Langues africaines et traduction automatique : exploration des outils existants et stratégies de pérennisation dans l'environnement numérique

Emmanuel Tapon

Mots-clés : Traduction automatique ; langues minorisées ; décolonisation ; revitalisation linguistique ; langues africaines.

Keywords : Machine translation ; minoritized languages ; decolonization ; language revitalization ; African languages.

Résumé

La traduction automatique (TA) constitue un levier technologique essentiel pour la préservation et la promotion des langues africaines minorisées dans l'environnement numérique. Cet article évalue l'impact de la TA sur ces langues en examinant trois dimensions complémentaires : son application dans les processus de documentation linguistique, sa contribution au développement de ressources numériques et son potentiel comme outil de revitalisation linguistique. Notre analyse met en évidence les obstacles majeurs liés à la représentativité des données, aux considérations éthiques du développement technologique et aux risques d'impérialisme linguistique. Face à ces enjeux, nous proposons un cadre méthodologique pour une TA collaborative, critique et décoloniale, qui s'appuie sur des approches participatives, des infrastructures ouvertes et une implication accrue des institutions africaines. L'article souligne l'importance d'intégrer la TA dans une stratégie globale de préservation de la diversité linguistique africaine.

Abstract

Machine translation (MT) is a key technological tool for the preservation and promotion of minoritized African languages in the digital environment. This study sets out to evaluate MT's impact on these languages by examining three complementary dimensions: its application to linguistic documentation processes, its contribution to digital resource development, and its potential as a language revitalization tool. Our analysis highlights major obstacles related to data representativeness, ethical considerations in technological development, and risks of linguistic imperialism. In response to these challenges, we propose a methodological framework for a collaborative, critical, and decolonial MT, based on participatory approaches, open infrastructure, and greater involvement of African institutions. The study emphasizes the importance of integrating MT into a comprehensive strategy for preserving Africa's linguistic diversity.

Pour citer cet article

Tapon, E. (2026). Langues africaines et traduction automatique : exploration des outils existants et stratégies de pérennisation dans l'environnement numérique. *Facteurs humains : revue en sciences humaines et sociales de l'Université Laval*, *3*(1), 195-218. https://doi.org/10.62920/3wr56492

© Emmanuel Tapon, 2026. Publié par *Facteurs humains : revue en sciences humaines et* sociales *de l'Université Laval*. Ceci est un article en libre accès, diffusé sous licence Attribution 4.0 International (CC BY 4.0).



Liste des abréviations

AfroLID : African Language Identification (Outil d'identification des langues africaines)

BLEU : Bilingual Evaluation Understudy (Métrique d'évaluation de la traduction automatique)

IA : Intelligence artificielle

LID : Language Identification (Identification de la langue)

METEOR : Metric for Evaluation of Translation with Explicit ORdering (Métrique d'évaluation de la traduction automatique)

MT : Machine Translation (Traduction automatique)

NLP : Natural Language Processing (Traitement automatique du langage naturel)

TA : Traduction automatique

TALN : Traitement automatique du langage naturel

TAN : Traduction automatique neuronale

Introduction

L'Afrique est le continent qui abrite la plus grande diversité linguistique au monde, avec plus de 2 000 langues parlées, soit près d'un tiers des langues vivantes de la planète (Eberhard et al., 2024). Cependant, cette richesse linguistique est menacée par de multiples facteurs, tels que la domination des langues héritées de la colonisation dans l'éducation et l'administration, les dynamiques d'urbanisation et de migration qui favorisent les langues véhiculaires ou encore la marginalisation économique et politique des communautés linguistiques minorisées (Djité, 2008). Face à ces défis, la documentation systématique des langues africaines – comprenant la collecte de données linguistiques, l'enregistrement des pratiques orales, la description des structures grammaticales et lexicales ainsi que la compilation des savoirs culturels associés – et leur préservation, en particulier pour les langues peu dotées et en danger[8], apparaissent comme un enjeu crucial pour la sauvegarde du patrimoine culturel et de la diversité linguistique du continent.

Dans cet article, nous distinguons les **langues minorisées** (marginalisées par des politiques linguistiques) et les **langues peu dotées** (disposant de peu de ressources numériques pour le développement technologique). Ces catégories se recoupent largement dans le contexte africain, mais ne sont pas identiques : une langue peut avoir un nombre important de locutrices et locuteurs, mais être technologiquement peu dotée, et vice-versa.

[8] Une langue est dite en danger lorsqu'elle est surtout utilisée par la génération des parents et leurs ascendants (UNESCO, 2003).

L'essor spectaculaire de l'intelligence artificielle (IA), qui connaît un tournant décisif depuis 2022 avec l'émergence de modèles de langage généralistes comme ChatGPT (OpenAI, 2022) et GPT-4 (OpenAI, 2023), a profondément transformé le paysage technologique mondial. Cette révolution, portée principalement par des entreprises américaines et chinoises, reproduit et amplifie les asymétries géopolitiques au détriment des pays en développement et de leurs langues (Bender et al., 2021). Les modèles entraînés sur des corpus dominés par l'anglais et quelques langues européennes perpétuent les exclusions linguistiques héritées de la colonisation (Blasi et al., 2022), ce qui soulève des questions essentielles sur la représentation équitable de la diversité linguistique mondiale (Joshi et al., 2021).

La traduction automatique (TA), en tant que composante historique et stratégique de l'IA appliquée au langage (Koehn, 2020), s'inscrit pleinement dans cette dynamique. Elle partage avec les technologies d'IA générative les mêmes biais structurels : surreprésentation des langues dominantes, marginalisation des langues peu dotées et concentration des capacités de développement dans les pays industrialisés (Nekoto et al., 2020). Cependant, la TA présente une particularité : elle constitue le point d'entrée privilégié des langues minorisées dans l'écosystème numérique, contrairement aux autres applications d'IA qui présupposent déjà une présence numérique consolidée (Joshi et al., 2021). Cette position stratégique explique pourquoi nous concentrons notre analyse sur la TA. Si les défis communs à l'ensemble des langues peu dotées (manque de données, biais algorithmiques) demeurent (Joshi et al., 2020), ils se posent avec une acuité particulière pour les langues africaines, qu'il s'agisse du besoin pressant de corpus parallèles fiables ou de la modélisation de leurs particularités typologiques, comme le marquage tonal ou les constructions verbales en série (Adebara et Abdul-Mageed, 2022, p. 3816-3817). La TA offre toutefois des perspectives prometteuses pour la documentation linguistique et la revitalisation culturelle. Le terme TA exclusivement sera donc utilisé dans la suite de ce travail pour désigner ce sous-domaine spécifique de l'IA où se joue une bataille décisive pour l'inclusion numérique des langues africaines.

Dans ce contexte, l'avancée des technologies de TA, de plus en plus simples à utiliser et accessibles au grand public (Ballier et Zimina-Poirot, 2023, p. 166), offre de nouvelles perspectives pour appuyer les efforts de documentation, de création de ressources et de revitalisation des langues africaines minorisées à l'ère numérique. En effet, les progrès récents en matière de TA, fondés sur des modèles d'apprentissage profond, ont considérablement amélioré la qualité et la fluidité des traductions pour de nombreuses paires de langues (Koehn, 2020, p. 40). Cependant, la plupart de ces progrès n'ont bénéficié qu'aux quelques langues majoritaires et bien dotées en ressources numériques, laissant de côté la grande majorité des langues africaines minorisées (Adebara et Abdul- Mageed, 2022, p. 1 ; Nekoto et al., 2020, p. 1). Celles-ci font partie du groupe des langues peu dotées, un statut qui freine leur accès aux avancées majeures en matière de TA. Par ailleurs, comme le soulignent Orife et al. (2020), malgré l'élaboration de stratégies prometteuses pour surmonter le problème de la faible disponibilité des ressources, les langues africaines n'ont pas encore bénéficié d'une application substantielle de ces avancées. Cette situation est particulièrement préoccupante dans la mesure où « jusqu'à 10 % des langues africaines – en particulier celles parlées par les petites communautés linguistiques – risquent de disparaître au cours des cent prochaines années » (UNESCO, 2010, p. 21). Par conséquent, il semble nécessaire de mettre au point rapidement des outils de TA pour contribuer à leur préservation. Ce défi s'adresse particulièrement à la communauté africaine des praticiens et praticiennes du traitement automatique du langage naturel (TALN) et appelle à une mobilisation accrue des ressources et des compétences pour la protection de la diversité linguistique. L'extension de la TA aux langues africaines

s'inscrit ainsi dans une démarche plus large de décolonisation numérique et de valorisation de la diversité linguistique. Elle revêt par conséquent une dimension culturelle et sociale fondamentale.

Cet article vise à explorer le rôle potentiel de la TA dans la préservation de ces langues, en analysant à la fois les possibilités et les défis que présente le développement de technologies langagières en contexte africain. Dans une première partie, nous dresserons un état des lieux de la TA dans le contexte des langues minorisées, en soulignant les avancées récentes et les obstacles spécifiques rencontrés. Nous examinerons ensuite, au moyen d'exemples de projets et d'initiatives en cours, comment la TA peut contribuer à la documentation des langues africaines minorisées et favoriser leur revitalisation et leur réappropriation grâce à la création de ressources linguistiques. Dans une troisième partie, nous discuterons des enjeux et des limites de la TA pour les langues minorisées, en mettant l'accent sur les questions de représentativité des données, d'équité dans le développement technologique et d'incidence sur les écosystèmes linguistiques. Enfin, nous proposerons des pistes de réflexion pour favoriser l'émergence d'une TA décoloniale et participative, au service de l'autonomisation des communautés linguistiques africaines, en insistant sur l'importance des collaborations interdisciplinaires, des approches critiques et des infrastructures ouvertes.

