Réhabiliter l'écriture Ajami : un levier technologique pour l'alphabétisation en Afrique

Samy Ouzerrout Idriss Saadallah
Université d'Orléans, France
samy.ouzerrout@etu.univ-orleans.fr, idriss.saadallah@univ-orleans.fr

RÉSUMÉ

Cet article explore l'écriture Ajami, système basé sur l'alphabet arabe historiquement utilisé pour transcrire les langues africaines, comme levier technologique d'alphabétisation et d'inclusion numérique en Afrique subsaharienne et au Maghreb. Nous présentons la création d'AjamiXTranslit, un corpus multilingue de paires de textes Latin–Ajami et de manuscrits annotés, accompagné d'une plateforme collaborative d'enrichissement par des locuteurs natifs. À partir de ces données, nous développons des modèles de translittération automatique et de reconnaissance optique de caractères (OCR) adaptés à la diversité graphique de l'Ajami. L'article discute les défis techniques (variabilité manuscrite, absence de standardisation), linguistiques (transcriptions phonétiques hétérogènes) et sociaux (acceptabilité, accessibilité) de cette réintégration scripturale. Nos travaux s'inscrivent dans une démarche transdisciplinaire alliant traitement automatique des langues, sciences sociales et politiques éducatives, en vue de préserver un patrimoine scriptural menacé et de renforcer l'accès équitable au savoir dans des contextes digraphiques.

ABSTRACT

Ajami as a Tool for Literacy and Linguistic Inclusion

This article explores the use of Ajami script—a writing system based on the Arabic alphabet historically used to transcribe African languages—as a technological lever for literacy and digital inclusion in Sub-Saharan Africa and the Maghreb. We present the creation of AjamiXTranslit, a multilingual corpus composed of Latin–Ajami text pairs and annotated manuscripts, supported by a collaborative platform enabling enrichment by native speakers. Using these resources, we develop automatic transliteration models and optical character recognition (OCR) systems adapted to the graphic diversity of Ajami. The article discusses the technical (manuscript variability, lack of standardization), linguistic (heterogeneous phonetic transcriptions), and social (acceptability, accessibility) challenges associated with re-integrating this script into contemporary digital practices. Our work adopts a transdisciplinary approach, combining natural language processing, social sciences, and language policy, to both preserve a threatened scriptural heritage and improve equitable access to knowledge in digraphic contexts.

MOTS-CLÉS : Écriture Ajami, Translittération, Reconnaissance Optique de Caractères (OCR), Langues peu dotées, Illettrisme fonctionnel, Traitement multilingue du langage naturel, Inclusion numérique, corpus à faibles ressources, Patrimoine linguistique, Documentation des langues, Langues africaines, Traitement intergraphique, Technologies linguistiques inclusives, Digraphie, Technologies pour l'alphabétisation.

KEYWORDS: Ajami Script, Transliteration, Optical Character Recognition (OCR), Under-resourced

Languages, Functional Illiteracy, Multilingual Natural Language Processing, Digital Inclusion, Low-resource Corpora, Linguistic Heritage, Language Documentation, African Languages, Cross-script Processing, Inclusive Language Technologies, Digraphia, Literacy Technologies.

ARTICLE : **Soumis à** RECITAL-RJCRI.

1 Introduction

L'Afrique subsaharienne et le Maghreb font face à des taux d'analphabétisme parmi les plus élevés au monde, en particulier dans les zones rurales où l'accès à l'éducation formelle reste limité ce qui contribue à faire de cette région l'une des plus touchées au monde. Dans de nombreux pays, la forte diversité linguistique — comme celle observée dans les langues Fula — complique encore davantage l'apprentissage de la lecture et de l'écriture, y compris au sein d'un même groupe ethnique (National African Language Resource Center, 2015). Les systèmes éducatifs privilégient généralement l'enseignement en langue coloniale (français, anglais ou portugais), souvent au détriment des langues locales. Par ailleurs, l'alphabétisation dans ces langues repose presque exclusivement sur l'alphabet latin, alors même que d'autres traditions graphiques continuent d'exister et d'être maîtrisées par certaines communautés (Priest & Hosken, 2010).

Notre travail se concentre sur la redécouverte et la valorisation de l'écriture *Ajami*, un système basé sur l'alphabet arabe et historiquement utilisé pour transcrire diverses langues africaines (Souag, 2009). La littérature scientifique a documenté l'usage de l'Ajami dans le mandingue (Vydrine, 1998; Vydrin, 2014a), le peul, le soninké, le wolof, et même le tamazight (Priest & Hosken, 2010; Ngom, 2016, 2018; Callahan, 2022). Loin d'être marginal, ce système a longtemps constitué un vecteur de communication savante et de transmission du savoir, notamment en Afrique de l'Ouest (Mc Laughlin, 2017). Toutefois, à partir de l'époque coloniale, l'Ajami a été progressivement évincé par l'introduction de l'alphabet latin dans les systèmes éducatifs formels (Burssens & van Bulck, 1935; Education Department of Gold Coast, 1930; International Institute of African Languages and Cultures (IIALC), 1927; Institut international des langues et civilisations africaines (IILCA), 1930; Ward, 1936; Westermann, 1927), une tendance qui s'est poursuivie après les indépendances (UNESCO, 1980; Lacroix, 1968; Mann & Dalby, 1987; UNESCO, 1981).

Ce recul institutionnel contraste avec la persistance de l'alphabet arabe dans les pratiques religieuses. Celui-ci reste enseigné à travers l'apprentissage du Coran, dans des milliers d'écoles coraniques à travers la région (Cissé, 2006; Clark, 2007; Mc Laughlin, 2017; Ngom, 2010; Warren-Rothlin, 2012a). Cette situation produit un paradoxe sociolinguistique : une large part de la population sait lire l'alphabet arabe, mais sans comprendre l'arabe moderne ou classique. Conformément aux définitions de l'UNESCO, ces individus sont classés comme analphabètes fonctionnels (for Statistics, 2025).

