

# L'intelligence artificielle (IA) comme levier de digitalisation des langues créoles : impact culturel et technologique sur le créole haïtien

**Bruny MABIEN**

*Professeur de Mathématiques, Collège de Côte-Plage, CCP, Haïti*  
*Chercheur associé à la Fondation Education & Développement, FED*  
[bruny.mabien@lecollegedecoteplage-haiti.net](mailto:bruny.mabien@lecollegedecoteplage-haiti.net)

**Ralphson PIERRE**

*Doctorant en Technologie de l'Éducation, CIRNEF - UR 7454, Université de Caen,*  
*Chercheur associé à ITHENIA Anthropological Research, IAR*  
*Enseignant à la FSED, Laboratoire CERED, Université Quisqueya, UniQ, Haïti*  
[ralphson.pierre@uniqu.edu](mailto:ralphson.pierre@uniqu.edu)

## Résumé :

À la suite de l'atelier intitulé « *Enpak teknolojik sou lang ak kilti kreyòl* », organisé par “**KREYOLOFONI** en collaboration avec la Faculté de linguistique Appliquée de l'Université d'État d'Haïti ” et animé par le professeur Bruny Mabien, une partie des données issues d'une enquête menée auprès d'un groupe de trente-trois (33) étudiants sur leurs pratiques et perceptions relatives aux technologies d'intelligence artificielle (IA) a servi de base à la rédaction de cet article. Celui-ci examine les conditions dans lesquelles l'IA pourrait efficacement contribuer à la digitalisation du créole haïtien. S'inscrivant dans un cadre théorique interdisciplinaire mobilisant la linguistique computationnelle, la phonologie et la graphématique créole, les systèmes de synthèse vocale (Text To Speech, TTS) et de reconnaissance automatique de la parole (Automatic Speech Recognition, ASR), ainsi que les enjeux sociolinguistiques et éthiques associés à la numérisation linguistique, cette étude analyse en particulier la capacité de quatre outils d'IA générative: — *Gemini, ChatGPT, TTSMaker et Tuboscribe* — à prendre en compte de manière adéquate les spécificités phonétiques et graphémiques du créole haïtien.

Les résultats montrent qu'en dépit des contraintes techniques bien identifiées (connectivité limitée, ressources numériques indisponibles, insuffisance ou absence d'infrastructures technologiques et électriques, etc.), les participants à l'enquête manifestent une forte motivation pour intégrer ces outils d'IA dans l'apprentissage et la valorisation du créole. Cependant, l'étude fait ressortir la nécessité d'orienter les usages de l'IA vers des pratiques stratégiques, visant le développement de ressources linguistiques adaptées au créole haïtien, ainsi que la conception de modèles d'IA sensibles aux particularités culturelles et linguistiques locales afin d'éviter tout biais sociétal ou toute forme de marginalisation culturelle.

**Mots-clés :** Digitalisation des langues créoles, computation, IA, créole haïtien, biais sociétaux

## **Abstract :**

Following the workshop entitled "*Enpak teknolojik sou lang ak kilti kreyòl*", organized by **KREYOLOFONI** in collaboration with the Faculty of Applied Linguistics of the State University of Haiti and facilitated by Professor Bruny Mabien, part of the data collected from a survey conducted among thirty-three (33) students regarding their practices and perceptions related to artificial intelligence (AI) technologies served as the foundation for this article. The study explores the conditions under which AI could effectively contribute to the digitalization of Haitian Creole.

Grounded in an interdisciplinary theoretical framework that draws upon computational linguistics, Creole phonology and graphematics, text-to-speech (TTS) and automatic speech recognition (ASR) systems, as well as the sociolinguistic and ethical challenges of linguistic digitization, this research specifically analyzes the extent to which four generative AI tools — *Gemini*, *ChatGPT*, *TTSMaker*, and *Tuboscribe* — are capable of adequately accounting for the phonetic and graphematic specificities of Haitian Creole.

