

Quand l'IA ne parle pas notre langue : Enjeux techniques et sociolinguistiques de la reconnaissance automatique de l'arabe algérien

Sarra, BELGHOUAR¹ 

Université Amar Têlidji, Laghouat / Algérie
Laboratoire des Sciences du langage
s.belghouar@lagh-univ.dz

Souad, KHOUDOUR²

Ecole Normale Supérieure de Constantine / Algérie
benhcine.souad@ensc.dz

Reçu : 15/08/2025,

Accepté: 09/12/2025,

Publié: 22/12/2025

When AI Does Not Speak Our Language: Technical and Sociolinguistic Challenges in Automatic Speech Recognition of Algerian Arabic

Abstract: This article investigates the technical and sociolinguistic challenges of integrating Algerian Arabic into Automatic Speech Recognition and Natural Language Processing systems, particularly in contexts involving code-switching with French. Significant dialectal variation, the absence of orthographic standardization, and the overrepresentation of Modern Standard Arabic in training corpora limit these systems' ability to accurately reflect our linguistic reality. This situation requires addressing both technical issues (phonetic, prosodic, and script diversity) and an inclusion imperative to prevent the invisibilization of dialectal varieties. To this end, we adopt a theoretical framework combining computational linguistics and digital sociolinguistics, and propose a methodology based on the collaborative construction and annotation of representative corpora that incorporate regional variants, code-switching, and scriptural diversity, with the aim of developing fairer models adapted to the Algerian context.

Keywords: Algerian Arabic, dialectal variation, automatic speech recognition, natural language processing, linguistic inclusion.

Résumé : Nous examinons dans cet article les enjeux techniques et sociolinguistiques liés à l'intégration de l'arabe algérien dans les systèmes de reconnaissance automatique de la parole et de traitement automatique du langage, notamment dans les situations de code-switching avec le français. La forte variation dialectale, l'absence de standardisation orthographique et la surreprésentation de l'arabe standard dans les corpus d'entraînement limitent la capacité de ces systèmes à refléter fidèlement notre réalité langagière. Cette situation impose de relever des défis techniques (diversité phonétique, prosodique et graphique) et un impératif d'inclusion pour éviter l'invisibilisation des variétés dialectales. Pour y répondre, nous mobilisons un cadre théorique associant linguistique computationnelle et sociolinguistique numérique, et proposons une méthodologie fondée

sur la constitution et l'annotation collaborative de corpus représentatifs intégrant variantes régionales, alternances codiques et spécificités scripturales, en vue de développer des modèles plus équitables et adaptés au contexte algérien.

Mots-clés : Arabe algérien, variation dialectale, reconnaissance automatique de la parole, traitement automatique du langage, inclusion linguistique.

Introduction

Ces dernières années, nous avons été témoins d'une transformation profonde dans le domaine des technologies de l'information, portée par les avancées de l'Intelligence Artificielle (IA) et du traitement automatique du langage (TAL¹). L'IA ne se limite plus à exécuter des instructions programmées: elle est désormais capable de comprendre, d'analyser et de produire notre langage naturel. Des applications comme la reconnaissance vocale, la traduction automatique et les assistants conversationnels, jadis réservées à la science-fiction, font désormais partie intégrante de notre quotidien, présentes dans nos téléphones, nos enceintes connectées et nos véhicules.

Ces progrès reposent sur des années de recherche et sur l'exploitation massive de données, notamment à travers des modèles neuronaux end-to-end² tels que Whisper d'OpenAI (Radford *et al.*, 2023), capables de transcrire la parole même en présence de bruit de fond ou d'un accent marqué. En 2024, on estimait à 8,4 milliards le nombre d'appareils dotés de capacités de reconnaissance vocale dans le monde (Juniper Research, 2020).

Face à cette expansion, une question s'impose : toutes les langues, tous les accents et toutes les variétés dialectales bénéficient-ils du même traitement ? Si les systèmes actuels traitent aisément l'anglais standard ou l'arabe classique, ils rencontrent de sérieuses difficultés face aux parlers régionaux, aux usages hybrides et aux alternances de code.

Le cas algérien illustre particulièrement bien cette problématique : notre paysage linguistique est marqué par la cohabitation de l'arabe algérien, décliné en multiples variantes régionales, du tamazight avec ses différentes variétés et du français utilisé très fréquemment en alternance linguistique (code-switching). Ce phénomène, profondément ancré dans nos échanges quotidiens, se manifeste par des passages fluides d'une langue à l'autre, parfois à l'intérieur d'une même phrase.

Pour les systèmes de TAL, l'arabe algérien pose des défis majeurs : absence de norme orthographique, coexistence d'écritures arabe et latine, grande variation phonétique et lexicale selon les régions. Un même mot peut ainsi être écrit et prononcé différemment d'un locuteur à l'autre. Ces variations, combinées au code-switching, compliquent la

¹ Ensemble des méthodes et technologies permettant aux ordinateurs de traiter le langage humain.

² Architecture qui apprend directement à effectuer une tâche (ex. transcription) sans étapes intermédiaires distinctes.

reconnaissance automatique et risquent de conduire à une invisibilisation numérique des usages linguistiques locaux. Lorsque les modèles sont entraînés principalement sur l'arabe standard ou sur des dialectes étrangers, les pratiques langagières algériennes se retrouvent sous-représentées, ce qui renforce les hiérarchies linguistiques existantes et marginalise certaines variétés.

