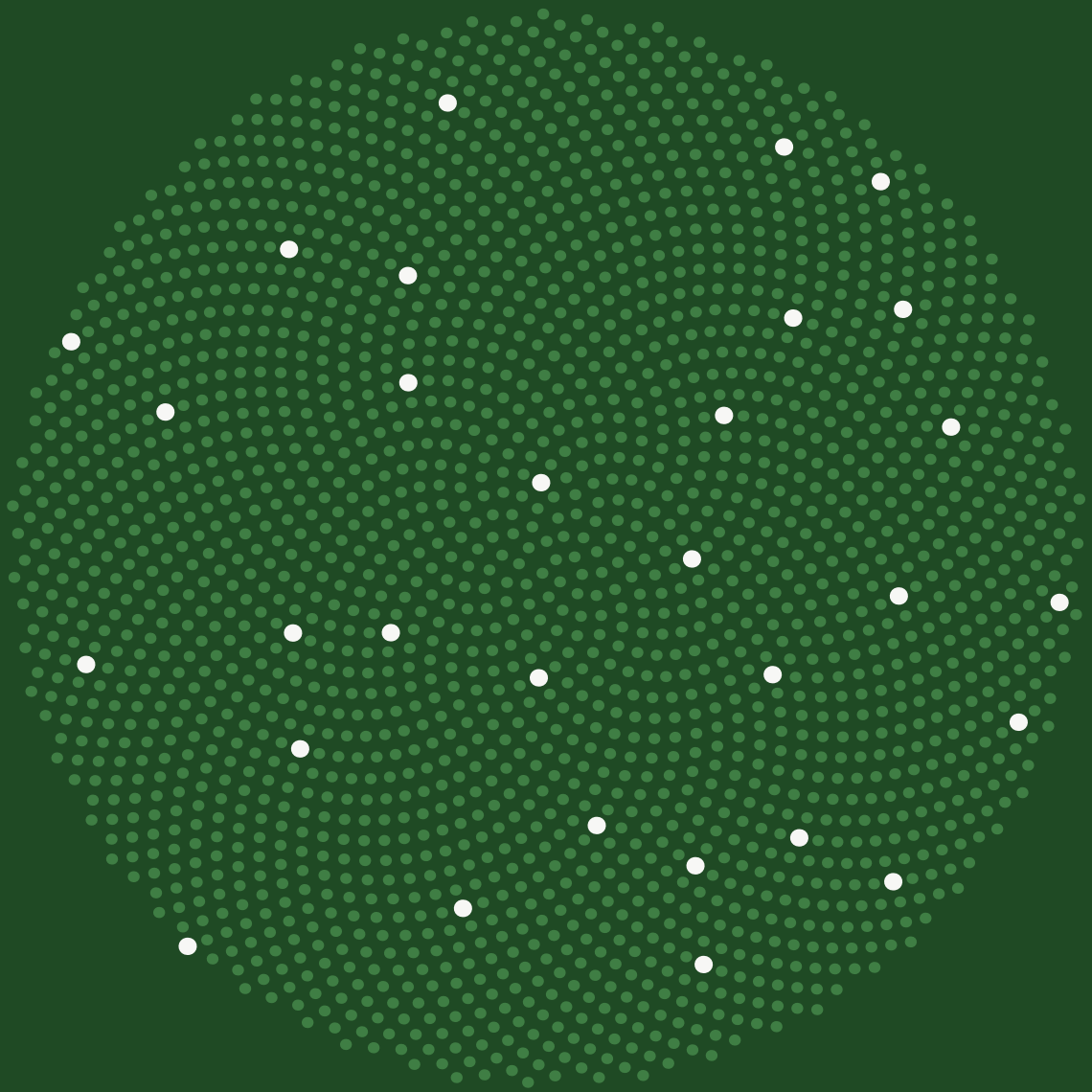
Teaching the machine Africa's languages

The gap lies in the data, and data can be built



2,213 languages in the Glottolog 5.4 Africa macro-area, one dot per language

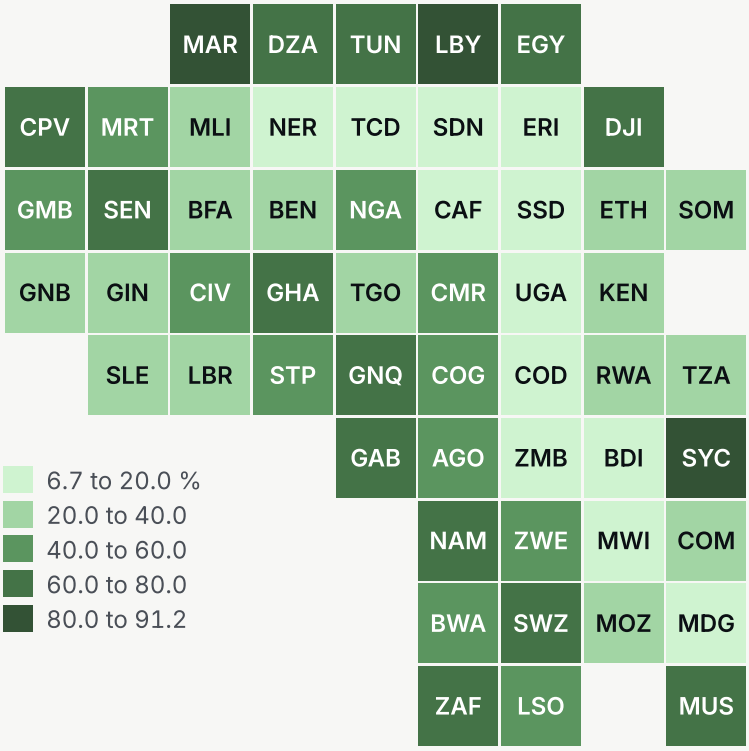
30 African languages recognised by Common Crawl's CLD2 detector: together, 0.053% of the pages in the CC-MAIN-2026-43 archive; English, 42.56%

ሀ ለ ሐ መ ሠ ረ ሰ ⴰ ⵀ ⵣ ⵍ ⴻ ⴼ ⵣ ⵄ ⵋ ⵛ ⵏ ⵟ ⵢ ⵕ ⵖ ⵜ ⵡ ⵅ ⵙ ⵃ ⵊ ⵥ ⵚ ⵎ ⵓ ⵔ ⵗ ⵞ ⵘ ⵉ ⵇ ⵍ ⵒ ⵝ ⵤ ⵓ ⵏ ⵕ ⵜ ⵖ

FIGURE 1

In 35 of 54 African countries, less than half the population uses the Internet

Share of the population that used the Internet in the last three months, %, latest year available (2024 for 51 countries, 2017 to 2020 for Sudan, South Sudan and Eritrea)



How to read: One square per country; the darker the shade, the higher the use. Nigeria reaches 41.2% in 2024, Ethiopia 21.9%, Kenya 35.0%.

Source: ITU, World Telecommunication/ICT Indicators Database, via the World Bank, World Development Indicators (updated 11 November 2026)

Planned permanent address: youthshareinitiatives.org/r/digital-palaver-ia-langues-africaines-en/carte-internet

Data available now: data.worldbank.org/indicator/IT.NET.USER.ZS

THE FINDINGS

01
Africa holds 2,213 of the 7,685 spoken languages recorded by Glottolog; in the ITU Africa region, 36% of people use the Internet in 2025.

02
In October 2026, thirty African languages account for 0.053% of the pages in Common Crawl, and English for 42.56%.

03
In the 2025 AfroBench evaluation, the four best closed models lose 13.2 to 19.8 points in African languages, and an open model up to 40.1.

04
Kinyarwanda has more validated hours of speech than French in Common Voice; Hausa has 4.35.

05
The African Union and UNESCO call for artificial intelligence that serves African languages; budgets and archives remain to be built.

1 The next users will speak their own languages

Africans who discover artificial intelligence in the coming years will meet it first in their own languages. These languages are many. Glottolog, the linguists' open reference database, records

7,685 languages spoken as a first language worldwide, 2,213 of them in its Africa macro-area.[1] Several count tens of millions of speakers, from Lingala (40 million) to Amharic (57 million) and Hausa (77 million); for Swahili, estimates range from 71 million to "more than 200 million".[2]

These speakers are going online fast, and very unevenly. According to the International Telecommunication Union (ITU), 66% of people aged 10 and over in its Africa region owned a mobile phone in 2024. In 2025, 36% of the region's population use the Internet, against a world average of 74%.[3] Country by country (**figure 1**), use ranges from 6.7% in South Sudan in 2019 to 91.2% in Morocco in 2024, and 35 of 54 countries remain under half their population.