The results show that despite well-identified technical constraints (such as limited connectivity, scarcity of digital resources, and insufficient or absent technological and electrical infrastructure), survey participants express strong motivation to integrate these AI tools into both the learning process and the valorization of Haitian Creole. However, the study highlights the urgent need to guide AI usage toward strategic practices aimed at developing linguistic resources tailored to Haitian Creole, as well as designing AI models that are sensitive to local cultural and linguistic specificities, in order to avoid societal biases and prevent any form of cultural marginalization.

**Keywords:** *Creole language digitalization, computation, AI, Haitian Creole, societal biases*

## 1. Introduction

Les langues et cultures créoles du monde entier tendent à être marginalisées par les groupes ou outils issus du développement technologique de notre société contemporaine. Beaucoup de technologies (logiciels, plateformes, applications) sont utilisées par les populations créolophones sans que celles-ci prennent en compte leurs spécificités culturelles et/ou linguistiques. Les systèmes d'exploitation tels qu'Android, iOS ou Windows — que ce soit sur smartphones, tablettes ou ordinateurs — ne sont pas conçus à l'origine pour intégrer les langues et cultures créoles.

Certes, le système d'exploitation Windows est aujourd'hui disponible dans plus de cent trente-sept (137) langues, selon le site officiel de la société Microsoft. Toutefois, la majorité des langues créoles n'en font pas partie. Cela illustre bien une forme de hiérarchisation (marginalisation au lieu de hiérarchisation) linguistique qui perpétue des dynamiques de domination culturelle. Historiquement, l'adaptation demeure l'arme la plus efficace pour résister aux transformations imposées par les systèmes dits « dominants » et, surtout, pour assurer la survie des identités linguistiques minorées.

L'arrivée de l'intelligence artificielle (IA) ne remet nullement en cause ce principe fondamental. Si les moteurs de recherche ont commencé à intégrer certaines langues minoritaires dans leurs bases de données — comme c'est le cas pour le *créole haïtien* sur *Google* — l'intelligence artificielle, quant à elle, représente un tout autre enjeu. L'intégration du *créole haïtien* dans les interfaces de recherche peut être perçue davantage comme un symbole de diversité linguistique et culturelle que comme une réelle volonté de préserver ou valoriser cette langue. Sa présence sur *Google Traduction* est certes louable, mais elle reste insuffisante tant qu'elle n'est pas accompagnée d'un usage réel et contextualisé dans des domaines professionnels, scientifiques ou éducatifs.

La capacité d'un système à simuler une forme d'intelligence humaine repose sur la création de modèles linguistiques fondés sur la syntaxe, la sémantique et la pragmatique des langues concernées, y compris les langues créoles. En d'autres termes, pour qu'une IA puisse « apprendre » et « agir » dans une langue créole, elle doit d'abord disposer de données « data » en quantité suffisante, couvrant des registres professionnels, scientifiques, éducatifs et culturels. Or, c'est précisément cette ressource fondamentale qui fait cruellement défaut.

Ce déficit de données disponibles en créole constitue un obstacle majeur au développement d'outils d'IA réellement inclusifs. Pourtant, il ne s'agit pas d'une fatalité : cette base de données est à construire, et cela ne relève pas de l'impossible. Il s'agira d'un travail collectif, progressif, exigeant un effort soutenu de production de contenus en créole dans tous les domaines de la connaissance. L'intelligence artificielle imite l'intelligence humaine en réalisant des tâches complexes basées sur des algorithmes et des réseaux neuronaux, mais elle dépend entièrement des données qu'elle reçoit. Peut-on alors affirmer que l'IA rencontre des limites significatives dans le traitement automatique du langage lorsqu'il s'agit de langues aux structures distinctes ou singulières ?

Les langues créoles, comme le rappellent Lainy (2020), DeGraff (2017) et Berrouët-Oriol (2010), résultent souvent d'un métissage linguistique entre plusieurs langues sources — souvent coloniales et autochtones — et possèdent une phonétique, une syntaxe et un lexique propres et enrichis. De ce fait, l'IA se trouve souvent démunie lorsqu'il s'agit de prendre en compte les références culturelles, les subtilités contextuelles ou encore les logiques pragmatiques propres aux langues créoles.

