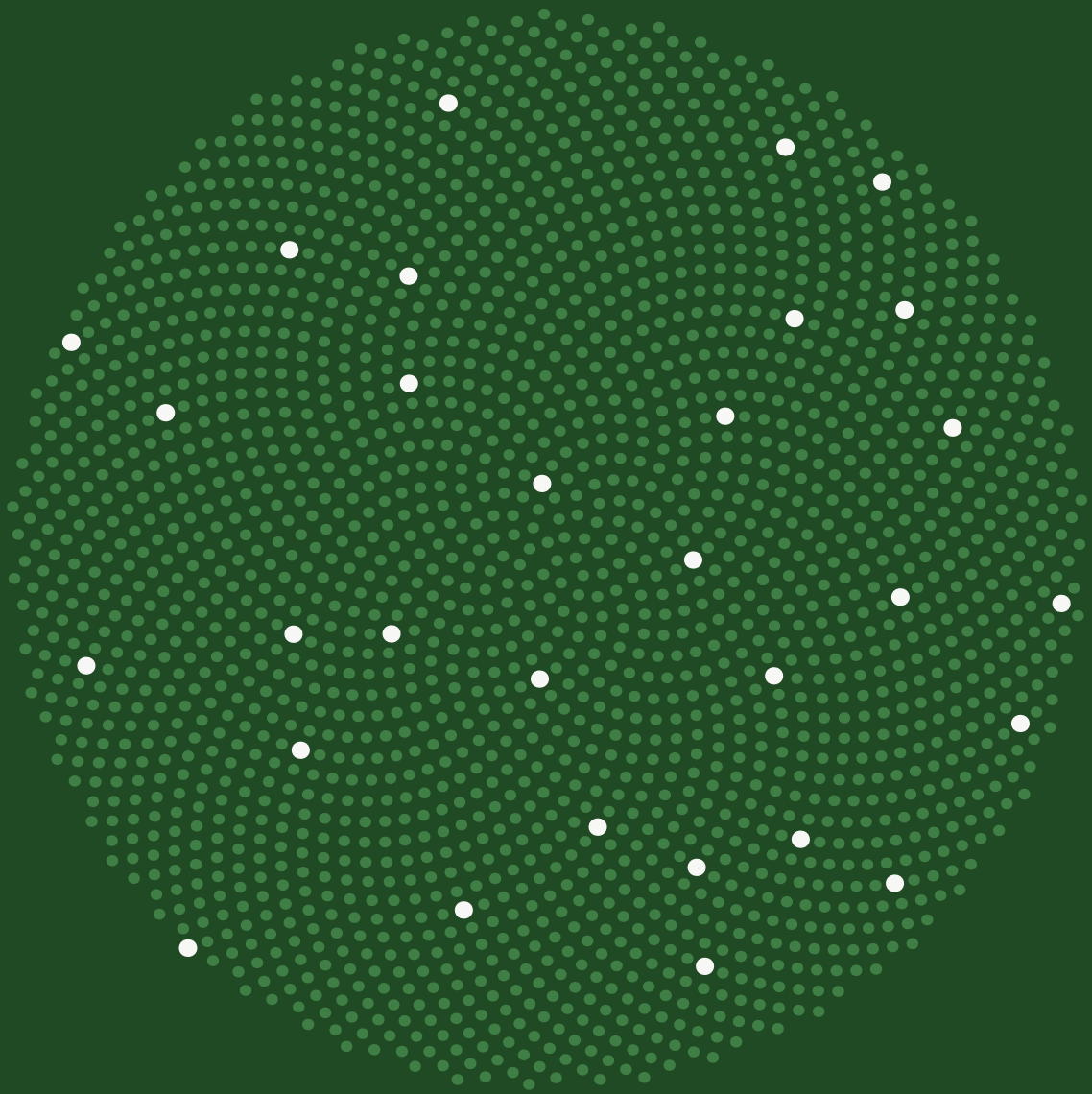
Apprendre à la machine les langues de l’Afrique

L’écart se mesure dans les données, et les données se construisent



2 213 langues de la macro-aire Afrique de Glottolog 5.4, un point par langue

30 langues africaines que reconnaît le détecteur CLD2 de Common Crawl : ensemble, 0,053 % des pages de l’archive CC-MAIN-2026-43 ; l’anglais, 42,56 %