Notre recherche vise à évaluer la capacité des systèmes actuels - Open source et commerciaux- à reconnaître et traiter l'arabe algérien, en particulier dans les contextes de code-switching avec le français. Pour ce faire, nous avons structuré notre démarche autour de trois axes :

1. La constitution d'un corpus multidialectal d'arabe algérien et de français, issu de médias audiovisuels, de réseaux sociaux et de conversations spontanées enregistrées dans différentes régions (Est, Ouest, Centre, Sud).
2. L'évaluation comparative de modèles de reconnaissance vocale et de traduction automatique sur ce corpus, en analysant leur tolérance à la variation régionale et au mélange linguistique.
3. L'analyse sociolinguistique des erreurs de transcription et de traduction afin de mettre en évidence les biais et limitations structurelles.

Les travaux antérieurs (Abdelali *et al.*, 2024 ; Elnagar *et al.*, 2021 ; Shaalan *et al.*, 2019 ; Hussein *et al.*, 2022 ; Alsayadi *et al.*, 2022 ; Rahman *et al.*, 2024 ; Djanibekov *et al.*, 2025 ; Özyilmaz *et al.*, 2025) confirment que la performance des systèmes diminue nettement face aux dialectes sous-représentés, principalement en raison du manque de corpus annotés et de la diversité orthographique. Nous formulons l'hypothèse d'un biais structurel découlant de la surreprésentation de l'arabe classique et de certains dialectes arabe. Ce biais se manifeste par : une baisse de précision dans la reconnaissance de l'arabe algérien, notamment en contexte de code-switching ; une disparité de traitement entre les régions ; et des difficultés accrues liées à l'absence de norme écrite, qui entravent la prise en compte des particularités lexicales et phonétiques.

Ainsi, cette problématique, située à la croisée de la linguistique computationnelle et de la sociolinguistique numérique, ne se limite pas à un défi technique. Elle constitue également un enjeu de justice linguistique, en invitant à repenser le développement de modèles et d'algorithmes, à construire des corpus représentatifs et à réfléchir à la place que doivent occuper les variétés dialectales algériennes dans l'espace numérique, afin de garantir leur reconnaissance et leur valorisation dans un environnement technologique mondialisé.

I.Cadre théorique

Le traitement automatique du langage (TAL) s'appuie sur deux piliers essentiels : la linguistique computationnelle et la sociolinguistique numérique. Ces deux approches

permettent de comprendre et d'intégrer la complexité linguistique qui se manifeste notamment dans les dialectes algériens.

1. Vers une linguistique computationnelle adaptée aux réalités algériennes

Dans la lignée des technologies de reconnaissance vocale (ASR³), la linguistique computationnelle nous offre des méthodes et des modèles capables de transcrire, analyser et interpréter la parole. Nous vivons aujourd'hui une époque où l'IA transforme profondément notre rapport au langage, notamment grâce aux architectures neuronales end-to-end, capables d'apprendre directement à partir de données audio sans passer par des étapes intermédiaires.

Cependant, la diversité des dialectes arabes, l'absence de standardisation orthographique et la rareté des corpus annotés freinent encore le développement de systèmes ASR véritablement performants dans nos contextes (Rahman *et al.*, 2024). Nous faisons face à un paradoxe : les modèles sont plus puissants que jamais, mais nos dialectes restent sous-représentés. De nouvelles approches émergent néanmoins, comme les modèles *weakly supervised*⁴ entraînés sur 15 000 heures de données mêlant arabe standard et dialectes, atteignant des performances remarquables malgré l'absence de transcriptions manuelles (Salhab *et al.*, 2025)

Ces avancées s'accompagnent d'efforts pour mieux couvrir la diversité dialectale.

Un travail présenté à ACL⁵ 2025 a dévoilé des modèles ASR couvrant 17 pays arabophones, capables de gérer plusieurs dialectes et variations de code-switching, avec des gains notables sur des données réelles (Djanibekov *et al.*, 2025).

Pour nous, la question n'est donc plus de savoir si la linguistique computationnelle peut reconnaître nos dialectes, mais comment nous pouvons la guider pour qu'elle le fasse avec justesse, en respectant notre diversité linguistique.

2. Sociolinguistique numérique et enjeux de la variation dialectale en Algérie

La sociolinguistique numérique nous invite à regarder nos langues à travers le prisme du numérique. Elle interroge la manière dont nos variétés sont représentées, valorisées ou, parfois, marginalisées dans les environnements technologiques. Pour nous, cela ne relève pas seulement d'une question technique : c'est aussi une question d'identité et de reconnaissance culturelle. Dans le cas des dialectes algériens, elle nous aide à comprendre que l'intégration dans les systèmes d'Intelligence Artificielle ne dépend pas uniquement de modèles puissants, mais aussi d'une volonté de préserver notre patrimoine linguistique.

³ Automatic Speech Recognition : technologie qui convertit la parole humaine en texte écrit.

⁴ De supervision faible.

⁵ Association for Computational Linguistics : Association internationale spécialisée dans la recherche et le développement en traitement automatique du langage (TAL).

L'un des concepts centraux est celui de l'invisibilisation linguistique: lorsqu'une variété est absente des corpus d'entraînement, elle devient «inexistante» pour les systèmes de traitement automatique du langage. Cette absence, comme le rappellent Alsharhan *et al.* (2020), n'est pas neutre : elle consolide les hiérarchies linguistiques déjà présentes, où l'arabe standard bénéficie d'un prestige institutionnel tandis que les dialectes, pourtant langue première de millions de locuteurs, sont perçus comme moins légitimes. Nous devons donc être conscients que chaque absence de nos variétés dans les données d'entraînement contribue à effacer une partie de notre voix collective.

En Algérie, cette hiérarchie se double d'une pluralité interne : l'arabe algérien se décline en variantes régionales marquées (Ouest, Centre, Est, Sud) et entretient un contact constant avec le français. Nous passons d'une langue à l'autre avec aisance, souvent dans la même phrase, pratiquant ce code-switching devenu une norme sociale. Or, comme le soulignent Shaalan *et al.* (2019), les systèmes TAL dominants sont souvent entraînés sur des données monolingues stabilisées, ce qui les rend inadaptés à nos usages hybrides.

