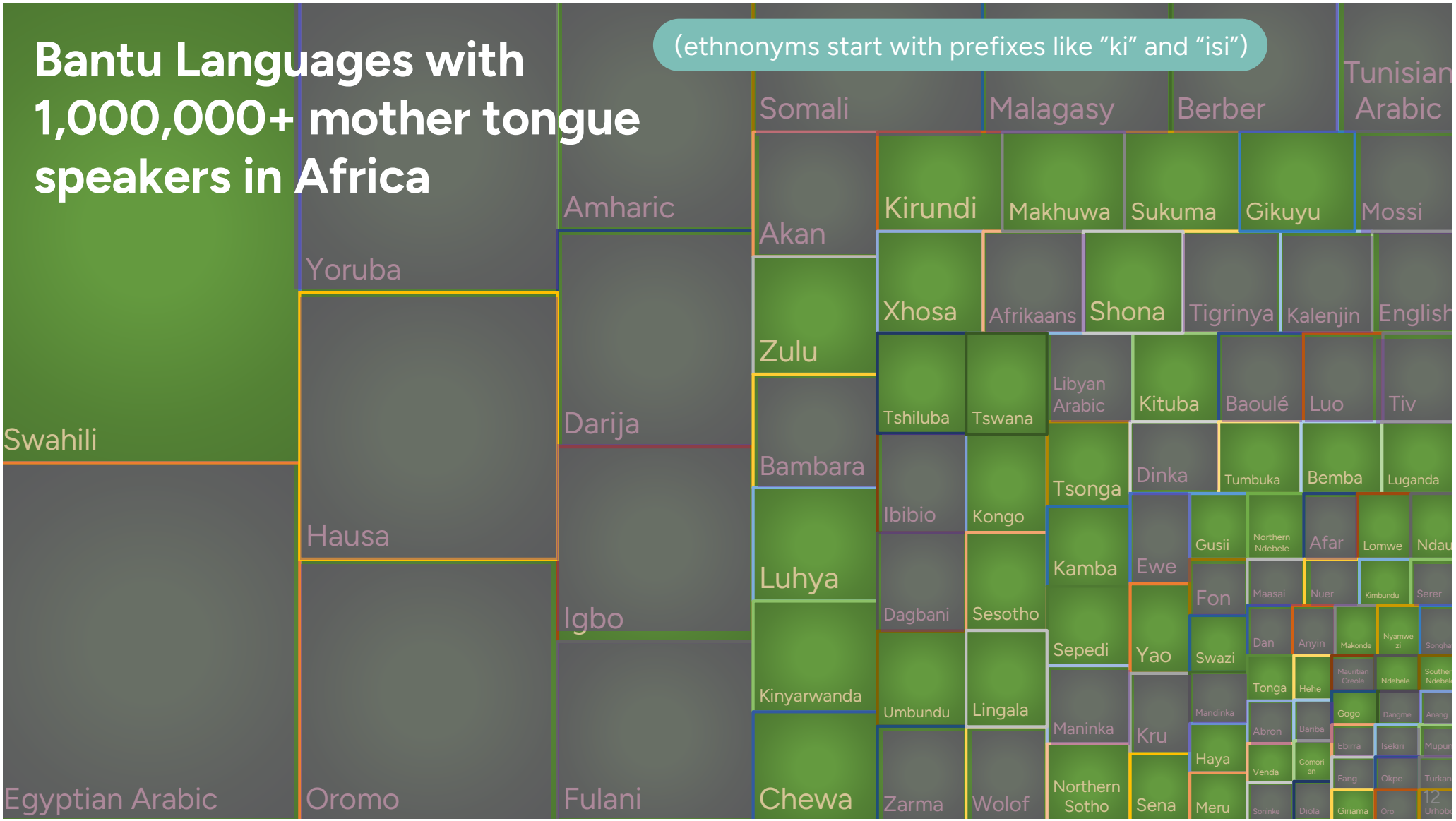



100 Languages with 1,000,000+ mother tongue speakers in Africa
Swahili
Egyptian Arabic
Yoruba
Hausa
Oromo
Amharic
Darija
Igbo
Fulani
Somali
Malagasy
Berber
Tunisian Arabic
Akan
Kirundi
Makhuwa
Sukuma
Gikuyu
Mossi
Zulu
Xhosa
Afrikaans
Shona
Tigrinya
Kalenjin
English
Bambara
Tshiluba
Tswana
Libyan Arabic
Kituba
Baoulé
Luo
Tiv
Luhya
Ibibio
Kongo
Tsonga
Dinka
Tumbuka
Bemba
Luganda
Kinyarwanda
Dagbani
Sesotho
Kamba
Ewe
Gusii
Northern Ndebele
Afar
Lomwe
Ndau
Umbundu
Lingala
Sepedi
Yao
Fon
Maasai
Nuer
Kimbundu
Serer
Chewa
Zarma
Wolof
Maninka
Kru
Swazi
Dan
Anyin
Makonde
Nyamwezi
Songha
Northern Sotho
Sena
Mandinka
Tonga
Hehe
Mauritian Creole
Ndebele
Haya
Abron
Bariba
Gogo
Dangme
Meru
Venda
Comorian
Ebirra
Isekiri
Fang
Okpe
Soninke
Diola
Giriama
Oro
10

African Languages Classification

this map gives a general idea

on the ground, territories overlap


Bantu Languages with 1,000,000+ mother tongue speakers in Africa
(ethnonyms start with prefixes like "ki" and "isi")
Swahili
Egyptian Arabic
Yoruba
Hausa
Oromo
Amharic
Darija
Igbo
Fulani
Somali
Malagasy
Berber
Tunisian Arabic
Akan
Kirundi
Makhuwa
Sukuma
Gikuyu
Mossi
Zulu
Xhosa
Afrikaans
Shona
Tigrinya
Kalenjin
English
Bambara
Tshiluba
Tswana
Libyan Arabic
Kituba
Baoulé
Luo
Tiv
Luhya
Ibibio
Kongo
Tsonga
Dinka
Tumbuka
Bemba
Luganda
Kinyarwanda
Dagbani
Sesotho
Kamba
Ewe
Gusii
Northern Ndebele
Afar
Lomwe
Ndau
Umbundu
Lingala
Sepedi
Yao
Fon
Maasai
Nuer
Kimbundu
Serer
Chewa
Zarma
Wolof
Maninka
Kru
Swazi
Dan
Anyin
Makonde
Nyamwezi
Songha
Northern Sotho
Sena
Haya
Tonga
Hehe
Mauritian Creole
Ndebele
Mandinka
Meru
Abron
Bariba
Gogo
Dangme
Anang
Venda
Comorian
Ebirra
Isekiri
Fang
Okpe
Soninke
Dioula
Giriama
Oro
12

Bantu Languages

-ntu (ba**NTU**) is root for “person”