Two gaps point to the next Internet users. Young people lead the way: in 2024, 53% of people aged 15 to 24 in the region used the Internet, against 34% of the rest of the population. The continent is young, with a median age of 19.0 years in 2023. Rural areas follow at a distance, with 23% of rural people online against 57% of urban dwellers.[4]

> Africans who discover artificial intelligence in the coming years will meet it first in their own languages.
>
> THE DIGITAL PALAVER, YOUTHSHARE INITIATIVES

The most serious objection fits in one sentence: Africans who are online already read and write English, French, Arabic or Portuguese, and digital services can make do with those languages. It fits the minority online today. It fits far less well the people who will join them. According to an estimate cited by UNESCO, up to 40% of the world's population lacks access to education in a language they speak or understand. The difficulty peaks where linguistic diversity is greatest, in sub-Saharan Africa as in Asia and the Pacific.[5]

Radio shows the other side. The continent's leading source of news, it informs 59% of citizens in 38 countries surveyed in 2024-2025 every day or a few times a week. It reaches women, rural residents, the poorest and the least educated better than any other medium; among 18- to 35-year-olds, social media already lead it, at 58% against 55%.[6] We draw a lesson from this for voice assistants. A technology that asks only that people listen and speak reaches those whom writing keeps at a distance, provided it speaks their language.

1 Glottolog 5.4 (November 2026), open data under a CC BY 4.0 licence: languoids at "language" level, excluding sign languages, pidgins and artificial, mixed, unattested or unclassifiable languages; our count matches the total of 7,685 published by Glottolog.

2 Adelani et al. (2025), table 9 (first- and second-language speakers); United Nations General Assembly, resolution 78/312 (2024). Mozilla Foundation (2021) gives an estimated 100 million Swahili speakers.

3 ITU, Facts and Figures 2025. The ITU Africa region comprises 44 sub-Saharan countries; the ITU counts the ten other African countries, from Mauritania to the Comoros, among the Arab States.

4 ITU (April 2025), 2024 data; UN, World Population Prospects 2024, processed by Our World in Data, for the median age.

5 UNESCO (2016), policy paper 24, after Walter and Benson (2012).

6 Afrobarometer, Dispatch 1173 (April 2026).

2 The Web that feeds the models speaks English first

A large language model learns what it reads, and much of what it reads comes from the Web. GPT-3, described in 2020, drew 60% of its training examples from Common Crawl, a public archive of the Web. In its corpus, 93% of the words were English.[7] Four years later, Meta's team describes for Llama 3 a mix in which "multilingual" data, those in languages other than English, make up about 8% of the tokens.[8]

The latest language count from Common Crawl, based on the October 2026 archive, measures this imbalance. Of 2.17 billion HTML pages, English occupies 923 million, or 42.56%, and French 4.76%. Swahili, the first African language on the list, has 215,652 pages, or 0.0099%. The thirty African languages that the detector recognises total 1,148,503 pages, 0.053% of the archive.[9] For every page in Swahili, the archive holds more than 4,000 in English.

> A large language model learns what it reads, and much of what it reads comes from the Web.
>
> THE DIGITAL PALAVER, YOUTHSHARE INITIATIVES

Sixteen African languages are compared in **table 1**, and demography explains only part of the pattern. Lingala, spoken by 40 million people, has 22 million characters of web text in the MADLAD-400 corpus. Kinyarwanda, spoken by a quarter as many people, has 749 million. Hausa and Swahili each exceed 120,000 Wikipedia articles; Oromo, a language of 37 million people, has 1,972. English dwarfs the rest: 9 trillion characters in MADLAD-400, more than 14,000 times the volume of Hausa.

METHOD

Measuring a language's place in the data

Common Crawl identifies the main language of each HTML page with the CLD2 detector. Thirty African languages appear in its count; Ewe, Fula and Bambara escape it. MADLAD-400 (2023) gathers documents drawn from Common Crawl, filtered and then checked by sampling: for each language, two of the authors inspected twenty documents. Wikipedia counts its content pages, short or long. Each measure is partial; their convergence grounds the finding.

7 Brown et al. (2020), table 2.2 and section 3.3.

8 Llama Team, AI @ Meta (2024), sections 3.1.1 and 3.1.2.

9 Common Crawl, CC-MAIN-2026-43 archive; sum of the pages of the thirty African languages recognised by CLD2, authors' calculation.