1 TA et langues minorisées : état des lieux

1.1 Progrès récents en matière de TA neuronale

Au cours de la dernière décennie, le domaine de la TA a connu une véritable révolution avec l'avènement des modèles neuronaux basés sur l'apprentissage profond (Zhang et Zong, 2020, p. 2029). Cette évolution marque un tournant décisif par rapport aux approches précédentes, fondées sur des règles linguistiques ou des statistiques. Comme la TA statistique, la TA neuronale s'entraîne sur de vastes corpus de textes parallèles, c'est-à-dire des textes dont chaque phrase source est alignée avec sa traduction dans la langue cible. Toutefois, contrairement à sa devancière qui traduisait par segments de quelques mots, la TA neuronale traite des phrases entières en modélisant les relations sémantiques et syntaxiques entre les mots. Le perfectionnement des architectures de réseaux neuronaux artificiels (Bowker, 2024) a ainsi conduit à une amélioration considérable de la qualité des traductions. Dans certains cas, les résultats atteignent un niveau proche de celui des traducteurs humains (Hassan et al., 2018, p. 1), ce qui représente une avancée remarquable dans le domaine. Cette progression s'est traduite non seulement par une amélioration de la précision des traductions, mais également par un élargissement du champ des applications possibles de la TA. Au-delà de la simple traduction de documents, ces progrès ont ouvert de nouvelles possibilités. Par exemple, la traduction automatique de sites web, d'applications et de contenus multimédias divers permet désormais un accès élargi à l'information dans différentes langues, bien que la qualité demeure perfectible. Cet accès favorise une diffusion plus large et plus rapide des connaissances à l'échelle mondiale, réduisant ainsi les barrières linguistiques dans divers domaines tels que l'éducation, la recherche scientifique et le commerce international.

Un autre domaine d'application prometteur est la mise au point d'outils d'aide à la traduction pour les professionnels et les apprenants de langues (Barysevich et Costaris, 2021). Ces outils, qui s'appuient sur les avancées de la TA neuronale, peuvent assister les traductrices et traducteurs humains en leur proposant des suggestions de traduction, en vérifiant la cohérence terminologique ou en fournissant des analyses

contextuelles. Pour les apprenants de langues, ces outils offrent de nouvelles possibilités d'immersion linguistique et d'apprentissage interactif. Cette piste pourrait être explorée une fois que des avancées considérables auront été enregistrées dans la documentation des langues africaines minorisées, afin d'en accroître le nombre de locutrices et locuteurs. Les retombées de ces avancées s'étendent également au domaine de la localisation et de l'internationalisation des produits et services. Les entreprises peuvent désormais adapter plus facilement leurs offres à différents marchés linguistiques, ce qui favorise une mondialisation inclusive et diversifiée.

Malgré les progrès considérables de la TA, plusieurs défis persistent, notamment en ce qui concerne la traduction de nuances culturelles et sociolinguistiques ou d'expressions idiomatiques (Naveen et Trojovský, 2024, p. 8). Ces défis stimulent la recherche, dont les avancées pourraient transformer davantage notre façon de communiquer par-delà les barrières linguistiques. Cependant, les progrès actuels ne profitent pas également à toutes les langues : pour les langues peu dotées, le développement de systèmes de TA performants se heurte à des obstacles particuliers.

1.2 Défis spécifiques de la TA pour les langues peu dotées

La grande majorité des langues du monde restent largement sous-représentées dans le domaine de la TA. Cette fracture numérique linguistique touche particulièrement les langues minorisées et peu dotées (Mager et al., 2018 ; Nekoto et al., 2020). Dans le contexte africain, moins de 2 % des langues sont intégrées dans les systèmes de TA majeurs (Martinus et Abbott, 2019). Malgré des progrès notables, c'est aussi moins de 2 % des langues africaines qui sont prises en charge de façon fiable par les grands modèles de langage (Hussen et al., 2025). La situation n'est pas très différente dans les Amériques où très peu de technologies ont été développées pour prendre en charge les langues autochtones, beaucoup d'entre elles étant d'ailleurs menacées d'extinction (Mager et al., 2018, p. 1), même si des efforts de recherche structurés ont pris forme ces dernières années autour d'initiatives comme AmericasNLP, qui organise depuis 2021 des tâches partagées de traduction automatique, portant en 2025 sur 13 langues autochtones (de Gibert et al., 2025). Selon la typologie de Joshi et al. (2020), la majorité des quelque 7 000 langues recensées dans le monde (Eberhard et al., 2024) demeurent exclues ou marginalement représentées dans les technologies de TA ; seule une infime fraction d'entre elles bénéficie d'une prise en charge réelle par des plateformes de premier ordre, comme Google Translate ou Microsoft Translator, avec une qualité très inégale.

Ce fossé technologique s'explique en grande partie par la rareté des données numériques disponibles pour entraîner les modèles de TA dans ces langues. En effet, contrairement aux langues dominantes, comme l'anglais ou le français, pour lesquelles il existe de vastes corpus de textes parallèles produits notamment par des institutions internationales ou des médias, la plupart des langues minorisées ne disposent que de peu de ressources écrites numérisées ou n'en disposent simplement pas (Nekoto et al., 2020, p. 1).

Cette situation est particulièrement préoccupante en Afrique, où la grande majorité des 2 000 langues sont parlées par des peuples qui ont historiquement privilégié la transmission orale de leur culture et de leurs savoirs, ce qui se traduit par une faible standardisation écrite et une présence très limitée sur le Web. La préservation des traditions orales est en effet tributaire du pouvoir de la mémoire de générations d'êtres humains qui se succèdent (Vansina, 2006, p. 1). De plus, ces langues présentent souvent des caractéristiques linguistiques complexes, telles qu'une morphologie riche (c'est-à-dire de nombreuses variations de forme des

mots selon leur fonction grammaticale) ou une syntaxe éloignée de celle des langues indo-européennes, qui rendent difficile l'application directe des modèles de TA conçus initialement pour ces dernières et qui nécessitent d'autres outils plus adaptés (Enguehard et Kané, 2004, p. 10). Enfin, la grande variation dialectale au sein de ces langues ainsi que le manque de ressources linguistiques de base, comme des dictionnaires ou des grammaires de référence, compliquent davantage la tâche de construction de systèmes de TA robustes et adaptés aux réalités locales. Au-delà de ces défis techniques, cette situation s'enracine dans des contraintes économiques structurelles rarement explicitées, mais déterminantes. Le développement de systèmes de TA pour les langues minorisées se heurte à une logique de marché implacable : ces langues, souvent parlées par des populations qui maîtrisent déjà les langues véhiculaires héritées de la colonisation, ne sont pas considérées comme économiquement viables par les entreprises technologiques. Cette réalité économique explique pourquoi, malgré l'ancienneté relative de la TA, l'investissement privé demeure faible pour ces langues, une réalité qui contribue à perpétuer leur marginalisation numérique. À cette dimension s'ajoute un enjeu écologique croissant : l'empreinte énergétique considérable des modèles de TA neuronale met en doute leur durabilité dans un contexte africain où les ressources limitées pourraient être prioritairement allouées à d'autres besoins de développement fondamentaux. Cependant, il convient de reconnaitre que l'émergence du standard Unicode[9] pour le codage des caractères spéciaux permet de lever progressivement les obstacles à l'écriture de ces langues (Enguehard et Kané, 2004, p. 15). Cette évolution s'est concrétisée à travers plusieurs initiatives au cours des deux dernières décennies. Bodomo et al. (2006) ont mis au point un clavier Unicode pour le dagaaré et le twi, deux langues ghanéennes, en utilisant des représentations binaires et hexadécimales pour codifier leurs graphèmes complexes et leurs signes diacritiques. Plus récemment, l'exemple du yoruba est venu confirmer cette tendance : alors que les schémas de codage traditionnels, dont l'ASCII et même Unicode, ne permettaient pas de représenter correctement les signes diacritiques et les points souscrits caractéristiques de cette langue, Ajao et al. (2020) ont proposé l'utilisation de représentations binaires et hexadécimales pour codifier les caractères yoruba complexes en un seul point de code Unicode, ce qui facilite leur transformation en entités HTML et leur intégration dans des claviers numériques fonctionnels. Ces initiatives démontrent comment Unicode constitue désormais un outil essentiel pour l'inclusion numérique des langues africaines et leur utilisation effective dans les interfaces humain-machine.

1.3 Initiatives de TA en faveur des langues africaines

Face aux multiples défis liés à la faible disponibilité des ressources numériques pour les langues africaines, une dynamique encourageante se dessine depuis quelques années. Une communauté croissante de recherche et de pratique, associant expertises africaines et collaborations avec des partenaires d'autres régions du monde, s'est mobilisée pour créer des ressources et des outils de TA adaptés à la diversité linguistique du continent.