𐓤 𐓛 𐓱 𐓜 𐓢 𐓻 𐓢 𐓪 𐓘 𐓧 𐓯 𐓘 𐓮 𐓪 𐓘 𐓣 𐓘 𐓵 𐓘 𐓲 𐓪 𐓘 𐓴 𐓘 𐓷 𐓘 𐓸 𐓘 𐓹 𐓘 𐓺 𐓘 𐓻 𐓘 𐓼 𐓘 𐓽 𐓘 𐓾 𐓘 𐓿 𐓘 𐔀 𐔁 𐔂 𐔃 𐔄 𐔅 𐔆 𐔇 𐔈 𐔉 𐔊 𐔋 𐔌 𐔍 𐔎 𐔏 𐔐 𐔑 𐔒 𐔓 𐔔 𐔕 𐔖 𐔗 𐔘 𐔙 𐔚 𐔛 𐔜 𐔝 𐔞 𐔟 𐔠 𐔡 𐔢 𐔣 𐔤 𐔥 𐔦 𐔧 𐔨 𐔩 𐔪 𐔫 𐔬 𐔭 𐔮 𐔯 𐔰 𐔱 𐔲 𐔳 𐔴 𐔵 𐔶 𐔷 𐔸 𐔹 𐔺 𐔻 𐔼 𐔽 𐔾 𐔿 𐕀 𐕁 𐕂 𐕃 𐕄 𐕅 𐕆 𐕇 𐕈 𐕉 𐕊 𐕋 𐕌 𐕍 𐕎 𐕏 𐕐 𐕑 𐕒 𐕓 𐕔 𐕕 𐕖 𐕗 𐕘 𐕙 𐕚 𐕛 𐕜 𐕝 𐕞 𐕟 𐕠 𐕡 𐕢 𐕣 𐕤 𐕥 𐕦 𐕧 𐕨 𐕩 𐕪 𐕫 𐕬 𐕭 𐕮 𐕯 𐕰 𐕱 𐕲 𐕳 𐕴 𐕵 𐕶 𐕷 𐕸 𐕹 𐕺 𐕻 𐕼 𐕽 𐕾 𐕿 𐖀 𐖁 𐖂 𐖃 𐖄 𐖅 𐖆 𐖇 𐖈 𐖉 𐖊 𐖋 𐖌 𐖍 𐖎 𐖏 𐖐 𐖑 𐖒 𐖓 𐖔 𐖕 𐖖 𐖗 𐖘 𐖙 𐖚 𐖛 𐖜 𐖝 𐖞 𐖟 𐖠 𐖡 𐖢 𐖣 𐖤 𐖥 𐖦 𐖧 𐖨 𐖩 𐖪 𐖫 𐖬 𐖭 𐖮 𐖯 𐖰 𐖱 𐖲 𐖳 𐖴 𐖵 𐖶 𐖷 𐖸 𐖹 𐖺 𐖻 𐖼 𐖽 𐖾 𐖿 𐗀 𐗁 𐗂 𐗃 𐗄 𐗅 𐗆 𐗇 𐗈 𐗉 𐗊 𐗋 𐗌 𐗍 𐗎 𐗏 𐗐 𐗑 𐗒 𐗓 𐗔 𐗕 𐗖 𐗗 𐗘 𐗙 𐗚 𐗛 𐗜 𐗝 𐗞 𐗟 𐗠 𐗡 𐗢 𐗣 𐗤 𐗥 𐗦 𐗧 𐗨 𐗩 𐗪 𐗫 𐗬 𐗭 𐗮 𐗯 𐗰 𐗱 𐗲 𐗳 𐗴 𐗵 𐗶 𐗷 𐗸 𐗹 𐗺 𐗻 𐗼 𐗽 𐗾 𐗿 𐘀 𐘁 𐘂 𐘃 𐘄 𐘅 𐘆 𐘇 𐘈 𐘉 𐘊 𐘋 𐘌 𐘍 𐘎 𐘏 𐘐 𐘑 𐘒 𐘓 𐘔 𐘕 𐘖 𐘗 𐘘 𐘙 𐘚 𐘛 𐘜 𐘝 𐘞 𐘟 𐘠 𐘡 𐘢 𐘣 𐘤 𐘥 𐘦 𐘧 𐘨 𐘩 𐘪 𐘫 𐘬 𐘭 𐘮 𐘯 𐘰 𐘱 𐘲 𐘳 𐘴 𐘵 𐘶 𐘷 𐘸 𐘹 𐘺 𐘻 𐘼 𐘽 𐘾 𐘿 𐙀 𐙁 𐙂 𐙃 𐙄 𐙅 𐙆 𐙇 𐙈 𐙉 𐙊 𐙋 𐙌 𐙍 𐙎 𐙏 𐙐 𐙑 𐙒 𐙓 𐙔 𐙕 𐙖 𐙗 𐙘 𐙙 𐙚 𐙛 𐙜 𐙝 𐙞 𐙟 𐙠 𐙡 𐙢 𐙣 𐙤 𐙥 𐙦 𐙧 𐙨 𐙩 𐙪 𐙫 𐙬 𐙭 𐙮 𐙯 𐙰 𐙱 𐙲 𐙳 𐙴 𐙵 𐙶 𐙷 𐙸 𐙹 𐙺 𐙻 𐙼 𐙽 𐙾 𐙿 𐚀 𐚁 𐚂 𐚃 𐚄 𐚅 𐚆 𐚇 𐚈 𐚉 𐚊 𐚋 𐚌 𐚍 𐚎 𐚏 𐚐 𐚑 𐚒 𐚓 𐚔 𐚕 𐚖 𐚗 𐚘 𐚙 𐚚 𐚛 𐚜 𐚝 𐚞 𐚟 𐚠 𐚡 𐚢 𐚣 𐚤 𐚥 𐚦 𐚧 𐚨 𐚩 𐚪 𐚫 𐚬 𐚭 𐚮 𐚯 𐚰 𐚱 𐚲 𐚳 𐚴 𐚵 𐚶 𐚷 𐚸 𐚹 𐚺 𐚻 𐚼 𐚽 𐚾 𐚿 𐛀 𐛁 𐛂 𐛃 𐛄 𐛅 𐛆 𐛇 𐛈 𐛉 𐛊 𐛋 𐛌 𐛍 𐛎 𐛏 𐛐 𐛑 𐛒 𐛓 𐛔 𐛕 𐛖 𐛗 𐛘 𐛙 𐛚 𐛛 𐛜 𐛝 𐛞 𐛟 𐛠 𐛡 𐛢 𐛣 𐛤 𐛥 𐛦 𐛧 𐛨 𐛩 𐛪 𐛫 𐛬 𐛭 𐛮 𐛯 𐛰 𐛱 𐛲 𐛳 𐛴 𐛵 𐛶 𐛷 𐛸 𐛹 𐛺 𐛻 𐛼 𐛽 𐛾 𐛿 𐜀 𐜁 𐜂 𐜃 𐜄 𐜅 𐜆 𐜇 𐜈 𐜉 𐜊 𐜋 𐜌 𐜍 𐜎 𐜏 𐜐 𐜑 𐜒 𐜓 𐜔 𐜕 𐜖 𐜗 𐜘 𐜙 𐜚 𐜛 𐜜 𐜝 𐜞 𐜟 𐜠 𐜡 𐜢 𐜣 𐜤 𐜥 𐜦 𐜧 𐜨 𐜩 𐜪 𐜫 𐜬 𐜭 𐜮 𐜯 𐜰 𐜱 𐜲 𐜳 𐜴 𐜵 𐜶 𐜷 𐜸 𐜹 𐜺 𐜻 𐜼 𐜽 𐜾 𐜿 𐝀 𐝁 𐝂 𐝃 𐝄 𐝅 𐝆 𐝇 𐝈 𐝉 𐝊 𐝋 𐝌 𐝍 𐝎 𐝏 𐝐 𐝑 𐝒 𐝓 𐝔 𐝕 𐝖 𐝗 𐝘 𐝙 𐝚 𐝛 𐝜 𐝝 𐝞 𐝟 𐝠 𐝡 𐝢 𐝣 𐝤 𐝥 𐝦 𐝧 𐝨 𐝩 𐝪 𐝫 𐝬 𐝭 𐝮 𐝯 𐝰 𐝱 𐝲 𐝳 𐝴 𐝵 𐝶 𐝷 𐝸 𐝹 𐝺 𐝻 𐝼 𐝽 𐝾 𐝿 𐞀 𐞁 𐞂 𐞃 𐞄 𐞅 𐞆 𐞇 𐞈 𐞉 𐞊 𐞋 𐞌 𐞍 𐞎 𐞏 𐞐 𐞑 𐞒 𐞓 𐞔 𐞕 𐞖 𐞗 𐞘 𐞙 𐞚 𐞛 𐞜 𐞝 𐞞 𐞟 𐞠 𐞡 𐞢 𐞣 𐞤 𐞥 𐞦 𐞧 𐞨 𐞩 𐞪 𐞫 𐞬 𐞭 𐞮 𐞯 𐞰 𐞱 𐞲 𐞳 𐞴 𐞵 𐞶 𐞷 𐞸 𐞹 𐞺 𐞻 𐞼 𐞽 𐞾 𐞿 𐟀 𐟁 𐟂 𐟃 𐟄 𐟅 𐟆 𐟇 𐟈 𐟉 𐟊 𐟋 𐟌 𐟍 𐟎 𐟏 𐟐 𐟑 𐟒 𐟓 𐟔 𐟕 𐟖 𐟗 𐟘 𐟙 𐟚 𐟛 𐟜 𐟝 𐟞 𐟟 𐟠 𐟡 𐟢 𐟣 𐟤 𐟥 𐟦 𐟧 𐟨 𐟩 𐟪 𐟫 𐟬 𐟭 𐟮 𐟯 𐟰 𐟱 𐟲 𐟳 𐟴 𐟵 𐟶 𐟷 𐟸 𐟹 𐟺 𐟻 𐟼 𐟽 𐟾 𐟿 𐠀 𐠁 𐠂 𐠃 𐠄 𐠅 𐠆 𐠇 𐠈 𐠉 𐠊 𐠋 𐠌 𐠍 𐠎 𐠏 𐠐 𐠑 𐠒 𐠓 𐠔 𐠕 𐠖 𐠗 𐠘 𐠙 𐠚 𐠛 𐠜 𐠝 𐠞 𐠟 𐠠 𐠡 𐠢 𐠣 𐠤 𐠥 𐠦 𐠧 𐠨 𐠩 𐠪 𐠫 𐠬 𐠭 𐠮 𐠯 𐠰 𐠱 𐠲 𐠳 𐠴 𐠵 𐠶 𐠷 𐠸 𐠹 𐠺 𐠻 𐠼 𐠽 𐠾 𐠿 𐡀 𐡁 𐡂 𐡃 𐡄 𐡅 𐡆 𐡇 𐡈 𐡉 𐡊 𐡋 𐡌 𐡍 𐡎 𐡏 𐡐 𐡑 𐡒 𐡓 𐡔 𐡕 𐡖 𐡗 𐡘 𐡙 𐡚 𐡛 𐡜 𐡝 𐡞 𐡟 𐡠 𐡡 𐡢 𐡣 𐡤 𐡥 𐡦 𐡧 𐡨 𐡩 𐡪 𐡫 𐡬 𐡭 𐡮 𐡯 𐡰 𐡱 𐡲 𐡳 𐡴 𐡵 𐡶 𐡷 𐡸 𐡹 𐡺 𐡻 𐡼 𐡽 𐡾 𐡿 𐢀 𐢁 𐢂 𐢃 𐢄 𐢅 𐢆 𐢇 𐢈 𐢉 𐢊 𐢋 𐢌 𐢍 𐢎 𐢏 𐢐 𐢑 𐢒 𐢓 𐢔 𐢕 𐢖 𐢗 𐢘 𐢙 𐢚 𐢛 𐢜 𐢝 𐢞 𐢟 𐢠 𐢡 𐢢 𐢣 𐢤 𐢥 𐢦 𐢧 𐢨 𐢩 𐢪 𐢫 𐢬 𐢭 𐢮 𐢯 𐢰 𐢱 𐢲 𐢳 𐢴 𐢵 𐢶 𐢷 𐢸 𐢹 𐢺 𐢻 𐢼 𐢽 𐢾 𐢿 𐣀 𐣁 𐣂 𐣃 𐣄 𐣅 𐣆 𐣇 𐣈 𐣉 𐣊 𐣋 𐣌 𐣍 𐣎 𐣏 𐣐 𐣑 𐣒 𐣓 𐣔 𐣕 𐣖 𐣗 𐣘 𐣙 𐣚 𐣛 𐣜 𐣝 𐣞 𐣟 𐣠 𐣡 𐣢 𐣣 𐣤 𐣥 𐣦 𐣧 𐣨 𐣩 𐣪 𐣫 𐣬 