TABLE 1

The four lowest scores go to languages with fewer than 50 million characters of web text

Language	Speakers, millions	Web text, million characters	Wikipedia articles	Common Crawl pages	GPT-4o score (AfriMMLU)
Swahili	71-106	2,400	134,816	215,652	77.4
Kinyarwanda	10	749	11,273	71,437	64.2
Hausa	77	630	121,976	87,352	67.2
Amharic	57	509	15,751	90,407	59.8
isiXhosa	19	287	2,604	22,514	70.2
chiShona	11	266	11,897	23,961	67.6
isiZulu	27	257	13,030	35,948	68.2
Igbo	31	251	49,320	35,342	67.2
Yoruba	46	239	41,371	37,545	61.2
Sesotho	13	227	2,256	18,537	67.4
Oromo	37	88	1,972	14,256	61.0
Luganda	11	48	5,900	8,426	52.8
Ewe	7	33	1,438	-	33.6
Twi	9	25	4,749	6,613	43.2
Lingala	40	22	5,346	12,901	61.0
Wolof	5	5	1,736	4,708	37.8

How to read: First- and second-language speakers; web text in the MADLAD-400 corpus; GPT-4o score on AfriMMLU questions, out of 100, where chance would give 25. The Common Crawl detector recognises Akan, which includes Twi, and leaves Ewe aside.

Source: Adelani et al. (2025), IrokoBench, tables 3 and 9; Wikimedia, Wikipedia edition statistics, 20 November 2026; Common Crawl, CC-MAIN-2026-43 archive

Planned permanent address: youthshareinitiatives.org/r/digital-palaver-ia-langues-africaines-en/tab-langues

Data available now: aclanthology.org/2025.naacl-long.139

One objection holds that the Web reflects use: if Wolof has 4,708 pages in Common Crawl, Wolof speakers must write little online. Causality also runs the other way. A language that keyboards, spell-checkers and schools leave aside is written less on screen, and the detector itself recognises thirty African languages out of more than two thousand. Yet the archive exists beyond the Web, in manuscripts, newspapers and radio collections. Digitising it changes the count.

3 Models succeed where data abound

Evaluations published in 2025 confirm the link between data and performance, language by language. IrokoBench, a benchmark translated by professionals and designed by researchers from the Masakhane network, put sixteen large models through three tasks in seventeen African languages.[10] On knowledge questions, known as AfriMMLU, GPT-4o answers 87.4% of questions correctly in English, 83.2% in French, 77.4% in Swahili

10 Adelani et al. (2025), sections 3 and 4: logical inference, mathematics and knowledge questions; ten open and six proprietary models; averages over sixteen African languages.

and 33.6% in Ewe. Each question offers four answers: in Ewe, the model barely exceeds the 25% that chance would give. The last column of **table 1** gives all sixteen languages.

The order of languages broadly follows that of the data. The four lowest scores go to Ewe, Wolof, Twi and Luganda. These four are among the five languages with fewer than 50 million characters of web text. Swahili, the African language with the most web text, obtains the best African score. The authors themselves link the lowest performance to the languages with the least data online.[11] Lingala is a reminder that other factors play a part: 22 million characters, and a score of 61.0.

A second result weighs on daily use. Most models tested answer better when a machine first translates the question into English; GPT-4 Turbo and GPT-4o are among the rare exceptions. A service built on this detour imposes an intermediate language and depends on a machine translator whose quality also varies with data.

AfroBench extends the finding to 64 languages and fifteen tasks, with a light version of fourteen languages that tracks recent models (**figure 2**). The four best-ranked closed models score 79.7 to 85.1 points in English and 62.8 to 67.0 on average in African languages. Open-weight models, which African teams can adapt to their languages, show more scattered gaps: 26.9 to 40.1 points for the three in the figure, 10.3 to 33.2 for the others, which are smaller or weaker in English.[12]

0.053 %

of the HTML pages in the October 2026 Common Crawl archive belong to the thirty African languages its detector recognises

Source: Common Crawl, CC-MAIN-2026-43 archive

There remains the optimistic argument: each generation of models improves, and scale will eventually close the gap. The figures partly bear it out. In African languages, GPT-4.1, released in April 2025, outperforms GPT-4o, and Gemma 3 gains 9.1 points over Gemma 2. The AfroBench authors note this progress themselves, while judging it limited on knowledge tasks.[13]

The languages that machine translation has gained were won through targeted efforts. In 2022, Meta released NLLB-200, a translation model whose weights are freely available for non-commercial use and which, according to the company, supports 55 African languages “with high-quality results”. Widely used tools then covered fewer than 25. In June 2024, Google added 110 languages to its translator, about a quarter of them from Africa.[14] Scale helps. Attention to each language decides.