Le projet Masakhane (Okolo, 2023 ; Orife et al., 2020) se démarque comme l'une des initiatives les plus prometteuses dans ce domaine. Ce collectif de recherche participative illustre parfaitement l'approche collaborative nécessaire pour relever les défis de la TA en Afrique. En impliquant activement les locutrices et

[9] Unicode est un standard international d'encodage informatique qui attribue un identifiant numérique unique à chaque caractère, symbole et signe de ponctuation utilisé dans les systèmes d'écriture du monde entier. Mis au point depuis 1987, ce standard vise à permettre la représentation, le traitement et l'échange numériques de textes dans toutes les langues. Il tient compte des écritures contemporaines, historiques et spécialisées (The Unicode Consortium, 2023).

locuteurs et les communautés dans le processus, Masakhane vise à construire des jeux de données, c'est-à-dire des corpus annotés et mis en forme typiquement pour l'entraînement des modèles de TA, et des modèles de TA pour une cinquantaine de langues africaines. Cette approche ascendante permet non seulement de générer des ressources linguistiques précieuses, mais aussi de renforcer les capacités locales et de valoriser les connaissances endogènes. Plus récemment, cette dynamique de collaboration s'est institutionnalisée avec la création, en juillet 2025, du *Masakhane African Languages Hub*, qui étend cette approche communautaire à la constitution de jeux de données multimodaux (texte, parole, image) pour 40 langues africaines, avec le soutien de Google.org, du ministère britannique des Affaires étrangères, du Commonwealth et du Développement, du Centre de recherche pour le développement international (CRDI) et de la Fondation Gates (CRDI, 2025).

Dans le même ordre d'idées, le projet GhanaNLP (Okolo, 2023) a enregistré des avancées considérables en se concentrant sur les langues du Ghana. En combinant des méthodes automatiques et manuelles de collecte et d'annotation de données, cette initiative a permis de rassembler des ressources linguistiques et de développer des outils de TALN, y compris des systèmes de TA, pour 17 langues ghanéennes. Selon les communications de l'équipe, le projet poursuit son expansion avec l'ajout d'autres langues africaines et intègre progressivement des fonctionnalités de reconnaissance vocale automatique (NLP Ghana, 2023). Ce projet démontre l'importance d'une approche hybride alliant technologies de pointe et expertise humaine locale.

D'autres initiatives se concentrent sur le développement d'applications et de plateformes de TA destinées directement aux usagers des langues africaines. C'est le cas, par exemple, de Lanfrica, une initiative communautaire en ligne visant à résoudre les problèmes de faible visibilité, de manque de bancs d'essai publics et de partage limité des recherches existantes sur les langues africaines. Son objectif principal est de documenter les travaux de recherche, les résultats, les bancs d'essai et les projets sur les langues africaines, et de les rendre facilement accessibles à toutes et à tous (Emezue et Dossou, 2020). En octobre 2025, la plateforme a lancé sa version 2.0 et annoncé l'extension de sa mission au-delà des seules ressources linguistiques pour couvrir d'autres domaines de l'IA (santé, climat, agriculture), avec l'objectif de répertorier plus d'un million de ressources africaines d'ici la fin de 2025 (Lanfrica, 2025).

Ces initiatives et bien d'autres témoignent d'une prise de conscience croissante de l'importance de développer des solutions de TA adaptées aux contextes africains, au service des besoins et des usages locaux. Elles mettent en lumière le potentiel de collaboration entre personnes chercheuses, ingénieures, linguistes et communautés locales pour surmonter les obstacles liés au manque de ressources numériques. En favorisant l'appropriation locale des technologies de TA, ces projets contribuent non seulement à la préservation et à la promotion des langues africaines, mais aussi à l'autonomisation technologique du continent. Cependant, de nombreux défis persistent, notamment la pérennisation de ces initiatives, leur mise à l'échelle et leur intégration dans des politiques linguistiques plus larges. Ces enjeux sont fondamentaux pour consolider durablement la diversité linguistique et l'inclusion numérique en Afrique.

Par ailleurs, l'Initiative de traduction des langues africaines (ALTI), qui s'inscrit dans le cadre de la communication interculturelle et de la diplomatie publique, mérite également d'être mentionnée ici. Lancée par l'Université Howard aux États-Unis, cette initiative vise à combler le fossé linguistique entre les États-Unis et l'Afrique en traduisant des documents stratégiques dans des langues africaines majeures. Dans le contexte du Sommet des dirigeants États-Unis–Afrique de l'administration Biden, l'ALTI a entrepris la traduction de la

« Stratégie américaine envers l'Afrique subsaharienne » en amharique, en swahili, en yoruba et en zoulou. Cette démarche s'inscrit dans une approche de diplomatie linguistique (Kourliandsky, 2023) et de *soft power* (Nye, 2004), visant à favoriser une compréhension mutuelle profonde et à démocratiser l'accès à l'information politique. En mobilisant les ressources universitaires pour soutenir les objectifs diplomatiques, l'ALTI illustre le concept de « diplomatie de connaissance » (Knight, 2022), où les institutions éducatives jouent un rôle crucial dans la promotion du dialogue international et la construction de ponts culturels. Cette initiative souligne également l'importance croissante de la traduction dans la gouvernance mondiale et les relations internationales (Footitt et Kelly, 2012), en particulier dans le contexte des relations entre les États-Unis et l'Afrique. Comme on peut le constater, aussi louable soit-elle, cette initiative ne porte, à présent, que sur des langues considérées comme dotées, ce qui confirme que des efforts ciblés doivent être fournis en faveur des langues disposant jusqu'ici de peu de ressources.

2 Potentiel de la TA pour la préservation des langues africaines

2.1 Documentation et description linguistique

La TA peut constituer un levier important pour la préservation des langues africaines, notamment en facilitant leur documentation. Face à l'urgence de sauvegarder les nombreuses langues menacées d'extinction sur le continent, la TA peut offrir des outils précieux pour automatiser et accélérer certaines tâches de collecte, de transcription et d'annotation des données linguistiques (Anastasopoulos, 2019). Par exemple, des systèmes de TA peuvent être utilisés pour aligner automatiquement des enregistrements audio ou vidéo de parole spontanée avec leurs transcriptions, ce qui facilite la création de corpus oraux pour des langues disposant de peu de ressources écrites (Godard et al., 2018).

De même, la TA peut contribuer à la création de ressources parallèles pour des langues minorisées, mais cette approche nécessite une stratégie progressive. Bien que les systèmes de TA actuels soient encore limités pour ces langues, une approche itérative est possible : les premiers corpus parallèles, même de taille modeste, permettent de développer des systèmes de TA initiaux qui, bien qu'imparfaits, peuvent aider à la création de nouvelles ressources sous supervision humaine. Ces textes traduits, une fois validés et corrigés par des locutrices et locuteurs natifs, peuvent ensuite servir de base pour créer des corpus annotés avec différents niveaux d'information linguistique – à l'image du Texte Glosé Interlinéaire[10] (morphologique, syntaxique, sémantique) – essentiels pour la recherche en linguistique descriptive et typologique (Anastasopoulos, 2019, p. 2). Il existe des techniques permettant d'entraîner des systèmes de TA sur des corpus parallèles de faible densité associés à des corpus monolingues (Karakanta et al. 2018 ; Ranathunga et al., 2021, p. 5-6). Cette approche cyclique, où chaque amélioration du système facilite la création de nouvelles ressources, peut constituer un point de départ viable pour le développement progressif de corpus plus importants.

[10] De l'anglais *Interlinear Glossed Text* (IGT). Format standard en linguistique documentaire qui présente, ligne par ligne, la transcription d'un énoncé, son analyse morphologique (gloses) et sa traduction libre dans une langue de grande diffusion.

2.2 Développement de ressources et outils de traitement automatique du langage

Au-delà de la documentation linguistique, la TA peut également jouer un rôle clé dans la mise au point d'autres ressources et outils relevant du TALN pour les langues africaines minorisées, dans la mesure où certaines méthodes de TALN tirent parti des outils d'annotation déjà utilisés pour des langues mieux dotées (Bernhard et Soria, 2018, p. 8). En effet, à l'image des méthodes de projection d'annotations déjà mobilisées pour amorcer le développement d'outils dans les langues sous-dotées, les systèmes de TA offrent une stratégie pragmatique pour générer rapidement un volume initial important de données textuelles. Même bruitées, c'est-à-dire contenant des erreurs ou des éléments parasites, ces données peuvent servir de base pour entraîner des modèles de langage, des analyseurs morphologiques ou des outils d'extraction terminologique. Leur utilité réside dans leur capacité à fournir la masse critique nécessaire au démarrage de projets de traduction de langues sous-dotées, qui serait autrement difficile ou très longue à rassembler manuellement. Les données générées peuvent ensuite être raffinées (par post-édition, filtrage ou combinaison avec des corpus plus petits mais de haute qualité) pour améliorer progressivement la performance des outils développés.

Des progrès notables ont été réalisés dans le développement de systèmes de traduction automatique neuronale (TAN) pour des langues africaines. Abbott et Martinus (2018) ont mis en œuvre un modèle de transformeur pour ces langues, qui a largement surpassé les architectures existantes de TA statistique pour la traduction de l'anglais vers le setswana sud-africain. Le transformeur (ou *transformer* en anglais) est un modèle d'apprentissage profond décrit pour la première fois en 2017 dans un article intitulé « Attention Is All You Need » (Vaswani et al., 2017). Il s'agit d'une architecture de réseau neuronal qui traite les séquences de texte en parallèle plutôt que séquentiellement, permettant ainsi de mieux établir des dépendances à longue distance dans la langue. Le mécanisme d'attention sur lequel il s'appuie permet au modèle de se concentrer sur différentes parties de la phrase source lors de la génération de la traduction, améliorant ainsi la qualité des traductions produites. Dans leurs travaux, Martinus et Abbott (2019, p. 1-2) ont étendu la recherche en entraînant des modèles de TAN de l'anglais vers cinq langues sud-africaines au moyen de deux architectures différentes : le modèle convolutionnel séquence-à-séquence et le transformeur. Leurs résultats[11] démontrent le potentiel prometteur des modèles de TAN pour les langues africaines.