Ce paradoxe révèle un enjeu encore peu exploité : l'écriture Ajami, qui permet de transcrire les langues locales à l'aide d'un alphabet déjà familier, pourrait jouer un rôle décisif dans les politiques d'alphabétisation et d'inclusion numérique. Cependant, plusieurs obstacles limitent sa diffusion : absence de normalisation, rareté des ressources numériques, et domination de l'alphabet latin dans la production textuelle contemporaine. Aujourd'hui, la majorité des documents en langues africaines sont écrits en alphabet latin, inaccessibles à ceux qui maîtrisent uniquement l'alphabet arabe.

Dans ce contexte, nous proposons une approche technologique et linguistique intégrée. Nous dévelop-

pons une ressource multilingue combinant translittération automatique entre les alphabets latin et Ajami, reconnaissance optique de caractères (OCR) pour les manuscrits en écriture Ajami, alimentée par une plateforme collaborative ouverte aux locuteurs natifs pour l'enrichissement du corpus. Cette dernière permet à la fois d'annoter des documents historiques et d'intégrer des paires de translittération, renforçant ainsi la qualité des données tout en favorisant leur appropriation communautaire.

En articulant enjeux linguistiques, défis techniques et objectifs sociaux, cette recherche explore les moyens de réhabiliter l'écriture Ajami comme outil d'alphabétisation, de documentation linguistique et d'inclusion numérique. Elle s'inscrit dans une dynamique plus large de justice linguistique pour les langues peu dotées, en tirant parti des avancées du traitement automatique des langues (TAL) et des humanités numériques.

2 Ajami : un levier pour l'inclusion linguistique et la réduction de l'analphabétisation

L'Afrique subsaharienne présente effectivement l'un des taux d'alphabétisation les plus faibles au monde, en raison d'un accès limité à l'éducation, notamment en milieu rural. Selon les données de la Banque mondiale, le taux d'alphabétisation des adultes (15 ans et plus) dans cette région était de 65,0 % en 2018, bien en deçà de la moyenne mondiale (Direction Générale du Trésor, 2024). Dans plusieurs pays, une part significative des adultes ne sait ni lire ni écrire. Par exemple, le Tchad affiche un taux d'alphabétisation de 34,5 %, le Mali 35,5 %, le Niger 35,0 %, la Guinée 32,0 %, et le Sénégal 51,9 % (Atlasocio, 2024). Malgré une légère hausse ces dernières années, ces chiffres reflètent une réalité préoccupante où l'exclusion de l'écrit limite fortement l'accès au savoir et aux opportunités économiques. Par ailleurs, selon l'UNESCO, l'Afrique subsaharienne regroupe environ 26 % de la population adulte analphabète mondiale (Institut de statistique de l'UNESCO, 2016).

2.1 Une alphabétisation religieuse, mais un illettrisme fonctionnel

Le recensement de 2012 au Niger révèle que 20,07 % de la population est alphabétisée en français, 3,95 % en arabe, et 9,29 % en langues nationales utilisant l'alphabet latin (principalement le haoussa)v(Maga & Hamidou, 2015). Cependant, le pourcentage réel d'individus capables de lire et écrire l'alphabet arabe est bien supérieur à 3,95 %. En 1990, on estimait que le Niger comptait environ 40 000 écoles coraniques (Bank, 1999), soit plus du double du nombre d'écoles primaires recensées en 2018 (17 793 selon (Diallo, 2025)).

La situation est comparable au Sénégal, malgré l'un des taux d'alphabétisation les plus élevés de la région. En 2013, 37 % de la population était alphabétisée en français, et 13 % en langues nationales ou en arabe(Agence Nationale de la Statistique et de la Démographie (ANSD), 2013). Par ailleurs, une étude antérieure indique que 74,7 % des personnes utilisaient les caractères latins pour écrire les langues nationales, contre seulement 7,7 % pour les caractères arabes(Diop *et al.*, 1990), illustrant un usage marginalisé de l'écriture arabe. Toutefois, environ 60 % des Sénégalais ont appris à lire l'alphabet arabe dans leur jeunesse(NGUER *et al.*, 2020), un taux atteignant 75 % dans les villages(Bank, 1999), voire 90 % dans certaines régions à forte tradition religieuse comme le Futa Tooro.

Une dynamique similaire a été observée lors de notre enquête ethnographique dans la localité de

Gakoura, au Mali. Cette commune rurale, d'environ 6 000 habitants, présente un taux d'alphabétisation de 20 % en français contre 40 % en arabe. Environ 35 % de la population est mineure, et la commune compte 192 unités d'habitation principales (maisons/foyers).

Ces chiffres montrent qu'en intégrant la maîtrise de l'alphabet arabe, même sans compréhension de la langue, les taux d'alphabétisation seraient largement rehaussés. Cet apprentissage reste limité dans ses usages, car l'arabe est le plus souvent mobilisé à des fins religieuses ou rituelles, sans accès réel aux contenus compréhensibles. En revanche, les langues vernaculaires historiquement transcrites en caractères arabes (Ajami) sont parfaitement comprises.

2.2 Un système d'écriture sous-exploité pour les langues africaines

Paradoxalement, bien que l'alphabet arabe soit massivement enseigné pour des raisons religieuses dans les régions concernées, son usage reste marginal dans la production de contenus en langues africaines. Cela contribue à classer une grande partie de la population dans l'analphabétisme fonctionnel selon la définition de l'UNESCO(Institut de statistique de l'UNESCO, 2025). Ce constat révèle cependant un potentiel sous-exploité : si les langues vernaculaires, historiquement transcrites en Ajami, étaient plus largement diffusées dans ce système, les personnes maîtrisant l'alphabet arabe pourraient accéder à des textes dans leur propre langue, renforçant ainsi l'inclusion linguistique.