𐣭 𐣮 𐣯 𐣰 𐣱 𐣲 𐣳 𐣴 𐣵 𐣶 𐣷 𐣸 𐣹 𐣺 𐣻 𐣼 𐣽 𐣾 𐣿 𐤀 𐤁 𐤂 𐤃 𐤄 𐤅 𐤆 𐤇 𐤈 𐤉 𐤊 𐤋 𐤌 𐤍 𐤎 𐤏 𐤐 𐤑 𐤒 𐤓 𐤔 𐤕 𐤖 𐤗 𐤘 𐤙 𐤚 𐤛 𐤜 𐤝 𐤞 𐤟 𐤠 𐤡 𐤢 𐤣 𐤤 𐤥 𐤦 𐤧 𐤨 𐤩 𐤪 𐤫 𐤬 𐤭 𐤮 𐤯 𐤰 𐤱 𐤲 𐤳 𐤴 𐤵 𐤶 𐤷 𐤸 𐤹 𐤺 𐤻 𐤼 𐤽 𐤾 𐤿 𐥀 𐥁 𐥂 𐥃 𐥄 𐥅 𐥆 𐥇 𐥈 𐥉 𐥊 𐥋 𐥌 𐥍 𐥎 𐥏 𐥐 𐥑 𐥒 𐥓 𐥔 𐥕 𐥖 𐥗 𐥘 𐥙 𐥚 𐥛 𐥜 𐥝 𐥞 𐥟 𐥠 𐥡 𐥢 𐥣 𐥤 𐥥 𐥦 𐥧 𐥨 𐥩 𐥪 𐥫 𐥬 𐥭 𐥮 𐥯 𐥰 𐥱 𐥲 𐥳 𐥴 𐥵 𐥶 𐥷 𐥸 𐥹 𐥺 𐥻 𐥼 𐥽 𐥾 𐥿 𐦀 𐦁 𐦂 𐦃 𐦄 𐦅 𐦆 𐦇 𐦈 𐦉 𐦊 𐦋 𐦌 𐦍 𐦎 𐦏 𐦐 𐦑 𐦒 𐦓 𐦔 𐦕 𐦖 𐦗 𐦘 𐦙 𐦚 𐦛 𐦜 𐦝 𐦞 𐦟 𐦠 𐦡 𐦢 𐦣 𐦤 𐦥 𐦦 𐦧 𐦨 𐦩 𐦪 𐦫 𐦬 𐦭 𐦮 𐦯 𐦰 𐦱 𐦲 𐦳 𐦴 𐦵 𐦶 𐦷 𐦸 𐦹 𐦺 𐦻 𐦼 𐦽 𐦾 𐦿 𐧀 𐧁 𐧂 𐧃 𐧄 𐧅 𐧆 𐧇 𐧈 𐧉 𐧊 𐧋 𐧌 𐧍 𐧎 𐧏 𐧐 𐧑 𐧒 𐧓 𐧔 𐧕 𐧖 𐧗 𐧘 𐧙 𐧚 𐧛 𐧜 𐧝 𐧞 𐧟 𐧠 𐧡 𐧢 𐧣 𐧤 𐧥 𐧦 𐧧 𐧨 𐧩 𐧪 𐧫 𐧬 𐧭 𐧮 𐧯 𐧰 𐧱 𐧲 𐧳 𐧴 𐧵 𐧶 𐧷 𐧸 𐧹 𐧺 𐧻 𐧼 𐧽 𐧾 𐧿 𐨀 𐨁 𐨂 𐨃 𐨄 𐨅 𐨆 𐨇 𐨈 𐨉 𐨊 𐨋 𐨌 𐨍 𐨎 𐨏 𐨐 𐨑 𐨒 𐨓 𐨔 𐨕 𐨖 𐨗 𐨘 𐨙 𐨚 𐨛 𐨜 𐨝 𐨞 𐨟 𐨠 𐨡 𐨢 𐨣 𐨤 𐨥 𐨦 𐨧 𐨨 𐨩 𐨪 𐨫 𐨬 𐨭 𐨮 𐨯 𐨰 𐨱 𐨲 𐨳 𐨴 𐨵 𐨶 𐨷 𐨸 𐨹 𐨺 𐨻 𐨼 𐨽 𐨾 𐨿 𐩀 𐩁 𐩂 𐩃 𐩄 𐩅 𐩆 𐩇 𐩈 𐩉 𐩊 𐩋 𐩌 𐩍 𐩎 𐩏 𐩐 𐩑 𐩒 𐩓 𐩔 𐩕 𐩖 𐩗 𐩘 𐩙 𐩚 𐩛 𐩜 𐩝 𐩞 𐩟 𐩠 𐩡 𐩢 𐩣 𐩤 𐩥 𐩦 𐩧 𐩨 𐩩 𐩪 𐩫 𐩬 𐩭 𐩮 𐩯 𐩰 𐩱 𐩲 𐩳 𐩴 𐩵 𐩶 𐩷 𐩸 𐩹 𐩺 𐩻 𐩼 𐩽 𐩾 𐩿 𐪀 𐪁 𐪂 𐪃 𐪄 𐪅 𐪆 𐪇 𐪈 𐪉 𐪊 𐪋 𐪌 𐪍 𐪎 𐪏 𐪐 𐪑 𐪒 𐪓 𐪔 𐪕 𐪖 𐪗 𐪘 𐪙 𐪚 𐪛 𐪜 𐪝 𐪞 𐪟 𐪠 𐪡 𐪢 𐪣 𐪤 𐪥 𐪦 𐪧 𐪨 𐪩 𐪪 𐪫 𐪬 𐪭 𐪮 𐪯 𐪰 𐪱 𐪲 𐪳 𐪴 𐪵 𐪶 𐪷 𐪸 𐪹 𐪺 𐪻 𐪼 𐪽 𐪾 𐪿 𐫀 𐫁 𐫂 𐫃 𐫄 𐫅 𐫆 𐫇 𐫈 𐫉 𐫊 𐫋 𐫌 𐫍 𐫎 𐫏 𐫐 𐫑 𐫒 𐫓 𐫔 𐫕 𐫖 𐫗 𐫘 𐫙 𐫚 𐫛 𐫜 𐫝 𐫞 𐫟 𐫠 𐫡 𐫢 𐫣 𐫤 𐫥 𐫦 𐫧 𐫨 𐫩 𐫪 𐫫 𐫬 𐫭 𐫮 𐫯 𐫰 𐫱 𐫲 𐫳 𐫴 𐫵 𐫶 𐫷 𐫸 𐫹 𐫺 𐫻 𐫼 𐫽 𐫾 𐫿 𐬀 𐬁 𐬂 𐬃 𐬄 𐬅 𐬆 𐬇 𐬈 𐬉 𐬊 𐬋 𐬌 𐬍 𐬎 𐬏 𐬐 𐬑 𐬒 𐬓 𐬔 𐬕 𐬖 𐬗 𐬘 𐬙 𐬚 𐬛 𐬜 𐬝 𐬞 𐬟 𐬠 𐬡 𐬢 𐬣 𐬤 𐬥 𐬦 𐬧 𐬨 𐬩 𐬪 𐬫 𐬬 𐬭 𐬮 𐬯 𐬰 𐬱 𐬲 𐬳 𐬴 𐬵 𐬶 𐬷 𐬸 𐬹 𐬺 𐬻 𐬼 𐬽 𐬾 𐬿 𐭀 𐭁 𐭂 𐭃 𐭄 𐭅 𐭆 𐭇 𐭈 𐭉 𐭊 𐭋 𐭌 𐭍 𐭎 𐭏 𐭐 𐭑 𐭒 𐭓 𐭔 𐭕 𐭖 𐭗 𐭘 𐭙 𐭚 𐭛 𐭜 𐭝 𐭞 𐭟 𐭠 𐭡 𐭢 𐭣 𐭤 𐭥 𐭦 𐭧 𐭨 𐭩 𐭪 𐭫 𐭬 𐭭 𐭮 𐭯 𐭰 𐭱 𐭲 𐭳 𐭴 𐭵 𐭶 𐭷 𐭸 𐭹 𐭺 𐭻 𐭼 𐭽 𐭾 𐭿 𐮀 𐮁 𐮂 𐮃 𐮄 𐮅 𐮆 𐮇 𐮈 𐮉 𐮊 𐮋 𐮌 𐮍 𐮎 𐮏 𐮐 𐮑 𐮒 𐮓 𐮔 𐮕 𐮖 𐮗 𐮘 𐮙 𐮚 𐮛 𐮜 𐮝 𐮞 𐮟 𐮠 𐮡 𐮢 𐮣 𐮤 𐮥 𐮦 𐮧 𐮨 𐮩 𐮪 𐮫 𐮬 𐮭 𐮮 𐮯 𐮰 𐮱 𐮲 𐮳 𐮴 𐮵 𐮶 𐮷 𐮸 𐮹 𐮺 𐮻 𐮼 𐮽 𐮾 𐮿 𐯀 𐯁 𐯂 𐯃 𐯄 𐯅 𐯆 𐯇 𐯈 𐯉 𐯊 𐯋 𐯌 𐯍 𐯎 𐯏 𐯐 𐯑 𐯒 𐯓 𐯔 𐯕 𐯖 𐯗 𐯘 𐯙 𐯚 𐯛 𐯜 𐯝 𐯞 𐯟 𐯠 𐯡 𐯢 𐯣 𐯤 𐯥 𐯦 𐯧 𐯨 𐯩 𐯪 𐯫 𐯬 𐯭 𐯮 𐯯 𐯰 𐯱 𐯲 𐯳 𐯴 𐯵 𐯶 𐯷 𐯸 𐯹 𐯺 𐯻 𐯼 𐯽 𐯾 𐯿 𐰀 𐰁 𐰂 𐰃 𐰄 𐰅 𐰆 𐰇 𐰈 𐰉 𐰊 𐰋 𐰌 𐰍 𐰎 𐰏 𐰐 𐰑 𐰒 𐰓 𐰔 𐰕 𐰖 𐰗 𐰘 𐰙 𐰚 𐰛 𐰜 𐰝 𐰞 𐰟 𐰠 𐰡 𐰢 𐰣 𐰤 𐰥 𐰦 𐰧 𐰨 𐰩 𐰪 𐰫 𐰬 𐰭 𐰮 𐰯 𐰰 𐰱 𐰲 𐰳 𐰴 𐰵 𐰶 𐰷 𐰸 𐰹 𐰺 𐰻 𐰼 𐰽 𐰾 𐰿 𐱀 𐱁 𐱂 𐱃 𐱄 𐱅 𐱆 𐱇 𐱈 𐱉 𐱊 𐱋 𐱌 𐱍 𐱎 𐱏 𐱐 𐱑 𐱒 𐱓 𐱔 𐱕 𐱖 𐱗 𐱘 𐱙 𐱚 𐱛 𐱜 𐱝 𐱞 𐱟 𐱠 𐱡 𐱢 𐱣 𐱤 𐱥 𐱦 𐱧 𐱨 𐱩 𐱪 𐱫 𐱬 𐱭 𐱮 𐱯 𐱰 𐱱 𐱲 𐱳 𐱴 𐱵 𐱶 𐱷 𐱸 𐱹 𐱺 𐱻 𐱼 𐱽 𐱾 𐱿 𐲀 𐲁 𐲂 𐲃 𐲄 𐲅 𐲆 𐲇 𐲈 𐲉 𐲊 𐲋 𐲌 𐲍 𐲎 𐲏 𐲐 𐲑 𐲒 𐲓 𐲔 𐲕 𐲖 𐲗 𐲘 𐲙 𐲚 𐲛 𐲜 𐲝 𐲞 𐲟 𐲠 𐲡 𐲢 𐲣 𐲤 𐲥 𐲦 𐲧 𐲨 𐲩 𐲪 𐲫 𐲬 𐲭 𐲮 𐲯 𐲰 𐲱 𐲲 𐲳 𐲴 𐲵 𐲶 𐲷 𐲸 𐲹 𐲺 𐲻 𐲼 𐲽 𐲾 𐲿 𐳀 𐳁 𐳂 𐳃 𐳄 𐳅 𐳆 𐳇 𐳈 𐳉 𐳊 𐳋 𐳌 𐳍 𐳎 𐳏 𐳐 𐳑 𐳒 𐳓 𐳔 𐳕 𐳖 𐳗 𐳘 𐳙 𐳚 𐳛 𐳜 𐳝 𐳞 𐳟 𐳠 𐳡 𐳢 𐳣 𐳤 𐳥 𐳦 𐳧 𐳨 𐳩 𐳪 𐳫 𐳬 𐳭 𐳮 𐳯 𐳰 𐳱 𐳲 𐳳 𐳴 𐳵 𐳶 𐳷 𐳸 𐳹 𐳺 𐳻 𐳼 𐳽 𐳾 𐳿 𐴀 𐴁 𐴂 𐴃 