Centre for International
Governance Innovation

Mémoire No 216 — Novembre 2025

L'IA et le torchage des données linguistiques en Afrique : remédier au manque de ressources

Ife Adebara

Points principaux

- → À cause de données linguistiques limitées, les langues africaines sont sous-représentées dans les systèmes fondés sur l'intelligence artificielle (IA), ce qui exclut des millions de personnes de la participation numérique dans leur langue maternelle.
- → Des facteurs comme la complexité multilingue, les politiques dominantes sur les langues étrangères, un appui institutionnel faible et le manque d'infrastructure numérique contribuent à la classification à faibles ressources des langues africaines.
- → À l'instar du traitement du gaz, le « torchage des données linguistiques » montre bien la négligence systémique et la gestion médiocre des données sur les langues africaines qui ont provoqué une sous-collecte des données, une conservation déficiente et une utilisation limitée de ces langues dans l'IA.
- → Pour remédier à ce problème, il faut élaborer des politiques qui intègrent les langues africaines dans les programmes numériques nationaux, appuient la production des documents d'appoint, permettent de financer les projets connexes et favorisent le développement inclusif et collaboratif de l'IA.
- → La documentation communautaire, l'utilisation d'outils libres et la reconnaissance croissante de la diversité linguistique de l'IA constituent des voies de développement prometteuses.

Introduction

Construits à partir de techniques fondées sur l'apprentissage en profondeur, les systèmes d'IA modernes nécessitent des quantités massives de données pour fonctionner efficacement afin de produire des résultats réalistes qui reflètent les modèles et les structures au sein des données de formation. Dans le cas des technologies linguistiques, ces données sont tirées des nouvelles, d'ouvrages, de blogues, de publications sur les médias sociaux et d'autres plateformes numériques qui hébergent du contenu linguistique. Cependant, seul un petit nombre de langues possèdent suffisamment de données pour favoriser le développement de technologies de l'IA solides. En dépit des millions de personnes qui les parlent, la majorité des langues africaines n'ont que peu de ressources, ce qui signifie qu'il leur manque les données nécessaires pour construire des modèles d'IA solides. C'est pourquoi nombre d'Africains sont privés d'outils numériques, de ressources en ligne et de services suscités par l'IA, parce que ceux-ci ne sont pas disponibles dans leurs langues maternelles (Adams et coll., 2024). Le Graphique 1 montre le manque de diversité culturelle et linguistique de l'IA dans les cadres gouvernementaux, les mesures gouvernementales et les entités non étatiques de l'Afrique par rapport aux autres régions.

Selon Pratik Joshi et coll. (2020), plus de 95 % des langues africaines sont classées « à la traîne », c'est-à-dire parmi les langues pour lesquelles il est

À propos de l'auteur

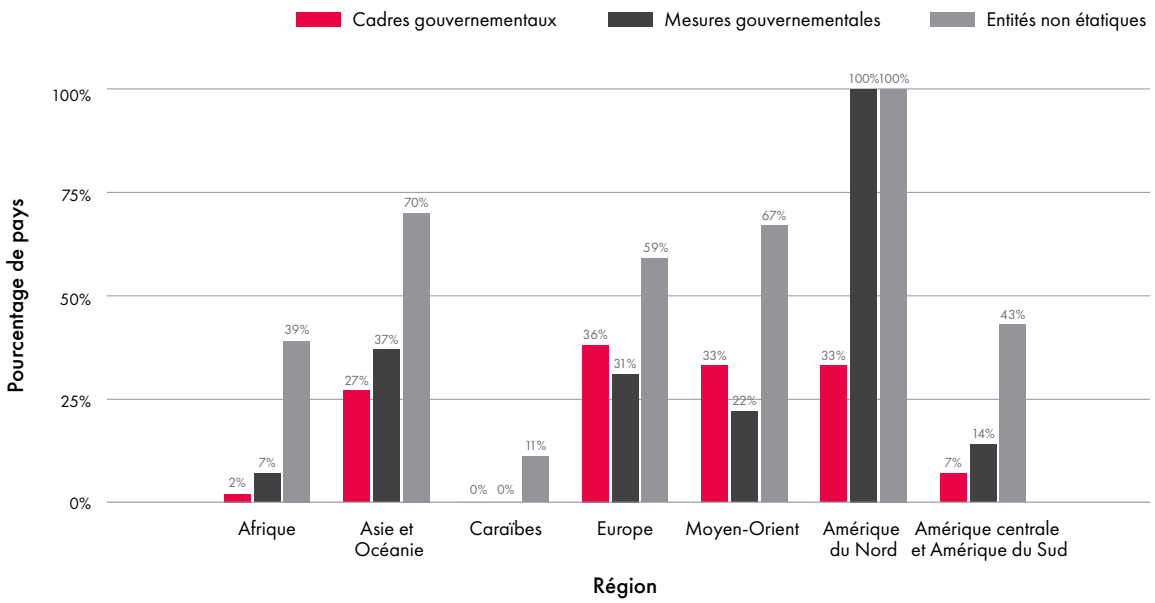
Ife Adebara est une chercheuse dont le travail intègre le traitement de la langue naturelle à la préservation et au développement des langues africaines. Elle détient un doctorat en linguistique de l'Université de la Colombie-Britannique ainsi qu'une maîtrise en informatique de l'Université Simon Fraser et de l'Université de Birmingham. Forte de plus de 10 années d'expérience en IA multilingue, elle se spécialise dans la création de technologies linguistiques inclusives pour les langues autochtones et faibles en ressources. Elle axe sa recherche sur la conservation éthique des données, le développement de modèles et la politique linguistique afin de faire progresser l'équité linguistique dans le contexte de l'IA. Elle dirige le développement d'AfroLID, de Serengeti et de Cheetah, des modèles novateurs à l'appui de plus de 500 langues et dialectes africains. Ife est la co-fondatrice et l'agente principale des technologies d'EqualyzAI, une entreprise qui a inventé l'IA individuelle à partir des séries de données linguistiques les plus inclusives d'Afrique.