Le numérique amplifie ces défis. Sur les réseaux sociaux, l'arabe algérien s'écrit souvent en caractères latins (Arabizi), avec une orthographe instable façonnée par nos pratiques informelles. Ce choix graphique, que nous adoptons souvent par commodité, devient pourtant un obstacle pour l'IA. Cette diversité scripturale rend difficile toute normalisation, mais témoigne d'une créativité que Messaoudi (2020) considère comme un signe de vitalité linguistique. Toutefois, cette pluralité reste un défi technique majeur pour les systèmes d'IA.

Ces obstacles techniques ont des conséquences sociolinguistiques profondes : un assistant vocal qui ne reconnaît pas notre dialecte exclut de fait ses locuteurs d'un accès complet aux services numériques, ce qui soulève une question essentielle de justice linguistique (Mattioda, 2024). Si nous voulons que nos communautés soient incluses, nous devons veiller à ce que nos langues soient présentes dans les corpus qui façonnent les IA.

Face à cette situation, des initiatives émergent. Le Workshop on Arabic Corpus Linguistics (WACL-4), organisé dans le cadre de COLING 2025, insiste sur la nécessité de corpus représentatifs intégrant la variation dialectale et le code-switching. Des projets comme le corpus *Casablanca* documentent systématiquement la diversité dialectale arabe, y compris nos variétés algériennes, et mettent ces ressources à la disposition des chercheurs (Talafha *et al.*, 2024).

Ainsi, la sociolinguistique numérique nous rappelle que la technologie ne doit pas seulement comprendre nos mots, mais aussi reconnaître nos identités linguistiques. Pour le cas algérien, cela implique de considérer nos dialectes non comme un problème technique, mais comme un patrimoine que nous devons préserver et valoriser.

3. Corpus dialectaux algériens : constitution, annotation et normalisation graphique

La constitution de corpus représentatifs du dialecte algérien constitue une étape centrale dans la mise en place de systèmes de reconnaissance automatique de la parole (ASR) et de traitement automatique du langage naturel⁶ (NLP) adaptés à notre réalité linguistique. Les approches récentes montrent l'importance des méthodes collaboratives (crowdsourcing⁷) pour collecter des données issues de locuteurs variés, couvrant plusieurs régions et registres d'usage (Hamed *et al.*, 2025). Nous savons que ces méthodes permettent d'intégrer des formes orales et écrites authentiques, y compris celles produites dans des environnements numériques comme les réseaux sociaux.

L'annotation d'un tel corpus nécessite des protocoles adaptés aux phénomènes de code-switching arabe algérien/français, omniprésents dans nos échanges quotidiens. Comme le souligne Messaoudi (2020), il ne s'agit pas seulement de transcrire fidèlement la parole, mais aussi de documenter les alternances linguistiques, les variantes régionales et les choix orthographiques que nous faisons au quotidien.

Sur le plan graphique, un défi majeur réside dans la reconnaissance des écritures non standardisées. En contexte numérique, l'arabe algérien s'écrit fréquemment en caractères latins (arabizi), souvent avec des chiffres remplaçant certaines consonnes (ex. 3, 7, 9). Cette pluralité orthographique, loin d'être un obstacle insurmontable, nous offre au contraire la possibilité de l'exploiter grâce à des approches de translittération automatique et à des modèles basés sur les caractères ou sur les unités sous-lexicales capables de gérer la variabilité scripturale (Abainia, 2020). L'objectif est de faire de la normalisation graphique un levier pour améliorer la qualité de l'ASR et du NLP tout en respectant la richesse expressive de notre dialecte.

4. Ressources linguistiques algériennes : fondements pour la reconnaissance automatique de la parole

Dans notre démarche pour mieux appréhender la variation dialectale algérienne et développer des technologies linguistiques inclusives, nous nous sommes appuyés sur plusieurs corpus de référence, chacun présentant des atouts spécifiques pour capturer notre diversité langagière.

Sawt El-Djazair constitue, à ce jour, l'un des corpus les plus aboutis dédiés à l'arabe algérien. Conçu dans une perspective médico-légale, il rassemble environ 50 heures de parole multi-dialectale, couvrant douze sous-dialectes algériens, dans des contextes variés allant des appels cellulaires aux communications VoIP telles que WhatsApp ou

⁶ Natural Language Processing : Traitement automatique du langage naturel : discipline de l'IA qui permet aux machines de comprendre et de produire du texte.

⁷ Production participative: collecte de données ou d'annotations via la contribution d'un grand nombre de personnes, souvent en ligne.

Viber (Zergat *et al.*, 2023). Nous pouvons reconnaître dans ce corpus une richesse rare, car il a été structuré pour garantir une couverture géographique et contextuelle étendue : en modes spontané, formel et informel

Le second corpus, *ALGASD* (Algerian Arabic Speech Database, 2010), bien que centré sur l'arabe standard moderne (MSA), présente l'intérêt d'avoir été recueilli auprès de 300 locuteurs algériens issus de onze régions du pays. Selon Roua-Hamdani (2010), l'ensemble a été construit selon un protocole rigoureux inspiré de TIMIT : des phrases phonétiquement équilibrées, des informations précises sur les locuteurs (âge, genre, niveau d'instruction) et des données qui rendent ce corpus particulièrement utile comme point de référence pour évaluer les modèles ASR dans notre contexte linguistique et régional.

Enfin, le projet *Casablanca* (2024) incarne un effort communautaire multilatéral remarquable. Ce corpus couvre huit variétés arabes dialectales dont l'algérien, avec des annotations riches incluant la transcription, le genre des locuteurs, l'identification dialectale, et surtout les phénomènes de code-switching (Talafha *et al.*, 2024). Notre intérêt dans ce projet est double : il offre une base multidialectale d'une rare qualité, tout en permettant de confronter nos usages linguistiques - y compris lorsque langues et registres s'entremêlent - à des modèles de pointe.