Dans le domaine des langues créoles, Ogueji et Ahia (2019) ont réalisé un travail pionnier en constituant le plus grand corpus connu de pidgin nigérian et ont entraîné les premiers modèles de traduction entre l'anglais et cette langue. Face à l'absence de données parallèles, ils ont utilisé des techniques de TAN non supervisées. Cette approche permet d'entraîner des modèles de traduction en utilisant uniquement des textes monolingues dans chaque langue, sans nécessiter de traductions préexistantes (Lample et al., 2018). Le système apprend à traduire en découvrant automatiquement les correspondances entre les deux langues à partir de leurs structures et distributions statistiques similaires. Par exemple, en identifiant que certains mots ou expressions apparaissent dans des contextes similaires dans les deux langues, le système peut inférer qu'ils sont probablement des traductions l'un de l'autre (Artetxe et al., 2019). Ces avancées illustrent le

[11] Les résultats varient beaucoup selon la langue traduite : ils sont excellents pour le xitsonga (score de qualité proche de 50 sur 100), mais très faibles pour l'isiZulu (à peine 3 sur 100) – un écart que les auteurs expliquent surtout par la quantité de données disponibles pour chaque langue (Martinus et Abbott, 2019, p. 4-5). Dans l'ensemble, le modèle transformeur a donné de meilleurs résultats que le modèle convolutionnel séquence-à-séquence.

potentiel de la TAN dans le développement de ressources documentaires pour les langues africaines minorisées. La facilité d'accès relative des corpus comparables plutôt que de corpus parallèles (Déjean et Gaussier, 2002, p. 3) constitue un argument important en faveur de l'adoption de cette approche. Le principal défi à relever sera cependant de travailler à rendre disponibles des corpus monolingues de haute qualité pour permettre une extraction optimale des données d'entraînement.

2.3 Réappropriation des langues et promotion de leur vitalité

La TA offre des perspectives prometteuses pour la revitalisation et la valorisation des langues africaines minorisées, allant au-delà de ses applications en recherche et en ingénierie linguistique. En facilitant la traduction rapide et économique de contenus numériques dans ces langues, la TA peut accroître de façon substantielle leur présence en ligne, élargissant leurs domaines d'usage au-delà des sphères vernaculaires (échanges familiaux, rituels locaux, etc.). Cette visibilité numérique renforce leur légitimité symbolique auprès des jeunes générations, favorisant ainsi une réappropriation durable qui consolide la transmission intergénérationnelle (Fishman, 1991, p. 87-91). Cette accessibilité accrue peut renforcer l'identité culturelle des populations locutrices et valoriser les savoirs locaux (Cru, 2014, p. 285 ; Osborn, 2010), tout en fournissant des ressources pédagogiques essentielles pour l'enseignement de ces langues.

Cependant, plusieurs obstacles limitent la portée de cette approche, le premier étant la qualité de la traduction en raison du manque de données d'entraînement (Nekoto et al., 2020). Il existe aussi un risque de standardisation excessive au détriment des variantes dialectales (Bird, 2020). De plus, l'implication active des communautés locales dans le développement et l'utilisation des outils de TA demeure insuffisante et mérite d'être renforcée. Enfin, cette approche doit surmonter le « linguicisme » ambiant que Skutnabb-Kangas (1988, p. 41) définit comme un ensemble d'idéologies et de structures servant à légitimer, mettre en œuvre et reproduire une répartition inégale du pouvoir et des ressources entre des groupes définis sur la base de leur langue maternelle. Ce phénomène tend à marginaliser les communautés locutrices des langues minorisées en renforçant davantage les langues déjà dominantes (Guo et Beckett, 2007). Les populations locutrices des langues ainsi minorisées n'ont généralement, dans leur lutte pour la survie, d'autre choix que de délaisser leur patrimoine linguistique au profit des langues véhiculaires. Dans une démarche visant à inverser cette tendance à l'abandon, plusieurs initiatives récentes ont cherché à impliquer directement les communautés locutrices dans des projets de traduction collaborative en ligne afin de produire des ressources linguistiques pour leurs propres langues. Par exemple, dans le cadre du projet AfroLID, une boîte à outils d'identification de langues basée sur des réseaux neuronaux artificiels a été mise au point pour 517 langues et variétés africaines (Adebara et al., 2022, p. 1). Ce projet exploite un ensemble de données Web multidomaines manuellement organisées à partir de 14 familles linguistiques utilisant 5 systèmes orthographiques. L'identification de la langue (Language Identification, LID) est un précurseur important pour le traitement du langage naturel, en particulier l'exploitation des données Web. Cette étape de reconnaissance linguistique est essentielle : elle permet aux systèmes de traitement automatique de sélectionner les modèles et les ressources linguistiques appropriés pour chaque langue identifiée et conditionne ainsi l'efficacité de toutes les tâches de traitement ultérieures, telles que la TA, l'analyse syntaxique et la reconnaissance vocale (Jauhiainen et al., 2019 ; Torfi et al., 2020). Dans le contexte des données Web, où les contenus multilingues abondent sans annotation explicite de la langue, les systèmes LID performants deviennent indispensables pour structurer et exploiter efficacement ces ressources textuelles (Apple Machine Learning Research, 2020). AfroLID vise à contribuer à la résolution du problème pressant du manque de couverture des technologies LID pour la majorité des

quelque 7 000 langues du monde, en se concentrant sur les langues africaines. Cette approche novatrice pourrait permettre non seulement d'améliorer les outils technologiques pour ces langues, mais aussi de renforcer l'autonomisation des communautés linguistiques dans le processus de développement et de validation des ressources.

3 Enjeux et limites de la TA pour les langues africaines minorisées

3.1 Représentativité et qualité des données d'entraînement

Le développement de systèmes de TA pour les langues africaines minorisées s'accompagne de certains défis qui vont au-delà des obstacles techniques classiques. Ils touchent à la fois la qualité des données d'entraînement, les considérations éthiques du développement technologique et les risques d'impact sur les écosystèmes linguistiques existants. Le manque de données représentatives pour entraîner les modèles de TA dans les langues africaines minorisées constitue un obstacle majeur qui peut conduire à des biais importants dans les systèmes de TA (Ji, 2023, p. 100), étant donné que le déséquilibre dans les données d'entraînement peut conduire les modèles de TA à reproduire et à amplifier certains stéréotypes ou préjugés sociaux présents dans les textes sources. Bien que les modèles de TALN aient démontré leur efficacité dans diverses applications, ils ont tendance à propager, et parfois même à amplifier, les biais de genre présents dans les corpus textuels utilisés pour leur entraînement. Par exemple, une étude menée par Prates et al. (2020) sur Google Translate a révélé d'importants biais de genre dans la traduction de phrases neutres sur le plan du genre vers des langues dans lesquelles les pronoms sont genrés. Les personnes chercheuses ont constaté que le système avait tendance à traduire des mots anglais neutres comme « doctor » ou « engineer » en utilisant des pronoms masculins dans les langues cibles, tandis que « nurse » ou « baker » étaient systématiquement traduits avec des pronoms féminins, reproduisant ainsi des stéréotypes de genre plutôt que de proposer des traductions équilibrées (Prates et al., 2020, p. 63-65). Bien que l'étude des biais dans l'IA ne soit pas nouvelle, les méthodes visant à atténuer les biais de genre dans le TALN sont relativement récentes. Il est donc essentiel de poursuivre les recherches sur la reconnaissance et l'atténuation de ces biais, en particulier dans le contexte des langues africaines – dont la majorité présente des systèmes de classification nominale complexes très différents du genre grammatical masculin/féminin des langues indo-européennes. Ces langues, notamment celles du groupe Niger-Congo qui regroupe la plupart des langues d'Afrique subsaharienne, possèdent des systèmes de classes nominales pouvant comporter de 10 à 30 catégories. Les catégories sont organisées autour de critères sémantiques variés (animé/inanimé, humain/non-humain, forme des objets, mobilité, etc.) plutôt que sur le sexe biologique (Creissels, 2000). Ces systèmes de classification, qui se fondent sur d'autres principes que la distinction féminin/masculin, restent un champ de recherche pertinent dans le contexte actuel où les enjeux de représentation équitable dans la langue sont au cœur des débats.

Pour remédier à de tels biais, il est donc crucial de développer des méthodes de collecte et d'échantillonnage des données qui reflètent mieux la variation sociolinguistique des langues traitées et la diversité de leurs locutrices et locuteurs. Cette démarche peut passer par l'utilisation de techniques d'apprentissage actif (Ambati et al., 2010) ou d'augmentation de données (Fadaee et al., 2017) pour enrichir les corpus

d'entraînement avec des exemples plus variés et équilibrés. Une autre piste prometteuse est celle de l'apprentissage antagoniste (Pryzant et al., 2020) qui repose sur l'entraînement conjoint d'un modèle de TA et d'un modèle discriminant chargé de détecter les biais dans les traductions produites afin de les atténuer.