presqu'impossible de construire des technologies linguistiques alimentées par l'IA en raison d'une insuffisance de données. Cette classification n'est pas fonction du nombre de personnes qui parlent la langue, mais plutôt de la disponibilité de texte numérisé. Les divergences deviennent plus claires lorsqu'on compare les populations d'orateurs avec la disponibilité des données. Des langues comme le catalan (cinq millions d'orateur), le finnois (10 millions) et le suédois (10,5 millions) sont classées à ressources élevées grâce à une vaste documentation numérique. Cependant, les langues africaines comme l'amharique (37 millions), l'igbo (30 millions) et le swahili (80 millions) manquent encore de ressources en dépit de leurs grandes populations d'orateurs. Cette situation met en lumière une lacune systémique dans la représentation linguistique, où l'accès à de la documentation numérique historique et les décisions politiques dictent si une langue prospère dans le contexte de l'IA ou en est exclue.

La rareté des données linguistiques n'est pas un simple enjeu technique : elle est modelée par un problème socio-économique, politique et infrastructurel profondément enraciné. Plusieurs facteurs contribuent à l'absence de séries de données complètes pour les langues africaines.

- **Les politiques gouvernementales et la négligence institutionnelle :** Nombre de gouvernements africains post-coloniaux ont historiquement accordé la priorité aux langues européennes (anglais, français, portugais) dans les secteurs de l'éducation, de la gouvernance et des médias, ce qui a limité le soutien formel envers les langues autochtones.
- **Les lacunes infrastructurelles numériques :** Les langues africaines sont rarement intégrées dans les technologies conventionnelles, comme les claviers, les moteurs de recherche et les interfaces de médias sociaux, ce qui rend la collecte des textes numériques difficiles.
- **Les défis liés à la préservation des données :** Nombre de langues africaines ont de solides traditions orales, mais une documentation écrite limitée. Les ressources linguistiques existantes sont souvent entreposées dans des formats inaccessibles pour la formation en IA, comme des manuscrits tangibles, des enregistrements audio ou des articles scientifiques éparpillés.

Graphique 1 : La diversité culturelle et linguistique dans l'IA



Source : Adams et coll. (2024, 48).

Cette lacune est ce que nous désignons sous le terme de « torchage des données linguistiques » analogue au torchage dans l'industrie du pétrole. De la même façon que le torchage de gaz fait intervenir le gaspillage par combustion de ressources naturelles durant l'extraction du pétrole, le torchage des données linguistiques donne lieu à la collecte, à la préservation et à l'utilisation inadéquates de données linguistiques. En Afrique, ce problème se manifeste par l'égarement ou l'inaccessibilité de contenus linguistiques, la sous-utilisation d'archives, de la médiocrité des pratiques de numérisation et de la limitation de l'intégration des langues locales dans les infrastructures éducatives et technologiques. Non seulement ces pratiques dilapident les ressources linguistiques, mais elles creusent aussi le fossé numérique qui fait que les langues africaines demeurent marginalisées dans le paysage mondial de l'IA.

Dans ce mémoire, l'auteur explore comment les langues africaines ont été numériquement marginalisées en raison de pratiques de torchage des données linguistiques. Cette situation exclut systématiquement les langues africaines des progrès technologiques, ce qui réduit au silence leur présence dans le monde numérique suscité par l'IA. De la même façon qu'une coupure de courant déconnecte une communauté de ressources critiques, la négligence des données linguistiques exclut les langues africaines de l'écosystème numérique. Au fil du temps, cette exclusion contribue à l'anéantissement culturel et linguistique, ce qui limite les possibilités d'inclusion technologique. Cependant, les efforts visant à remédier à ces problèmes prennent de l'ampleur : dans toute l'Afrique, des chercheurs, des technologues et des décideurs suscitent des initiatives pour améliorer la disponibilité des données linguistiques et leur intégration dans les systèmes d'IA (Adebara et coll., 2022, 2023; Adebara, Elmadany et Abdul-Mageed, 2024; Adelani et coll., 2022). L'auteur de ce mémoire examine les facteurs qui contribuent à la faiblesse des ressources des langues africaines, met en lumière des histoires de réussite émergentes dont les acteurs ont remédié à la rareté des données, et offre des recommandations fructueuses aux décideurs, aux chercheurs et aux dirigeants industriels. Pour remédier à ce problème, il faut un effort concerté pour documenter et préserver les langues africaines et les intégrer dans les systèmes d'IA afin de veiller à ce que leur voix soit non seulement entendue, mais qu'elle puisse prospérer à l'ère numérique. En mettant en œuvre ces stratégies, il est possible de combler le fossé numérique et de veiller à ce que les langues africaines jouent un rôle pertinent dans l'avenir de l'IA.