Chacune de ces ressources nous permet ainsi d'aborder la reconnaissance des dialectes algériens sous plusieurs angles : le contexte forensique (*Sawt El-Djazair*), l'équilibre MSA / variation régionale (*ALGASD*) et la richesse multidialectale y compris le code-switching (*Casablanca*). Ensemble, elles forment un socle solide pour concevoir des technologies linguistiques plus justes et représentatives de notre réalité langagière.

II. Cadre pratique

1. Méthodologie

Afin de répondre à nos objectifs de recherche, nous avons adopté une approche méthodologique combinant analyse comparative, expérimentation empirique et analyse qualitative.

1.1. Sources et données

Notre étude repose sur deux types de sources complémentaires. D'une part, nous avons analysé plusieurs systèmes de reconnaissance vocale déjà existants, parmi lesquels Google Speech-to-Text⁸, Whisper (OpenAI) et Deepgram. Ce choix nous permet de comparer des approches commerciales et open source, et d'observer leurs performances respectives sur des données réelles en arabe algérien. D'autre part, nous avons constitué un corpus original intégrant différentes modalités : des vidéos issues

⁸ Conversion de la parole en texte par un système ASR.

de YouTube, sélectionnées pour leur diversité thématique et géographique ; des enregistrements oraux réalisés avec des locuteurs natifs représentant plusieurs régions d'Algérie, et des dialogues extraits de réseaux sociaux, reflétant l'usage spontané et non standardisé de notre dialecte.

1.2. Procédure de sélection et d'échantillonnage

La constitution de ce corpus a suivi un protocole rigoureux. Nous avons veillé à sélectionner des échantillons représentatifs des principales variantes régionales de l'arabe algérien : Est, Centre, Ouest et Sud. Cette diversité garantit que les résultats de l'évaluation ne soient pas biaisés en faveur d'une seule variété dialectale. Les échantillons incluent aussi bien des discours formels (conférences, interviews) que des échanges informels (discussions amicales, extraits humoristiques), afin de couvrir un large spectre de registres linguistiques.

1.3. Évaluation et analyse

Pour évaluer la précision de transcription, nous avons mesuré des indicateurs standard tels que le Word Error Rate (WER⁹) et le Character Error Rate (CER¹⁰), tout en tenant compte des spécificités de notre contexte linguistique, notamment le code-switching arabe/français. Nous avons également conduit une classification dialectale visant à déterminer la capacité des systèmes à identifier correctement la variante régionale des locuteurs.

Au-delà des mesures quantitatives, nous avons réalisé une analyse qualitative approfondie des erreurs. Celle-ci a porté sur la nature des fautes (substitution¹¹, omission¹², insertion¹³), la sensibilité des systèmes aux emprunts lexicaux au français, et les difficultés liées à la transcription de l'arabe algérien en caractères latins. Cette double approche, quantitative et qualitative, nous permet de mieux comprendre non seulement le niveau de performance des systèmes, mais aussi leurs limites structurelles dans la prise en compte de la richesse et de la variabilité de notre dialecte.

⁹ Taux d'erreurs par mot : mesure de performance d'un système ASR basée sur les substitutions, insertions et omissions.

¹⁰ Taux d'erreurs par caractère : indicateur de précision de transcription au niveau des caractères individuels.

¹¹ Remplacement d'un terme par un autre phonétiquement proche, sans prise en compte du contexte dialectal.

¹² Suppression d'éléments lexicaux, souvent en contexte de débit rapide ou de parole informelle.

¹³ Ajout d'éléments absents dans l'original, souvent motivé par des schémas syntaxiques appris.

2. Résultats

Sur la base de la méthodologie décrite, nous présentons ici des résultats détaillés issus de l'analyse comparative des systèmes de reconnaissance vocale (Google Speech-to-Text, Whisper d'OpenAI et Deepgram). Ces résultats intègrent des mesures quantitatives (WER et CER), une évaluation de la classification dialectale, ainsi qu'une analyse qualitative des erreurs, enrichie d'exemples concrets tirés de notre corpus. Pour rappel, le corpus totalise environ 50 heures d'audio, réparties équitablement entre les variantes régionales (Est, Centre, Ouest, Sud) et les registres (formel/informel). Les tests ont été menés sur un échantillon de 200 extraits audio (durée moyenne : 15 secondes), transcrits manuellement par des linguistes natifs pour servir de référence.

2.1 Mesures Quantitatives : Précision Globale et par Variante Régionale

Nous avons calculé les indicateurs WER et CER en tenant compte du code-switching (mélange arabe/français) et de la transcription en caractères latins, courante pour l'arabe algérien. Les résultats montrent une performance globale modérée, avec Whisper surpassant les autres systèmes grâce à son entraînement sur des données multilingues. Cependant, tous les systèmes peinent avec les variantes régionales moins représentées (Ouest et Sud), où le WER dépasse souvent 40 % en raison de la variabilité phonétique et lexicale.

Système	WER Global (%)	CER Global (%)	WER Région : Est (%)	Centre (%)	Ouest (%)	Sud (%)
Google Speech-to-Text	35.2	28.1	29.8	32.4	38.7	40.0
Whisper (OpenAI)	24.5	19.3	20.1	22.6	27.4	28.0
Deepgram	31.8	25.6	26.5	29.2	34.9	36.7

Tableau 1 : récapitulatif des performances moyennes des systèmes ASR

2.1.1. Interprétation

Whisper affiche le WER le plus bas (24.5 % global), particulièrement dans les régions Est et Centre, où les accents sont plus proches de l'arabe standard ou influencés par des médias diffusés. En revanche, pour l'Ouest (ex. : accents oranais) et le Sud (ex. :

variantes sahariennes avec emprunts touareg), les erreurs augmentent, soulignant un biais d'entraînement vers des données urbaines algéroises (Centre). Le CER est systématiquement inférieur au WER, indiquant que les erreurs portent plus sur les mots entiers que sur les caractères isolés, souvent dues à des substitutions lexicales.