11 Same source, section 4.1 and table 9.

12 Ojo et al. (2025), table 8.

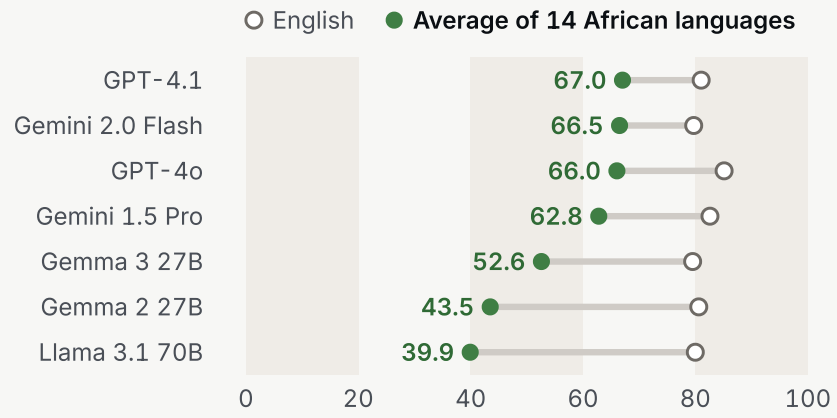
13 Ojo et al. (2025), table 8 and section 5.1; OpenAI, Introducing GPT-4.1 in the API (14 April 2025), for the release of GPT-4.1.

14 Meta AI (July 2022), weights under a CC BY-NC 4.0 licence; Google (27 June 2024).

FIGURE 2

Across seven models, the gap between English and African languages ranges from 13.2 to 40.1 points

Average score over seven tasks, out of 100, in English and in fourteen African languages, models evaluated in 2024 and 2025



How to read: The hollow circle gives the score in English, the solid dot the African average; Gemma and Llama are open-weight models.

Source: Ojo et al. (2025), AfroBench, Findings of ACL 2025, table 8

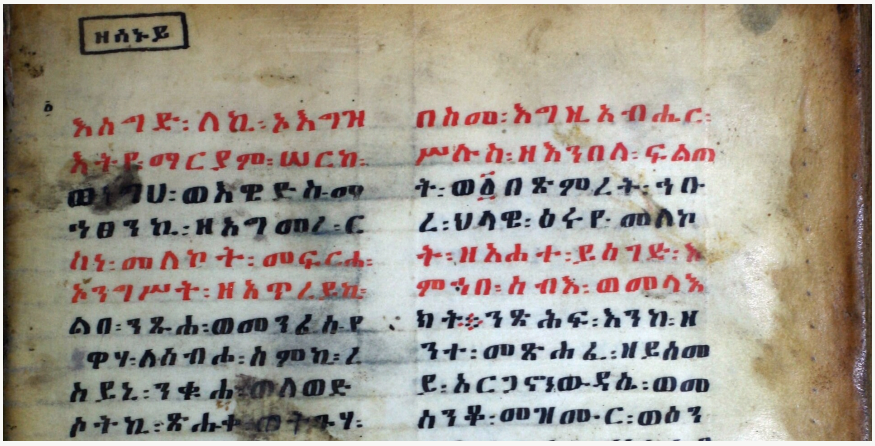
Planned permanent address: youthshareinitiatives.org/r/digital-palaver-ia-langues-africaines-en/fig-afrobench

Data available now: aclanthology.org/2025.findings-acl.976

4 Each new technology has had to learn Africa's languages

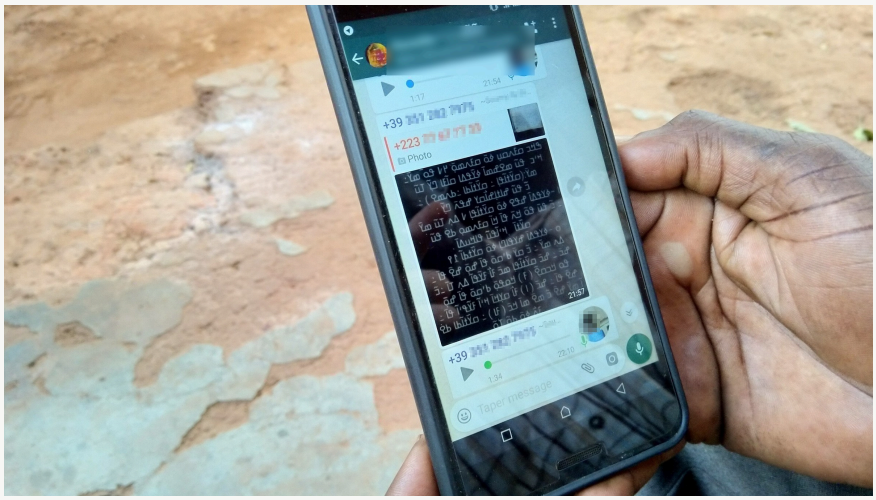
Today's gap extends a long history. Each communication technology has come to carry African languages once institutions and communities gave it the means. Writing offers the first chapter. The syllabary that evolved to write Ge'ez now serves to write Amharic and Tigre.