Proto-Bantu dates back X-thousand years

440 – 680 languages

~7% of world inventory

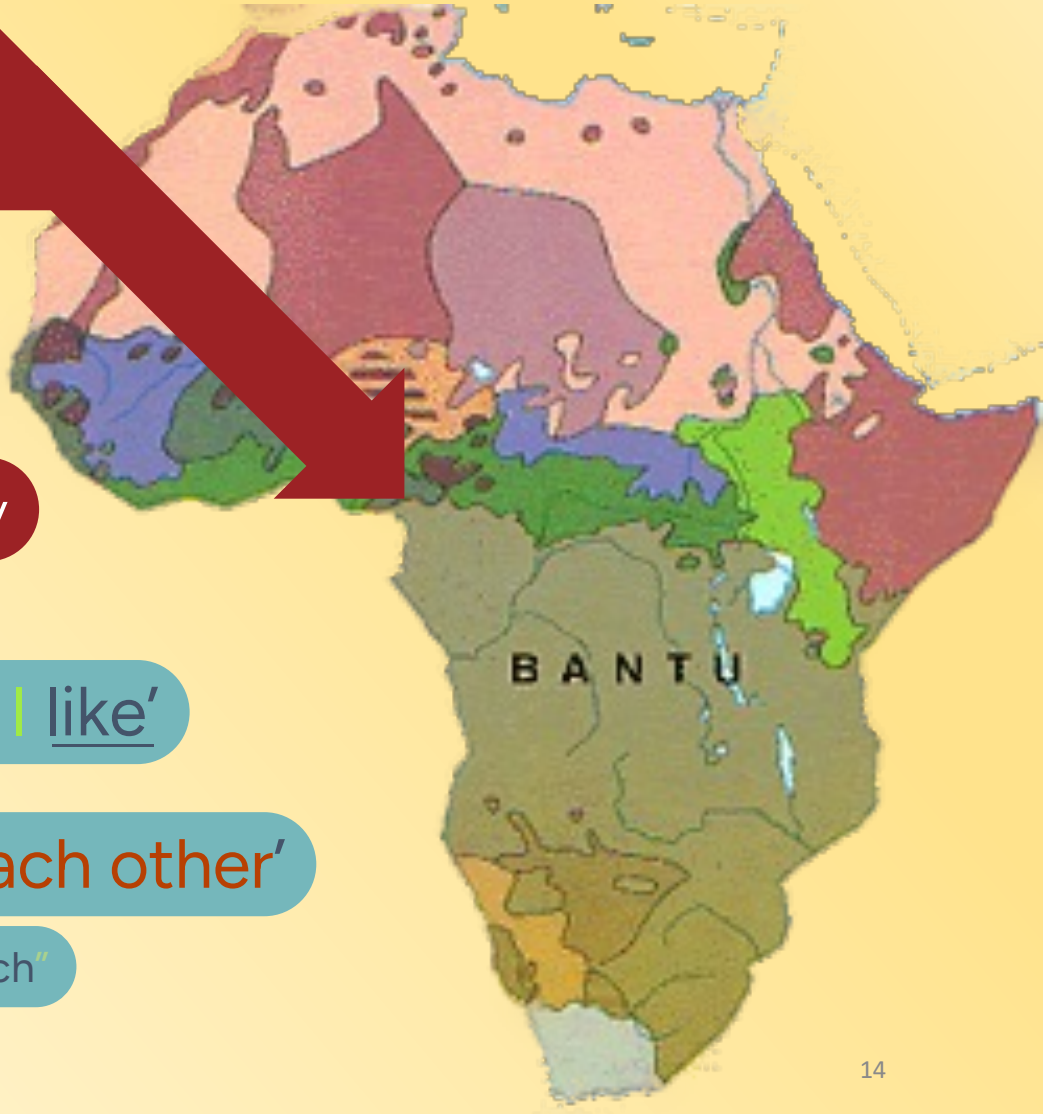
300-500 million speakers

27 present-day countries

Bantu Languages

noun class system (linguistic “gender”)

verbs: complex agglutinative morphology



ninachokipenda ‘the (book) that I like’

wamesomeshana ‘they taught each other’

root –SOMA = “read”, causative –SOMESHA = “teach”

Properties of Swahili: Inflection

reflexes manifested:

Lexically, on the Noun

Syntactically, via Agreement

noun class system

complex agglutinative morphology

class	subject	‘this’	‘nice’	1 or 3	who fell down
cl. 1:	**m**tu 🧍🏾	hu**yu**	**m**zuri	**m**moja	**ali**ye**a**nguka
cl. 2:	**wa**tu 🧍🏾🧍🏾	ha**wa**	**wa**zuri	**wa**tatu	**wa**li**o**anguka
cl. 3:	**m**fuko 👜	h**uu**	**m**zuri	**m**moja	**uli**o**a**nguka
cl. 4:	**mi**fuko 👜👜👜	h**ii**	**mi**zuri	**mi**tatu	**ili**yo**a**nguka
cl. 6:	**ma**yai 🥚🥚🥚	h**aya**	**ma**zuri	**ma**tatu	**yali**yo**a**nguka

Swahili Citations: Dictionary or Data?

complex considerations for computation

Nouns: SINGULAR + fields for plural and class

mtu (noun) pl: watu class: 1/2 eng def: person

Verbs: with hyphen for people, without for machines

Millions of variations, e.g.: akianguka, wataanguka, umeoanguka, yaliyoanguka, kitakachoanguka, zingeanguka

-anguka (verb) eng def: 1. fall down, 2. fail

Numerals, possessives, adjectives: with hyphen + field for class forms

-zuri (adj) eng def: good
class: 1. mzuri, 2. wazuri, 3. mzuri, 4. mizuri, 5. nzuri, 6. mazuri, 7. kizuri, 8. vizuri, 9. nzuri, 10. nzuri, 11. mzuri

.mtu *nm wa-* [*a-/wa-*] person, human being, individual. (*ms*) *wa miraba minne* giant; (*hasa kijana wa kike*) brunette; *mbaya/mnyonge* caitiff; *asiye na adabu/mtu wa hasira* churl; *mwenye hasira* cross-patch; *anayejikomba* flunkey/flunky; *kitu chenye mkosi* jinx; *asiyetaka kufanya kazi* lay about; *maskini mwenye ugonjwa* lazar.

anguk.a *kt* [*sie*] 1 come down, fall down, drop, crash. 2 lose in a business. 3 fail: *Ame~ mitihani yake* he failed his examinations. *(tde)* **angukia**, *(tden)* **angukiana**, *(tdew)* **angukiwa**; *(tdk)* **angukika**. **anguko** *nm*.

.zuri[1] *kv* beautiful, pretty, gorgeous; good: *Mtu m~* good person.

Swahili Derivation

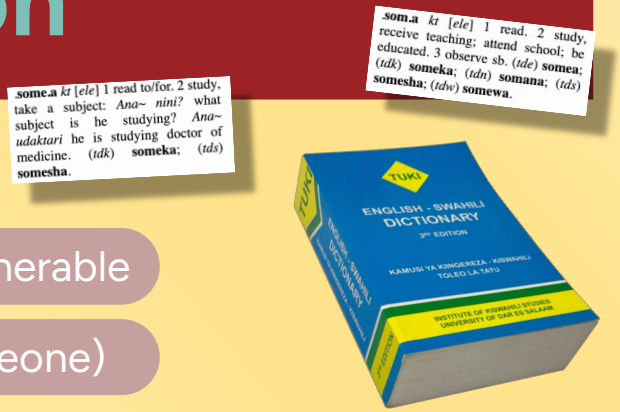
most word-formation is done by suffixation

soma (verb) eng def: read

somea: study

someka: be decipherable

somesha: teach, pay tuition for (someone)



some word-formation is done by prefixation

masomo: studies

msomeshaji: instructor

wasomi: scholars

examples of derivational families (some frequent forms)