Le cas du Niger illustre bien cette problématique. Bien que le haoussa soit la langue la plus parlée, seuls 10 % des Nigériens sont alphabétisés dans cette langue (Ndagi, 2011). Permettre l'accès au haoussa écrit en Ajami à des personnes sachant lire l'alphabet arabe permettrait d'augmenter ce taux, tout en ouvrant l'accès à des ressources numériques comme Wikipédia (disponible en haoussa latinisé) ou les outils de traduction automatique, encore peu développés pour l'Ajami et souvent dépourvus de fonctionnalités vocales (ASR).

L'un des principaux obstacles à cette réintroduction réside dans l'absence d'outils facilitant l'usage de l'Ajami dans les contextes modernes. La quasi-totalité des contenus écrits en langues africaines est aujourd'hui produite en alphabet latin, ce qui limite leur accessibilité pour ceux qui n'ont été alphabétisés qu'en arabe. La mise en place d'outils de translittération automatique entre l'alphabet latin et l'Ajami permettrait de combler cette lacune, en rendant ces textes accessibles à une population bien plus large.

Cependant, le développement de tels outils suppose l'existence de corpus linguistiques conséquents, capables de refléter la diversité des langues, des écritures et des contextes d'usage. Au-delà de la translittération textuelle, ces ressources doivent également intégrer la dimension manuscrite et visuelle propre à l'Ajami, notamment pour entraîner des systèmes OCR adaptés. C'est dans cette perspective que nous proposons une approche centrée sur la constitution de corpus multimodaux, rejoignant textes en écriture latine et Ajami, manuscrits numérisés et annotations croisées.

3 Enrichir les corpus multimodaux : intégrer la dimension texte-image

La constitution de corpus multimodaux représente aujourd'hui un axe central de la documentation linguistique (Adolphs & Carter, 2013; Taylor & Fauzi, 2024). Aller au-delà de la simple transcription

textuelle en intégrant des données audio et visuelles permet d'ancrer les langues dans leur réalité socioculturelle, tout en renforçant les usages pédagogiques et technologiques.

3.1 Au-delà des corpus audio : promouvoir l'écrit comme ressource d'alphabétisation

Les recherches récentes sur les langues africaines se sont concentrées sur la constitution de **corpus audio**, dans l'objectif de développer des modèles de *speech-to-text* (ASR) et de *text-to-speech* (TTS). Ces technologies sont essentielles pour améliorer l'accessibilité des contenus numériques pour les personnes analphabètes. Toutefois, comme le montre (Huettig & Pickering, 2019), ces outils ne permettent pas à eux seuls de lutter contre l'analphabétisme, dans la mesure où ils ne favorisent pas l'apprentissage de la lecture et de l'écriture, éléments pourtant cruciaux dans le développement des compétences cognitives liées à la compréhension du langage oral.

C'est pourquoi nous défendons une approche complémentaire centrée sur les **corpus écrits**, en particulier autour de deux dimensions : la **translittération Latin–Ajami** et la **reconnaissance optique des caractères** (OCR) appliquée aux manuscrits en Ajami. Ces corpus constituent un levier essentiel pour la réhabilitation numérique de ce système d'écriture, en rendant les documents historiques accessibles et en facilitant la création de contenus modernes adaptés aux usages contemporains.

3.2 Méthodologie de collecte des données

La constitution d'un corpus de qualité est un prérequis à la mise en œuvre d'outils de TAL performants. Nous avons structuré notre approche autour de deux types de ressources complémentaires.

Corpus de translittération Latin–Ajami : Ce corpus couvre plusieurs langues d'Afrique de l'Ouest et du Centre, dans l'objectif de former un modèle de translittération capable de généraliser au-delà des variantes dialectales. Contrairement aux approches fondées sur des règles linguistiques formelles (hadji M. Fall *et al.*, 2020), notre démarche vise à refléter les usages réels, tels qu'observés dans les manuscrits et pratiques communautaires.

Les systèmes fondés sur des règles présentent en effet plusieurs limites : une rigidité face à la variation phonétique (Kandybowicz, 2008, 2009), une faible tolérance aux mots rares ou aux exceptions, une dépendance à des conventions parfois peu connues de la population, et une incapacité à prendre en compte le contexte syntaxique ou discursif. En entraînant nos modèles sur des données directement issues de manuscrits, nous visons à respecter la diversité des pratiques graphiques traditionnelles et à produire une translittération fidèle et accessible.

Corpus OCR pour manuscrits Ajami : Le second volet de notre corpus porte sur la numérisation et l'annotation de documents historiques en Ajami. L'objectif est de développer un modèle OCR adapté à la morphologie spécifique de cette écriture.

Dans les communautés rurales, la majorité des personnes alphabétisées en Ajami l'ont été via la calligraphie manuscrite, et sont peu familiarisées avec des polices numériques modernes. Ainsi, la génération d'images typographiques en style manuscrit, à partir de texte numérique, pourrait

faciliter l'appropriation locale. Le corpus OCR, en permettant la reconnaissance automatique de ces manuscrits et la génération d'écriture manuscrite, constitue une étape clef dans la constitution de bibliothèques numériques et d'outils pédagogiques accessibles.

3.3 Vers un système intégré de translittération et de numérisation

Afin d'outiller durablement la réhabilitation numérique de l'Ajami, nous présentons ici deux contributions techniques principales.

AjamiXTranslit (TutlaytAI, 2025) : corpus de translittération Latin–Ajami & OCR Ce corpus contient :

- des paires de phrases en alphabet latin et en Ajami pour plusieurs langues africaines (Haoussa, Peul, Soninké, Wolof, Swahili, Old Kanembu, Tamazight) ;
- des annotations de manuscrits anciens, avec numérisation du texte Ajami et alignement phrase par phrase avec sa version en alphabet latin.