Par registre linguistique, les performances varient également :

- **Formel** (conférences, interviews) : WER moyen de 22.0 % (tous systèmes), car le discours est plus structuré et proche de l'arabe moderne standard (MSA).
- **Informel** (discussions, humoristiques) : WER moyen de 36.8 %, avec un pic à 45 % pour les extraits incluant du code-switching rapide.

2.2. Évaluation de la Classification Dialectale

Nous avons testé la capacité des systèmes à identifier la variante régionale via une tâche de classification intégrée (basée sur des modèles fine-tunés pour Whisper et Deepgram ; Google n'offre pas cette fonctionnalité native). Sur 100 échantillons labellisés manuellement, les taux d'exactitude sont les suivants :

Système	Exactitude Globale (%)	Exactitude par Région : Est (%)	Centre (%)	Ouest (%)	Sud (%)
Whisper (OpenAI)	68.0	75.0	72.0	65.0	60.0
Deepgram	55.0	62.0	58.0	52.0	48.0
Google	Non applicable	-	-	-	-

Tableau 2 : Taux d’exactitude des systèmes ASR selon les régions en Algérie

2.2.1. Exemples Illustratifs de Classification

N°	Système	Variante visée	Contexte énoncé	et	Résultat obtenu	Analyse
----	---------	----------------	-----------------	----	-----------------	---------

1	Whisper	Est – Sétifien	Interview formelle (YouTube) : "Lazem nqadrou el kbir"	Classifié Est, 85%	Succès, reconnaissance du /q/ guttural
2	Deepgram	Ouest – Oranais	"Wach rakoum dirou fi l'week-end ?" (informel)	Classifié Centre, 52%	Échec, influence du français et voyelles proches
3	Whisper	Centre – Algérois	"Rohna ndirou picnic fi el jardin"	Classifié Centre, 78%	Succès, accent urbain et code- switching
4	Deepgram	Sud – Laghouat	"El hadj rah ychouf el sahra"	Classifié Ouest	Échec, influences berbères communes

Tableau 3 : Exemples de succès et d'échecs de classification dialectale

2.2.2. Synthèse

Ces résultats montrent que les systèmes sont sensibles aux marqueurs phonétiques (affaiblissement des gutturales dans le Sud, allongements consonantiques dans l'Ouest, etc.), mais présentent des confusions lorsque les zones dialectales partagent des traits prosodiques ou lorsqu'il y a code-switching, entraînant une baisse d'exactitude estimée entre 10 % et 15 %.

2.3 Analyse Qualitative des Erreurs

L'analyse qualitative a porté sur 150 erreurs représentatives, classées par type (substitutions, omissions, insertions) et contexte (code-switching, transcription latine). Nous avons identifié des motifs récurrents : sensibilité aux emprunts français, confusion avec l'arabe standard, et difficultés avec les sons non représentés en latin (ex. : 'ain, ghayn). Voici une classification des erreurs avec exemples tirés du corpus :

2.3.1 Substitution (45 % des erreurs)

N°	Système & variante	Référence	Sortie système	Analyse
----	-----------------------	-----------	-------------------	---------

Quand l'IA ne parle pas notre langue : Enjeux techniques et sociolinguistiques de la reconnaissance automatique de l'arabe algérien

1	Google Centre (Algérois)	"Ana nchouf el film hada, mais khassni nkhallas el boulot d'abord."	"...nkhallas el boulet d'abord."	Confusion boulot→boulet par méconnaissance du terme français
2	Deepgram - Sud (Saharien)	"...rah ndourou fi dar."	"...rah ndourou fi dahr."	Substitution due à la difficulté de transcription des voyelles allongées
3	Whisper - Est (Sétifien)	"ntaybo el makrout ?"	"...el makroud ?"	Variation lexicale ignorée au profit d'une forme standardisée
4	Google Ouest (Tlemcénien)	"El economie rah t3awen el blad."	"El ikonomi rah t3awen el blad."	Transformation erronée de l'emprunt français en forme arabisée

Tableau 4 : Exemples d'erreurs de substitution des systèmes ASR selon les dialectes algériens

2.3.2 Omission (30 % des erreurs)

N°	Système & variante	Référence	Sortie système	Analyse
1	Whisper - Est (Annabi)	"...Rak t3raf el histoire ?"	"...Rak el histoire ?"	Perte du verbe <i>t3raf</i> , masqué par le code-switching
2	Deepgram - Centre (Blidéen)	"...w nrouhou l'Espagne."	"...nrouhou l'Espagne."	Suppression de la conjonction w
3	Google - Sud (Adrarais)	"...w ybi3ouh fi el marché."	"...ybi3ouh fi el marché."	Omission liée à l'allongement vocalique saharien

Tableau 5 : Exemples d'erreurs d'omission des systèmes ASR selon les dialectes algériens

2.3.3 Insertion (25 % des erreurs)

N°	Système & variante	Référence	Sortie système	Analyse
1	Google - Ouest (Chleuh influencé)	"...mais khir men el ancien."	"...mais khir men el ancien wahed."	Insertion d'un numéral pour se conformer à une structure arabe standard
2	Whisper - Est (Skikdien)	"Ana nkrah el froid."	"Ana nkrah el froid hada."	Ajout de <i>hada</i> pour "compléter" la phrase
3	Deepgram - Centre (Mitidjien)	"El gouvernement rah ydir reformes."	"El gouvernement rah ydir les reformes."	Influence du français par insertion d'un article

Tableau 6 : Exemples d'erreurs d'insertion des systèmes ASR selon les dialectes algériens

2.4. Sensibilité au code-switching

Dans 60 % des erreurs, un mélange arabe/français est impliqué. Par exemple, des termes comme "week-end", "boulot" ou "film" sont souvent substitués par des équivalents arabes (ex. : "film" → "*filim*" mal transcrit). Les systèmes Open source comme Whisper gèrent mieux cela (taux d'erreur 25 % vs. 40 % pour Google), grâce à un entraînement sur des corpus multilingues.