3.2 Incidences sur les écosystèmes linguistiques africains

Un dernier enjeu majeur du développement de la TA pour les langues africaines concerne ses incidences possibles sur les écosystèmes linguistiques et les rapports de pouvoir entre les langues. Malgré les efforts déployés pour développer des systèmes multilingues, la grande majorité des ressources et des outils de TA reste concentrée sur un petit nombre de langues dominantes, en particulier l'anglais, qui est par ailleurs utilisé dans bien des cas comme langue pivot (Paul et al., 2013). Cette situation fait craindre un renforcement de l'hégémonie de l'anglais comme langue globale et hypercentrale (de Swaan, 2001), au détriment de la diversité linguistique mondiale.

Dans le contexte africain, les systèmes de TA utilisent fréquemment l'anglais comme langue pivot – une langue intermédiaire qui sert de pont entre deux autres langues sans système de traduction directe. Cette approche trouve sa justification dans la contrainte imposée par les ressources disponibles : bien des langues africaines sont confrontées à un manque de données d'entraînement et de corpus parallèles. Cette situation contraint les développeurs à s'appuyer sur l'anglais comme intermédiaire en raison de son statut de langue officielle dans de nombreux pays africains et de la disponibilité généralement plus importante de ressources numériques dans cette langue. Les difficultés que pose le développement de modèles pour les langues africaines peu dotées illustrent bien ce problème (Tonja et al., 2024). Cette médiation est particulièrement fréquente pour des langues géographiquement ou typologiquement éloignées. L'éloignement géographique concerne des langues parlées dans des régions distantes du continent. Cette distance géographique limite les contacts historiques et les échanges linguistiques naturels. L'éloignement typologique, quant à lui, désigne les différences structurelles fondamentales entre les systèmes linguistiques : différences dans l'ordre des mots (par exemple, Sujet-Verbe-Objet par opposition à Verbe-Sujet-Objet), dans les systèmes tonaux (langues tonales par opposition à langues non tonales), dans les structures morphologiques (langues agglutinantes par opposition à langues isolantes) ou dans les systèmes de classes nominales.

L'exemple de la traduction du yoruba vers le swahili illustre précisément un cas de double éloignement géographique et typologique. Géographiquement, ces langues sont séparées par plus de 3 000 kilomètres entre l'Afrique de l'Ouest et l'Afrique de l'Est. Typologiquement, bien qu'elles appartiennent toutes deux à la famille Niger-Congo, elles présentent des différences structurelles considérables : le yoruba (branche Defoid) possède un système tonal complexe avec trois tons de base et un ordre Sujet-Verbe-Objet, tandis que le swahili (branche bantoue) est moins tonal et présente un système de classes nominales bantoues avec préfixes complexes. Un autre exemple serait la traduction entre l'amharique, une langue sémitique d'Éthiopie utilisant l'écriture guèze et un ordre Sujet-Objet-Verbe, et le wolof, langue atlantique du Sénégal utilisant l'alphabet latin. Ces deux dernières langues cumulent éloignements géographique, génétique (familles linguistiques complètement différentes) et typologique.

L'approche par langue pivot offre l'avantage d'exploiter les vastes corpus et modèles existants en anglais. Elle rend ainsi possible la traduction même lorsqu'il n'existe pas de données parallèles suffisantes entre les langues africaines concernées. Par exemple, pour traduire du yoruba vers le swahili, l'on effectue d'abord une

traduction du yoruba vers l'anglais, puis de l'anglais vers le swahili (Martinus et Abbott, 2019). Si cette approche permet d'exploiter les ressources plus abondantes en anglais, elle présente deux risques distincts. D'une part, un mauvais entraînement des systèmes peut conduire à l'imposition des structures linguistiques anglaises aux langues africaines traduites, générant des traductions inexactes qui ne respectent pas les spécificités grammaticales des langues cibles. D'autre part, même avec un entraînement approprié, l'usage systématique de l'anglais comme langue pivot peut conduire à une forme d'« impérialisme linguistique » (Phillipson, 2013) où les modèles conceptuels et sémantiques de la langue dominante influencent progressivement la perception et l'expression des réalités locales dans les langues africaines.

Pour contrer ces effets pervers, il semble donc important de favoriser le développement de systèmes de TA décentralisés qui permettent une communication directe entre les langues africaines, sans passer systématiquement par l'anglais ou d'autres langues européennes (Nekoto et al., 2020). Pour ce faire, il est nécessaire de mettre l'accent sur la constitution de corpus de données parallèles ou comparables entre ces langues ainsi que sur l'adaptation des architectures et des méthodes de TA à leurs propriétés typologiques. Pour que cette adaptation technique se fasse efficacement, une forte implication des communautés linguistiques locales s'avère nécessaire, tant dans la collecte des données que dans la validation des traductions produites. Il est également important que les chercheuses et chercheurs, ainsi que les développeuses et développeurs africains, soient au cœur de ces initiatives pour garantir que les systèmes répondent aux besoins réels des populations utilisatrices.

À plus long terme, l'objectif devrait être de construire des écosystèmes linguistiques numériques résilients et autonomes soutenus par des technologies inclusives, où les langues minorisées peuvent coexister et interagir équitablement avec les langues dominantes, plutôt que de graviter autour d'elles comme de simples satellites (Calvet, 1999), présentes en ligne sans réelle vitalité numérique (Kornai, 2013). Concrètement, ces écosystèmes prendraient la forme de plateformes intégrées offrant une gamme complète de services : des outils de création de contenu multimédia dans les langues locales, des systèmes de traduction directe entre langues africaines (évitant le passage par l'anglais ou une autre langue pivot), des bibliothèques numériques alimentées par les communautés elles-mêmes et des espaces d'échange permettant aux locutrices et locuteurs de différentes langues de collaborer sur des projets communs. On pourrait imaginer, par exemple, un réseau collaboratif où des contenus produits en ewondo (langue bantoue du Cameroun), qu'il s'agisse de pratiques agricoles d'adaptation au climat, de guides de santé communautaire ou de matériel pédagogique, seraient traduits directement en luganda (langue bantoue d'Ouganda), enrichi au besoin d'annotations culturelles par l'expertise locale, puis adaptés en divers formats par des créatrices et créateurs de contenu de différentes communautés linguistiques, sans passer par une langue européenne. Il faut reconnaitre tout de même que le chemin qui mène à cet idéal reste long, en raison de la forte fragmentation linguistique du continent : si une telle mise en réseau directe existe déjà pour certaines paires de langues, elle demeure difficilement généralisable à l'ensemble des langues africaines.

Ces écosystèmes devraient favoriser non seulement la traduction entre les langues, mais aussi la création et le partage de contenus originaux dans chaque langue, à travers des plateformes intégrant des fonctionnalités adaptées aux spécificités culturelles : reconnaissance vocale optimisée pour les langues tonales, interfaces prenant en compte les systèmes d'écriture locaux, outils de collaboration respectant les protocoles traditionnels de transmission du savoir et mécanismes de validation communautaire garantissant la qualité et l'authenticité des contenus. À l'ère de l'IA, ces écosystèmes pourraient également intégrer des assistants

conversationnels entraînés sur les corpus de langues africaines, capables de répondre aux questions dans ces langues tout en respectant les codes culturels et les nuances linguistiques locales. Bien que de tels corpus restent encore largement à constituer, cette approche permettrait de faire de l'IA un vecteur d'autonomisation plutôt que de domination, en servant les communautés linguistiques plutôt qu'en les assimilant aux modèles dominants. Toutefois, une vigilance particulière serait nécessaire pour éviter que l'IA ne génère des contenus artificiels qui pourraient dénaturer ou fabriquer des éléments culturels, d'où l'importance des mécanismes de validation communautaire mentionnés précédemment.

4 Pistes pour une TA au service des langues minorisées

4.1 Approches critiques et décoloniales

Il importe de développer des approches critiques et réflexives du développement technologique qui remettent en question les présupposés et les biais des modèles dominants de la TA et du TALN en général (Bird, 2020). En effet, plusieurs travaux (Birhane, 2021 ; Mohamed et al., 2020) démontrent que ces modèles sont souvent imprégnés d'une vision universaliste des langues et des connaissances, qui tend à effacer les spécificités et les épistémologies locales au profit de catégories et de méthodes standardisées.

Pour contrer cette tendance et favoriser une TA véritablement décoloniale, il faut donc réévaluer les fondements mêmes des approches actuelles, en s'interrogeant, par exemple, sur la pertinence des notions de « langue », de « corpus » ou encore de « qualité » de traduction en contexte minoritaire, à l'image de la déconstruction des postulats eurocentriques de la traduction que propose Chagnon (2019). En matière d'évaluation de la qualité, les approches dominantes en TA privilégient souvent des métriques techniques (BLEU, METEOR) qui ne reflètent pas nécessairement l'adéquation culturelle et communicative des traductions. Ces métriques automatiques, bien qu'utiles pour l'évaluation technique, présentent des limitations importantes dans la détection d'erreurs critiques qui peuvent entraîner de graves incompréhensions du message, particulièrement dans les contenus à dimension affective (Saadany et Orăsan, 2021). Dans ce contexte, la théorie du skopos (Reiss et Vermeer, 2013), qui définit la qualité d'une traduction par sa capacité à remplir sa fonction dans le contexte d'usage visé, offre une perspective pertinente, car elle permet de dépasser les critères purement linguistiques pour intégrer les dimensions pragmatique et culturelle. Cette approche fonctionnelle suggère que la conceptualisation même de la « langue » en TA devrait évoluer : plutôt que de traiter les langues comme des entités homogènes et stables, il conviendrait d'intégrer les continuums linguistiques, les variations dialectales et la diversité des pratiques communicatives au sein des communautés linguistiques. Les corpus traditionnels, basés sur l'écrit standardisé, devraient s'étendre aux formes orales et aux variations locales. Les critères d'évaluation de qualité pourraient privilégier l'acceptabilité culturelle et l'efficacité communicative dans les contextes spécifiques d'utilisation.