Une plateforme collaborative d'enrichissement :

- Une interface permettant aux locuteurs natifs et chercheurs d'enrichir le corpus via la saisie directe de translittérations et transcription Ajami.
- Un module d'annotation visuelle pour associer des segments d'image de manuscrits à leur transcription textuelle.

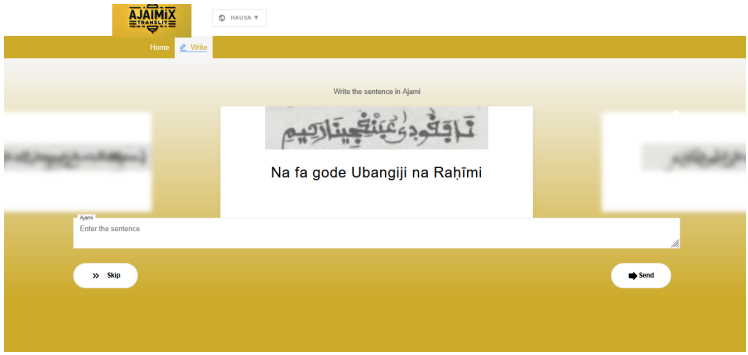

FIGURE 1 – Plateforme pour la transcription collaborative en Ajami.

Ces deux ressources visent à moderniser l'écriture Ajami en la rendant interopérable avec les outils numériques actuels. En combinant la translittération automatique et l'OCR, notre approche contribue à réduire l'analphabétisme fonctionnel tout en préservant un pan essentiel du patrimoine linguistique africain.

3.4 Expérimentations préliminaires sur l'OCR Ajami

Pour valider la faisabilité de l'OCR Ajami à partir de notre corpus, nous avons constitué un petit corpus de 264 échantillons issus de manuscrits Hausa en Ajami et de leurs transcriptions, notamment à partir des ressources du projet (BUA, 2020).

Protocole expérimental. Nous avons entraîné le modèle TrOCR (Li *et al.*, 2022), performant pour la reconnaissance manuscrite, bien qu'il ne soit pas initialement conçu pour la graphie arabe. Les entraînements ont été réalisés sur GPU T4 via Google Colab, en utilisant un batch de taille 8.

Résultats avec 264 échantillons. Les résultats sont présentés dans le tableau 1.

Le taux d'erreur caractère (CER) obtenu est de 85,6 %. Ce résultat, bien que élevé, s'explique par la taille éxtrémement réduite du corpus et l'absence de pré-entraînement spécifique à l'écriture arabe.

Impact d'un accroissement modeste du corpus. Nous avons ensuite enrichi le corpus avec 111 échantillons supplémentaires, portant le jeu d'entraînement à 375 échantillons. Les performances, résumées dans le tableau 1, montrent un gain de 10 points sur le CER, qui passe à 74 %.

	264 ÉCHANTILLONS				**375 ÉCHANTILLONS**			
Époque	1	2	3	4	1	2	3	4
Training Loss	4.71	3.58	2.45	1.40	-	9.25	5.91	4.98
Validation Loss	5.49	5.56	6.12	6.52	7.92	6.57	5.80	5.66
CER	85.6%				74%			

TABLE 1 – Comparaison des résultats OCR pour les deux corpus (264 et 375 échantillons)

Conclusion expérimentale. Cette amélioration, obtenue malgré un corpus encore très limité, souligne l'importance de constituer un corpus plus conséquent pour l'OCR Ajami. Elle valide la pertinence de notre démarche et la nécessité de la plateforme collaborative AjamiXTranslit pour renforcer la collecte et l'annotation à large échelle.

4 L'illettrisme fonctionnel en contexte digraphique

L'illettrisme fonctionnel désigne une situation dans laquelle un individu, bien qu'ayant appris à lire ou écrire dans un système donné, demeure incapable d'accéder à l'information écrite dans un autre système graphique couramment utilisé pour sa langue. Ce phénomène est particulièrement accentué dans les contextes de digraphie, où une même langue est véhiculée par plusieurs systèmes d'écriture sur un même territoire. La digraphie, en multipliant les normes graphiques, peut constituer une barrière supplémentaire à l'accès au savoir, notamment dans un monde de plus en plus numérisé (Baraka, 2024).

Au Sénégal, ce phénomène reste relativement marginal : le wolof y est principalement écrit en alphabet latin, et l'usage parallèle de l'Ajami reste limité (NGUER *et al.*, 2020). En revanche, au Maghreb,

la digraphie affecte massivement l'arabe dialectal (darija) et les langues berbères, engendrant des formes spécifiques d'exclusion linguistique et numérique.

4.1 Problématique de la digraphie au Maghreb

L'arabe dialectal constitue un exemple paradigmatique d'instabilité graphique. Bien que l'alphabet arabe en soit historiquement l'écriture de référence, l'alphabet latin est massivement utilisé dans les échanges informels, en particulier via des formes hybrides comme l'Arabizi. Cette situation s'explique par plusieurs facteurs :

— L'essor d'Internet dans les années 2000, période durant laquelle les caractères arabes étaient peu accessibles sur les claviers et navigateurs, a favorisé la généralisation du *chat alphabet* (Arabizi) dans les usages numériques.
— L'omniprésence de l'alphabet latin dans l'espace public (signalétique, publicité) et dans les interactions quotidiennes, en particulier chez les jeunes générations.
— L'alphabétisation d'une part importante de la population en français ou en anglais, renforçant l'usage spontané de la graphie latine.