𐴄 𐴅 𐴆 𐴇 𐴈 𐴉 𐴊 𐴋 𐴌 𐴍 𐴎 𐴏 𐴐 𐴑 𐴒 𐴓 𐴔 𐴕 𐴖 𐴗 𐴘 𐴙 𐴚 𐴛 𐴜 𐴝 𐴞 𐴟 𐴠 𐴡 𐴢 𐴣 𐴤 𐴥 𐴦 𐴧 𐴨 𐴩 𐴪 𐴫 𐴬 𐴭 𐴮 𐴯 𐴰 𐴱 𐴲 𐴳 𐴴 𐴵 𐴶 𐴷 𐴸 𐴹 𐴺 𐴻 𐴼 𐴽 𐴾 𐴿 𐵀 𐵁 𐵂 𐵃 𐵄 𐵅 𐵆 𐵇 𐵈 𐵉 𐵊 𐵋 𐵌 𐵍 𐵎 𐵏 𐵐 𐵑 𐵒 𐵓 𐵔 𐵕 𐵖 𐵗 𐵘 𐵙 𐵚 𐵛 𐵜 𐵝 𐵞 𐵟 𐵠 𐵡 𐵢 𐵣 𐵤 𐵥 𐵦 𐵧 𐵨 𐵩 𐵪 𐵫 𐵬 𐵭 𐵮 𐵯 𐵰 𐵱 𐵲 𐵳 𐵴 𐵵 𐵶 𐵷 𐵸 𐵹 𐵺 𐵻 𐵼 𐵽 𐵾 𐵿 𐶀 𐶁 𐶂 𐶃 𐶄 𐶅 𐶆 𐶇 𐶈 𐶉 𐶊 𐶋 𐶌 𐶍 𐶎 𐶏 𐶐 𐶑 𐶒 𐶓 𐶔 𐶕 𐶖 𐶗 𐶘 𐶙 𐶚 𐶛 𐶜 𐶝 𐶞 𐶟 𐶠 𐶡 𐶢 𐶣 𐶤 𐶥 𐶦 𐶧 𐶨 𐶩 𐶪 𐶫 𐶬 𐶭 𐶮 𐶯 𐶰 𐶱 𐶲 𐶳 𐶴 𐶵 𐶶 𐶷 𐶸 𐶹 𐶺 𐶻 𐶼 𐶽 𐶾 𐶿 𐷀 𐷁 𐷂 𐷃 𐷄 𐷅 𐷆 𐷇 𐷈 𐷉 𐷊 𐷋 𐷌 𐷍 𐷎 𐷏 𐷐 𐷑 𐷒 𐷓 𐷔 𐷕 𐷖 𐷗 𐷘 𐷙 𐷚 𐷛 𐷜 𐷝 𐷞 𐷟 𐷠 𐷡 𐷢 𐷣 𐷤 𐷥 𐷦 𐷧 𐷨 𐷩 𐷪 𐷫 𐷬 𐷭 𐷮 𐷯 𐷰 𐷱 𐷲 𐷳 𐷴 𐷵 𐷶 𐷷 𐷸 𐷹 𐷺 𐷻 𐷼 𐷽 𐷾 𐷿 𐸀 𐸁 𐸂 𐸃 𐸄 𐸅 𐸆 𐸇 𐸈 𐸉 𐸊 𐸋 𐸌 𐸍 𐸎 𐸏 𐸐 𐸑 𐸒 𐸓 𐸔 𐸕 𐸖 𐸗 𐸘 𐸙 𐸚 𐸛 𐸜 𐸝 𐸞 𐸟 𐸠 𐸡 𐸢 𐸣 𐸤 𐸥 𐸦 𐸧 𐸨 𐸩 𐸪 𐸫 𐸬 𐸭 𐸮 𐸯 𐸰 𐸱 𐸲 𐸳 𐸴 𐸵 𐸶 𐸷 𐸸 𐸹 𐸺 𐸻 𐸼 𐸽 𐸾 𐸿 𐹀 𐹁 𐹂 𐹃 𐹄 𐹅 𐹆 𐹇 𐹈 𐹉 𐹊 𐹋 𐹌 𐹍 𐹎 𐹏 𐹐 𐹑 𐹒 𐹓 𐹔 𐹕 𐹖 𐹗 𐹘 𐹙 𐹚 𐹛 𐹜 𐹝 𐹞 𐹟 𐹠 𐹡 𐹢 𐹣 𐹤 𐹥 𐹦 𐹧 𐹨 𐹩 𐹪 𐹫 𐹬 𐹭 𐹮 𐹯 𐹰 𐹱 𐹲 𐹳 𐹴 𐹵 𐹶 𐹷 𐹸 𐹹 𐹺 𐹻 𐹼 𐹽 𐹾 𐹿 𐺀 𐺁 𐺂 𐺃 𐺄 𐺅 𐺆 𐺇 𐺈 𐺉 𐺊 𐺋 𐺌 𐺍 𐺎 𐺏 𐺐 𐺑 𐺒 𐺓 𐺔 𐺕 𐺖 𐺗 𐺘 𐺙 𐺚 𐺛 𐺜 𐺝 𐺞 𐺟 𐺠 𐺡 𐺢 𐺣 𐺤 𐺥 𐺦 𐺧 𐺨 𐺩 𐺪 𐺫 𐺬 𐺭 𐺮 𐺯 𐺰 𐺱 𐺲 𐺳 𐺴 𐺵 𐺶 𐺷 𐺸 𐺹 𐺺 𐺻 𐺼 𐺽 𐺾 𐺿 𐻀 𐻁 𐻂 𐻃 𐻄 𐻅 𐻆 𐻇 𐻈 𐻉 𐻊 𐻋 𐻌 𐻍 𐻎 𐻏 𐻐 𐻑 𐻒 𐻓 𐻔 𐻕 𐻖 𐻗 𐻘 𐻙 𐻚 𐻛 𐻜 𐻝 𐻞 𐻟 𐻠 𐻡 𐻢 𐻣 𐻤 𐻥 𐻦 𐻧 𐻨 𐻩 𐻪 𐻫 𐻬 𐻭 𐻮 𐻯 𐻰 𐻱 𐻲 𐻳 𐻴 𐻵 𐻶 𐻷 𐻸 𐻹 𐻺 𐻻 𐻼 𐻽 𐻾 𐻿 𐼀 𐼁 𐼂 𐼃 𐼄 𐼅 𐼆 𐼇 𐼈 𐼉 𐼊 𐼋 𐼌 𐼍 𐼎 𐼏 𐼐 𐼑 𐼒 𐼓 𐼔 𐼕 𐼖 𐼗 𐼘 𐼙 𐼚 𐼛 𐼜 𐼝 𐼞 𐼟 𐼠 𐼡 𐼢 𐼣 𐼤 𐼥 𐼦 𐼧 𐼨 𐼩 𐼪 𐼫 𐼬 𐼭 𐼮 𐼯 𐼰 𐼱 𐼲 𐼳 𐼴 𐼵 𐼶 𐼷 𐼸 𐼹 𐼺 𐼻 𐼼 𐼽 𐼾 𐼿 𐽀 𐽁 𐽂 𐽃 𐽄 𐽅 𐽆 𐽇 𐽈 𐽉 𐽊 𐽋 𐽌 𐽍 𐽎 𐽏 𐽐 𐽑 𐽒 𐽓 𐽔 𐽕 𐽖 𐽗 𐽘 𐽙 𐽚 𐽛 𐽜 𐽝 𐽞 𐽟 𐽠 𐽡 𐽢 𐽣 𐽤 𐽥 𐽦 𐽧 𐽨 𐽩 𐽪 𐽫 𐽬 𐽭 𐽮 𐽯 𐽰 𐽱 𐽲 𐽳 𐽴 𐽵 𐽶 𐽷 𐽸 𐽹 𐽺 𐽻 𐽼 𐽽 𐽾 𐽿 𐾀 𐾁 𐾂 𐾃 𐾄 𐾅 𐾆 𐾇 𐾈 𐾉 𐾊 𐾋 𐾌 𐾍 𐾎 𐾏 𐾐 𐾑 𐾒 𐾓 𐾔 𐾕 𐾖 𐾗 𐾘 𐾙 𐾚 𐾛 𐾜 𐾝 𐾞 𐾟 𐾠 𐾡 𐾢 𐾣 𐾤 𐾥 𐾦 𐾧 𐾨 𐾩 𐾪 𐾫 𐾬 𐾭 𐾮 𐾯 𐾰 𐾱 𐾲 𐾳 𐾴 𐾵 𐾶 𐾷 𐾸 𐾹 𐾺 𐾻 𐾼 𐾽 𐾾 𐾿 𐿀 𐿁 𐿂 𐿃 𐿄 𐿅 𐿆 𐿇 𐿈 𐿉 𐿊 𐿋 𐿌 𐿍 𐿎 𐿏 𐿐 𐿑 𐿒 𐿓 𐿔 𐿕 𐿖 𐿗 𐿘 𐿙 𐿚 𐿛 𐿜 𐿝 𐿞 𐿟 𐿠 𐿡 𐿢 𐿣 𐿤 𐿥 𐿦 𐿧 𐿨 𐿩 𐿪 𐿫 𐿬 𐿭 𐿮 𐿯 𐿰 𐿱 𐿲 𐿳 𐿴 𐿵 𐿶 𐿷 𐿸 𐿹 𐿺 𐿻 𐿼 𐿽 𐿾 𐿿 𑀀 𑀁 𑀂 𑀃 𑀄 𑀅 𑀆 𑀇 𑀈 𑀉 𑀊 𑀋 𑀌 𑀍 𑀎 𑀏 𑀐 𑀑 𑀒 𑀓 𑀔 𑀕 𑀖 𑀗 𑀘 𑀙 𑀚 𑀛 𑀜 𑀝 𑀞 𑀟 𑀠 𑀡 𑀢 𑀣 𑀤 𑀥 𑀦 𑀧 𑀨 𑀩 𑀪 𑀫 𑀬 𑀭 𑀮 𑀯 𑀰 𑀱 𑀲 𑀳 𑀴 𑀵 𑀶 𑀷 𑀸 𑀹 𑀺 𑀻 𑀼 𑀽 𑀾 𑀿 𑁀 𑁁 𑁂 𑁃 𑁄 𑁅 𑁆 𑁇 𑁈 𑁉 𑁊 𑁋 𑁌 𑁍 𑁎 𑁏 𑁐 𑁑 𑁒 𑁓 𑁔 𑁕 𑁖 𑁗 𑁘 𑁙 𑁚 𑁛 𑁜 𑁝 𑁞 𑁟 𑁠 𑁡 𑁢 𑁣 𑁤 𑁥 𑁦 𑁧 𑁨 𑁩 𑁪 𑁫 𑁬 𑁭 𑁮 𑁯 𑁰 𑁱 𑁲 𑁳 𑁴 𑁵 𑁶 𑁷 𑁸 𑁹 𑁺 𑁻 𑁼 𑁽 𑁾 𑁿 𑂀 𑂁 𑂂 𑂃 𑂄 𑂅 𑂆 𑂇 𑂈 𑂉 𑂊 𑂋 𑂌 𑂍 𑂎 𑂏 𑂐 𑂑 𑂒 𑂓 𑂔 𑂕 𑂖 𑂗 𑂘 𑂙 𑂚 𑂛 𑂜 𑂝 𑂞 𑂟 𑂠 𑂡 𑂢 𑂣 𑂤 𑂥 𑂦 𑂧 𑂨 𑂩 𑂪 𑂫 𑂬 𑂭 𑂮 𑂯 𑂰 𑂱 𑂲 𑂳 𑂴 𑂵 𑂶 𑂷 𑂸 𑂹 𑂺 𑂻 𑂼 𑂽 𑂾 𑂿 𑃀 𑃁 𑃂 𑃃 𑃄 𑃅 𑃆 𑃇 𑃈 𑃉 𑃊 𑃋 𑃌 𑃍 𑃎 𑃏 𑃐 𑃑 𑃒 𑃓 𑃔 𑃕 𑃖 𑃗 𑃘 𑃙 𑃚 𑃛 𑃜 𑃝 𑃞 𑃟 𑃠 𑃡 𑃢 𑃣 𑃤 𑃥 𑃦 𑃧 𑃨 𑃩 𑃪 𑃫 𑃬 𑃭 𑃮 𑃯 𑃰 𑃱 𑃲 𑃳 𑃴 𑃵 𑃶 𑃷 𑃸 𑃹 𑃺 𑃻 𑃼 𑃽 𑃾 𑃿 𑄀 𑄁 𑄂 𑄃 𑄄 𑄅 𑄆 𑄇 𑄈 𑄉 𑄊 𑄋 𑄌 𑄍 𑄎 𑄏 𑄐 𑄑 𑄒 𑄓 𑄔 𑄕 𑄖 𑄗 𑄘 𑄙 𑄚