L'Afrique multilingue : une épée à double tranchant pour le développement de l'IA

L'Afrique est un continent multilingue très complexe où se côtoient plus de 2000 langues et dialectes, soit environ un tiers de toutes les langues parlées dans le monde (Hammarström, 2018). C'est pourquoi nombre d'Africains naviguent, dès leur petite enfance, entre de multiples langues. Prenons, par exemple, un enfant qui grandit à Addis Ababa, la capitale de l'Éthiopie. Il parle peut-être l'oromo chez lui avec ses parents (langue maternelle), l'amharique dans son quartier et au marché local (langue de son environnement immédiat), puis l'anglais, qui est probablement sa principale langue d'instruction à l'école secondaire. Chaque langue représente un aspect différent de son identité et sert à une fonction sociale distincte. La situation peut être encore plus compliquée. Prenons une famille qui vit à Bauchi, au Nigeria : la mère vient du groupe ethnique haba et parle le kilba, sa langue maternelle, tandis que le père est idoma et a pour langue principale l'idoma. Cette famille vit principalement dans le Nord du Nigeria où l'on parle surtout le hausa, mais, dans les casernes militaires où elle réside, leur enfant est exposé au pidgin nigérian. Depuis sa naissance, cet enfant est donc immergé dans quatre langues différentes avant d'arriver à l'école primaire où il sera exposé à l'anglais comme cinquième langue.

Le paysage linguistique des communautés africaines créée des modèles uniques d'expertise linguistique. Ce même enfant, qui grandit à Bauchi, au Nigeria, par exemple, navigue entre plusieurs domaines linguistiques pour lesquels il acquière différentes compétences, et s'inscrit pour chacune des cinq langues dans son répertoire. Bien qu'il puisse activement utiliser l'idoma et le kilba dans son milieu familial, il peut avoir de la difficulté à utiliser ces langues lorsqu'il n'est pas chez lui. Simultanément, il peut utiliser le pidgin nigérian pour ses interactions avec ses pairs, mais pas avec les anciens, tout en perfectionnant plus largement son anglais et son hausa pour les contextes formels et institutionnels. La nature fluide des compétences multilingues signifie que la capacité d'un orateur d'utiliser chaque langue dépend souvent du contexte et est propre à chaque domaine. Cet enfant peut, en effet, être incapable de communiquer efficacement dans ces cinq langues hors de leur contexte d'utilisation établi, ce qui le confronte à des problèmes lorsqu'il tente d'utiliser des langues qu'il parle à la maison dans des contextes formels ou, inversement, il rencontre des difficultés à exprimer des concepts intimes ou culturels dans une langue réservée aux milieux institutionnels.

En Afrique, cette réalité multilingue est à la fois un défi et une occasion de développement de l'IA. D'un côté, la nature fluide et contextuelle de l'utilisation de la langue complique le développement de modèles fondés sur l'IA qui reflètent avec précision la diversité linguistique. D'un autre côté, l'expertise multilingue de l'Afrique offre un avantage unique : la capacité des gens de naviguer entre plusieurs langues, dialectes et registres constitue une occasion de concevoir des modèles d'IA plus souples et plus adaptables capables de gérer la diversité linguistique du monde réel. En accordant la priorité au développement de l'IA soucieux d'inclure les langues, l'Afrique peut créer des systèmes d'IA inédits qui reflètent mieux le multilinguisme humain, non seulement pour son continent, mais aussi pour le monde entier.

L'héritage colonial et les lacunes politiques : les gardiens invisibles des données linguistiques

Dans toute l'Afrique, en réponse au paysage multilingue, la politique linguistique dominante consiste à adopter une langue étrangère pour les affaires commerciales et, dans certains cas, quelques langues africaines autochtones reçoivent parfois un statut officiel au niveau régional ou national ou pour l'éducation (Petzell, 2012; Foster, 2021; Ouane et Glanz, 2010). Le Graphique 2, le Graphique 3 et le Graphique 4 montrent l'utilisation des langues officielles en Afrique. Au Nigeria, l'anglais est la langue officielle, tandis que seulement trois langues autochtones sur 512 y sont officiellement reconnues comme des langues régionales. Au Ghana, l'anglais est

également la langue officielle, mais 10 des 73 langues autochtones du pays sont aussi utilisées comme des langues institutionnelles; le swahili est la seule langue autochtone officielle de la Tanzanie sur 118 autres langues en sus de l'anglais; au Kenya, 12 langues sur 61 ont un statut officiel; en Afrique du Sud, seulement 12 langues autochtones sur 20 sont des langues institutionnelles (Adebara et Abdul-Mageed, 2022).