Les langues berbères sont également concernées par une triple digraphie (Ajami, latin, tifinagh).
— En Algérie, l'enseignement formel utilise principalement l'alphabet latin, mais des usages résiduels de l'Ajami subsistent dans certaines régions. Il est même proposé comme première écriture dans des manuels destinés aux enfants ayant des difficultés à maîtriser l'alphabet latin (Wikipédia, 2025).
— Au Maroc, l'Ajami reste pratiqué dans des contextes traditionnels, tandis que l'alphabet latin connaît un développement accru depuis la mise en place de la standardisation graphique.
— Le tifinagh, bien qu'érigé en alphabet officiel, demeure un symbole identitaire plus qu'un outil fonctionnel : son usage effectif dans la vie courante reste très limité (INALCO, 2025).

Cette coexistence de graphies non normalisées pose de réels défis à l'interopérabilité linguistique et à la construction d'outils numériques inclusifs pour les communautés maghrébines.

4.2 Importance d'un modèle de translittération unifié

Dans un tel contexte, l'élaboration d'un modèle de translittération unifié apparaît comme une condition essentielle pour garantir un accès équitable à l'information écrite. L'absence de standardisation de la graphie latine du darija engendre une fragmentation des usages : chaque locuteur peut adopter des conventions personnelles selon son bagage éducatif ou sa communauté.

Nous proposons de développer un modèle de translittération entraîné sur un corpus large et hétérogène, intégrant des variantes dialectales et des graphies informelles. Contrairement aux approches monodialectales, notre modèle vise à :

— Prendre en compte la diversité des écritures du darija (arabizi, arabes scripturaux, variations régionales).
— Générer des translittérations cohérentes, adaptées aux contraintes éducatives et numériques.
— Intégrer des représentations de l'écriture berbère en Ajami et en alphabet latin pour une couverture linguistique complète.

Ce modèle pourrait faciliter la mutualisation de ressources éducatives, la traduction automatique

dialectale, ainsi que la navigation intergraphique dans les interfaces numériques (moteurs de recherche, applications éducatives).

4.3 Limites des modèles monodialectaux et validation expérimentale

Nous avons évalué le modèle *Transliteration-Moroccan-Darija* (atl, b), entraîné spécifiquement pour la conversion du darija marocain écrit en Arabizi vers l'alphabet arabe (Younes *et al.*, 2020). L'évaluation a été conduite sur trois corpus dialectaux du Maghreb :

— **ATAM** (atl, a), pour le dialecte marocain ;
— **TArC** (Zribi *et al.*, 2020), pour le dialecte tunisien ;
— **DzNER** (Dahou & Cheragui, 2023), pour le dialecte algérien.

Les performances ont été mesurées via le score BLEU et le taux d'erreur de caractère (CER), comme résumé dans le tableau 2.

Dialecte	Score BLEU	Taux d'erreur (CER)
Marocain	7.11	65%
Tunisien	0.16	194%
Algérien	0.03	255%

TABLE 2 – Évaluation du modèle monodialectal Transliteration-Moroccan-Darija sur différents dialectes maghrébins

Ces résultats révèlent une faible performance pour le domaine du dialecte pour lequel le modèle a été entrainé. De plus pour les dialectes ne faisant pas partie du domaine dialectal du modèle initial, le modèle échoue considérablement à généraliser aux autres variantes, confirmant la nécessité de construire une ressource unifiée pour l'ensemble du Maghreb.

Nous envisageons donc d'enrichir le corpus avec une labellisation dialectale fine, permettant à la fois de renforcer les performances de translittération et d'améliorer leur couplage avec les outils de traduction automatique ou d'indexation linguistique.

5 Défis techniques, linguistiques et sociaux de la translittération Ajami–Latin et de l'OCR Ajami

L'intégration de l'écriture Ajami dans les technologies de traitement automatique des langues soulève une série de défis interdépendants, à la croisée des dimensions technique, linguistique et sociale. Dans cette section, nous analysons ces enjeux à partir de notre expérience de terrain et des expérimentations menées sur le corpus AjamiXTranslit.

5.1 Défis techniques

Sur le plan technique, deux grands axes se distinguent : la translittération automatique et la reconnaissance optique des caractères (OCR).

D'une part, l'absence de normalisation de l'écriture Ajami complique l'entraînement des modèles. Un système efficace doit être capable d'intégrer la diversité des graphies régionales, tout en conservant une cohérence dans la conversion vers l'alphabet latin. Cela nécessite des approches de type *many-to-one* ou *context-aware* basées sur des réseaux neuronaux, capables de capturer les correspondances contextuelles entre graphies et sons.

D'autre part, le développement de systèmes OCR adaptés à l'Ajami est entravé par plusieurs facteurs : la rareté des corpus annotés, la variabilité stylistique des écritures manuscrites, et la complexité morphographique des lettres arabes manuscrites. De nombreuses variantes calligraphiques — parfois très éloignées de l'écriture arabe imprimée — imposent la collecte de données spécifiques et l'adaptation des modèles de vision existants.

5.2 Défis linguistiques

L'un des principaux défis linguistiques concerne la grande variabilité des conventions orthographiques de l'Ajami. Contrairement aux systèmes d'écriture standardisés, l'Ajami s'est développé de manière empirique, au sein de communautés orales ou religieuses, entraînant une forte hétérogénéité des pratiques graphiques (Ziadah, 2009; Warren-Rothlin, 2012b; Vydrin, 2014b; Mumin & Versteegh, 2014). Ainsi, une même langue peut être transcrite de façons très différentes selon les régions ou les contextes.

Un autre défi concerne la prise en compte de la dimension phonétique. L'Ajami sert souvent à transcrire des langues orales sans tradition écrite normalisée. Par conséquent, les conventions orthographiques dépendent étroitement des prononciations régionales. Un système de translittération efficace doit donc intégrer une représentation des régularités phonologiques pour garantir une conversion fidèle à la réalité linguistique (Warren-Rothlin, 2010).