2.4.1.Limites structurelles

L'analyse révèle un biais vers l'arabe standard et les accents urbains, avec une sous-représentation des variants ruraux/Sud. La transcription latine pose des défis supplémentaires, car les systèmes sont optimisés pour l'alphabet arabe ou latin standard, menant à des confusions phonétiques (ex. : "kh" vs. "ch"). Ces résultats soulignent la nécessité de fine-tuning¹⁴ spécifique pour l'arabe algérien.

En conclusion, bien que Whisper offre les meilleures performances, aucun système n'atteint une précision satisfaisante (>80 %) pour un usage quotidien en contexte dialectal varié. Ces résultats ouvrent des pistes pour des améliorations, comme l'intégration de corpus locaux plus diversifiés.

¹⁴ Ajustement fin : réentraînement ciblé d'un modèle pré-entraîné pour l'adapter à un domaine ou à une tâche spécifique.

3. Discussion : Décrypter les défis cachés de l'IA face aux dialectes algériens

Les résultats quantitatifs et qualitatifs que nous avons obtenus mettent en évidence des écarts notables dans les performances des systèmes de reconnaissance vocale appliqués à l'arabe algérien. Le Word Error Rate (WER) global varie de 24,5 % à 35,2 %, tandis que l'exactitude de classification dialectale ne dépasse pas 68 %. Ces limites, loin d'être anecdotiques, s'inscrivent dans un paysage plus vaste de défis structurels propres aux langues dialectales. Dans le cas de l'arabe algérien, nous faisons face à une triple contrainte : une variabilité régionale prononcée, un code-switching fréquent avec le français, et l'absence d'une norme orthographique stabilisée. Au-delà des chiffres, ces écarts révèlent comment l'IA, malgré ses progrès, reproduit des biais systémiques liés au caractère faiblement doté en ressources de nos dialectes. Nous présentons ci-dessous les raisons techniques et sociolinguistiques qui éclairent ces difficultés, en mobilisant des analyses issues de la littérature récente sur l'ASR en arabe.

3.1. Origines techniques des écarts : entre rareté des corpus et absence de norme

Les limites techniques constituent un obstacle majeur. Tout d'abord, la rareté et le manque de diversité des corpus d'entraînement pour les dialectes algériens réduisent la capacité des modèles à capter la richesse phonétique des variantes régionales (Est, Ouest, Sud). Alors que l'anglais ou l'arabe standard moderne (MSA) disposent de millions d'heures annotées, nos dialectes atteignent rarement ce volume et sont souvent biaisés vers des accents urbains (tels que l'algérois). Les travaux récents montrent que des approches de pré-entraînement multicorpus et de fine-tuning sur données dialectales permettent des gains significatifs, mais exigent tout de même des ressources dédiées pour les variantes rurales et sahariennes (Talafha *et al.*, 2023 ; Yunpeng *et al.*, 2024).

Nous constatons que l'absence de standardisation orthographique aggrave les substitutions observées (45 % de nos erreurs). La littérature souligne le rôle néfaste de la romanisation hétérogène et de l'absence de conventions de transcription pour l'émergence d'erreurs lexicales et prosodiques dans les systèmes ASR. De plus, le code-switching franco-arabe - très répandu en Algérie - augmente la fréquence d'omissions et d'insertions si les modèles n'ont pas été entraînés explicitement sur des corpus mixtes (Talafha *et al.*, 2023 ; Yunpeng *et al.*, 2024).

Enfin, le déficit en données annotées est un frein majeur : l'annotation manuelle, coûteuse et exigeante en expertise dialectologique, limite les initiatives à grande échelle. Des travaux récents explorent des stratégies de supervision faible (weakly

supervised / pseudo-labelling¹⁵) et montrent qu'avec plusieurs milliers d'heures faiblement annotées, on peut considérablement améliorer les performances, notamment en combinant architectures Conformer et apprentissage faiblement supervisé (Bhogale *et al.*, 2024).

3.2. Facteurs sociolinguistiques : quand l'invisibilité institutionnelle façonne la technologie

Les dynamiques sociolinguistiques jouent un rôle déterminant. En Algérie, la langue institutionnelle demeure le MSA, tandis que la *darija* est généralement confinée à l'oralité informelle ; il en résulte une sous-représentation systématique des dialectes dans les ressources linguistiques utilisées pour entraîner l'IA. Cette invisibilité institutionnelle engendre des reconfigurations hiérarchiques au sein des jeux de données : les accents urbains sont sur-représentés et mieux reconnus que les variantes rurales, phénomène observé dans nos mesures par des WER supérieurs de l'ordre de 10-15 % pour le Sud et l'Ouest (Hussein *et al.*, 2022).

La diglossie arabe, avec la valorisation du MSA, conduit à une priorité de collecte et d'annotation des formes «hautes»; or cette hiérarchisation sociolinguistique se traduit en biais algorithmiques dès l'étape d'entraînement. Pour corriger ces effets, les études recommandent d'intégrer des démarches communautaires et participatives lors de la constitution de corpus (Hussein *et al.*, 2022).