South of the Sahara, scholars wrote their languages in Arabic characters, in the form known as Ajami. A Boston University project funded by the British Library has just digitised 52,415



Detail of a page from a fifteenth-century Ethiopian prayer book, in Ge'ez script, in black and red ink.

Miami University Libraries, public domain, via Wikimedia Commons (cropped detail).



An N’Ko class on WhatsApp in 2019: the alphabet created in 1949 now travels on mobile phones.

Coleman Donaldson, CC BY-SA 4.0, via Wikimedia Commons (cropped and reduced, conversation header blurred).

pages of Arabic and Ajami manuscripts produced in Fuuta Jalon, in Guinea, and in Senegal. Their authors lived between the eighteenth and the twenty-first century.[15]

Other scripts were born on the continent for particular languages. Vai appeared in Liberia in the 1830s, Bamum in Cameroon between 1896 and 1910. N’Ko dates from 1949 for the Manding languages, Garay from 1961 for Wolof, Adlam from the late 1980s for Fula.

Each entered machines through the same door: its encoding in Unicode, the standard that assigns a number to every character. Ethiopic has been there since 1999, Tifinagh since 2005, N’Ko since 2006, Vai since 2008, Bamum since 2009, Adlam since 2016 and Garay since 2024.[16] N’Ko now has its own Wikipedia edition, with 1,622 articles.

Language policies weighed as much as technologies. As early as 1963, the Charter of the Organization of African Unity named African languages, “if possible”, among the working languages of the organisation, alongside English and French; Arabic and Portuguese were added later.[17] South Africa’s 1996 Constitution recognised eleven official languages, and South African Sign Language became the twelfth in 2026. The text acknowledges “the historically diminished use and status” of indigenous languages, which the state must elevate through “practical and positive measures”. Morocco made Tamazight an official language of the state in its 2011 Constitution.[18]

In 2003, UNESCO already recommended measures so that “all cultures can express themselves and have access to cyberspace in all languages, including indigenous ones”.[19] Artificial intelligence raises the question again on a larger scale.

15 Unicode Standard 18.0, chapter 19; Boston University, Readers in Ajami (27 May 2026), a project funded by the British Library’s Endangered Archives Programme.

16 Unicode Standard 18.0, chapter 19, and the DerivedAge file: versions 3.0 (1999), 4.1 (2005), 5.0 (2006), 5.1 (2008), 5.2 (2009), 9.0 (2016) and 16.0 (2024). N’Ko Wikipedia: 1,622 articles on 20 November 2026.

17 OAU Charter (25 May 1963), article XXIX, original text (United Nations Treaty Series, vol. 479); the African Union’s version adds Arabic and Portuguese.

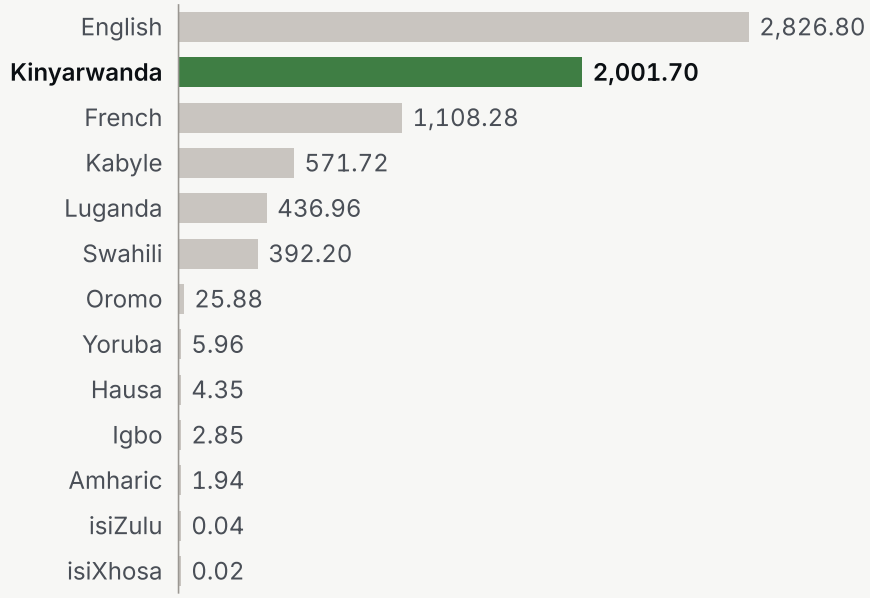
18 Constitution of the Republic of South Africa (1996), section 6; Constitution Eighteenth Amendment Act (2023), in force on 26 June 2026 (Government Gazette 54901); Constitution of the Kingdom of Morocco (2011), article 5.