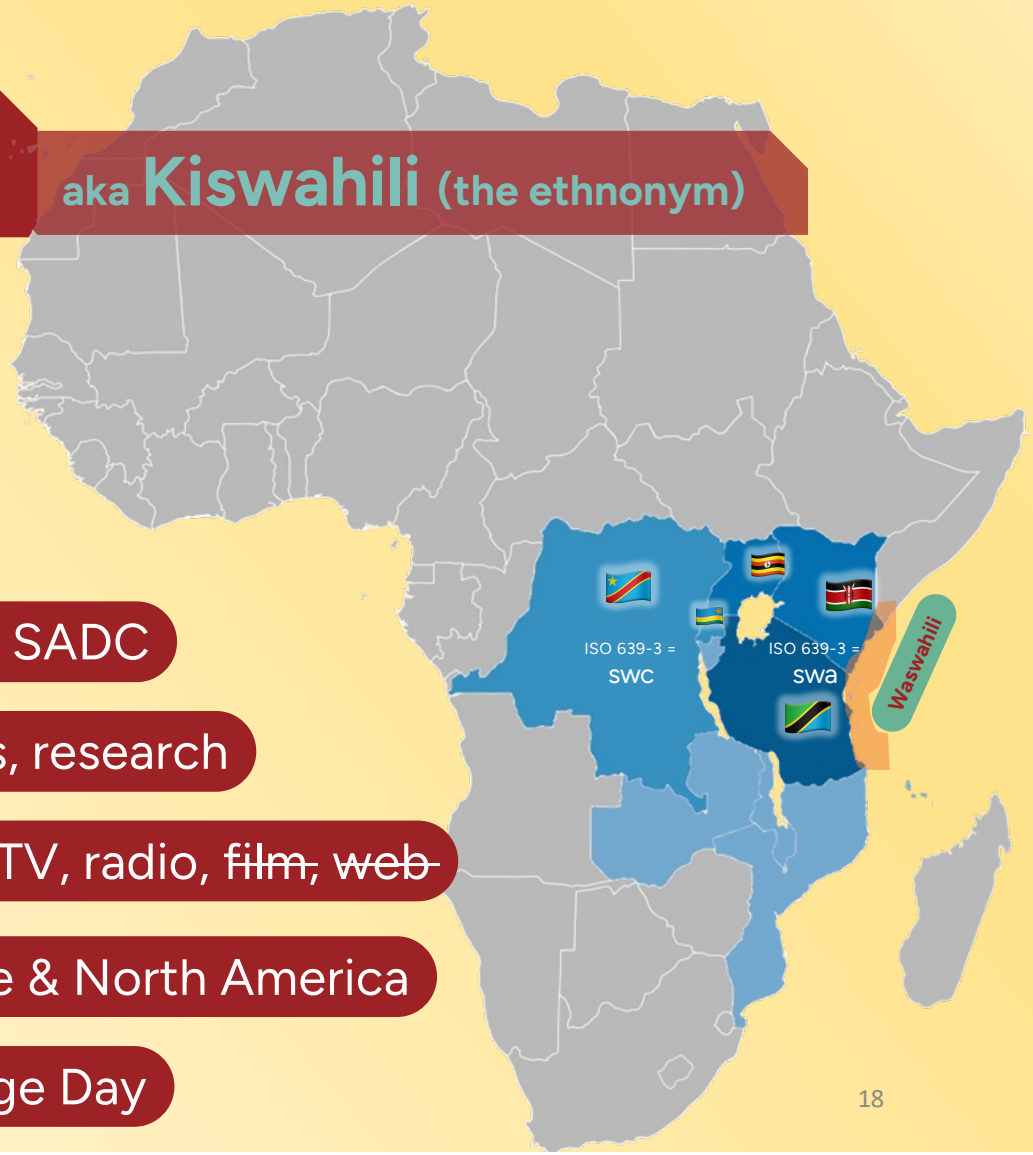
anga, angama, angika, angua, anguka, angukia, angukiana, anguko, angusha, jiangusha, angulia, kiangushio, mwangushi, mwanguko

tuma, tumia, tumika, tumikia, tumikisha, tumikiana, tumisha, tumilia, tumana, mtume, mtumishi, mtumwa, mtumwaji, tume, utume, utumi, utumishi, utumwa, matumizi, matumishi

Swahili Language

aka **Kiswahili** (the ethnonym)

- **mother tongue** of millions
- **daily language** of tens of millions
- **trade language** of 100 – 200,000,000
- **official language** of African Union, EAC, & SADC
- grammars, dictionaries, teaching materials, research
- poetry (400 years), literature, newspapers, TV, radio, ~~film~~, ~~web~~
- taught at many universities in Europe & North America
- July 7th is UNESCO World Kiswahili Language Day



Resources for Swahili

Dictionaries ("kamusi" in Swahili)

- Kamusi GOLD (Global Online Living Dictionary) – 60,000+ entries (currently offline)
- Africanlanguages.com – 15,000 entries
- FreeDict – 1,500 entries
- Africanlanguages.com – 15,000 entries
- Swahili-Polish dictionary – 7,000 entries
- TUKI (University of Dar es Salaam)– semi-available online

Resources for Swahili

Corpora

ACALAN Kiswahili Corpus - NEW a small collection of literary texts; unknown size, unknown sources

Helsinki Corpus of Swahili - about 25 million words available in the Language Bank of Finland

- The Not Annotated Version is available for download
- The Annotated Version is available in the concordance service
- The corpus is divided into subcorpora
  - Old material (~1952-2003)
  - New material (2004–2015)

Global Voices – used as UD development corpus

Various new datasets for AI and Speech Recognition

Resources for Swahili

Tools

AINI - morphological analyser (1989), produced the lemma list of a text with the indication of the main morphological categories; not available anymore

SALAMA Swahili Language Manager made of SWATWOL a two-level morphological analyser and SWACGP a morphological disambiguator and syntactic mapper; not freely available

Segmenter for Swahili verbs - based on the FreeDict Swahili-English dictionary

Swahili Wikipedia – 80,000 articles, but very problematic

Resources for Swahili

Language Boards

- National Swahili Council of Tanzania - BAKITA
- Chama cha Kiswahili cha Taifa Kenya - CHAKITA
- Chama cha Kiswahili cha Afrika Mashariki - CHAKAMA
- Chama cha Kiswahili cha Afrika - CHAKA
- Chama cha Wanafunzi wa Kiswahili Afrika Mashariki - CHAWAKAMA
- East African Kiswahili Commision - EAKC
- Baraza la Kiswahili la Zanzibar - BAKIZA
- Baraza la Kiswahii la Kenya - BAKIKE
- Baraza la Kiswahili la Uganda - BAKIU

the African Language Tech space

Academic

- SADiLaR
- Masakhane
- African Universities – Nairobi, Addis Ababa. Accra
- & Universities – Random researchers

Corporate

- Google – L10n Translate Bard AI
- Microsoft – Office L10n Bing MT + AI
- Meta – Facebook L10n SeamlessMT (AI)
- Small African endeavors

Philanthropic

- Lacuna Fund – AI focus
- FAIR Forward – AI focus
- IDRC - defunct
- & – non-existent