LES CONSTATS

01
L'Afrique compte 2 213 des 7 685 langues parlées recensées par Glottolog ; dans la région Afrique de l'UIT, 36 % de la population utilise Internet en 2025.

02
En octobre 2026, trente langues africaines réunissent 0,053 % des pages de Common Crawl, l'anglais 42,56 %.

03
Dans l'évaluation AfroBench de 2025, les quatre meilleurs modèles fermés cèdent de 13,2 à 19,8 points dans les langues africaines, un modèle ouvert jusqu'à 40,1.

04
Le kinyarwanda compte plus d'heures de voix validées que le français dans Common Voice ; le haoussa en compte 4,35.

05
L'Union africaine et l'UNESCO demandent une intelligence artificielle au service des langues africaines ; budgets et archives restent à bâtir.

1 Les prochains utilisateurs parleront leurs langues

Les Africains qui découvriront l'intelligence artificielle dans les années qui viennent l'aborderont d'abord dans leurs propres langues. Ces langues sont nombreuses. Glottolog, la base ouverte de référence des linguistes, recense 7 685 langues parlées comme langue première dans le monde, dont 2 213 dans sa macro-aire Afrique.[1] Plusieurs réunissent des dizaines de millions de locuteurs, du lingala (40 millions) à l'amharique (57 millions) et au haoussa (77 millions) ; pour le swahili, les estimations vont de 71 millions à « plus de 200 millions ».[2]

Ces locuteurs se connectent vite, et de façon très inégale. Selon l'Union internationale des télécommunications (UIT), 66 % des habitants de sa région Afrique âgés de 10 ans et plus possédaient un téléphone mobile en 2024. En 2025, 36 % de la population de la région utilise Internet, quand la moyenne mondiale atteint 74 %.[3] Pays par pays (**figure 1**), l'usage va de 6,7 % au Soudan du Sud en 2019 à 91,2 % au Maroc en 2024, et 35 pays sur 54 restent sous la moitié de leur population.

> « Les Africains qui découvriront l'intelligence artificielle dans les années qui viennent l'aborderont d'abord dans leurs propres langues.
>
> THE DIGITAL PALAVER, YOUTHSHARE INITIATIVES

1 Glottolog 5.4 (novembre 2026), données ouvertes sous licence CC BY 4.0 : langues de niveau « langue », hors langues des signes, pidgins, langues artificielles, mixtes, inattestées ou inclassables ; notre calcul retrouve le total de 7 685 publié par Glottolog.

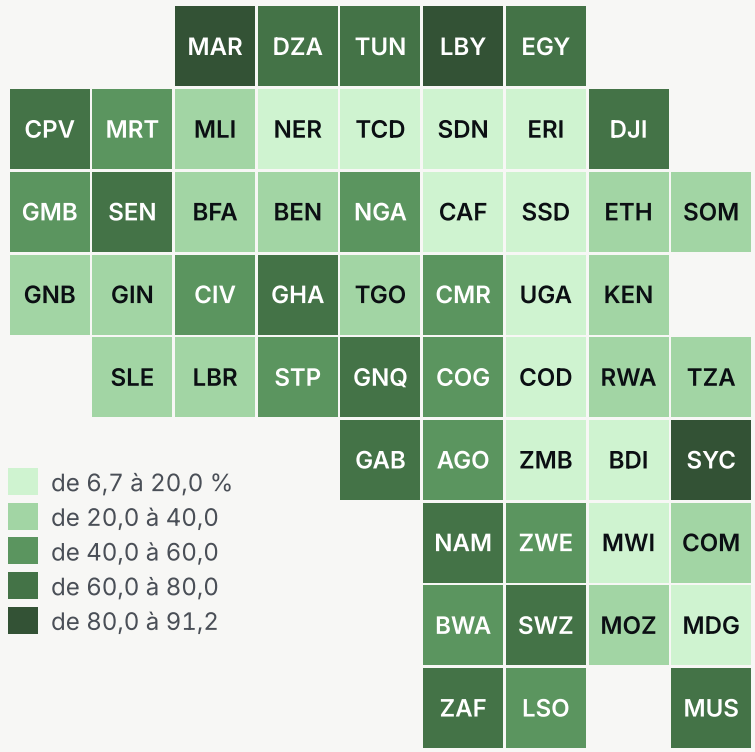
2 Adelani et al. (2025), tableau 9 (locuteurs de langue première et seconde) ; Assemblée générale des Nations unies, résolution 78/312 (2024). Mozilla Foundation (2021) retient environ 100 millions de locuteurs du swahili.

3 UIT, Facts and Figures 2025. La région Afrique de l'UIT réunit 44 pays d'Afrique subsaharienne ; l'UIT range les dix autres pays africains, de la Mauritanie aux Comores, parmi les États arabes.

FIGURE 1

Dans 35 pays africains sur 54, moins de la moitié de la population utilise Internet

Part de la population qui a utilisé Internet au cours des trois derniers mois, en %, dernière année disponible (2024 pour 51 pays, 2017 à 2020 pour le Soudan, le Soudan du Sud et l'Érythrée)



Lecture : Une case par pays, plus la teinte est soutenue, plus l'usage est élevé ; le Nigéria atteint 41,2 % en 2024, l'Éthiopie 21,9 %, le Kenya 35,0 %.