Même lorsque des langues autochtones obtiennent un statut officiel parallèlement à des langues étrangères, elles jouent souvent un rôle symbolique et non pas fonctionnel. Par exemple, bien que l'Union africaine reconnaisse le kiswahili comme l'une de ses langues officielles, son site Web et ses publications sur papier officielles demeurent en anglais et en français, pas en kiswahili. Ce modèle s'étend aux systèmes de l'éducation : même dans les régions où des langues autochtones sont utilisées, leur rôle se limite généralement à l'éducation de la petite enfance, souvent en parallèle d'une langue étrangère et non pas comme moyen d'instruction unique (Petzell,

Graphique 2 : Langues officielles de l'Afrique

Source : Auteur.

Graphique 3 : Langues officielles de l'éducation en Afrique

Source : Auteur.

Remarque : « Aucune » désigne les pays où il n'y a pas de langues officielles distinctes pour l'éducation. Dans ces pays, les langues utilisées pour l'éducation sont les mêmes que celles utilisées officiellement.

Graphique 4 : Langues régionales officielles en Afrique

Source : Auteur.

Remarqu : « Aucune » désigne les pays où il n'y a pas de langues régionales officielles distinctes. Dans ces pays, les langues officielles sont utilisées régionalement.

2012; Foster, 2021; Ouane et Glanz, 2010). De l'école secondaire à l'université, on utilise des langues étrangères comme principal moyen d'instruction. Cette structure politique a des conséquences directes sur le développement de l'IA. Premièrement, elle crée une lacune en matière de connaissances des langues autochtones : les orateurs peuvent exceller dans la pratique parlée de leur langue maternelle, mais ne pas avoir les compétences nécessaires à l'écrit pour élaborer un contenu numérique. Deuxièmement, l'utilisation écrasante de langues étrangères dans les documents gouvernementaux officiels exclut les langues autochtones du discours gouvernemental, ce qui limite l'accès à une information publique cruciale en langues autochtones. Comme la majorité des modèles d'IA, surtout ceux axés sur le traitement de la langue naturelle (TLN), dépendent de données textuelles écrites à grande échelle, la domination des langues étrangères dans le domaine de l'éducation non seulement éradique les langues autochtones des sphères formelles, mais les empêche aussi de prospérer dans les technologies numériques suscitées par l'IA, ce qui renforce leur marginalisation continue de la technologie.

Les médias et le fossé numérique : où sont les langues autochtones africaines?

Les médias sur papier, ou le rétrécissement de la sphère des langues autochtones

Le faible taux de connaissance des langues africaines a des répercussions directes sur les ventes et la lecture des journaux, ce qui restreint les publications en langues autochtones à une poignée de langues très répandues. Par exemple, sur les 11 journaux publiés en Ouganda, seuls quatre sont publiés dans des langues autochtones dominantes, tandis que les sept autres utilisent l'anglais (Lugalambi, Mwesige et Bussiek, 2010). On observe dans toute l'Afrique ce schéma où seules les langues ayant un statut officiel ou sont parlées par de grandes populations d'orateurs assurent la survie des journaux. Des pays comme le Ghana et l'Eswatini ont perdu tous leurs journaux en langues autochtones, tandis qu'au Nigeria, de nombreuses tentatives de maintenir ce type de publications ont échoué (ResCue et Agbozo, 2021; Mthembu et Lunga, 2020; Fosu, 2024; Onyenankeya, 2022). En dépit de quelques rares publications ayant réussi à survivre, comme *Alaroye* au Nigeria et *Isolezwe* (un quotidien dont les ventes s'élèvent à plus de 100 000 exemplaires) en Afrique du Sud et la presse amharique en Éthiopie, les journaux en langue autochtone ont souvent de la difficulté à survivre (Tshabangu et Salawu, 2022; Salawu, 2020).

De plus, nombre de journaux en langues africaines opèrent comme des filiales de médias en langue étrangère. Au Nigeria, d'anciens éditeurs anglophones comme Daily Sketch Press Ltd. et Concord Press publiaient des journaux en yoruba, en hausa et en igbo, mais la majorité de ces éditions autochtones ont maintenant disparu (Salawu, 2020; Tshabangu et Salawu, 2022; Onyenankeya, 2022). De même, Perskoporasie (Perskor), un éditeur d'Afrique du Sud, publiait à une certaine époque *Imvo Zabantsundu*, un journal en isiXhosa, qui n'existe cependant plus aujourd'hui. Au-delà des journaux, la publication d'ouvrages en langues africaines diminue également, car davantage d'auteurs choisissent des langues coloniales pour stimuler leur potentiel de vente. Dans les situations dans lesquelles des ouvrages en langues autochtones existent, ils sont rarement disponibles en format numérique, ce qui limite leur accessibilité pour les applications fondées sur l'IA. En raison de l'absence en ligne d'ouvrages et de journaux en langues autochtones, il manque aux modèles fondés sur l'IA des donnés sur papier de grande qualité pour la formation, ce qui renforce l'exclusion numérique.