La complexité s'accroît encore lorsqu'il s'agit de couvrir plusieurs langues (peul, wolof, haoussa, soninké, etc.), chacune ayant développé ses propres adaptations de l'alphabet arabe.

5.3 Défis sociaux et acceptabilité

La réussite d'un projet de réhabilitation de l'Ajami repose également sur des facteurs sociaux et culturels. L'adoption des outils de translittération ou d'OCR dépend de leur acceptabilité locale, de leur accessibilité, mais aussi de la manière dont ils s'insèrent dans les usages quotidiens.

Malgré une riche tradition d'usage de l'Ajami dans de nombreuses communautés, son déclin historique — au profit de l'alphabet latin — a contribué à marginaliser cette graphie. Dans ce contexte, toute tentative de réhabilitation implique un travail de sensibilisation, de formation, et de réintégration progressive dans les sphères éducatives, religieuses et administratives. La réussite passe notamment par la co-construction des outils avec les communautés concernées.

Enfin, les contraintes d'infrastructures technologiques dans certaines régions d'Afrique de l'Ouest et du Nord (Shanahan & Bahia, 2024) rendent nécessaire la conception d'outils accessibles hors ligne, à faible consommation énergétique, ou adaptables à des dispositifs mobiles peu puissants. L'ergonomie (interface adaptée, écriture facilitée, visualisation intuitive) joue également un rôle central dans l'appropriation effective des outils proposés.

6 Perspectives futures et applications transdisciplinaires

Les travaux présentés ouvrent la voie à de multiples applications, au croisement de l'éducation, de la préservation du patrimoine, des politiques linguistiques et de l'ingénierie des langues. Les ressources développées — corpus AjamiXTranslit, plateforme collaborative, modèles de translittération et d'OCR — offrent des opportunités concrètes pour renforcer l'accès au savoir, la valorisation culturelle et l'inclusion numérique.

6.1 Numérisation et éducation

Un des objectifs majeurs de la translittération Ajami–Latin et de l'OCR Ajami est de contribuer à l'alphabétisation des populations maîtrisant l'alphabet arabe mais considérées comme analphabètes selon les critères internationaux. En permettant à ces individus de lire et d'écrire dans leur langue maternelle, avec une graphie familière, ces outils réduisent les barrières d'accès à l'éducation formelle et aux contenus pédagogiques.

L'intégration de ces technologies dans des plateformes éducatives — en ligne ou hors ligne — permettrait de proposer des environnements multimodaux (écrit, oral, interactif) pour l'apprentissage des langues africaines. Des applications mobiles pourraient ainsi offrir des exercices de lecture, des outils de dictée en Ajami, ou encore des modules de traduction interactive. L'objectif est de rendre l'apprentissage accessible, contextualisé et culturellement adapté.

6.2 Préservation et valorisation du patrimoine écrit en Ajami

La numérisation des manuscrits Ajami constitue une opportunité unique de sauvegarder un pan essentiel du patrimoine linguistique africain. Grâce aux technologies d'OCR, ces documents manuscrits — souvent inaccessibles ou en voie de dégradation — peuvent être rendus lisibles, indexés, traduits, et diffusés à grande échelle.

Ces ressources pourraient alimenter des bibliothèques numériques, des bases de données linguistiques, ou des corpus d'études pour les chercheurs en linguistique, histoire, sociologie, ou anthropologie. En intégrant le patrimoine Ajami dans les outils contemporains d'édition et de documentation, on réinscrit cette tradition graphique dans une dynamique d'usage et de valorisation.

6.3 Impact sur les politiques linguistiques et technologiques

L'intégration de l'Ajami dans les technologies linguistiques pourrait également avoir des effets sur les politiques linguistiques nationales et régionales. Une reconnaissance officielle de son rôle éducatif, patrimonial et communicationnel pourrait encourager son inclusion dans les programmes scolaires, les administrations locales, ou les médias communautaires.

En outre, les avancées méthodologiques présentées ici pourraient être transposées à d'autres systèmes d'écriture faiblement dotés, comme le N'Ko, le Garay ou le Tifinagh. Cette démarche de généralisation renforcerait l'inclusivité des technologies de traitement automatique des langues, en élargissant leur champ d'application aux écritures minorées ou marginalisées.

7 Discussion et conclusion

Les résultats de notre étude mettent en lumière le potentiel considérable de l'écriture Ajami dans la lutte contre l'illettrisme fonctionnel en Afrique subsaharienne et au Maghreb. En exploitant un alphabet déjà maîtrisé par une large partie de la population, mais sous-utilisé à des fins linguistiques et éducatives, la translittération Latin–Ajami ainsi que les outils OCR apparaissent comme des moyens permettant l'inclusion linguistique.

Cette approche offre une réponse innovante à une problématique souvent abordée sous l'angle des systèmes éducatifs occidentaux, en réhabilitant un système d'écriture historiquement local. Elle s'inscrit dans la continuité des travaux de (Warren-Rothlin, 2010; Ngom, 2018; Mc Laughlin, 2017) qui ont souligné l'importance de valoriser les pratiques scripturales endogènes. De plus, elle rejoint les appels récents en TAL à mieux inclure les langues peu dotées dans les architectures NLP (Shanahan & Bahia, 2024), notamment en tenant compte des réalités sociales, technologiques et linguistiques locales.

Nos résultats suggèrent que les approches multimodales, couplant textes, images et pratiques collaboratives, peuvent dépasser les limites des modèles monodialectaux ou centrés uniquement sur l'oralité. La construction de corpus holistiques, combinant corpus de translittération et OCR, pourrait ainsi s'étendre à d'autres systèmes digraphiques, comme ceux du Maghreb, et contribuer à des projets de standardisation linguistique.