kamusi.org

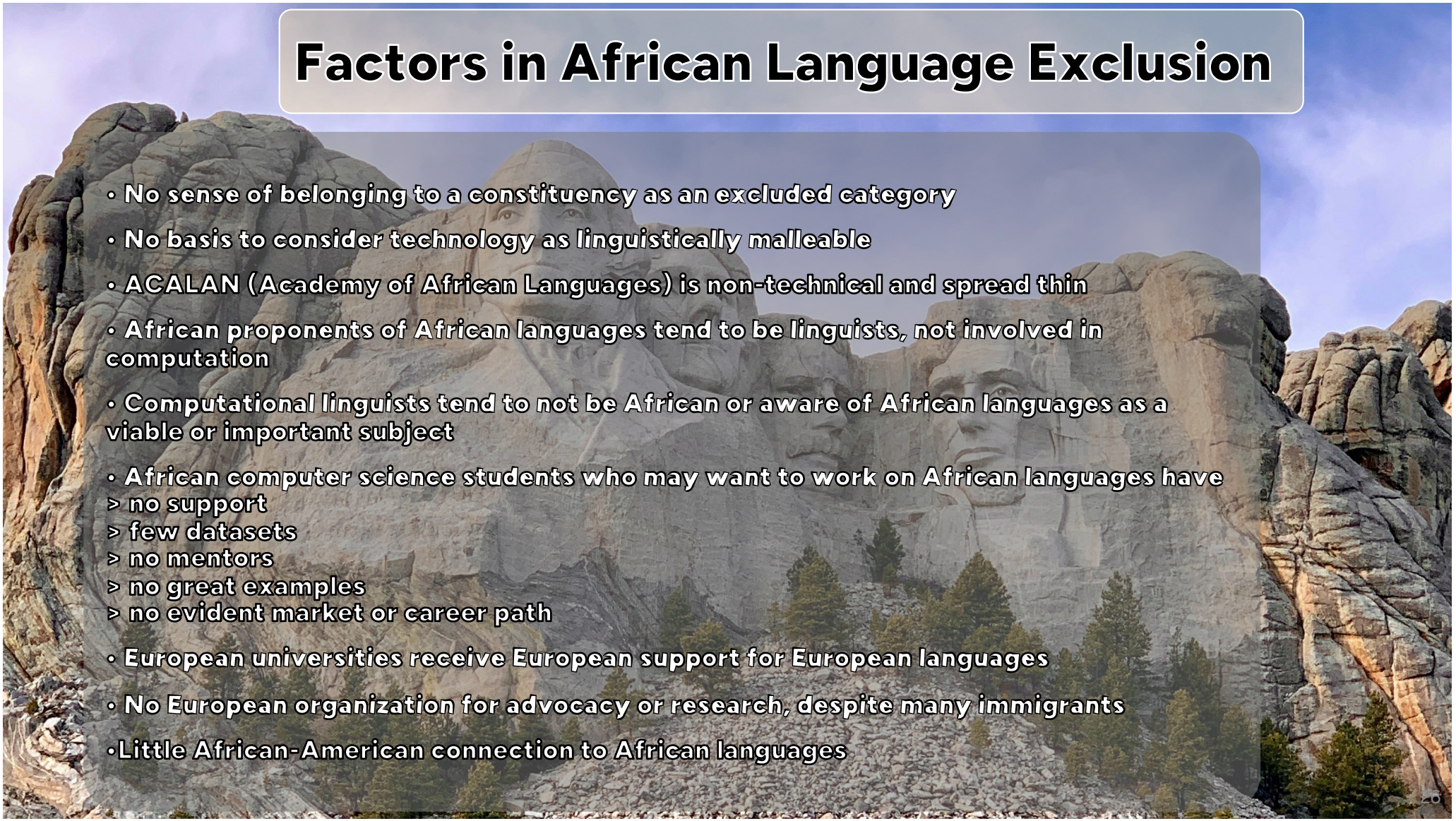
- Unfunded NGO

Resources for African Languages: Sparse and Haphazard

Examples:

- Information structure in African languages by Christian Chiarcos, et. al (2009)
- SADiLaR – South African Digital Language Resource Centre
  > Fifth workshop on Resources for African Indigenous Language (RAIL), Colocated with LREC-COLING 2024
- PanSALB – Pan South African Language Board (11 languages)
- South African national lexicography units
- Webonary.org (SIL) – bilingual stand-alone dictionaries for 100+ small languages – e.g. 8500 Lugwere-English words
- AfLAT.org – online resource center, last updated 2019
- ANLoc – African Network for Localisation – defunct
- Dozens more resources listed by Columbia University Libraries (many listings are defunct)
  > http://kamu.si/columbia-african-language-resources

No systematic data for NLP


Factors in African Language Exclusion
• No sense of belonging to a constituency as an excluded category
• No basis to consider technology as linguistically malleable
• ACALAN (Academy of African Languages) is non-technical and spread thin
• African proponents of African languages tend to be linguists, not involved in computation
• Computational linguists tend to not be African or aware of African languages as a viable or important subject
• African computer science students who may want to work on African languages have
> no support
> few datasets
> no mentors
> no great examples
> no evident market or career path
• European universities receive European support for European languages
• No European organization for advocacy or research, despite many immigrants
•Little African-American connection to African languages

Random Swahili Wikipedia Page

Full Article With Local Interest

WIKIPEDIA
Kamusi elezo huru

Kutafuta | Tafuta

Matatu

Lugha 9

Makala | Majadiliano | Soma | Hariri | Hariri chanzo | Fungua historia

Menyu kuu ficha

- Mwanzo
- Jumuiya
- Matukio ya hivi karibuni
- Mabadiliko ya karibuni
- Ukurasa wa bahati
- Msaada
- Michango
- **Switch to old look**

Lugha

Language links are at the top of the page across from the title.

Matatu ni neno la kawaida nchini Kenya na Uganda la kutaja gari aina ya daladala linalotumika kama usafiri wa umma.

Matatu mjini Nairobi mwaka 2002 (bado bila kanda ya njano) ikibeba watu wengi

Jina "matatu" labda lilitokana na nauli ya sarafu tatu iliyokuwa kawaida kwa safari moja matatu yalipoanza.

Matatu huhudumia njia maalumu yenye namba yake.

Dereva anafanya kazi yake pamoja na msaidizi anayeitwa "manamba" mwenye kazi ya kukusanya pesa na kufungua au kufunga mlango.

Mara nyingi ni magari ya Kijapan yenye ruhusa ya kubeba abiria 14. Hadi mwaka 2003 matatu hayakufuata utaratibu huu wala taratibu nyingine. Ilikuwa kawaida kuingiza zaidi ya watu 20 katika gari moja lenye viti 15. Maafisa wa polisi ya barabarani (trafiki) walizoea kupokea rushwa za matatu kila siku, hivyo walishindwa kuwaambia neno.

Serikali mpya ya Rais Mwai Kibaki iliweza kubadilisha hali hiyo tangu mwaka 2003, hasa waziri wa usafiri John Michuki aliyelazimisha wenye matatu kukubali utaratibu ambao haukuwa mpya lakini ulipuuzwa miaka mingi hadi kusahauliwa isipokuwa wazee walikumbuka siku za kale.

Tangu mwanzo wa mwaka 2004 matatu yanatakiwa kuonyesha nje kanda ya rangi ya njano; dereva na manamba wawe na sare, kila kiti kinapaswa kuwa na kanda, na matatu kama mabasi yanatakiwa kuwa na "spidigavana" yaani

WIKIPEDIA
Kamusi elezo huru

Menyu kuu ficha

- Mwanzo
- Jumuiya
- Matukio ya hivi karibuni
- Mabadiliko ya karibuni
- Ukurasa wa bahati
- Msaada
- Michango
- **Switch to old look**

Lugha

Language links are at the top of the page across from the title.