Concrètement, cette approche nécessite le développement d'architectures de TA plus flexibles et adaptables, conçues pour être progressivement modifiables et enrichies par les communautés locutrices, avec un niveau d'expertise technique minimal et un accompagnement adapté. Il convient parallèlement de promouvoir des approches multimodales de la TA intégrant non seulement les données textuelles, mais aussi des signaux vocaux – dont l'efficacité est déjà démontrée pour les langues dont l'usage est majoritairement oral – et explorant potentiellement des canaux gestuels ou visuels. Dans cette perspective, les recherches sur la

traduction dictée et la postédition dictée (Zapata et Quirion, 2016 ; Zapata et al., 2017) offrent des pistes opérationnelles prometteuses, dans la mesure où elles permettent à la population utilisatrice d'interagir vocalement avec les systèmes de TA. Cette modalité contourne partiellement les barrières liées à l'écrit et s'appuie sur la primauté des traditions orales. L'interaction multimodale avancée combinant geste, parole et regard (Herbig, 2022 ; Herbig et al., 2020 ; Teixeira et al., 2019), bien que techniquement prometteuse, relève encore largement de la recherche expérimentale et nécessite des validations en contexte réel. Si elles sont déployées de manière fiable et accessible, ces technologies pourraient permettre aux communautés locutrices de langues minorisées de recourir à la traduction au moyen de la parole ou du geste et ainsi de dépendre moins d'une alphabétisation préalable dans des systèmes d'écriture parfois récents. Sur cette base technique en évolution, il devient possible de proposer des méthodes et des critères d'évaluation plus contextualisés et pluralistes, lesquels devraient intégrer :

7. les spécificités linguistiques et socioculturelles des langues cibles,
8. les rétroactions directes des communautés concernées (Adebara et al., 2022),
9. une prise en compte explicite des limites infrastructurelles et techniques des environnements de déploiement.

4.2 Infrastructures technologiques ouvertes

Pour soutenir ces approches collaboratives et critiques de la TA, il apparaît enfin essentiel de développer des infrastructures technologiques ouvertes et en libre accès, qui favorisent l'autonomisation et la résilience des communautés linguistiques minorisées et qui tiennent compte du contexte actuel de ces langues (Emezue et Dossou, 2021). En effet, face aux risques de dépendance et de perte de contrôle liés à l'utilisation d'outils et de services propriétaires, souvent conçus et dirigés depuis les pays industrialisés, il semble crucial de promouvoir des solutions de substitution ouvertes et adaptées aux contextes locaux. Cette orientation suppose la création de plateformes et de logiciels en libre accès pour la collecte, le traitement et la diffusion des données linguistiques. Ces infrastructures doivent être conçues pour un déploiement et une maintenance simplifiés, exiger peu de ressources techniques et financières tout en favorisant l'appropriation locale des technologies.

Outre ces outils, il importe également de mettre en place des infrastructures de gouvernance et de régulation des données linguistiques qui garantissent la souveraineté et les droits des communautés contributrices (Adebara et al., 2022). Cette démarche peut s'appuyer sur la création d'institutions et de cadres juridiques spécifiques, à l'échelle nationale ou régionale, lesquels serviraient à définir les protocoles éthiques et les mécanismes de consentement, de traçabilité et de rétribution associés à l'utilisation de ces données (Carvallo, 2019). À terme, l'objectif devrait être de construire un écosystème technologique et linguistique décentralisé et résilient, où les communautés minorisées ont la capacité et la légitimité de créer et de gouverner leurs propres ressources et outils numériques, en fonction de leurs valeurs et de leurs aspirations.

Conclusion

Au regard de ces analyses, le chemin vers une TA qui sert véritablement la diversité linguistique africaine sans marginaliser les langues minorisées reste long et parsemé d'embûches. Néanmoins, loin de représenter une simple perte de temps, cet enjeu engage la pérennité des langues concernées. La TA recèle des atouts majeurs

pour appuyer les efforts de documentation, de création de ressources et de revitalisation des langues africaines minorisées à l'ère numérique. Par sa capacité à produire rapidement des ressources linguistiques parallèles et monolingues, à doter le traitement automatique de ces langues d'outils susceptibles de favoriser et d'accroître leur usage et leur visibilité dans les espaces numériques, elle peut effectivement en renforcer la vitalité et la résilience face à la mondialisation.

Il reste toutefois indispensable de reconnaître les tensions inhérentes à la coexistence entre idéaux émancipateurs et contraintes économiques et institutionnelles pour élaborer des stratégies à la fois réalistes et transformatrices. Outre les défis techniques et méthodologiques, des obstacles politiques, économiques et idéologiques devront être surmontés, notamment ceux liés aux rapports de pouvoir structurant les écosystèmes linguistiques et technologiques actuels. Parmi ces obstacles, les contraintes économiques structurelles représentent un facteur limitant souvent négligé. En raison de la logique de marché qui gouverne le développement technologique, l'investissement dans des systèmes de TA pour les langues minorisées semble peu rentable, spécialement lorsque leurs locutrices et locuteurs maîtrisent déjà les langues dominantes. Cette réalité économique explique principalement pourquoi, malgré des décennies d'existence de la TA, les outils dédiés aux langues africaines minorisées restent rares ou embryonnaires. À cette dimension s'ajoute un enjeu écologique majeur : l'empreinte énergétique massive des modèles de TA suscite des interrogations quant à leur durabilité éthique, particulièrement dans un contexte où les ressources énergétiques africaines pourraient être prioritairement mobilisées pour d'autres besoins de développement.

Pour exploiter pleinement le potentiel de ces technologies, il faut dépasser les approches techniques et universalistes jusqu'ici dominantes et bâtir des modèles de TA véritablement collaboratifs, centrés sur les communautés linguistiques minorisées. Cette transformation exige de remettre en cause les logiques économiques qui tendent à marginaliser ces langues et d'élaborer des solutions technologiques écologiquement viables. Une vigilance particulière s'impose pour que les progrès de la TA ne se réalisent pas aux dépens des langues minorisées et des communautés les plus vulnérables, mais contribuent à une plus grande équité linguistique à l'échelle du continent africain.

Cette démarche implique des partenariats interdisciplinaires et participatifs qui associent expertise technologique et savoirs locaux, ainsi qu'une réflexion critique sur les biais des modèles dominants. Elle commande aussi de promouvoir des infrastructures ouvertes et décentralisées qui garantissent l'autonomie des communautés dans la création et le contrôle de leurs ressources numériques, en tenant compte des impératifs de durabilité environnementale. Des mesures concrètes s'imposent à cet égard :

- L'association systématique des communautés linguistiques au développement et à l'évaluation des systèmes de TA, au moyen de comités consultatifs et de processus de validation participatifs ;
- La conception prioritaire d'outils pour les langues les plus menacées, indépendamment de leur rentabilité immédiate ;
- La formation de personnel de recherche-développement issu des communautés concernées pour une maîtrise locale des technologies ;
- L'établissement de cadres éthico-juridiques protégeant les droits linguistiques et la propriété intellectuelle communautaire ;
- L'élaboration de modèles économiques novateurs (financement public, coopération internationale, économie sociale) affranchis des logiques de rentabilité immédiate ;

- L'exploration d'approches technologiques moins énergivores qui privilégient l'efficacité énergétique et les infrastructures décentralisées.

Ces mesures exigent sans doute des arbitrages complexes entre développement technologique et préservation culturelle, entre standardisation et diversification, entre modèles d'innovation globaux et quête d'autonomie locale. Les arbitrages nécessitent un renforcement décisif des collaborations entre toutes les parties prenantes de la TA pour les langues africaines – chercheuses et chercheurs, ingénieures et ingénieurs, linguistes, mouvements militants, instances et communautés. Il importe de fédérer une communauté d'expertise autour de ces défis afin de mutualiser les ressources, de capitaliser les expériences et de faire émerger des solutions innovantes adaptées aux réalités locales. L'engagement des institutions régionales dotées de pouvoirs décisionnels, comme l'Union africaine et les organisations sous-régionales, sera déterminant.

Un autre levier essentiel pour assurer la pérennité de ces dynamiques réside dans le renforcement des capacités de formation et de recherche en TA/TALN au sein des institutions africaines, en vue de préparer une nouvelle génération de chercheuses et chercheurs ainsi que de praticiennes et praticiens porteurs d'approches collaboratives critiques. Cette formation nécessitera des programmes d'enseignement interdisciplinaires (informatique, linguistique, anthropologie, sciences sociales) et des partenariats de recherche entre pays industrialisés et pays en développement, le but étant de favoriser la construction collaborative et le partage des savoirs et des innovations.