Radio et télévision : une domination des langues étrangères qui perdure

En Afrique, la radio demeure l'une des formes de communication les plus accessibles et les plus influentes : des milliers de stations sont réparties sur tout le continent (Molale et Mpofu, 2023; Conroy-Krutz et Koné, 2022; Brooke, 2024). Seulement au Cameroun, il y a plus de 280 stations; le Ghana en compte 354, l'Ouganda 258 et le Mali plus de 300 (Cheo, Chie et Menguie, 2023; ; National Communications Authority, 2023; Myers et Harford, 2020; Myers, 2009). La disponibilité de la radio la rend indispensable, surtout dans les régions rurales, car elle ne nécessite ni littératie ni approvisionnement constant en électricité. Bien que la radio soit diffusée plus largement en langues autochtones que les journaux et la télévision (la télévision utilisant uniquement

les principales langues autochtones), les langues étrangères dominent encore les ondes. Dans de nombreux cas, les programmes en langue autochtone se voient attribuer une brève plage horaire, tandis que les émissions diffusées aux heures de grande écoute sont en anglais, en français ou en portugais. Même lorsque du contenu télévisé ou radiodiffusé en langue autochtone est disponible, il est rarement archivé ou numérisé, ce qui en complique l'utilisation pour une formation fondée sur l'IA ou la recherche linguistique.

Un autre obstacle majeur est que nombre de stations n'ont pas de présence en ligne, et celles qui offrent une transmission numérique accordent souvent la priorité aux programmes en langues étrangères, tandis que le contenu en langue autochtone demeure hors ligne. De plus, les lois locales sur la diffusion exige des stations qu'elles archivent leur contenu pour une brève période uniquement. Au Rwanda[1] et au Nigeria (National Broadcasting Commission, 2016), par exemple, les stations doivent conserver un contenu médiatique pendant trois mois et 90 jours, respectivement, et peuvent le détruire ensuite à leur discrétion. Vu le financement limité disponible pour l'entreposage des données, de précieuses émissions en langues autochtones sont vite perdues, ce qui creuse d'autant plus le fossé numérique.

Médias sociaux : les plateformes numériques et la marginalisation linguistique

Les plateformes de médias sociaux permettent à leurs utilisateurs de créer et de partager du contenu dans leur propre langue. Cependant, la majorité des plateformes n'offrent que peu, voire pas d'aide pour les langues africaines. Par exemple, Facebook de Meta soutient 112 langues, dont 11 langues africaines, mais nombre d'entre elles ne sont appuyées qu'en partie, et l'interface de la plateforme n'est pas entièrement localisée dans aucune d'elles. Par contraste, Instagram, LinkedIn et X soutiennent 32, 36 et 34 langues, respectivement, dont aucune n'est africaine. Ce manque d'appui empêche les personnes qui parlent une langue autochtone de participer entièrement aux sphères numériques, ce qui limite leur capacité de s'exprimer, d'obtenir de l'information et de contribuer aux conversations mondiales en ligne. De façon plus critique, cette lacune contribue à la sous-représentation des identités culturelles et linguistiques africaines dans les sphères numériques, ce qui renforce la marginalisation linguistique et accélère potentiellement l'abandon des langues autochtones.

Même quand les utilisateurs connaissent bien les langues autochtones, l'absence de claviers en langue africaine complique la rédaction en ligne (Adebara et Abdul- Mageed, 2022). En effet, nombre de langues africaines dépendent des signes diacritiques pour donner le ton de la conversation, déterminer la longueur des voyelles, entre autres caractéristiques, mais les claviers réguliers ne facilitent pas la saisie de ces caractères (ResCue et Agbozo, 2021). Or, sans la diacritique, on risque une ambiguïté sémantique substantielle, comme on peut le constater en yoruba : igbá (*calabash, panier*), igba (200), ìgbà (temps), ìgbá (œuf de jardin) et igbà (corde). De même, en akan, les distinctions entre les tonalités grammaticales influent sur les significations d'un verbe : Ama dá ha « *Ama dort généralement ici* » et Ama dà ha « *Ama est en train de dormir ici* » (Adebara et Abdul-Mageed, 2022). C'est pourquoi la présence des langues autochtones sur les médias sociaux est minime : les utilisateurs y interagissent principalement avec des « j'aime » et des « partager » et non par des commentaires écrits (Sunday et coll., 2018; Molale et Mpofu, 2023; ResCue et Agbozo, 2021). Lorsque des commentaires écrits sont adoptés, la diacritique en est souvent absente et le mélange des codes y règne (Molale et Mpofu, 2023; Yevudey, 2018).

Les répercussions du fossé numérique sur l'IA : recommandations pour le changement

En Afrique, la domination des langues étrangères à titre de langues officielles a des répercussions substantielles sur le développement de l'IA.

- **Une présence numérique limitée réduit les données de formation.** Bien que les modèles de TLN se fondent sur le texte à grande échelle et les séries de données parlées, la majorité des langues autochtones sont soit indisponibles en ligne, soit elles ne sont jamais numérisées.

1 Loi No 02/2013 du 08/02/2013 régissant les médias, en ligne : <https://rwandalii.org/akn/rw/act/law/2013/2/eng%402013-03-11>.

- **Le manque de transcription et d'archivage affaiblit les capacités de l'IA.** Sans transcription systématique ni préservation des contenus en langues autochtones, les technologies d'IA demeurent sous-développées, ce qui restreint d'autant plus les applications de l'IA pour les langues africaines.