Source : UIT, World Telecommunication/ICT Indicators Database, via la Banque mondiale, World Development Indicators (mise à jour du 11 novembre 2026)

Adresse permanente prévue : youthshareinitiatives.org/r/digital-palaver-ia-langues-africaines-fr/carte-internet

Données disponibles : data.worldbank.org/indicator/IT.NET.USER.ZS

Deux écarts désignent les prochains internautes. Les jeunes ouvrent la marche : en 2024, 53 % des 15-24 ans de la région utilisaient Internet, contre 34 % du reste de la population. Le continent est jeune, avec un âge médian de 19,0 ans en 2023. Les campagnes suivent à distance, avec 23 % de ruraux en ligne pour 57 % de citadins.[4]

L'objection la plus sérieuse tient en une phrase : les Africains connectés lisent et écrivent déjà l'anglais, le français, l'arabe ou le portugais, et les services numériques peuvent s'en contenter. Elle décrit bien la minorité aujourd'hui en ligne. Elle décrit mal celles et ceux qui la rejoindront. Selon une estimation que cite l'UNESCO, jusqu'à 40 % de la population mondiale reste à l'écart d'un enseignement dans une langue qu'elle parle ou comprend. La difficulté culmine là où la diversité linguistique est la plus forte, en Afrique subsaharienne comme en Asie-Pacifique.[5]

La radio montre l'autre versant. Premier média d'information du continent, elle informe chaque jour ou plusieurs fois par semaine 59 % des citoyens de 38 pays enquêtés en 2024-2025. Elle atteint mieux que tout autre média les femmes, les ruraux, les plus pauvres et les moins scolarisés ; chez les 18-35 ans, les

4 UIT (avril 2025), données 2024 ; ONU, World Population Prospects 2024, traité par Our World in Data, pour l'âge médian.

5 UNESCO (2016), document d'orientation 24, d'après Walter et Benson (2012).

réseaux sociaux la devancent déjà, avec 58 % contre 55 %.[6] Nous y voyons une leçon pour les assistants vocaux. Une technique qui demande seulement d'écouter et de parler rejoint celles et ceux que l'écrit tient à distance, à condition de parler leur langue.

2 Le Web qui nourrit les modèles parle d'abord anglais

Un grand modèle de langue apprend ce qu'il lit, et ce qu'il lit vient en grande partie du Web. Le modèle GPT-3, décrit en 2020, tirait 60 % de ses exemples d'entraînement de Common Crawl, une archive publique du Web. Dans son corpus, 93 % des mots étaient anglais.[7] Quatre ans plus tard, l'équipe de Meta décrit pour Llama 3 un mélange où les données « multilingues », celles des langues autres que l'anglais, pèsent environ 8 % des jetons.[8]

Le dernier relevé des langues de Common Crawl, établi sur l'archive d'octobre 2026, mesure ce déséquilibre. Sur 2,17 milliards de pages HTML, l'anglais en occupe 923 millions, soit 42,56 %, et le français 4,76 %. Le swahili, première langue africaine de la liste, réunit 215 652 pages, soit 0,0099 %. Les trente langues africaines que reconnaît le détecteur totalisent 1 148 503 pages, soit 0,053 % de l'archive.[9] Pour une page en swahili, l'archive en contient plus de 4 000 en anglais.

> « Un grand modèle de langue apprend ce qu'il lit, et ce qu'il lit vient en grande partie du Web.
>
> THE DIGITAL PALAVER, YOUTHSHARE INITIATIVES

Le **tableau 1** rassemble ces mesures pour seize langues africaines, et la démographie y explique une part seulement des écarts. Le lingala, parlé par 40 millions de personnes, compte 22 millions de caractères de texte Web dans le corpus MADLAD-400. Le kinyarwanda, quatre fois moins parlé, en compte 749 millions. Le haoussa et le swahili dépassent chacun 120 000 articles de Wikipédia ; l'oromo, langue de 37 millions de personnes, en réunit 1 972. L'anglais écrase l'ensemble : 9 000 milliards de caractères dans MADLAD-400, plus de 14 000 fois le volume du haoussa.

6 Afrobarometer, dépêche n° 1173 (avril 2026).

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

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

9 Common Crawl, archive CC-MAIN-2026-43 ; somme des pages des trente langues africaines reconnues par CLD2, calcul des auteurs.

TABLEAU 1

Les quatre scores les plus faibles reviennent à des langues qui comptent moins de 50 millions de caractères de texte Web

Langue	Locuteurs, millions	Texte Web, millions de caractères	Articles Wikipédia	Pages Common Crawl	Score de GPT-4o (AfriMMLU)
Swahili	71-106	2 400	134 816	215 652	77,4
Kinyarwanda	10	749	11 273	71 437	64,2
Haoussa	77	630	121 976	87 352	67,2
Amharique	57	509	15 751	90 407	59,8
Xhosa	19	287	2 604	22 514	70,2
Shona	11	266	11 897	23 961	67,6
Zoulou	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
Éwé	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

Lecture : Locuteurs de langue première et seconde ; texte Web du corpus MADLAD-400 ; score de GPT-4o aux questions AfriMMLU, sur 100, où le hasard donnerait 25. Le détecteur de Common Crawl reconnaît l'akan, dont le twi fait partie, et laisse l'éwé de côté.

Source : Adelani et al. (2025), IrokoBench, tableaux 3 et 9 ; Wikimedia, statistiques des éditions de Wikipédia, 20 novembre 2026 ; Common Crawl, archive CC-MAIN-2026-43

Adresse permanente prévue : youthshareinitiatives.org/r/digital-palaver-ia-langues-africaines-fr/tab-langues

Données disponibles : aclanthology.org/2025.naacl-long.139

MÉTHODE

Mesurer la place d'une langue dans les données

Common Crawl identifie la langue principale de chaque page HTML avec le détecteur CLD2. Trente langues africaines figurent dans son relevé ; l'éwé, le peul ou le bambara lui échappent. MADLAD-400 (2023) réunit des documents tirés de Common Crawl, filtrés puis contrôlés par sondage : pour chaque langue, deux des auteurs ont examiné vingt documents. Wikipédia compte ses pages de contenu, courtes ou longues. Chaque mesure reste partielle ; leur convergence fonde le constat.

On objecte que le Web reflète l'usage : si le wolof réunit 4 708 pages dans Common Crawl, les wolophones écriraient peu en ligne. La causalité court aussi dans l'autre sens. Une langue que les claviers, les correcteurs et l'école laissent de côté s'écrit moins à l'écran, et le détecteur lui-même reconnaît trente langues africaines sur plus de deux mille. L'archive existe pourtant hors du Web, dans les manuscrits, les journaux et les fonds des radios. La numériser change le compte.

3 Les modèles réussissent là où les données abondent

Les évaluations publiées en 2025 confirment le lien entre données et performance, langue par langue. IrokoBench, un banc d'essai traduit par des professionnels et conçu par des chercheurs du réseau Masakhane, a soumis seize grands modèles à trois tâches en dix-sept langues africaines.[10] Sur les questions de connaissances, appelées AfriMMLU, GPT-4o répond juste à 87,4 % des questions en anglais, à 83,2 % en français, à 77,4 % en swahili et à 33,6 % en éwé. Chaque question propose quatre réponses : en éwé, le modèle dépasse à peine les 25 % du hasard. La dernière colonne du **tableau 1** donne les seize langues.

L'ordre des langues suit, dans l'ensemble, celui des données. Les quatre scores les plus faibles reviennent à l'éwé, au wolof, au twi et au luganda. Ces quatre langues comptent parmi les cinq qui disposent de moins de 50 millions de caractères de texte Web. Le swahili, première langue africaine par le volume de texte, obtient le meilleur score africain. Les auteurs relient eux-mêmes les performances les plus basses aux langues qui ont le moins de données en ligne.[11] Le lingala rappelle que d'autres facteurs jouent : 22 millions de caractères, et un score de 61,0.

Un second résultat pèse sur l'usage quotidien. La plupart des modèles testés répondent mieux quand une machine traduit d'abord la question en anglais ; GPT-4 Turbo et GPT-4o font partie des rares exceptions. Un service bâti sur ce détour impose une langue intermédiaire et dépend d'un traducteur automatique dont la qualité varie, elle aussi, avec les données.

AfroBench élargit le constat à 64 langues et quinze tâches, avec une version allégée de quatorze langues qui suit les modèles récents (**figure 2**). Les quatre modèles fermés les mieux classés obtiennent de 79,7 à 85,1 points en anglais et de 62,8 à 67,0 en moyenne dans les langues africaines. Les modèles aux poids ouverts, que des équipes africaines peuvent adapter à leurs langues, montrent des écarts plus dispersés : de 26,9 à 40,1 points pour les trois de la figure, de 10,3 à 33,2 pour les autres, plus petits ou moins forts en anglais.[12]

Reste l'argument optimiste : chaque génération de modèles progresse, et la taille finira par combler l'écart. Les chiffres lui donnent en partie raison. Dans les langues africaines, GPT-4.1, publié en avril 2025, dépasse GPT-4o, et Gemma 3 gagne 9,1 points sur Gemma 2. Les auteurs d'AfroBench notent eux-mêmes ce progrès, qu'ils jugent limité sur les tâches de connaissances.[13]

Les langues que la traduction automatique a gagnées l'ont été par des efforts ciblés. En 2022, Meta a publié NLLB-200, un modèle de traduction aux poids librement téléchargeables pour

10 Adelani et al. (2025), sections 3 et 4 : inférence logique, calcul et questions de connaissances ; dix modèles ouverts et six modèles propriétaires ; moyennes calculées sur seize langues africaines.