4. Solutions émergentes : transformer les défis en leviers d'innovation

4.1. Identification dialectale hiérarchique et caractéristiques prosodiques

Dans notre approche, l'identification dialectale hiérarchique (approches prosodiques hiérarchisées) nous apparaît comme un outil efficace pour réduire les confusions interrégionales. Des travaux proposant des architectures hiérarchiques ou multimodales qui priorisent des marqueurs prosodiques (intonation, rythme, allongements, présence de gutturales) montrent des gains de précision substantiels par rapport à des classifieurs plats (Dhouib *et al.*, 2022).

4.2. Fine-tuning de modèles larges, apprentissage faiblement supervisé et pré-entraînement mixte

Le fine-tuning de modèles pré-entraînés multilingues (notamment Whisper) sur des corpus dialectaux réduit fréquemment le WER de manière notable. Nous constatons, à la lumière des études expérimentales récentes, des améliorations de l'ordre de 20-30 %

¹⁵ Étiquetage automatique : annotation automatique de données non étiquetées pour l'entraînement de modèles.

dans des scénarios faiblement dotés en ressources lorsqu'un modèle généraliste est adapté sur des données ciblées (Yunpeng *et al.*, 2024 ; Bhogale *et al.*, 2024).

4.3. Corpus communautaires et annotation collaborative : Casablanca & Sawt El-Djazair

Des initiatives communautaires récentes fournissent des modèles exemplaires : *Casablanca* est un dataset¹⁶ multidialectal de grande échelle, collecté et annoté de manière communautaire pour plusieurs dialectes arabes (y compris l'algérien) (Talafta *et al.*, 2024). Nous nous intéressons également à *Sawt El-Djazair*, qui illustre un corpus algérien orienté vers des usages forensiques et la variabilité de session (Zergat *et al.*, 2023).

5. Perspectives

En nous projetant vers l'avenir, nous envisageons des perspectives stratégiques pour surmonter les limitations actuelles des systèmes ASR face aux dialectes algériens. Ces orientations, ancrées dans nos résultats et les apports récents, visent à renforcer l'équité linguistique en promouvant des ressources durables et des collaborations inclusives (Dhouib *et al.*, 2022). Nous structurons ces perspectives autour de la constitution de corpus, de l'annotation collaborative, des modèles adaptatifs et des partenariats interdisciplinaires, en intégrant des initiatives comme *Sawt El-Djazair* et *Casablanca* pour enrichir notre écosystème (Mokeddem *et al.*, 2023 ; Zergat *et al.*, 2023).

La constitution de corpus représentatifs et équilibrés est primordiale pour corriger les biais observés dans nos évaluations, où les variants Ouest et Sud sont sous-représentés. Nous proposons de développer ou étendre des corpus comme *Sawt El-Djazair* et *Casablanca*. Idéalement crowdsourcés, ces corpus devraient couvrir plusieurs régions algériennes (Est, Centre, Ouest, Sud) et registres (formels/informels), en annotant non seulement les transcriptions mais aussi les traits prosodiques, linguistiques et instances de code-switching, pour mieux capturer la dynamique de notre dialecte (Talafta *et al.*, 2023).

L'annotation collaborative via crowdsourcing offre une voie démocratisée pour enrichir ces ressources, en impliquant des locuteurs natifs dans la labellisation de données audio issues de YouTube ou réseaux sociaux. Cette approche, inspirée de projets comme *Casablanca* (Talafta *et al.*, 2024), permet de générer des annotations riches même avec des budgets limités, en réduisant la dépendance à des experts rares et en favorisant une représentativité sociolinguistique authentique (Hussein *et al.*, 2022).

¹⁶ Jeu de données : collection organisée d'exemples ou d'échantillons servant à entraîner, valider ou tester un modèle.

Les modèles multilingues adaptatifs, capables de gérer le code-switching, constituent une perspective technique clé. Nous préconisons un pré-entraînement sur MSA et les dialectes, suivi d'un affinage sur données algériennes spécifiques, en explorant des architectures comme Conformer weakly supervised pour maximiser l'utilisation de données disponibles (Bhogale *et al.*, 2024 ; Yunpeng *et al.*, 2024).

Intégrer des systèmes de classification dialectale en amont, inspirés de HADID (Talaflha *et al.*, 2023), permettrait des interfaces adaptatives qui détectent et ajustent au dialecte en temps réel, améliorant ainsi la robustesse face aux mélanges franco-arabes observés dans 60 % de nos erreurs.

Enfin, la collaboration entre linguistes, informaticiens et communautés locales est essentielle pour développer des ressources respectueuses des réalités sociolinguistiques (Dhouib *et al.*, 2022). En impliquant des acteurs locaux dans la collecte et la validation, nous pouvons déconstruire les hiérarchies linguistiques et promouvoir une IA inclusive, alignée avec l'invisibilité institutionnelle de nos dialectes. Ces perspectives, en somme, préparent la voie vers une reconnaissance vocale qui honore la diversité de notre patrimoine algérien, en transformant les défis actuels en innovations durables.

Conclusion

Au terme de notre étude, nous constatons que les dialectes algériens demeurent structurellement sous-représentés dans les ressources, outils et modèles de reconnaissance automatique de la parole. Ce déficit de représentativité ne résulte pas uniquement de la complexité interne de ces variétés linguistiques - marquées par une diversité phonétique, prosodique et lexicale considérable - mais aussi de leur marginalisation historique dans les politiques linguistiques et technologiques. Ce constat soulève un enjeu fondamental : si nous n'agissons pas rapidement, une part significative de notre patrimoine linguistique risque d'être sous-documentée, voire invisibilisée, dans les écosystèmes numériques émergents.