19 UNESCO, recommendation adopted on 15 October 2003, at the 32nd session of the General Conference (29 September–17 October 2003).

FIGURE 3

Kinyarwanda has more validated hours of speech than French, while Hausa has 4.35

Hours of validated recordings in Common Voice, version 27.0, data cut-off 11 September 2026



How to read: Thirteen languages shown; a recording is validated when favourable votes prevail after at least two listens.

Source: Mozilla Common Voice, Scripted Speech 27.0, data cut-off 11 September 2026

Planned permanent address: youthshareinitiatives.org/r/digital-palaver-ia-langues-africaines-en/fig-common-voice

Data available now: github.com/common-voice/cv-dataset/blob/main/datasets/scripted-speech/cv-corpus-27.0-2026-09-11.json

One might object that this history also leans the other way. The European languages inherited from colonisation keep the school, the administration and print, and the market favours large languages. We read the trace of this legacy in the Web figures. History nonetheless traces a clear pattern: a language wins a technology when an institution or a community encodes it, teaches it or requires it. Technology supplies the tool. States, schools and communities of speakers decide its use.

5 Data can be built, and Africa has begun

The finding calls for an answer in data, and Africans are already building it. Kinyarwanda offers the clearest demonstration (figure 3). In version 27.0 of Common Voice, the speech database launched by Mozilla, it has 2,001.7 validated hours as of 11 September 2026, fourth among the database's 295 languages and ahead of French. Hausa, spoken by 77 million people, has 4.35.[20]

This lead has a history. After a hackathon held in Kigali by Mozilla and the German cooperation agency GIZ, the young Rwandan company Digital Umuganda mobilised contributors by drawing on umuganda, the Rwandan idea of community

20 Mozilla Common Voice, Scripted Speech 27.0: 29,295 validated hours in all; IrokoBench, table 9, for Hausa speakers.

cooperation. By September 2020, it had gathered 1,211 hours of Kinyarwanda from more than 420 people. It then developed Mbaza with GIZ and Rwanda's RISA agency, an assistant that delivered information on COVID-19 in Kinyarwanda through a phone menu, with its voice functions in preparation.[21] A public database, a local company, a cooperation partner and a collective tradition changed the scale of a language within a few years.

Other efforts are advancing across the continent. The Masakhane network aims to strengthen research on natural language processing for African languages "for Africans, by Africans": by 2020, it was publishing machine translation benchmarks for more than thirty languages. Lacuna Fund finances datasets that are locally developed and owned and open to all; 25 of the 31 datasets in its language catalogue name an African language or region.[22]

Data ownership raises the weightiest question. Voices and texts given by African volunteers will mainly feed the products of large foreign companies. The risk is real, and African legal scholars are answering it through law. The NOODL licence, developed in consultation with communities of speakers, builds on Creative Commons licences. It varies obligations according to the user's geographic and economic context, with preferential terms for users in the Global South.

It was first applied to a speech corpus in Dholuo, in Kenya. In April 2026, it remained an experiment: one licence applied to a single dataset.[23] Closing the data would keep these languages out of the models. Opening them to all on the same terms would benefit the best equipped first. Between the two lies a governance choice that African institutions can make.