11 Même source, section 4.1 et tableau 9.

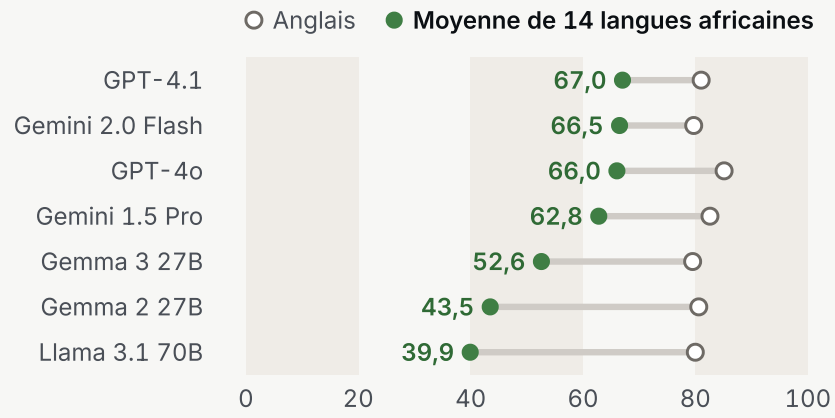
12 Ojo et al. (2025), tableau 8.

13 Ojo et al. (2025), tableau 8 et section 5.1 ; OpenAI, Introducing GPT-4.1 in the API (14 avril 2025), pour la sortie de GPT-4.1.

FIGURE 2

Parmi sept modèles, l'écart entre l'anglais et les langues africaines va de 13,2 à 40,1 points

Score moyen de sept tâches, sur 100, en anglais et dans quatorze langues africaines, modèles évalués en 2024 et 2025



Lecture : Le cercle creux donne le score en anglais, le point plein la moyenne africaine ; Gemma et Llama sont des modèles aux poids ouverts.

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

Adresse permanente prévue : youthshareinitiatives.org/r/digital-palaver-ia-langues-africaines-fr/fig-afrobench

Données disponibles : aclanthology.org/2025.findings-acl.976

tout usage hors commerce, qui prend en charge, selon l'entreprise, 55 langues africaines « avec des résultats de haute qualité ». Les outils courants en couvraient alors moins de 25. En juin 2024, Google a ajouté 110 langues à son traducteur, dont environ un quart venues d'Afrique.[14] La taille aide. L'attention portée à chaque langue décide.

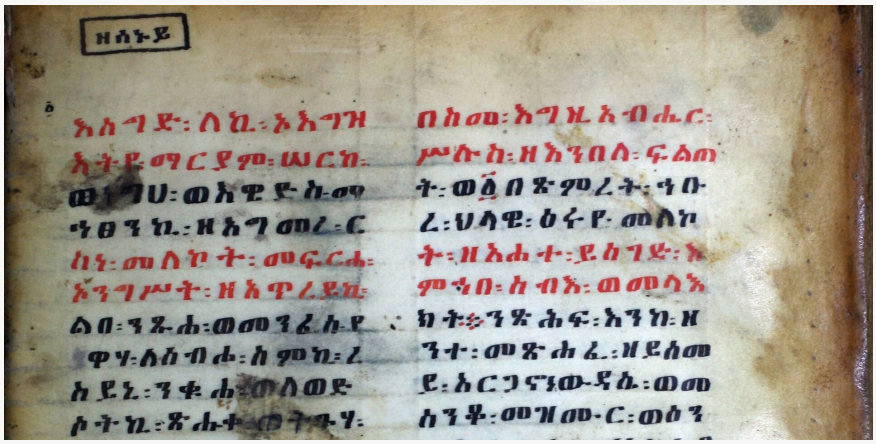
14 Meta AI (juillet 2022), poids sous licence CC BY-NC 4.0 ; Google (27 juin 2024).

4 Chaque technique nouvelle a dû apprendre les langues de l'Afrique

L'écart d'aujourd'hui prolonge une longue histoire. Chaque technique de communication a fini par porter les langues africaines, dès que des institutions et des communautés lui en ont donné les moyens. L'écriture en offre le premier chapitre. Le syllabaire né pour noter le guèze sert aujourd'hui à écrire l'amharique et le tigré.

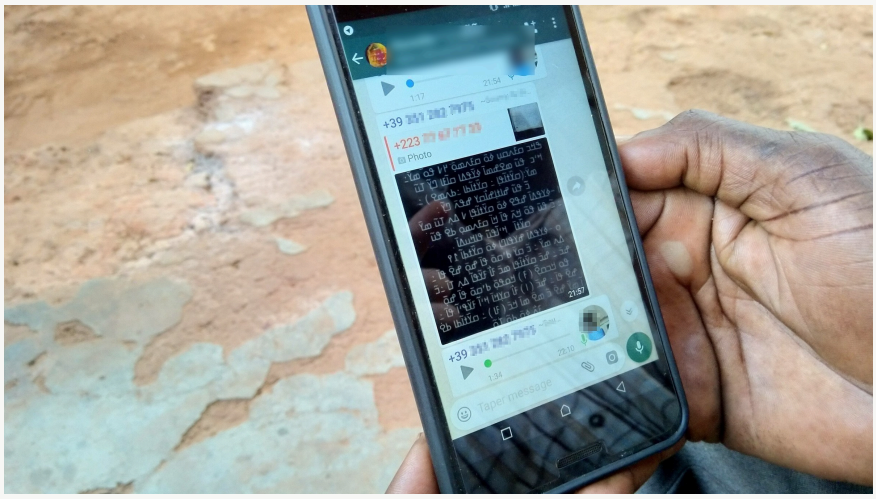
Au sud du Sahara, des lettrés ont écrit leurs langues en caractères arabes, sous la forme dite ajami. Un projet de l'université de Boston, financé par la British Library, vient de numériser 52 415 pages de manuscrits arabes et ajami produits au Fouta-Djalon, en Guinée, et au Sénégal. Leurs auteurs ont vécu du XVIIIe au XXIe siècle.[15]

15 Unicode Standard 18.0, chapitre 19 ; Boston University, Readers in Ajami (27 mai 2026), projet financé par l'Endangered Archives Programme de la British Library.



Détail d’une page d’un livre de prières éthiopien du XVe siècle, en écriture guèze, à l’encre noire et rouge.

Miami University Libraries, domaine public, via Wikimedia Commons (détail recadré).



Une classe de n’ko sur WhatsApp, en 2019 : l’alphabet créé en 1949 circule désormais sur les téléphones.

Coleman Donaldson, CC BY-SA 4.0, via Wikimedia Commons (image recadrée et réduite, en-tête de la conversation flouté).

D’autres écritures sont nées sur le continent pour des langues précises. Le vaï apparaît au Libéria dans les années 1830, le bamoun au Cameroun entre 1896 et 1910. Le n’ko date de 1949 pour les langues mandingues, le garay de 1961 pour le wolof, l’adlam de la fin des années 1980 pour le peul.

Chacune est entrée dans les machines par la même porte : son codage dans Unicode, le standard qui attribue un numéro à chaque caractère. L’éthiopien y figure depuis 1999, le tifinagh depuis 2005, le n’ko depuis 2006, le vaï depuis 2008, le bamoun depuis 2009, l’adlam depuis 2016 et le garay depuis 2024.[16] Le n’ko dispose aujourd’hui de sa propre édition de Wikipédia, forte de 1 622 articles.

Les politiques linguistiques ont pesé autant que les techniques. Dès 1963, la Charte de l’Organisation de l’unité africaine retenait comme langues de travail, « si possible », des

16 Unicode Standard 18.0, chapitre 19, et fichier DerivedAge : versions 3.0 (1999), 4.1 (2005), 5.0 (2006), 5.1 (2008), 5.2 (2009), 9.0 (2016) et 16.0 (2024). Wikipédia en n’ko : 1 622 articles le 20 novembre 2026.

langues africaines, avec l'anglais et le français ; l'arabe et le portugais s'y sont ajoutés par la suite.[17] La Constitution sud-africaine de 1996 reconnaissait onze langues officielles, et la langue des signes sud-africaine est devenue la douzième en 2026. Le texte constate « l'usage et le statut historiquement diminués » des langues autochtones, que l'État doit élever par des mesures « pratiques et positives ». Le Maroc a fait du tamazight une langue officielle de l'État dans sa Constitution de 2011.[18]

En 2003 déjà, l'UNESCO recommandait que toutes les cultures puissent accéder au cyberespace dans toutes les langues, y compris les langues autochtones.[19] L'intelligence artificielle repose la question à une échelle supérieure.

On objectera que cette histoire penche aussi de l'autre côté. Les langues européennes héritées de la colonisation gardent l'école, l'administration et l'imprimé, et le marché favorise les grandes langues. Nous lisons dans les chiffres du Web la trace de cet héritage. L'histoire dessine pourtant une tendance nette : une langue gagne une technique quand une institution ou une communauté la code, l'enseigne ou l'exige. La technique fournit l'outil. Les États, les écoles et les communautés de locuteurs en décident l'usage.