Pour autant, notre analyse révèle également que les progrès récents (2024-2025) dans le domaine - apprentissage faiblement supervisé, pré-entraînement multi-dialectal, fine-tuning adaptatif - constituent une fenêtre d'opportunité inédite. Ces avancées permettent de dépasser les limites des approches strictement supervisées, souvent dépendantes de grands corpus homogènes, et ouvrent la voie à des modèles plus robustes, capables d'intégrer la variabilité inhérente aux dialectes algériens. Leur mise en œuvre concertée, alliée à un effort national et international de constitution de corpus diversifiés, pourrait transformer notre rapport technologique à la langue : d'un simple outil de transcription, l'ASR deviendrait un instrument de préservation, de diffusion et de revalorisation de nos parlers.

En définitive, si les défis que nous avons identifiés - chevauchement dialectal, code-switching, lacunes phonétiques dans les représentations - sont abordés avec méthode et coopération, l'Intelligence Artificielle pourra jouer un rôle déterminant dans la

reconnaissance et la valorisation des dialectes algériens. Nous avons l'opportunité non seulement d'améliorer la performance des systèmes ASR, mais aussi de réaffirmer, à travers eux, la place légitime de nos dialectes dans le patrimoine culturel et scientifique mondial.

Liste des références

- ABDELALI A. et al., (2024), «*Larabench: Benchmarking arabic ai with large language models*», in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 487-520.
- ABAINIA K., (2020), «DZDC12: a new multipurpose parallel Algerian Arabizi–French code-switched corpus», *Language Resources and Evaluation*, 54(2), pp. 419-455.
- ALSAYADI H. A. et al., (2022), «Deep investigation of the recent advances in dialectal arabic speech recognition», *IEEE access*, vol. 10, pp. 57 063–57 079.
- ALSHARHAN E., RAMSAY A., (2020), «Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition», *Lang Resources & Evaluation*, 54, pp. 975–998.
- BHOGALE K.S. et al., (2024), «Empowering Low-Resource Language ASR via Large-Scale Pseudo Labeling», *arXiv preprint*, arXiv:2408.14026.
- DHOUIB Amira et al., (2022), «Arabic Automatic Speech Recognition: A Systematic Literature Review», *Applied Sciences*. 12. 8898. 10.3390/app12178898.
- DJANIBEKOV Amirbek et al., (2025), «*Dialectal Coverage And Generalization in Arabic Speech Recognition*», in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29490–29502, Vienna, Austria. Association for Computational Linguistics.
- ELNAGAR A. et al., (2021), «Systematic literature review of dialectal Arabic: identification and detection», *IEEE Access*, 9, pp.31010-31042.
- HAMED Injy et al., (2025), «*A Survey of Code-switched Arabic NLP: Progress, Challenges, and Future Directions*», in *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4561–4585, Abu Dhabi, UAE. Association for Computational Linguistics.
- HUSSEIN A., WATANABE S. et ALI A., (2022), «Reconnaissance vocale arabe par systèmes modulaires de bout en bout et humains», *Computer Speech & Language*, 71, 101272.
- MATTIODA M.M., (2024), «Intelligence artificielle et diversité linguistique: Quelle gestion équitable pour la garantie des droits linguistiques?», *Language Problems and Language Planning*, 48(2), pp.146-167.

- MESSAOUDI L., (2020), *Les Parlers arabes modernes du Maghreb*, ENS Éditions.
- MOKEDDEM A., ABDERRAHIM M., et BOUCHEMAL N., (2023), «The voice as a material clue: A new forensic Algerian corpus (Sawt El-Djazaïr)», *Language Resources and Evaluation*, 57, pp.1-21.
- ÖZYILMAZ Ö. T., COLER M., VALDENEGRO-TORO M., (2025), «Overcoming Data Scarcity in Multi-Dialectal Arabic ASR via Whisper Fine-Tuning», *arXiv preprint arXiv:2506.02627*.
- RADFORD A. *et al.*, (2023), «Robust Speech Recognition via Large-Scale Weak Supervision», in *International Conference on Machine Learning*. PMLR, pp. 28 492–28 518, iSSN: 2640-3498.
- RAHMAN A. *et al.*, (2024), «Arabic speech recognition: Advancement and challenges », *IEEE Access*, 12, pp. 39689-39716.
- ROUA-HAMDANI Ghania, SELOUANI Sid Ahmed et BOUDRAA Malika, (2010), «Algerian Arabic Speech Database (ALGASD): Corpus design and automatic speech recognition application», *Arabian Journal For Science And Engineering*, 35, pp. 157-166.
- SALHAB M., *et al.*, (2025), «Advancing Arabic Speech Recognition Through Large-Scale Weakly Supervised Learning», *arXiv preprint arXiv:2504.12254*.
- SHAALAN Khaled *et al.*, (2019), «Challenges in Arabic natural language processing», *Computational linguistics, speech and image processing for arabic language*, pp. 59-83.
- TALAFHA B., WAHEED A. *et* ABDUL-MAGEED M., (2023), «N-shot benchmarking of whisper on diverse arabic speech recognition», *arXiv preprint arXiv:2306.02902*.
- TALAFHA B. *et al.*, (2024), «Casablanca: Data and models for multidialectal arabic speech recognition», *arXiv preprint arXiv:2410.04527*.
- Juniper research, (2020), *Voice Assistant Market: Player Strategies, Monetisation & Market Size 2020-2024*. Disponible sur [\[https://www.juniperresearch.com/press/number-of-voice-assistant-devices-in-use/\]](https://www.juniperresearch.com/press/number-of-voice-assistant-devices-in-use/) (consulté le 12 /08/2025).
- YUNPENG Liu, XUKUI Yang *et* DAN Qu, (2024), «Exploration of Whisper fine-tuning strategies for low-resource ASR», *EURASIP Journal on Audio, Speech, and Music Processing*, 10.1186/s13636-024-00349-3.
- ZERGAT K.Y. *et al.*, (2023), «The voice as a material clue: a new forensic Algerian Corpus», *Multimedia Tools and Applications*, 82(19), pp.29095-29113.