Enfin, la pérennité de ces efforts sera tributaire de leur intégration dans une réflexion globale sur les politiques linguistiques et éducatives africaines ainsi que sur les modèles de développement socioéconomique sous-jacents. La sauvegarde des langues minorisées ne saurait se réduire à une dimension technique ; elle doit s'enraciner dans une vision politique et culturelle à long terme, qui fait de la diversité linguistique un levier de développement humain durable. Les initiatives de TA devraient ainsi s'articuler avec d'autres dynamiques comme les mouvements en faveur de l'éducation multilingue, la reconnaissance des langues minorisées ou les démarches de décolonisation des savoirs – au-delà de la seule reconnaissance officielle – en misant sur des actions concrètes de valorisation.

Au terme de cette réflexion, il semble se dégager un constat : plutôt que d'adapter la TA aux langues africaines, l'idéal serait de faire émerger une TA africaine – qui incarne les réalités, les besoins et les aspirations des communautés linguistiques du continent, tout en contribuant à leur émancipation. Il s'agit de bâtir une TA qui dépasse l'application passive de modèles exogènes pour les remettre en question, les adapter et les enrichir à partir des épistémologies locales. Cette démarche décoloniale et émancipatrice vise à préserver les langues africaines en tant que patrimoine culturel, souvent minorisé et vulnérabilisé, et à les revitaliser en tant que véhicules dynamiques de savoirs, d'identités et d'aspirations pour leurs communautés locutrices. En ce sens, la TA apparaît comme un levier possible de réappropriation et de revalorisation des langues minorisées, qui leur permettrait de s'affirmer comme des ressources stratégiques et légitimes dans les sociétés du savoir du XXI^{e} siècle (UNESCO, 2005), à condition que sa viabilité économique soit assurée et que son empreinte environnementale soit maîtrisée.

Références bibliographiques

Abbott, J. Z. et Martinus, L. (2018). Towards neural machine translation for African languages. *arXiv*. https://doi.org/10.48550/arXiv.1811.05467

Adebara, I. et Abdul-Mageed, M. (2022). Towards Afrocentric NLP for African languages: Where we are and where we can go. Dans S. Muresan, P. Nakov et A. Villavicencio (dir.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (*volume 1, p. 3814-3841). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.265

Adebara, I., Elmadany, A., Abdul-Mageed, M. et Alcoba Inciarte, A. (2022). AfroLID: A neural language identification tool for African languages. Dans Y. Goldberg, Z. Kozareva et Y. Zhang (dir.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (p. 1958-1981). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.128

Ajao, J. F., Babatunde, R. S., Asapetu, S. O. et Yusuff, S. R. (2020). Implementation of Yorùbá Unicode generation for an indigenous keyboard. *Technoscience Journal for Community Development in Africa*, *1*(1), 71-80. https://journals.kwasu.edu.ng/index.php/technoscience/article/view/13

Ambati, V., Vogel, S. et Carbonell, J. (2010, mai). *Active Learning and Crowd-Sourcing for Machine Translation*. Dans N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner et D. Tapias (dir.), *Actes de LREC 2010* (p. 2169-2174). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/244_Paper.pdf

Anastasopoulos, A. (2019). *Computational tools for endangered language documentation* [Thèse de doctorat, University of Notre Dame]. CurateND. https://curate.nd.edu/show/9019s18114q

Apple Machine Learning Research. (2020). Language identification from very short strings. *Apple Machine Learning Research*. https://machinelearning.apple.com/research/language-identification-from-very-short-strings

Artetxe, M., Labaka, G. et Agirre, E. (2019). An effective approach to unsupervised machine translation. Dans *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (p. 194-203). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1019

Ballier, N. et Zimina-Poirot, M. (2023). Littératie de la traduction automatique (TA) neuronale et traduction spécialisée : s'approprier les outils de la TA au travers de projets de recherche interdisciplinaires. Dans N. Froeliger, C. Larsonneur et G. Sofo (dir.), *Traduction humaine et traitement automatique des langues : vers un nouveau consensus ?* (p. 165-180). *Edizioni Ca'Foscari*. https://doi.org/10.30687/978-88-6969-762-3/011

Barysevich, A. et Costaris, C. (2021). Traducteurs automatiques neuronaux comme outil didactique/pédagogique : DeepL dans l'apprentissage du français langue seconde. *Nouvelle Revue Synergies Canada*, *14*, 1-16.

Bender, E. M., Gebru, T., McMillan-Major, A. et Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Dans *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (*p. 610-623). https://doi.org/10.1145/3442188.3445922

Bernhard, D. et Soria, C. (2018). Traitement automatique des langues peu dotées. *Traitement Automatique des Langues*, *59*(3), 7-14. https://aclanthology.org/2018.tal-3.1/

Bird, S. (2020). Decolonising speech and language technology. Dans D. Scott, N. Bel et C. Zong (dir.), *Proceedings of COLING 2020: The 28th International Conference on Computational Linguistics* (p. 3504-3519). https://doi.org/10.18653/v1/2020.coling-main.313

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Blasi, D., Anastasopoulos, A. et Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. Dans S. Muresan, P. Nakov et A. Villavicencio (dir.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers,* p. 5486-5505). https://doi.org/10.18653/v1/2022.acl-long.376

Bodomo, A., Anderson, J. et Dzahene-Quarshie, J. (2006). Digitizing African languages: challenges and opportunities in Ghana. Dans F. K. Ameka, A. Dench et N. Evans (dir.), *Catching Language: The Standing Challenge of Grammar Writing* (p. 417-440). Mouton de Gruyter.

Bowker, L. (2024). Lorsqu'on enseigne les technologies de la traduction, le choix des termes compte beaucoup. *ACET-LINET*. https://www.acet-linet.ca/post/choix-des-termes-en-enseignement

Carvallo, S. (2019). L'éthique de la recherche entre réglementation et réflexivité. *Revue d'anthropologie des connaissances*, *13*(2). https://doi.org/10.3917/rac.043.0299

Chagnon, K. (2019). Colonialisme, universalisme occidental et traduction. *Traduction, Terminologie, Rédaction*, *32*(1), 259-278. https://doi.org/10.7202/1068021ar

Centre de recherches pour le développement international (CRDI). (2025). *Le carrefour d'IA de Masakhane pour les langues*. https://idrc-crdi.ca/fr/ce-que-nous-faisons/projets-que-nous-soutenons/projet/le-carrefour-dia-de-masakhane-pour-les

Creissels, D. (2000). Typology. Dans B. Heine et D. Nurse (dir.), *African Languages: An Introduction* (*p.* 231-258). Cambridge University Press.

Cru, J. (2014). Language revitalisation from the ground up: Promoting Yucatec Maya on Facebook. *Journal of Multilingual and Multicultural Development*, *36*(3), 284-296. https://doi.org/10.1080/01434632.2014.921184

de Gibert, O., Pugh, R., Marashian, A., Vázquez, R., Ebrahimi, A., Denisov, P., Rice, E., Gow-Smith, E., Prieto, J. C., Robles, M., Manrique, R., Moreno Veliz, O., Lino Campos, Á., Coto-Solano, R., Alvarez, A., Agüero-Torales, M., Ortega, J. E., Chiruzzo, L., Oncevay, A., ... Mager, M. (2025). Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for Indigenous languages of the Americas. Dans *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)* (p. 134–152). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.americasnlp-1.16

de Swaan, A. (2001). *Words of the World: The Global Language System*. Polity.

Déjean, H. et Gaussier, É. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, *Alignement lexical dans les corpus multilingues*, 1-22. https://aclanthology.org/F13-1018.pdf

Djité, P. G. (2008). *The Sociolinguistics of Development in Africa*. Multilingual Matters. http://archive.org/details/sociolinguistics0000djit

Eberhard, D. M., Simons, G. F., et Fennig, C. D. (dir.). (2024). *Ethnologue: Languages of the World* (27e édition). SIL International. https://www.ethnologue.com/

Emezue, C. C. et Dossou, B. F. P. (2020). *Lanfrica: A participatory approach to documenting machine translation research on African languages* (arXiv:2008.07302). *arXiv*. https://doi.org/10.48550/arXiv.2008.07302

Emezue, C. C. et Dossou, B. F. P. (2021). MMTAfrica: Multilingual machine translation for African languages. *arXiv*. https://doi.org/10.48550/arXiv.2204.04306

Enguehard, C. et Kané, S. (2004). Langues africaines et communication électronique : développement de correcteurs orthographiques. Dans *Penser la Francophonie : concepts, actions et outils linguistiques : actes des premières Journées scientifiques communes des réseaux de chercheurs concernant la langue*. Éditions des archives contemporaines.

Fadaee, M., Bisazza, A. et Monz, C. (2017). Data augmentation for low-resource neural machine translation. Dans *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers, p. 567-573). https://doi.org/10.18653/v1/P17-2090

Fishman, J. A. (1991). *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Multilingual Matters.

Footitt, H. et Kelly, M. (dir.). (2012). *Languages at War: Policies and Practices of Language Contacts in Conflict*. Palgrave Macmillan.