Sans efforts proactifs pour numériser et préserver les langues autochtones et les intégrer dans les systèmes d'IA, ces langues risquent de continuer à être marginalisées à l'ère du numérique. C'est pourquoi il est essentiel de combler cette lacune pour veiller à ce que les technologies d'IA reflètent la diversité linguistique de l'Afrique au lieu de renforcer des inégalités historiques. Les recommandations suivantes mettent en lumière des stratégies clés pour remédier à ces problèmes :

- **Institutionnaliser les langues autochtones dans les domaines de l'éducation et de la gouvernance.** Dans l'ensemble de l'Afrique, on assiste à une forte corrélation entre le statut d'une langue officielle et sa disponibilité dans les médias numériques, surtout dans les pays où les langues locales sont utilisées dans les secteurs de l'éducation et de l'administration (Adebara et Abdul-Mageed, 2022). L'Éthiopie, la Tanzanie et l'Afrique du Sud font partie des quelques nations africaines dans lesquelles une ou plusieurs langues locales sont largement utilisées à des fins officielles et administratives, l'amharique, le swahili et le zoulou étant parmi certaines des langues les plus pourvues en ressources en Afrique. C'est pourquoi il est crucial d'intégrer entièrement les langues autochtones dans l'éducation et la gouvernance pour étendre les données linguistiques en Afrique. Le développement de contenus éducatifs, la documentation de l'information gouvernementale officielle et l'assurance que les procédures juridiques et les annonces des services publics sont accessibles en langues autochtones sont d'importantes mesures qu'ont prises les pays dont les politiques linguistiques reconnaissent les langues autochtones. Les pays qui n'ont pas ce genre de politique devraient envisager d'adopter des politiques similaires. La littératie est un processus à long terme qui dépend d'une éducation constante, et les politiques qui restreignent l'instruction en langue autochtone à quelques années uniquement dans les écoles sont inefficaces. Des taux plus élevés de littératie en langues africaines pourraient améliorer la paternité et la lecture d'œuvres ainsi que la viabilité commerciale d'ouvrages et de journaux dans ces langues, ce qui renforcerait leur présence dans les médias tant numériques qu'imprimés.

- **Accorder la priorité à la numérisation et à l'archivage en ligne des contenus en langues autochtones.** Les infrastructures numériques publiques jouent un rôle crucial dans la numérisation des langues africaines en offrant des plateformes libres, adaptables et interexploitables qui favorisent la diversité linguistique. Ces plateformes peuvent être issues de collaborations entre des gouvernements, des chercheurs et des entreprises technologiques. Les cadres politiques peuvent accélérer d'autant plus le progrès en incitant les diffuseurs et d'autres créateurs de contenus à conserver et à offrir des transcriptions libres en langues autochtones à des fins de formation en IA pour assurer un approvisionnement constant en données linguistiques de grande qualité. Les efforts de numérisation ne doivent pas se limiter aux nouveaux contenus, mais inclure aussi les archives historiques cloisonnées dans différentes institutions qui, au fil du temps, se révèlent des perspectives linguistiques et culturelles précieuses. De plus, l'investissement dans des claviers en langues autochtones et dans d'autres outils numériques améliorera d'autant plus l'accessibilité et la facilité d'utilisation, ce qui renforcera la présence numérique des langues africaines.

- **Promouvoir et financer la création et la conservation de contenus en langues autochtones.** Le financement d'initiatives qui appuient la production de journaux, d'ouvrages et de contenus radio- et télé-diffusés et transmis sur les médias sociaux en langues autochtones peut solliciter des créateurs et étendre substantiellement le bassin de ressources linguistiques. Il faut aussi encourager les organisations médiatiques à archiver les contenus en langues autochtones sur de longues périodes pour assurer leur accessibilité et leur préservation à long terme. De plus, il serait utile de lancer des initiatives communautaires pour assurer la diversité et la représentation des données linguistiques, car une langue est dynamique : bien qu'elle puisse avoir un vocabulaire limité, les manières dont ses mots sont associés pour former des phrases sont

illimitées. C'est pourquoi, pour obtenir une représentation linguistique exacte de l'utilisation d'une langue, il faut une vaste participation communautaire à la collecte des données. Sans cette diversité de contribution, les ressources linguistiques numérisée risquent de ne refléter qu'une sous-série restreinte de l'usage linguistique du paysage réel. Les efforts communautaires de collecte de données, de conservation et de développement des technologies de l'IA pour les langues africaines ont, du reste, déjà produit des résultats époustouflants (Adams et coll., 2024; Adelani et coll., 2022; Adebara et coll., 2022, 2023; Adebara, Elmadany et Abdul-Mageed, 2024). Des organisations comme l'African Languages Technology Initiative, Data Science Nigeria, Deep Learning Indaba and Masakhane ont réalisé un travail extensif de création de données et de conservation de l'IA en langues africaines. Des organisations comme EqualyzAI ont aussi développé Equalyz Crowd, une plateforme d'externalisation libre visant à faciliter la collecte, la création et l'enrichissement des données linguistiques en Afrique.