Les perspectives pour de futures recherches sont nombreuses. Un axe prioritaire consiste à affiner l'intégration du contexte phonologique et morphosyntaxique dans les modèles de translittération. Un second défi est de tester l'acceptabilité sociolinguistique des textes générés, en menant des enquêtes qualitatives sur la réception des contenus Ajami numérisés. Enfin, le développement de solutions déconnectées, accessibles hors-ligne, et la généralisation de l'annotation collaborative représentent des pistes concrètes pour une mise à l'échelle de ces outils.

En revalorisant une tradition graphique ancestrale au service de l'alphabétisation contemporaine, cette recherche ouvre la voie à une convergence entre technologie, patrimoine culturel et justice linguistique.

Références

Atam dataset. https://huggingface.co/datasets/atlasia/ATAM.

Transliteration-moroccan-darija. https://huggingface.co/atlasia/Transliteration-Moroccan-Darija.

(2020). Digitizing ajami : African written language. Consulté le 29 mai 2025.

ADOLPHS S. & CARTER R. (2013). *Spoken Corpus Linguistics : From Monomodal to Multimodal*. Routledge, 1 édition. Consulté en mars 2025, DOI : 10.4324/9780203526149.

AGENCE NATIONALE DE LA STATISTIQUE ET DE LA DÉMOGRAPHIE (ANSD) (2013). *Rapport définitif du Recensement Général de la Population, de l'Habitat, de l'Agriculture et de l'Élevage (RGPHAE) 2013*. Rapport interne, Agence Nationale de la Statistique et de la Démographie (ANSD). Consulté à la page 81.

ATLASOCIO (2024). Classement des États d'afrique selon le taux d'alphabétisation (adultes). Consulté en mars 2025.

BANK W. (1999). Education and koranic literacy in west africa. Published by the Africa Region's Knowledge and Learning Center. IK Notes 11, August 1999.

BARAKA K. (2024). Digital divide and social inequality. *International Journal of Humanity and Social Sciences*, **3**, 30–45. DOI : 10.47941/ijhss.2083.

BURSSENS A. & VAN BULCK G. (1935). De africa-spelling en de kongoleesche talen. *Kongo-Overzee*, **2**(2), 65–93.

CALLAHAN M. (2022). Unearthing a long ignored african writing system, one researcher finds african history, by africans. https://www.bu.edu/articles/2022/unearthing-a-long-ignored-african-writing-system/. Accessed on January 21, 2023.

CISSÉ M. (2006). Écrit et écriture en afrique de l'ouest. *Revue électronique internationale des sciences du langage Sud Langues*, (6).

CLARK S. (2007). *Alphabet and orthography statement for Fulfulde [FUB] Ajamiya (found in Nigeria, Cameroon, Chad and Central African Republic)*. Yaoundé : SIL.

DAHOU A. H. & CHERAGUI M. A. (2023). Dzner : A large algerian named entity recognition dataset. *Natural Language Processing Journal*, **3**, 100005. DOI : https ://doi.org/10.1016/j.nlp.2023.100005.

DIALLO A. B. (2025). À propos du système éducatif nigérien. *Revue Internationale d'Éducation de Sèvres*, **11302**.

DIOP B., FAYE A., SYLLA Y. & GUEYE A. (1990). *L'Impact des journaux en langues nationales sur les populations sénégalaises*. Dakar : Association des chercheurs sénégalais.

DIRECTION GÉNÉRALE DU TRÉSOR (2024). Le secteur de l'éducation en afrique subsaharienne. Consulté en mars 2025.

EDUCATION DEPARTMENT OF GOLD COAST (1930). *The New Script and Its Relation with the Languages of the Gold Coast*. Accra : Hertford.

FOR STATISTICS U. I. (2025). Literacy. Accessed : 2025-03-10.

HADJI M. FALL E., HADJI M. NGUER E., SOKHNA B. D., KHOULE M., MANGEOT M. & CISSE M. T. (2020). Digraphie des langues ouest africaines : Latin2ajami : un algorithme de translittération automatique. *arXiv preprint arXiv :2005.02827*.

HUETTIG F. & PICKERING M. J. (2019). Literacy advantages beyond reading : Prediction of spoken language. *Trends in Cognitive Sciences*, **23**(6), 464–475. DOI : 10.1016/j.tics.2019.03.008.

INALCO (2025). Le tifinagh, alphabet berbère : histoire et statut actuel. Consulté le 21 mars 2025.

INSTITUT DE STATISTIQUE DE L'UNESCO (2016). Les taux d'alphabétisation augmentent, mais des millions de personnes restent analphabètes. Consulté en mars 2025.

INSTITUT DE STATISTIQUE DE L'UNESCO (2025). Alphabétisation fonctionnelle. https://uis.unesco.org/fr/glossary-term/alphabetisation-fonctionnelle. Consulté en mars 2025.

INSTITUT INTERNATIONAL DES LANGUES ET CIVILISATIONS AFRICAINES (IILCA) (1930). *Orthographe pratique des langues africaines*. Londres et Paris : IILCA.

INTERNATIONAL INSTITUTE OF AFRICAN LANGUAGES AND CULTURES (IIALC) (1927). Practical orthography of african languages (supplément). *Le Maître phonétique*. October–December issue.

KANDYBOWICZ J. (2008). *The Grammar of Repetition : Nupe Grammar at the Syntax-phonology Interface*, volume 136 de *Linguistik Aktuell / Linguistics Today*. Amsterdam : John Benjamins Publishing. Consulté en mars 2025.

KANDYBOWICZ J. (2009). Embracing edges : syntactic and phono-syntactic edge sensitivity in nupe. *Natural Language & Linguistic Theory*, **27**(2), 305–344. Consulté en mars 2025, DOI : 10.1007/s11049-009-9064-6.

LACROIX P.-F. (1968). Transcription de langues africaines. *Journal de la Société des Africanistes*, **38**(2), 227–234. DOI : 10.3406/jafr.1968.1483.