5 Les données se construisent, et l'Afrique a commencé

Le constat appelle une réponse par les données, et des Africains la construisent déjà. Le kinyarwanda en donne la démonstration la plus nette (**figure 3**). Dans la version 27.0 de Common Voice, la base de voix lancée par Mozilla, il réunit 2 001,7 heures de voix validées au 11 septembre 2026, au quatrième rang des 295 langues de la base, devant le français. Le haoussa, parlé par 77 millions de personnes, en compte 4,35.[20]

Cette avance a une histoire. Après un hackathon organisé à Kigali par Mozilla et l'agence allemande de coopération GIZ, la jeune entreprise rwandaise Digital Umuganda a mobilisé des contributeurs en s'appuyant sur l'umuganda, l'idée rwandaise de coopération communautaire. En septembre 2020, elle avait recueilli 1 211 heures de kinyarwanda auprès de plus de 420 personnes. Elle a ensuite développé avec la GIZ et l'agence rwandaise RISA l'assistant Mbaza, qui délivrait des informations sur la covid-19 en kinyarwanda par un menu téléphonique, ses fonctions vocales étant en préparation.[21] Une base publique, une entreprise locale, un partenaire de coopération et une tradition collective ont changé l'échelle d'une langue en quelques années.

17 Charte de l'OUA (25 mai 1963), article XXIX, texte d'origine (Recueil des traités des Nations unies, vol. 479) ; la version de l'Union africaine ajoute l'arabe et le portugais.

18 Constitution de la République d'Afrique du Sud (1996), section 6 ; Constitution Eighteenth Amendment Act (2023), en vigueur le 26 juin 2026 (Government Gazette n° 54901) ; Constitution du Royaume du Maroc (2011), article 5.

19 UNESCO, recommandation adoptée le 15 octobre 2003, lors de la 32e session de la Conférence générale (29 septembre-17 octobre 2003).

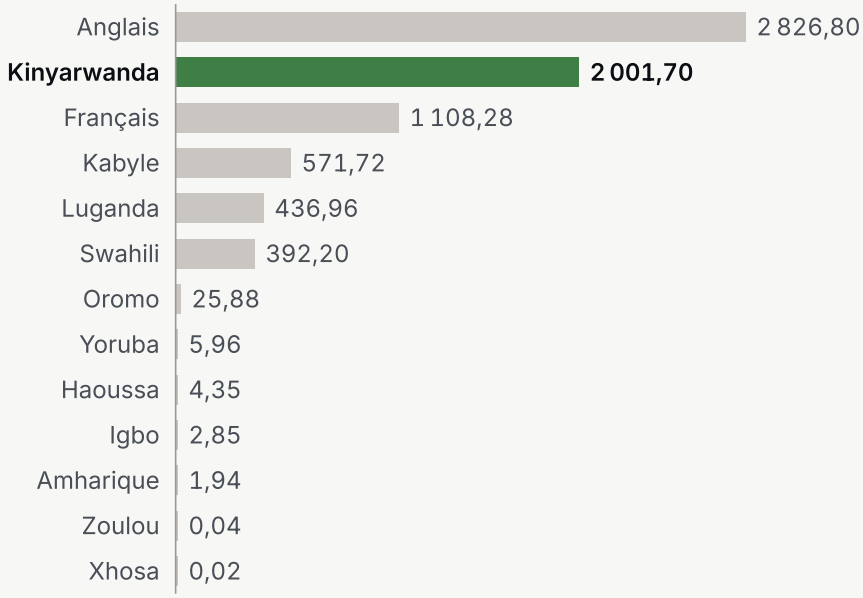
20 Mozilla Common Voice, Scripted Speech 27.0 : 29 295 heures validées au total ; IrokoBench, tableau 9, pour les locuteurs du haoussa.

21 Mozilla Foundation (16 septembre 2020) ; en mai 2021, plus de 1 700 heures et plus de 840 contributeurs (Mozilla Foundation, 25 mai 2021).

FIGURE 3

Le kinyarwanda réunit davantage d'heures de voix validées que le français, le haoussa en compte 4,35

Heures d'enregistrements validés dans Common Voice, version 27.0, données arrêtées au 11 septembre 2026



Lecture : Treize langues retenues ; un enregistrement est validé quand les avis favorables l'emportent après deux écoutes au moins.
Source : Mozilla Common Voice, Scripted Speech 27.0, données arrêtées au 11 septembre 2026
Adresse permanente prévue : youthshareinitiatives.org/r/digital-palaver-ia-langues-africaines-fr/fig-common-voice
Données disponibles : github.com/common-voice/cv-dataset/blob/main/datasets/scripted-speech/cv-corpus-27.0-2026-09-11.json

D'autres chantiers avancent à l'échelle du continent. Le réseau Masakhane veut renforcer la recherche sur le traitement automatique des langues africaines « pour les Africains, par les Africains » : dès 2020, il publiait des bancs d'essai de traduction pour plus de trente langues. Le Lacuna Fund finance des jeux de données conçus et détenus localement, ouverts à tous ; 25 des 31 jeux de son catalogue linguistique nomment une langue ou une région africaine.[22]

La propriété des données pose la question la plus lourde. Des voix et des textes donnés par des volontaires africains nourriront d'abord les produits de grandes entreprises étrangères. Le risque est réel, et des juristes africains y répondent par le droit. La licence NOODL, élaborée en consultation avec des communautés de locuteurs, reprend les licences Creative Commons. Elle module les obligations selon le contexte géographique et économique de l'utilisateur, avec des conditions préférentielles pour les pays du Sud.

Elle a d'abord servi à un corpus de parole en dholuo, au Kenya. En avril 2026, elle restait une expérience : une licence appliquée à un seul jeu de données.[23] Fermer les données tiendrait ces langues à l'écart des modèles. Les ouvrir à tous aux mêmes conditions profiterait d'abord aux mieux équipés. Entre les deux

22 Masakhane ; Nekoto et al. (2020) ; Lacuna Fund, catalogue consulté le 20 novembre 2026.

23 Okorie et Omino (2025), Law, Technology and Humans, 7 (3), juristes des universités de Pretoria et Strathmore qui ont conçu la licence ; Open Future (avril 2026), rapport commandé par la Fondation Mozilla avec le soutien de la GIZ.

se dessine un choix de gouvernance, que les institutions africaines peuvent trancher.

Les textes fixent déjà une direction. Adoptée à Accra en juillet 2024, la Stratégie continentale de l'Union africaine pour l'intelligence artificielle constate que le manque de données freine la reconnaissance vocale dans les langues africaines. Elle juge que la prolifération de modèles génératifs en anglais « menace les systèmes culturels africains et la diversité linguistique », et recommande d'investir dans les langues « locales et autochtones ».[24] La Recommandation de l'UNESCO sur l'éthique de l'intelligence artificielle (2021) invite les États à examiner l'effet de la traduction automatique et des assistants vocaux sur les langues, jusqu'au risque de disparition des langues menacées.[25]

À nos yeux, le verrou tient désormais aux institutions plus qu'à la technique. La technique apportera des micros et des modèles capables d'apprendre à partir de peu d'exemples. Les voix et les textes viendront des locuteurs. Le financement dans la durée, la conservation des archives et le choix des licences reviennent aux institutions. La formation de nouveaux auteurs revient à l'école.

24 Union africaine (2024), sections 1.4.4 et 2.4.4.1.

25 UNESCO (2021), paragraphes 94 et 95.

6 Ce que nous recommandons

01

Financer des corpus publics

Chaque État finance des corpus ouverts de textes et de voix dans ses langues, sous une licence choisie avec les locuteurs ; l'Académie africaine des langues en coordonne les normes.

POUR États, Union africaine **HORIZON** 2027-2030

02

Évaluer avant de déployer

Tout système d'intelligence artificielle d'un service public est évalué, résultats publiés, dans les langues de ses usagers, sur des bancs d'essai comme IrokoBench.

POUR Administrations **HORIZON** Dès 2027

03

Numériser l'archive écrite et sonore

Bibliothèques, universités et radios numérisent et transcrivent leurs fonds en langues africaines, des manuscrits ajami et guèzes aux archives radiophoniques.

POUR Ministères de la culture **HORIZON** 2027-2032

04 **Reconnaître la collecte de données**
Les universités créditent les étudiantes et étudiants qui enregistrent ou annotent des données dans leurs langues, et les rémunèrent quand ces données servent un produit.
POUR Universités **HORIZON** 2027-2029

05 **Écrire, enregistrer et exiger dans ses langues**
Les jeunes écrivent et enregistrent dans leurs langues sur Wikipédia et Common Voice, rejoignent Masakhane et réclament des services dans ces langues.
POUR Jeunes, associations **HORIZON** Dès aujourd'hui

7 Sources

Données consultées le 20 novembre 2026 ; citations de l'anglais traduites par nos soins.

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

Afrique du Sud (1996, 2023), *Constitution* et dix-huitième modification ; Maroc (2011), *Constitution*.

Afrobarometer (2026). Dépêche n° 1173, 27 avril 2026.

Assemblée générale des Nations unies (2024). Résolution 78/312.

Banque mondiale (2026). *World Development Indicators*.

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

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

Common Crawl (2026). Statistiques des langues, CC-MAIN-2026-43.

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

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

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). Annonce de NLLB-200.

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

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

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

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

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

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

UIT (2025). *Facts and Figures 2025* ; rapport régional Afrique.

UNESCO (2003, 2016, 2021). Recommandations ; document d'orientation 24.

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

Union africaine (2024). *Continental Artificial Intelligence Strategy* ; OUA (1963). *Charte*.

Wikimedia (2026). Statistiques des éditions de Wikipédia.

À PROPOS DE CETTE NOTE

Pour citer cette note : YouthShare Initiatives (2026). *Apprendre à la machine les langues de l'Afrique*. Note de preuves, The Digital Palaver, YouthShare Initiatives.

Licence : CC BY 4.0 · **Version** 1.1 · **Lire en ligne :** youthshareinitiatives.org/fr/publications/apprendre-a-la-machine-les-langues-de-lafrique