Conclusion

L'élimination du fossé numérique qui handicape les langues africaines est une entreprise complexe, mais nécessaire, et il y a des signes prometteurs qui montrent que ces langues peuvent jouer un rôle central dans la révolution de l'IA. Pour y parvenir, il faudra des efforts collaboratifs de la part des gouvernements, des décideurs, des experts linguistiques, des technologues et des communautés locales. La population jeune et dynamique de l'Afrique aspire à voir ses langues représentées dans la sphère numérique, et il est impératif que ces personnes soient habilitées à participer à des conversations mondiales dans la langue de leur choix. En mettant en œuvre les stratégies exposées dans ce mémoire, nous pouvons amorcer un avenir numérique plus inclusif, dans lequel les langues africaines prospèrent dans la sphère de l'IA. Finalement, pour obtenir une représentation linguistique et culturelle réelle, il faut non seulement appliquer les stratégies discutées dans cet article, mais aussi remettre en question les hypothèses qui orientent l'utilisation de l'IA et examiner d'un œil critique les fondements de l'incorporation des données linguistiques dans les technologies de l'IA.

Ouvrages cités

Adams, Rachel, Fola Adeleke, Ana Florido, Larissa Galdino de Magalhães Santo, Nicolás Grossman, Leah Junck et Kelly Stone. 2024. *Global Index on Responsible AI 2024*. Afrique du Sud : Global Center on AI Governance. https://girai-report-2024-corrected-edition.tiiny.site/.

Adebara, Ife et Muhammad Abdul-Mageed. 2022. « Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go. » Tiré des *Procédures de la 60e assemblée générale annuelle de l'Association for Computational Linguistics (Volume 1 : Long Papers)*, révisé par Smaranda Muresan, Preslav Nakov et Aline Villavicencio, 3814– 41. Dublin, Irlande : Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.265.

Adebara, Ife, AbdelRahim Elmadany et Muhammad Abdul-Mageed. 2024. « Cheetah: Natural Language Generation for 517 African Languages. » Tiré des *Procédures de la 62e assemblée générale annuelle de l'Association for Computational Linguistics (Volume 1 : Long Papers)*, révisé par Lun-Wei Ku, Andre Martins et Vivek Srikumar, 12798–823. Bangkok, Thaïlande : Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.691.

Adebara, Ife, AbdelRahim Elmadany, Muhammad Abdul-Mageed et Alcides Inciarte. 2022. « AfroLID: A Neural Language Identification Tool for African Languages. » Tiré des *procédures du congrès de 2022 sur les méthodes empiriques du traitement de la langue naturelle*, révisé par Yoav Goldberg, Zornitsa Kozareva et Yue Zhang, 1958–81. Abu Dhabi, Émirats arabes unis : Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.128.

———. 2023. « SERENGETI: Massively Multilingual Language Models for Africa. » Tiré des résultats de l'Association for Computational Linguistics: ACL 2023, révisé par Anna Rogers, Jordan Boyd-Graber et Naoaki Okazaki, 1498–537. Toronto, ON: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.97.

Adelani, David Ifeoluwa, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter et coll. 2022. « A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. » Tiré des *procédures du congrès de 2022 de la section nord-américaine de l'Association for Computational Linguistics: Human Language Technologies*, révisé par Marine Carpuat, Marie-Catherine de Marneffe et Ivan Vladimir Meza Ruiz, 3053–70. Seattle, WA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.223.

Brooke, Peter. 2024. « Radio in Africa: Past and Present. » *Journal of African Cultural Studies* 36 (1): 1–5. https://doi.org/10.1080/13696815.2023.2294814.

Cheo, Victor Ngu, Esther Phubon Chie et Yollande Menguie. 2023. « An Evaluation of Media Use of Indigenous Languages in Cameroon. » *The International Journal of African Language and Media Studies*: 30–46. www.rhycekerex.org/an-evaluation-of-media-use-of-indigenouslanguages-in-cameroon.html.

Conroy-Krutz, Jeffrey et Joseph Koné. 2023. « Promise and peril: In changing media landscape, Africans are concerned about social media but opposed to restricting access. » Dépêche No 509, 18 février. Ghana: Afrobarometer. www.afrobarometer.org/wp-content/uploads/2022/04/AD509-PAP7-Promise-and-peril-Africas-changing-medialandscape-Afrobarometer-dispatch-19feb22.pdf.

Foster, Danny S. 2021. « Language of Instruction in Rural Tanzania: A Critical Analysis of Parents' Discursive Practices and Valued Linguistic Capabilities. » Thèse de doctorat, Université de Bristol. https://researchinformation.bris.ac.uk/en/studentTheses/language-of-instruction-in-rural-tanzania.

Fosu, Modestus. 2024. « Language Choice and the Problematics of Ideology in the Pre- and Post-Independence Ghanaian Press: A Historical and Cultural Analysis. » *Journalism and Media* 5 (3): 1194–210. https://doi.org/10.3390/journalmedia5030076.