The texts already set a direction. Adopted in Accra in July 2024, the African Union's Continental Artificial Intelligence Strategy notes that a lack of data holds back speech recognition in African languages. It finds that the proliferation of generative models in English "threatens African cultural systems and linguistic diversity", and recommends investment in "local and indigenous" languages.[24] UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021) invites states to examine the effect of machine translation and voice assistants on languages, up to the risk that endangered languages disappear.[25]

In our view, the bottleneck now lies with institutions more than with technology. Technology will bring microphones and models able to learn from few examples. Voices and texts will come from speakers. Long-term funding, the care of archives and the choice of licences fall to institutions. Training new authors falls to schools.

21 Mozilla Foundation (16 September 2020); in May 2021, more than 1,700 hours and more than 840 contributors (Mozilla Foundation, 25 May 2021).

22 Masakhane; Nekoto et al. (2020); Lacuna Fund, catalogue consulted on 20 November 2026.

23 Okorie and Omino (2025), Law, Technology and Humans, 7 (3), legal scholars at the universities of Pretoria and Strathmore who designed the licence; Open Future (April 2026), a report commissioned by the Mozilla Foundation with support from GIZ.

24 African Union (2024), sections 1.4.4 and 2.4.4.1.

25 UNESCO (2021), paragraphs 94 and 95.

6 What we recommend

01

Fund public corpora

Each state funds open corpora of texts and voices in its languages, under a licence chosen with speakers; the African Academy of Languages coordinates their standards.

FOR States, African Union **HORIZON** 2027-2030

02

Evaluate before deploying

Every artificial intelligence system in a public service is evaluated, with published results, in its users' languages, on benchmarks such as IrokoBench.

FOR Administrations **HORIZON** From 2027

03

Digitise the written and sound archive

Libraries, universities and radio stations digitise and transcribe their holdings in African languages, from Ajami and Ge'ez manuscripts to radio archives.

FOR Ministries of culture **HORIZON** 2027-2032

04

Recognise data collection

Universities give credit to students who record or annotate data in their languages, and pay them when the data serve a product.

FOR Universities **HORIZON** 2027-2029

05

Write, record and demand in your languages

Young people write and record in their languages on Wikipedia and Common Voice, join Masakhane and ask for services in these languages.

FOR Young people, associations **HORIZON** Starting now

7 Sources

Data consulted on 20 November 2026; quotations keep the wording of their sources.

Adelani, D. I., et al. (2025). *IrokoBench*. NAACL 2025.

African Union (2024). *Continental Artificial Intelligence Strategy*; OAU (1963). *Charter*.

Afrobarometer (2026). Dispatch 1173, 27 April 2026.

Boston University (2026). *Readers in Ajami*, 27 May 2026.

Brown, T., et al. (2020). *Language Models are Few-Shot Learners*.

Common Crawl (2026). Language statistics, CC-MAIN-2026-43.

Google (2024). *110 new languages are coming to Google Translate*.

Hammarström, H., et al. (2026). *Glottolog 5.4*.

ITU (2025). *Facts and Figures 2025*; Africa regional report.

Kudugunta, S., et al. (2023). *MADLAD-400*. arXiv:2309.04662.

Lacuna Fund (2026). *Language Datasets*.

Llama Team, AI @ Meta (2024). *The Llama 3 Herd of Models*.

Meta AI (2022). Announcement of NLLB-200.

Mozilla (2020, 2021, 2026). Blog posts; *Common Voice, Scripted Speech 27.0*.

Nekoto, W., et al. (2020). Findings of EMNLP 2020, pp. 2144-2160.

Ojo, J., et al. (2025). *AfroBench*. Findings of ACL 2025.

Okorie, C. I., and Omino, M. (2025). *Law, Technology and Humans*, 7 (3).

OpenAI (2025). *Introducing GPT-4.1 in the API*, 14 April 2025.

Open Future (2026). *NOODL. An experiment in equitable data licensing*.

South Africa (1996, 2023), *Constitution* and Eighteenth Amendment; Morocco (2011), *Constitution*.

UNESCO (2003, 2016, 2021). Recommendations; policy paper 24.

Unicode Consortium (2026). *The Unicode Standard 18.0*, chapter 19.

United Nations General Assembly (2024). Resolution 78/312.

Wikimedia (2026). Wikipedia edition statistics.

World Bank (2026). *World Development Indicators*.

ABOUT THIS NOTE

How to cite this note: YouthShare Initiatives (2026). *Teaching the machine Africa's languages*. Evidence note, The Digital Palaver, YouthShare Initiatives.

Licence: CC BY 4.0 · **Version** 1.1 · **Read online:** youthshareinitiatives.org/en/publications/teaching-the-machine-africas-languages