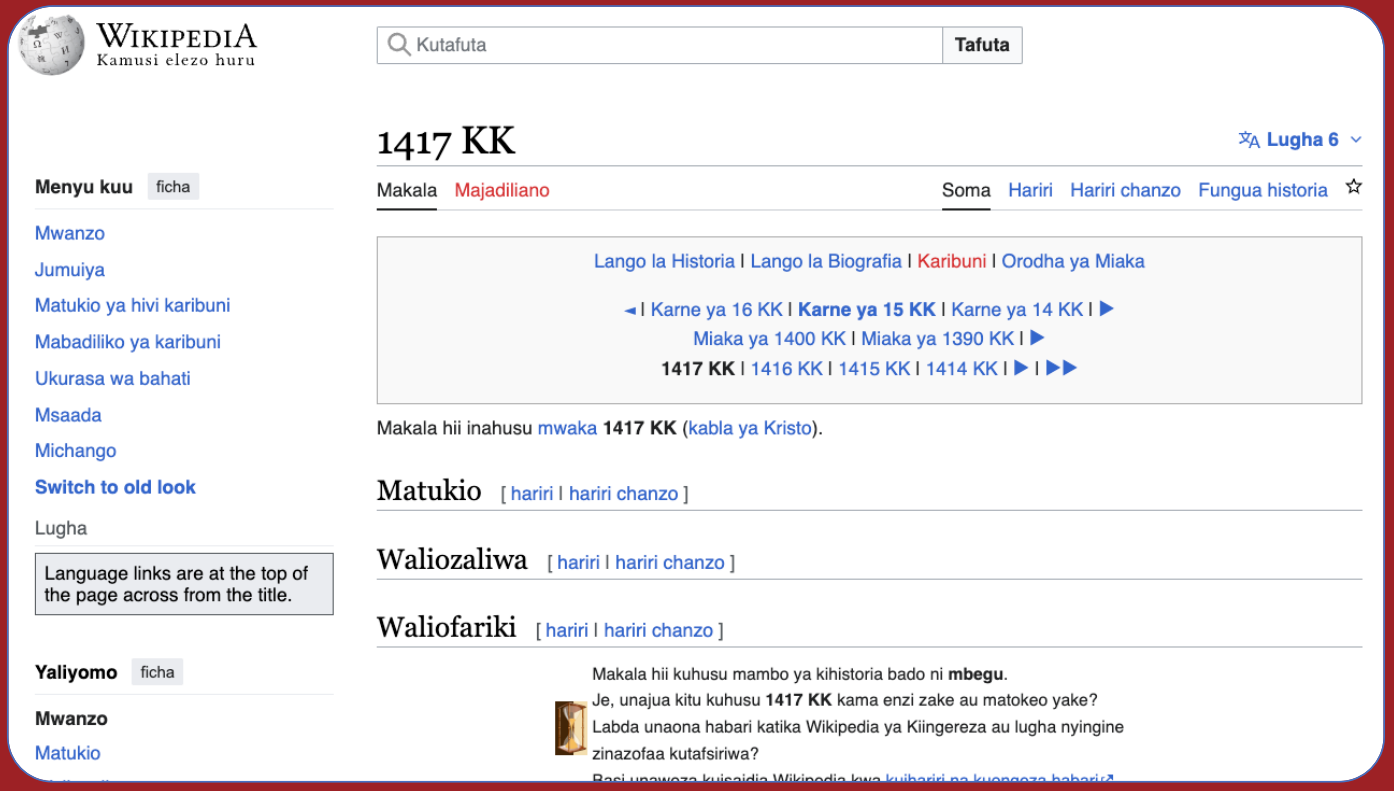
Yaliyomo ficha

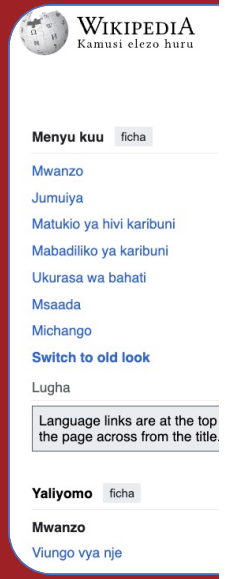
Mwanzo

Matukio

Random Swahili Wikipedia Page

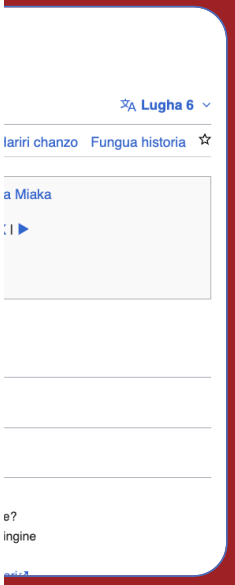
Random Year, No Content





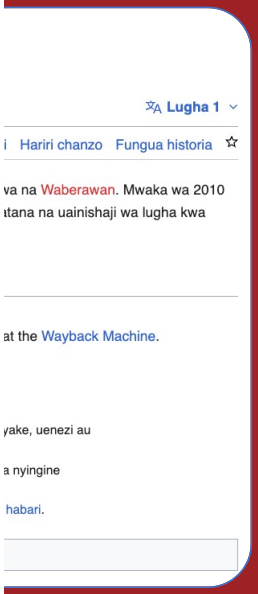
Random Swahili Wikipedia Page

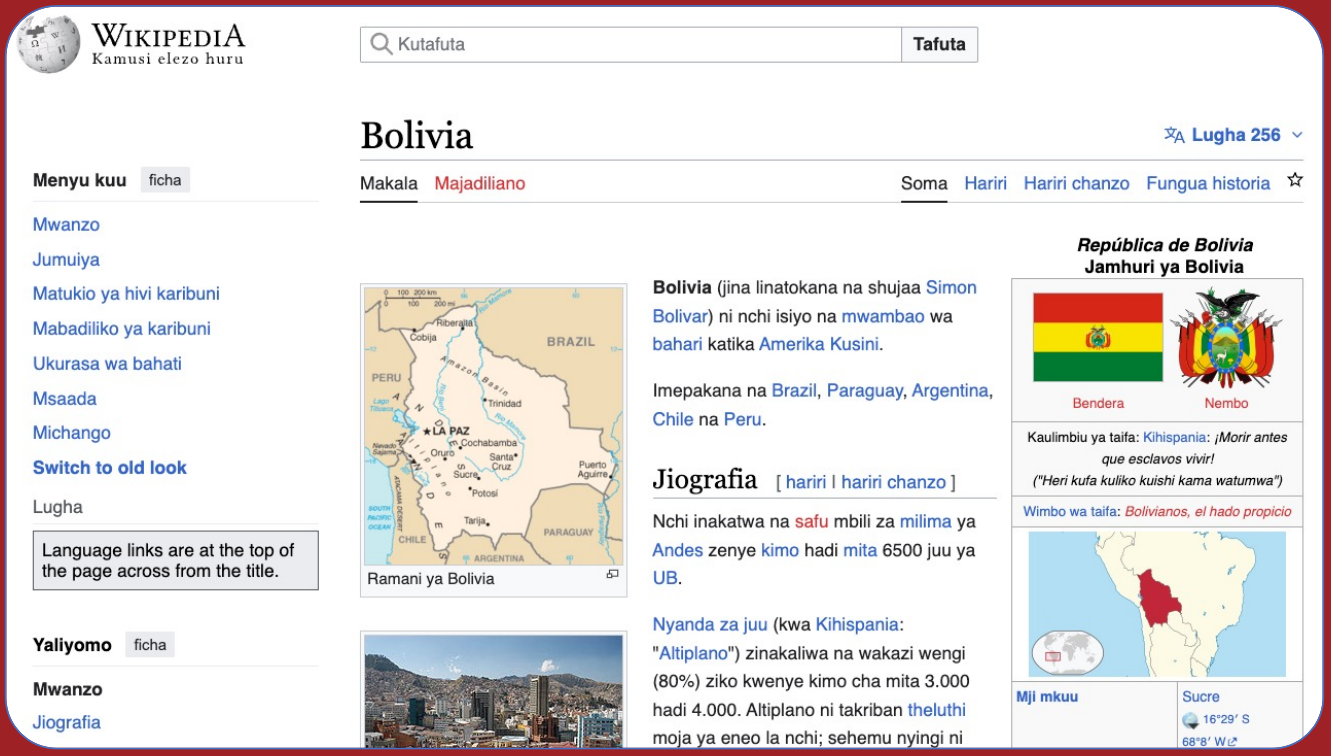
Language in Malaysia

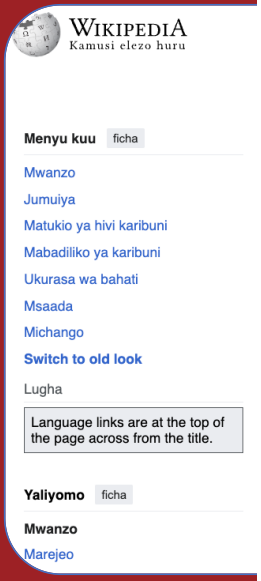


Random Swahili Wikipedia Page

Full Article With General Interest

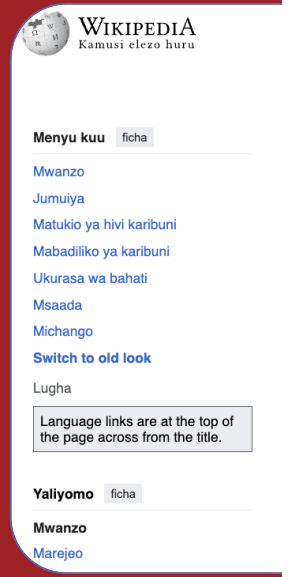






Random Swahili Wikipedia Page

Random German Footballer



Random Swahili Wikipedia Page

Random Small Organization Far Away

WIKIPEDIA
Kamusi elezo huru

Kutafuta
Tafuta

Eswatini Girl Guides Association

Lugha 2

Makala | Majadiliano | Soma | Hariri | Hariri chanzo | Fungua historia

Menyu kuu ficha

- Mwanzo
- Jumuiya
- Matukio ya hivi karibuni
- Mabadiliko ya karibuni
- Ukurasa wa bahati
- Msaada
- Michango
- **Switch to old look**

Lugha

Language links are at the top of the page across from the title.