Hammarström, Harald. 2018. « A survey of African languages. » Tiré de *The Languages and Linguistics of Africa*, révisé par Tom Güldemann, 1–57. Berlin, Allemagne : De Gruyter Mouton.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali et Monojit Choudhury. 2020. « The State and Fate of Linguistic Diversity and Inclusion in the NLP World. » Tiré de *Procédures de la 58e assemblée générale annuelle de l'Association for Computational Linguistics*, révisé par Dan Jurafsky, Joyce Chai, Natalie Schluter et Joel Tetreault, 6282–93. Association for Computational Linguistics, juillet. https://doi.org/10.18653/v1/2020.acl-main.560.

Lugalambi, George W., Peter G. Mwesige et Hendrik Bussiek. 2010. *Ouganda*. A Survey by the Africa Governance Monitoring and Advocacy Project, the Open Society Initiative for East Africa and the Open Society Media Program. Nairobi, Kenya : Open Society Initiative for East Africa. www.opensocietyfoundations.org/uploads/98954150-a6df-49eea680-d9d4a8cd07fe/uganda-publicbroadcasting-20100701.pdf.

Molale, Tshepang et Phillip Mpofu. 2023. « (Dis)continuities of African Language Radio on Social Media: The Case of South Africa's Motsweding FM and Radio Zimbabwe. » Tiré de *African Language Media*, révisé par Phillip Mpofu, Israel A. Fadipe et Thulani Tshabangu. Londres, R.-U. : Routledge. https://doi.org/10.4324/9781003350194.

Mthembu, Maxwell V. et Carolyne M. Lunga. 2020. « The extinctio of siSwati-language newspapers in the Kingdom of Eswatini. » Tiré de *African Language Media: Development, Economics and Management*, révisé par Abiodun Salawu. Londres, R.-U. : Routledge. https://doi.org/10.4324/9781003004738.

Myers, Mary. 2009. *Radio and Development in Africa: A Concept Paper*. Ottawa, ON: Centre de recherches pour le développement international. https://idl-bnc-idrc.dspacedirect.org/items/c805347c-9447-48f4-9bdc-b94031b655b4/full.

Myers, Mary et Nicola Harford. 2020. *Local Radio Stations in Africa: Sustainability or Pragmatic Viability?* Washington, DC: Center for International Media Assistance. Juin.www.cima.ned.org/publication/local-radio-stationsin-africa-sustainability-or-pragmatic-viability/.

National Broadcasting Commission. 2016. Nigeria Broadcasting Code. 6e édition. www.scribd.com/document/490616209/NBC-Code-6TH-EDITION.

National Communications Authority. 2023. *Liste des stations de radio VHF-FM autorisées du Ghana*. https://nca.org.gh/wp-content/uploads/2023/11/FM-LIST-Q2-2023.pdf.

Onyenankeya, Kevin. 2022. « Indigenous language newspapers and the digital media conundrum in Africa. » *Information Development* 38 (1): 83–96. https://doi.org/10.1177/0266666920983403.

Ouane, Adama et Christine Glanz. 2010. *How and why Africa should invest in African languages and multilingual education: An evidence- and practice-based policy advocacy brief*. Hambourg, Allemagne : Institut de l'UNESCO pour l'apprentissage tout au long de la vie. https://files.eric.ed.gov/fulltext/ED540509.pdf.

Petzell, Malin. 2012. « The linguistic situation in Tanzania. » *Moderna Språk* 106 (1): 136–44. https://doi.org/10.58221/mosp.v106i1.8233.

ResCue, Elvis et G. Edzordzi Agbozo. 2021. « Creating Translated Interfaces: The Representations of African Languages and Cultures in Digital Media. » Tiré de *Rethinking Language Use in Digital Africa: Technology and Communication in Sub-Saharan Africa*, révisé par Leketi Makalela et Goodith White, 51–72. Bristol, R.-U. : Multilingual Matters and Channel View Publications. https://doi.org/10.2307/jj.22730532.7.

Salawu, Abiodun, éd. 2020. *African Language Media: Development, Economics and Management*. Londres, R.-U. : Routledge. https://doi.org/10.4324/9781003004738.

Sunday, Oloruntola, Ayo Yusuff, Simon Godwin Iretomiwa, Vincent Adakole Obia et Samuel Ejiwunmi, éds. 2018. « Use of indigenous languages for social media communication: The Nigerian experience. » Tiré de *African Language Digital Media and Communication*, révisé par Abiodun Salawu. 1ère éd. Londres, R.-U.: Routledge.

Tshabangu, Thulani et Abiodun Salawu. 2022. « Indigenous language Media Research in Africa: Gains, Losses, Towards a New Research Agenda. » *African Journalism Studies* 43 (1): 1–16. https://doi.org/10.1080/23743670.2021.1998787.

Yevudey, Elvis. 2018. « The representation of African languages and cultures on social media: A case of Ewe in Ghana. » Tiré de *The Routledge Handbook of African Linguistics*, révisé par Augustine Agwuele et Adams Bodomo, 1ère éd. Londres, R.-U. : Routledge.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

Credits

Directrice, gestionnaire de programmes Dianna English
Gestionnaire de programmes Ifeoluwa Olorunnipa
Gestionnaire, publications Jennifer Goyder
Conception graphique Abhilasha Dewan

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

CIGI