Godard, P., Adda, G., Adda-Decker, M., Benjumea, J., Besacier, L., Cooper-Leavitt, J., Kouarata, G.-N., Lamel, L., Maynard, H., Mueller, M., Rialland, A., Stueker, S., Yvon, F. et Zanon-Boito, M. (2018). A very low resource language speech corpus for computational language documentation Experiments. *arXiv*. https://doi.org/10.48550/arXiv.1710.03501

Guo, Y. et Beckett, G. H. (2007). The hegemony of English as a global language: Reclaiming local knowledge and culture in China. *Convergence*, *40*(1-2), 117-131. https://globalization.anthro-seminars.net/wp-content/uploads/2016/11/guo-2007.pdf

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., ... Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv*. https://doi.org/10.48550/ARXIV.1803.05567

Herbig, N. (2022). *Multi-modal post-editing of machine translation* [Thèse de doctorat, Saarland University]. https://umtl.cs.uni-saarland.de/paper_preprints/PhD_Nico_Herbig.pdf

Herbig, N., Düwel, T., Pal, S., Meladaki, K., Monshizadeh, M., Krüger, A. et van Genabith, J. (2020). MMPE: A multi-modal interface for post-editing machine translation. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 1691-1702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.155

Hussen, K. Y., Sewunetie, W. T., Ayele, A. A., Imam, S. H., Alemu, E. N., Muhammad, S. H. et Yimam, S. M. (2025). The state of large language models for African languages: Progress and challenges. *arXiv*. https://doi.org/10.48550/arXiv.2506.02280

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., et Ljubešić, N. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research, 65*, 675-782. https://doi.org/10.1613/jair.1.11675

Ji, M. (2023). Cultural and linguistic bias of neural machine translation technology. Dans M. Ji, P. Bouillon et M. Selignan (dir.), *Translation Technology in Accessible Health Communication* (p. 100-128). Cambridge University Press. https://doi.org/10.1017/9781108938976.005

Joshi, P., Santy, S., Budhiraja, A., Bali, K. et Choudhury, M. (2021). The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv*. https://doi.org/10.48550/arXiv.2004.09095

Karakanta, A., Dehdari, J. et Van Genabith, J. (2018). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, *32*(1-2), 167-189. https://doi.org/10.1007/s10590-017-9203-5

Knight, J. (2022). *Knowledge diplomacy in international relations and higher education*. Springer. https://doi.org/10.1007/978-3-031-14977-1

Koehn, P. (2020). *Neural Machine Translation*. Cambridge University Press. https://doi.org/10.1017/9781108608480

Kornai, A. (2013). Digital language death. *PLoS ONE*, *8*(10), Article e77056. https://doi.org/10.1371/journal.pone.0077056

Kourliandsky, J.-J. (2023, 18 avril). Paradoxe de la francophonie : un espace organisé en quête de sens. *Fondation Jean Jaurès/GeoStrategia*. https://www.geostrategia.fr/paradoxe-francophonie-espace-organise-quete-de-sens/

Lample, G., Conneau, A., Denoyer, L. et Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. Dans *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*. https://openreview.net/forum?id=rkYTTf-AZ

Lanfrica. (2025, 23 octobre). Lanfrica V2 is here! A new era for African AI discovery. *Lanfrica Blog*. https://lanfrica.com/blog/lanfrica-v2-is-here-a-new-era-for-african-ai-discovery/

Mager, M., Gutierrez-Vasques, X., Sierra, G. et Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. Dans E. M. Bender, L. Derczynski et P. Isabelle (dir.), *Proceedings of COLING 2018: The 27th International Conference on Computational Linguistics* (p. 55-69). Association for Computational Linguistics. https://aclanthology.org/C18-1006/

Martinus, L. et Abbott, J. Z. (2019). A Focus on neural machine translation for African languages. *arXiv*. http://arxiv.org/abs/1906.05685

Mohamed, S., Png, M.-T. et Isaac, W. (2020). Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*(4), 659-684. https://doi.org/10.1007/s13347-020-00405-8

Naveen, P. et Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, *27*(10), 110878. https://doi.org/10.1016/j.isci.2024.110878

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., … Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. Dans T. Cohn, Y. He et Y. Liu (dir.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (p. 2144-2160). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195

NLP Ghana. (2023). Khaya African language translation AI adds Gurene, Kikuyu, Kimeru & Luo. *Medium*. https://nlpghana.medium.com/khaya-african-language-translation-ai-adds-gurene-kikuyu-kimeru-luo-4171de1273ce

Nye, J. S. (2004). *Soft Power: The Means to Success in World Politics*. Public Affairs.

Ogueji, K. et Ahia, O. (2019). PidginUNMT: Unsupervised neural machine translation from West African Pidgin to English. *arXiv*. https://doi.org/10.48550/arXiv.1912.03444

Okolo, C. (2023). *AI in the Global South: Opportunities and challenges towards more inclusive governance*. Brookings. https://www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance/

OpenAI. (2022). ChatGPT: Optimizing language models for dialogue. OpenAI. https://openai.com/blog/chatgpt

OpenAI. (2023). GPT-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., Ali, J. T., Abbott, J., Marivate, V., Kabongo, S., Meressa, M., Murhabazi, E., Ahia, O., van Biljon, E., Ramkilowan, A., Akinfaderin, A., Öktem, A., Akin, W., Kioko, G., … Bashir, A. (2020). Masakhane: Machine translation for Africa. *arXiv*. https://doi.org/10.48550/ARXIV.2003.11529

Osborn, D. (2010). *African languages in a digital age: Challenges and opportunities for indigenous language computing*. HSRC Press / International Development Research Centre.

Paul, M., Finch, A. et Sumita, E. (2013). How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing*, *12*(4), 1-17. https://doi.org/10.1145/2505126

Phillipson, R. (2013). *Linguistic Imperialism Continued*. Routledge. https://doi.org/10.4324/9780203857175

Prates, M. O. R., Avelar, P. H. C. et Lamb, L. (2020). Assessing gender bias in machine translation—A case study with Google Translate. *Neural Computing and Applications*, *32*(10), 6363-6381. https://doi.org/10.1007/s00521-019-04144-6

Pryzant, R., Chung, Y., Jurafsky, D. et Britz, D. (2020). JESC: Japanese-English subtitle corpus. *arXiv*. https://doi.org/10.48550/arXiv.1710.10639

Ranathunga, S., Lee, E.– S. A., Prifti Skenduli, M., Shekhar, R., Alam, M. et Kaur, R. (2023). Neural machine translation for low-resource languages: A Survey. *ACM Computing Survey*, *55* (11), 229. https://doi.org/10.1145/3567592

Reiss, K. et Vermeer, H. J. (2013). *Towards a general theory of translational action: Skopos theory explained* (traduit par C. Nord). St. Jerome Publishing

Saadany, H. et Orăsan, C. (2021). BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. Dans *Proceedings of the Translation and Interpreting Technology Online Conference* (p. 48-56). Association for Computational Linguistics. https://aclanthology.org/2021.triton-1.6/

Skutnabb-Kangas, T. (1988). Multilingualism and the education of minority children. Dans T. Skutnabb-Kangas et J. Cummins (dir.) *Minority Education: From Shame to Struggle* (p. 13-34). Multilingual Matters.

Teixeira, C. S. C., Moorkens, J., Turner, D., Vreeke, J. et Way, A. (2019). Creating a multimodal translation tool and testing machine translation integration using touch and voice. *Informatics*, *6*(1), 13. https://doi.org/10.3390/informatics6010013

The Unicode Consortium. (2023). *The Unicode Standard, Version 15.1.0*. The Unicode Consortium. https://www.unicode.org/versions/Unicode15.1.0/

Tonja, A. L., Dossou, B. F. P., Ojo, J., Rajab, J., Thior, F., Wairagala, E. P., Anuoluwapo, A., Moiloa, P., Abbott, J., Marivate, V., et Rosman, B. (2024). InkubaLM: A small language model for low-resource African languages. *arXiv*. https://doi.org/10.48550/arXiv.2408.17024

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., et Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. *arXiv*. https://doi.org/10.48550/arXiv.2003.01200

UNESCO. (2003). *Vitalité et disparition des langues* (publication CLT/CEI/DCE/ELP/PI/2003/1). https://unesdoc.unesco.org/ark:/48223/pf0000183699_fre

UNESCO. (2005). *Vers les sociétés du savoir : rapport mondial de l'UNESCO*. https://unesdoc.unesco.org/ark:/48223/pf0000141907

UNESCO. (2010). *Atlas des langues en danger dans le monde* (C. Moseley, dir.). Bibliothèque numérique UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000187026_fre

Vansina, J. (2006). *Oral Tradition: A Study in Historical Methodology*. Aldine Transaction. http://archive.org/details/oraltraditionstu0000vans

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. et Polosukhin, I. (2017, 12 juin). Attention is all you need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

Zapata, J., Castilho, S. et Moorkens, J. (2017). Translation dictation vs. post-editing with cloud-based voice recognition: A pilot experiment. Dans M. Yamada et M. Seligman (dir.), *Proceedings of Machine Translation Summit XVI: Commercial MT users and translators track* (p. 116-129). Asia-Pacific Association for Machine Translation. https://aclanthology.org/2017.mtsummit-commercial.13

Zapata, J. et Quirion, J. (2016). La traduction dictée interactive et sa nécessaire intégration à la formation des traducteurs. *Babel*, *62*(4), 531-551. https://doi.org/10.1075/babel.62.4.01zap

Zhang, J. et Zong, C. (2020). Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, *63*(10), 2028-2050. https://doi.org/10.1007/s11431-020-1632-x