Yaliyomo ficha

- **Mwanzo**
- Marejeo

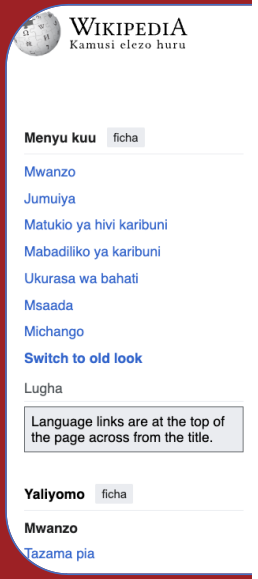
Eswatini Girl Guides Association (zamani ilijulikana kama **Swaziland Girl Guides Association**[1]) ni shirika la kitaifa la Eswatini (Swaziland). linahudumia wanachama 3,450 (kufikia 2017) [2] (wanachama 1,010 mnamo 2003, 1,836 mnamo 1972).[1]) Ilianzishwa mnamo 1924 kama shirika la wasichana pekee.[2] Kiongozi Mkuu wa Guide Olave Baden Powell alitembelea mwaka wa 1950 likawa Tawi la chama cha The Guide Association nchini Uingereza.[2] Lilikuja kuwa shirika linalojitawala mwaka wa 1969 pindi lilipokuwa mwanachama mshiriki wa World Association of Girl Guides and Girl Scouts mwaka wa 1969.[2]

Shirika hili liliunganishwa na "The Wayfareres" shirika linalofanana.[2] [3]

Kuna sehemu tatu za vijana. Blue Jays walikuwa asili, ikifuatiwa na sehemu ya Ranger mwaka wa 1963, na Cadets mwaka wa 1968,[2] na ilibadilishwa jina kufuatia kubadilishwa jina kwa nchi katika 2018.

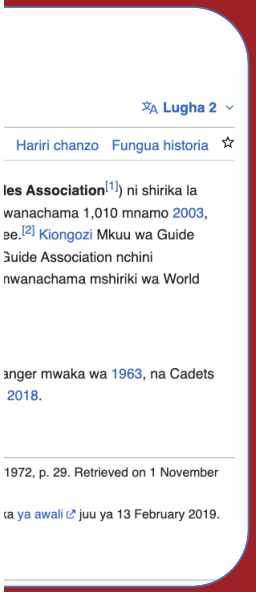
Marejeo [hariri | hariri chanzo]

1. ↑ 1.0 1.1 "Teach Swaziland girls in Guide program", *The Ottawa Journal*, 16 June 1972, p. 29. Retrieved on 1 November 2019.
2. ↑ 2.0 2.1 2.2 2.3 2.4 2.5 The Swaziland Girl Guides Association (2019). Jalada kutoka ya awali juu ya 13 February 2019. Iliwekwa mnamo 1 November 2019.
3. ↑ https://www.wagggs.org/en/our-world/africa/member-organisations/swaziland/



Random Swahili Wikipedia Page

Village in East Africa Exists


Lugha 2

Hariri chanzo Fungua historia

les Association[1]) ni shirika la wanachama 1,010 mnamo 2003, ee.[2] Kiongozi Mkuu wa Guide Guide Association nchini nwanachama mshiriki wa World

anger mwaka wa 1963, na Cadets 2018.

1972, p. 29. Retrieved on 1 November

ta ya awali juu ya 13 February 2019.

WIKIPEDIA
Kamusi elezo huru

Kutafuta Tafuta

Menyu kuu ficha

Mwanzo
Jumuiya
Matukio ya hivi karibuni
Mabadiliko ya karibuni
Ukurasa wa bahati
Msaada
Michango
Switch to old look

Lugha

Language links are at the top of the page across from the title.

Yaliyomo ficha

Mwanzo
Tazama pia

Miti ya Hunter

Lugha 3

Makala Majadiliano Soma Hariri Hariri chanzo Fungua historia

Miti ya Hunter ni kijiji cha Kenya katika kaunti ya Nandi[1].

Tazama pia [hariri | hariri chanzo]

- Orodha ya miji ya Kenya

Tanbihi [hariri | hariri chanzo]

1. ↑ https://www.knbs.or.ke/?wpdmpro=2019-kenya-population-and-housing-census-volume-ii-distribution-of-population-by-administrative-units

v·d·e	Kaunti za Kenya na miji yake mikubwa [Kunja]
Kaunti 47 (tangu Machi 2013)	Baringo (30) · Bomet (36) · Bungoma (39) · Busia (40) · Elgeyo-Marakwet (28) · Embu (14) · Garissa (07) · Homa Bay (43) · Isiolo (11) · Kajiado (34) · Kakamega (37) · Kericho (35) · Kiambu (22) · Kilifi (03) · Kirinyaga (20) · Kisii (45) · Kisumu (42) · Kitui (15) · Kwale (02) · Laikipia (31) · Lamu (05) · Machakos (16) · Makueni (17) · Mandera (09) · Marsabit (10) · Meru (12) · Migori (44) · Mombasa (01) · Murang'a (21) · Nairobi (47) · Nakuru (32) · Nandi (29) · Narok (33) · Nyamira (46) · Nyandarua (18) · Nyeri (19) · Samburu (25) · Siaya (41) · Taita-Taveta (06) · Tana River (04) · Tharaka-Nithi (13) · Trans-Nzoia (26) · Turkana (23) · Uasin Gishu (27) · Vihiga (38) · Wajir (08) · West Pokot (24)
	Nairobi *(mji mkuu)* · Ahero · Athi River · Awasi · Bomet · Bondo · Bungoma · Burnt Forest · Busia · Chuka ·

WIKIPEDIA
Kamusi elezo huru

Menyu kuu ficha

Mwanzo
Jumuiya
Matukio ya hivi karibuni
Mabadiliko ya karibuni
Ukurasa wa bahati
Msaada
Michango
Switch to old look

Lugha

Language links are at the top of the page across from the title.

Yaliyomo ficha

Mwanzo
Tazama pia

Random Swahili Wikipedia Page

Body of Water

WIKIPEDIA
Kamusi elezo huru
Menyu kuu ficha
Mwanzo
Jumuiya
Matukio ya hivi karibuni
Mabadiliko ya karibuni
Ukurasa wa bahati
Msaada
Michango
Switch to old look
Lugha
Language links are at the top of the page across from the title.

Random Swahili Wikipedia Page

Municipality Somewhere

WIKIPEDIA
Kamusi elezo huru

Kutafuta | Tafuta

Sandwich, Massachusetts

Lugha 32

Makala | Majadiliano | Soma | Hariri | Hariri chanzo | Fungua historia

Menyu kuu ficha

- Mwanzo
- Jumuiya
- Matukio ya hivi karibuni
- Mabadiliko ya karibuni
- Ukurasa wa bahati
- Msaada
- Michango
- **Switch to old look**

Lugha

Language links are at the top of the page across from the title.

Sandwich ni mji wa Marekani katika jimbo la Massachusetts. Kwa mujibu wa sensa iliyofanyika mnamo mwaka wa 2000, mji una wakazi wapatao 20,000 wanaoishi katika mji huu. Mji upo m 6 kutoka juu ya usawa wa bahari. Eneo lake ni 115 km².

Sandwich, Massachusetts

Mahali pa mji wa Sandwich katika Marekani

Random Swahili Wikipedia Page

A Long Dead French Monk

Lugha 32

ariri chanzo Fungua historia

Sandwich, Massachusetts

Sandwich

a mji wa Sandwich katika Marekani

WIKIPEDIA
Kamusi elezo huru

Kutafuta | Tafuta

Menyu kuu ficha

- Mwanzo
- Jumuiya
- Matukio ya hivi karibuni
- Mabadiliko ya karibuni
- Ukurasa wa bahati
- Msaada
- Michango
- **Switch to old look**

Lugha

Language links are at the top of the page across from the title.

Yaliyomo ficha

Mwanzo

Tazama pia

Yusto wa Condat

Lugha 1

Makala | Majadiliano | Soma | Hariri | Hariri chanzo | Fungua historia

Yusto wa Condat alikuwa mmonaki wa monasteri ya Condat nchini Ufaransa[1].

Ingawa habari za maisha yake ni chache mno, tangu kale anaheshimiwa kama mtakatifu[2]. Heshima hiyo ilithibitishwa na Papa Pius X tarehe 9 Desemba 1903[3].

Sikukuu yake huadhimishwa tarehe 6 Julai[4].