LI M., LV T., CHEN J., CUI L., LU Y., FLORENCIO D., ZHANG C., LI Z. & WEI F. (2022). Trocr : Transformer-based optical character recognition with pre-trained models.

MAGA H. I. & HAMIDOU O. (2015). *La dynamique de l'alphabétisation au Niger : que nous apprennent les données censitaires et administratives ?* Rapport interne, Observatoire démographique et statistique de l'espace francophone (ODSEF), Université Laval.

MANN M. & DALBY D. (1987). *A Thesaurus of African Languages : A Classified and Annotated Inventory of the Spoken Languages of Africa with an Appendix on Their Written Representation*. London : Hans Zell Publishers.

MC LAUGHLIN F. (2017). Ajami writing practices in atlantic-speaking africa. In F. LÜPKE, Éd., *The Atlantic Languages*. Oxford University Press.

MUMIN M. & VERSTEEGH K. (2014). *The Arabic Script in Africa : Studies in the Use of a Writing System*. Leiden, The Netherlands : Brill. DOI : 10.1163/9789004256804.

NATIONAL AFRICAN LANGUAGE RESOURCE CENTER (2015). Pulaar. http://yumpu.com/en/document/read/43874132/pulaar-national-african-language-resource-center. Published on July 11, 2015.

NDAGI M. U. (2011). A thematic exposition of the nupe ajami manuscript heritage of northern nigeria. *Islamic Africa*. Consulté en mars 2025, DOI : 10.5192/21540993020111.

NGOM F. (2010). Ajami scripts in the senegalese speech community. *Journal of Arabic and Islamic Studies*, **10**, 1–23.

NGOM F. (2016). *Muslims beyond the Arab World*. Oxford University Press. DOI : 10.1093/acprof :oso/9780190279868.001.0001.

NGOM F. (2018). Ajami literacies of west africa. In E. A. ALBAUGH & K. M. DE LUNA, Éds., *Oxford Scholarship Online*, volume 1. Oxford University Press. DOI : 10.1093/oso/9780190657543.001.0001.

NGUER E. H. M., BAO D. S., FALL Y. A. & KHOULE M. (2020). Digraph of senegal's local languages : Issues, challenges and prospects of their transliteration. *arXiv preprint*.

PRIEST L. A. & HOSKEN M. (2010). Proposal to add arabic script characters for african and asian languages. https://www.unicode.org/L2/L2010/10288r-arabic-proposal.pdf. Submitted on August 4, 2010.

SHANAHAN M. & BAHIA K. (2024). *The State of Mobile Internet Connectivity Report 2024*. Rapport interne, GSMA. Contributors : Abi Gleek, Melle Tiel Groenestege, Claire Sibthorpe, Anne Shannon Baxter, Eleanor Sarpong, Harry Fernando Aquije Ballon, Anna-Noémie Ouattara Boni. Consulté le 11 mars 2025.

SOUAG L. (2009). Ajami in west africa. Accessed : 2025-03-10.

TAYLOR S. & FAUZI F. (2024). Multimodal sentiment analysis for the malay language : New corpus using cnn-based framework. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted, DOI : 10.1145/3703445.

TUTLAYTAI (2025). Ajamixtranslit. https://huggingface.co/datasets/TutlaytAI/AjamiXTranslit. Dataset disponible sur Hugging Face.

UNESCO (1980). *Alphabet africain de référence*. Paris : UNESCO, Secteur de la Culture et de la Communication.

UNESCO (1981). *African Languages : Proceedings of the Meeting of Experts on the Transcription and Harmonization of African Languages, Niamey (Niger), 17–21 July 1978*. Paris : UNESCO.

VYDRIN V. (2014a). Ajami script for the mande languages. In M. MUMIN & K. VERSTEEGH, Éds., *The Arabic Script in Africa : Studies in the Use of a Writing System*, p. 199–224. Brill. DOI : 10.1163/9789004256804_011.

VYDRIN V. (2014b). *Ajami Scripts for Mande Languages*, p. 199 – 224. Brill : Leiden, The Netherlands. DOI : 10.1163/9789004256804_011.

VYDRINE V. (1998). Sur l'écriture mandingue et mandé en caractères arabes (mandinka, bambara, soussou, mogofin). *Mandenkan*, (33).

WARD I. C. (1936). Problems of orthography in the congo belge—'de "africa"-spelling ende kongoleesche talen.'. *Africa*. DOI : 10.1017/S0001972000006112.

WARREN-ROTHLIN A. (2010). West african scripts and arabic-script orthographies in socio-political context. *The Arabic Script in Africa : Diffusion, Usage, Diversity and Dynamics of a Writing System*, **8 [860]**, 261–89. Consulted March the 11th of 2025.

WARREN-ROTHLIN A. (2012a). Arabic script in modern nigeria. In R. BLENCH & S. MCGILL, Éds., *Advances in Minority Language Research in Nigeria*, volume 1 de *Kay Williamson Educational Foundation African languages monographs*. Köln : Rüdiger Köppe Verlag.

WARREN-ROTHLIN A. (2012b). Arabic script in modern nigeria.

WESTERMANN D. (1927). *A Common Script for Twi, Fante, Gã and Ewe*. Accra : Government Printer.

WIKIPÉDIA (2025). Alphabet arabe berbère - algérie. [Consulté le 11 mars 2025].

YOUNES J., SOUISSI E., ACHOUR H. & FERCHICHI A. (2020). Language resources for maghrebi arabic dialects' nlp : a survey. *Language Resources and Evaluation*, **54**(4), 1079–1142.

ZIADAH M. (2009). What are the ajami ? *The UNESCO Courier*, **8 [860]**, 10. Consulté le 11 mars 2025.

ZRIBI I., ELLOUZE M. *et al.* (2020). Tarc : Incrementally and semi-automatically collecting a tunisian arabish corpus. In *Proceedings of WANLP 2020*.