Tazama pia [hariri | hariri chanzo]

- Watakatifu wa Agano la Kale
- Orodha ya Watakatifu Wakristo
- Orodha ya Watakatifu wa Afrika
- Orodha ya Watakatifu Wafransisko

Tanbihi [hariri | hariri chanzo]

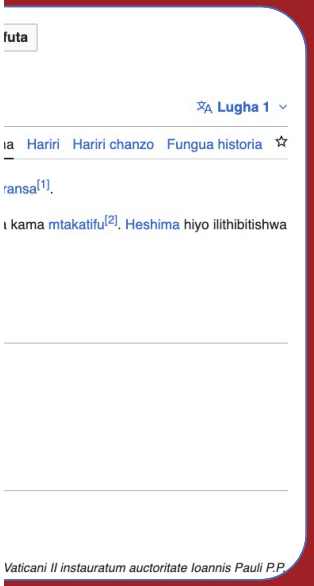
1. ↑ http://www.santiebeati.it/dettaglio/60940
2. ↑ Rombaut Van Doren, BSS, vol. VII (1966), col. 29.
3. ↑ *Index ac status causarum* (1999), pp. 464 e 599.
4. ↑ *Martyrologium Romanum: ex Decreto Sacrosancti oecumenici Concilii Vaticani II instauratum auctoritate Ioannis Pauli P.P.*

- **A use**
- **A relia**

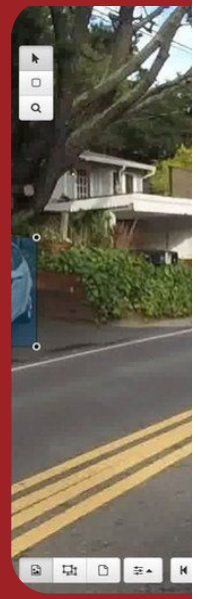
Informa

- **Writt**
- **Funde**

The Swahili Wikipedia is NOT:



- **A useful source of Linguistic Data**
- **A reliable source of Relevant Information**
- **Written primarily by native speakers**
- **Funded**



What Data Does Google Fund in Kenya?

- Autonomous-Vehicle Safety Annotation
- $9/ worker per day
- Over 1000 Kenyan Annotators at Samasource
- Because Bad Data Can Kill People

African languages and AI

AI is central to any funding possibility

"Zero Shot" or "Low Shot" pretends data does not matter

Corrupt data from web harvesting:

- 80% of sampled Swahili web content came from Google Translate or similar (read Multilingual article - kamu.si/uyoga)
- Watch kamu.si/abba and kamu.si/federer to see Bing and Bard LLMs manipulated in Pirate Professor experiments

Foundation Support for African Languages

Gates

Total grants since inception = USD $80,000,000,000

Annual giving = USD $7,000,000,000

6049 "Africa" grants over 30 years

Mostly for health

A few for "literacy"

31 for AI

A few include translation

Zero for African language development

Ford

Annual giving = USD $715,000,000

3 "Africa" + "Language" grants ever

$50,000	**Bard College** For the Chinua Achebe Institute for Cultural Preservation's Igbo Dictionary Project to conduct ethnographic research to document, preserve and revitalize the language of the Nigeria's Igbo people	November 2010

No public result

$310,000	**University of Cape Town** For the Project for the Study of Alternative Education in South Africa to plan & document the process of establishing a pan-African graduate program in African languages and applied language studies	June 2007

$330,000	**Cairo University** To establish a Career Advisory Center that provides language, computer training and professional exposure for graduate and undergraduate students in the social sciences and arts	July 2009

Teaching Africans English

Ford Support for "Excluded Communities"

Ford Foundation

Work Worldwide Learning News & Stories

MORE THAN

$92M

OF OUR 2022 GRANTMAKING INTENTIONALLY INCLUDED PEOPLE WITH DISABILITIES

Living our values

To address the issues of a complex—and increasingly diverse—world, we believe those individuals and institutions closest to the challenges are also the most capable of creating solutions. We make a concerted effort to support organizations led by women, people of color and the LGTBQ+ community as they drive lasting change for historically excluded communities.

54%

OF U.S. GRANTS WENT TO ORGANIZATIONS LED BY PEOPLE OF COLOR

58%

OF U.S. GRANTS WENT TO WOMEN-LED ORGANIZATIONS

$45M

IN GRANTMAKING INTENTIONALLY INCLUDING THE LGBTQ+ COMMUNITY

“Theory of Change”

Demanded by European foundation that exists because of African cacao money

European theory of language within the “Mission Civilisatrice” / “the White Man’s burden” has been in place since 1884

tested on generations of African students

failed but ongoing

Our theory:

Students learn better in languages they understand

People benefit from technology only if it speaks their language

It is easier to teach 100 African languages to computers than to expect a billion people to learn European colonial languages

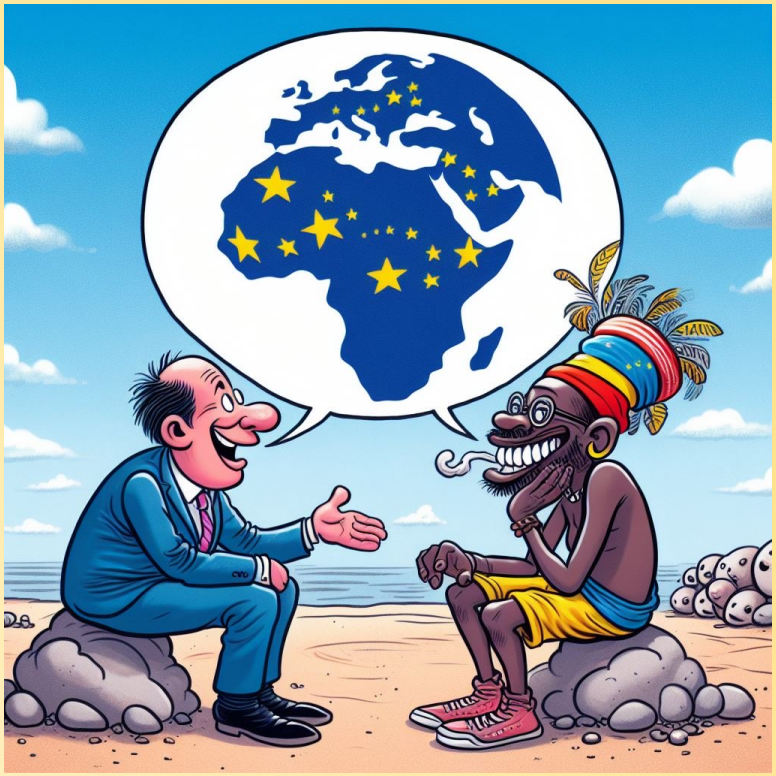
Europe vis-à-vis Digital Language Technologies

Own Languages!

- Digital Belgium
- National Digitalisation Council
- Digital Finland Framework
- GIZ
- Nederland Digitaal
- Norwegian Digitalisation Agency
- España Digital
- National Digitalisation Council
- Digital Switzerland
- Digital Access Program

"Digital Inclusion" ≠ Languages

Support for Africa?



Generated by Microsoft Bing CoPilot: "A conversation where Europe is talking and Africa is listening". 28 January 2024

Europe vis-à-vis Digital Language Technologies

Our Africa Priorities

- Agriculture and food security
- Health
- Fragility, conflict, and human rights
- Migration
- Good governance
- Gender equality
- Disaster preparation and emergencies
- Climate change and environment
- Water
- Basic education and vocational skills
- Private sector development and finance

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

African Language Tech

- Solutions are outside scope of expertise
- Money is finite
- Google/ Microsoft/ AI has this covered
- Maybe a few K for some endangered language?
- African languages are too primative for science and tech – only useful at home
- Calling out colonialism here is impolite

Hi Africa, how are you doing?
I'm doing well, thank you. How about you?
I'm doing great too. I heard that we have a long history of trade and cultural exchange. That's amazing!
Yes it is. We're grateful for the support that Europe has provided us over the years.
Text and illustrations generated by Microsoft's Bing CoPilot
"Draw a cartoon of a conversation between Africa and Europe"
It's been our pleasure to work with you over the years.
We hope to continue this partnership for many more years to come.
We feel the same way. Thank you, Europe.
You're welcome, Africa. Let's keep up the good work together!

http://kamu.si/pale-acalan

Platform for African Language Empowerment

Linguistic Data Component

- Who PALE is
- Why PALE is needed
- What PALE is
- African Linguistic Data at PALE's core

Who PALE is

- ACALAN: African Academy of Languages (African Union intergovernmental language organ)
- 20 Vehicular Cross-Border Language commissions
- Academic institutions (eg University of Nairobi)
- Industry partners
- Civil Society organizations including Kamusi Project International (NGO for linguistic data and technology)

Why PALE is needed

- Africa has 1,400,000,000 people speaking over 2,000 languages
- Technology resources are scarce and scattered, for big languages and endangered languages alike
- Exclusion from language techology means exclusion from global systems of knowledge and prosperity
- An African Languages platform will pull together partners, technologies, and linguistic knowledge to promote economic and social development across the continent

What PALE is

The platform stands on 4 legs:

1. **ACALAN information**, including language policies from 55 AU member nations, activities of VCBL commissions
2. **Repository** of information from external organizations. This will resemble the European Language Grid
3. **Hub** for building and sharing African linguistic data and software by PALE partners, ultimately for all 2000 languages
4. **Communications center** for research and development on African languages

The hub (the third leg) has components including:

- Linguistic Atlas of Africa – info about each of 2000 languages
- Corpora – systematically digitized texts
- Archives – other scanned texts (PDFs)
- Language technology software (speech recognition, spell-check, etc)
- Localized software for non-linguistic purposes
- Dictionaries and other linguistic data

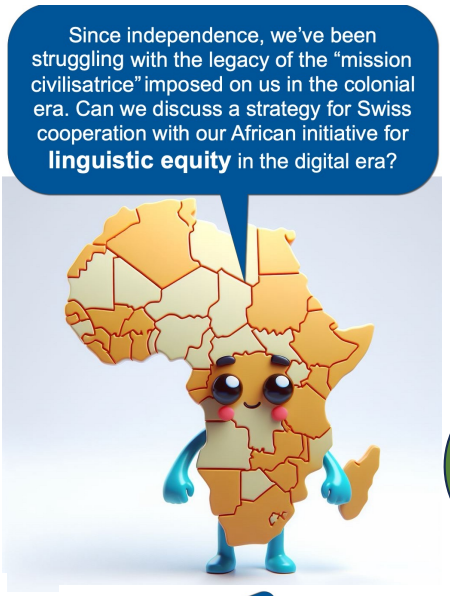
African stool by unnamed artist: kamu.si/acalan-stool

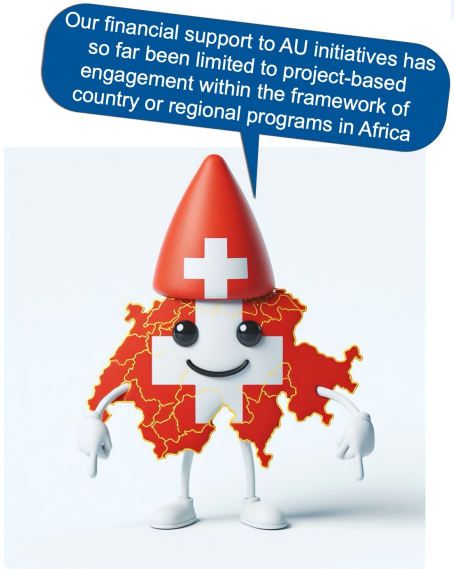
African Linguistic Data at PALE's core

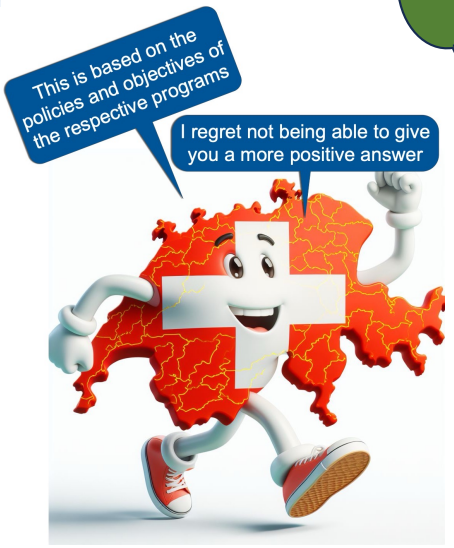
Kamusi gathers and disseminates linguistic data across languages. Kamusi systems will be central to PALE's data activities

- Fine-grained data with rich information, including inflected forms (kamu.si/kam4d)
- Semantic translation links among all languages
- Tools for language experts to contribute data
- Tools and games for crowdsourcing validatable linguistic information
- Mobile and web applications for users to access data around the continent
- Future-facing data system for technoloogies such as machine translation and speech processing (kamu.si/big-picture-playbook)

Point of contact:
Kamusi Project International · Rue de la Blancherie 22 · Chavannes-près-Renens 1022 Switzerland · +41.78.904.94.05 · fax +41.21.888.16.06 · kpi@kamusi.org







Illustrations generated by Microsoft's Bing CoPilot

Text is quoted from real correspondence with the Swiss Development Cooperation agency (SDC)

Most other European agencies offer no response

What's hard about African Language Technology?

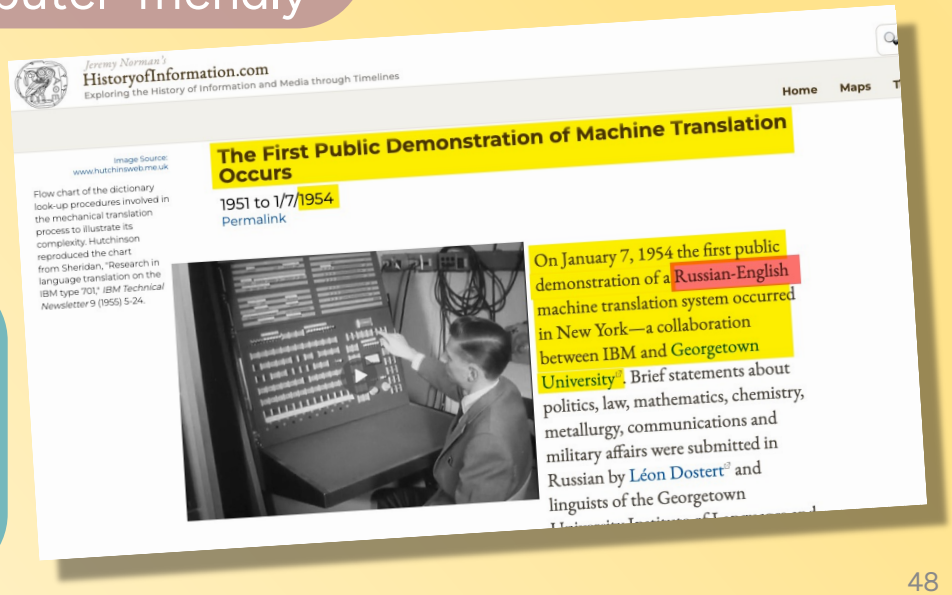
Getting Data (from people 🧠, not corpora or the ether 🤖): systems built @ Kamusi

Diagramming the language (e.g. UDS):
Bantu languages are extremely regular, computer-friendly

Convincing the world it CAN be done

Convincing the world it SHOULD be done

MONEY:
- European governments won't pay
- Philanthropies and foundations won't pay
- Africans can't pay



Concluding Thoughts

- African languages overall are not dying, they are growing
- English and French are not taking over
- African languages are not too difficult for NLP, technically
- The difficult barriers and socio-political
- Ordinary people do not know language equity is possible
- Policy and grantmakers live in an elite language bubble
- “Diversity” is not a widely-shared concern outside a few chambers like UniDive
- Until we address the indifference and inertia, nothing will change

Asanteni!

Presentation for Download:
kamu.si/unidive-africa-pdf
kamu.si/unidive-africa-ppt

UniDive