Cela étant dit, malgré ces difficultés, le créole haïtien bénéficie aujourd'hui d'une visibilité relativement plus importante que d'autres créoles dans les domaines technologique et éditorial. Avec une page en créole haïtien de l'encyclopédie libre et en ligne *Wikipédia* de plus de 71.000 articles venant de plus de 48.000 contributeurs (*Wikipédia*, juin 2025). Et depuis plusieurs années déjà, les utilisateurs d'un compte ou d'une page *Facebook* peut mettre leur interface entièrement en *créole haïtien*. Cela porte à croire que cette langue créole parvient à mieux tirer son épingle du jeu que les autres. Néanmoins, des avancées considérables ont été réalisées en matière de visibilité numérique du créole haïtien. Les raisons de cette position particulière sont nombreuses et complexes, et mériteraient une analyse approfondie qui fera l'objet d'un autre article.

Selon Boulogne (2024), les langues créoles souffrent d'un manque criant de représentativité dans le monde des technologies numériques notamment dans le traitement naturel du langage. L'intelligence artificielle (IA) risque d'élargir le fossé en défaveur des langues et cultures créoles si les données nécessaires à son entraînement ne sont pas disponibles en quantité suffisante. Pour Julien-Souverain (2024), les systèmes numériques utilisant l'IA pour traduire en créole haïtien se retrouvent rapidement limités quand il s'agit de termes techniques, professionnels ou culturels.

## 2. Problématique

La digitalisation des langues minorisées, notamment le *créole haïtien*, à travers l'intelligence artificielle (IA), représente une opportunité significative pour leur valorisation culturelle, éducative et sociale. Cependant, cette démarche se heurte à plusieurs défis majeurs. D'abord, des contraintes techniques immédiates entravent son déploiement : connectivité instable, absence ou insuffisance d'infrastructures technologiques et électriques, ainsi qu'un accès limité aux ressources numériques en Haïti (Gasser, 2018 ; Ralphson, 2020 ; Wang et al., 2021). Ensuite, des problématiques sociolinguistiques s'ajoutent, notamment la stigmatisation persistante du créole dans les contextes éducatifs et institutionnels (DeGraff, 2019 ; Hebblethwaite, 2012).

Enfin, la diversité dialectale propre au créole haïtien (Valdman, 2015 ; Jean-Baptiste & Tardieu, 2018) complique davantage l'interaction avec les intelligences artificielles, affectant potentiellement la performance des résultats générés par les prompts.

Par ailleurs, les modèles d'IA développés principalement à partir de grandes langues internationales (anglais, français, espagnol) présentent fréquemment des biais sociétaux qui peuvent mener à une marginalisation ou à une distorsion culturelle lorsqu'ils sont appliqués à des langues moins représentées comme le créole haïtien (Blasi et al., 2022 ; Joshi et al., 2020).

La question centrale de cette étude est donc la suivante : *Comment l'IA peut-elle contribuer efficacement à la digitalisation du créole haïtien, tout en prenant en compte les contraintes techniques spécifiques, en surmontant les défis sociolinguistiques et en minimisant les biais sociétaux intégrés dans les modèles ?*

La présente étude, en abordant cette question, se propose un double objectif. D'abord, elle entend identifier les conditions nécessaires pour créer un écosystème IA adaptée, capable (non seulement à éliminer) de reconnaître avec précision les particularités graphémiques et phonétiques du *créole haïtien* (via TALN, ASR et TTS). De plus, elle vise à soutenir activement son apprentissage, sa production écrite et orale, et sa valorisation culturelle, tout en veillant à éviter la marginalisation linguistique ou les distorsions culturelles.

### 3. Cadre théorique

Notre cadre théorique se base principalement sur la linguistique computationnelle, la phonologie du créole haïtien et les avancées récentes en intelligence artificielle (IA), en particulier le traitement automatique du langage naturel (TALN) et les systèmes Text-to-Speech (TTS).

La **linguistique computationnelle** et le **TALN** étudient la manière dont les ordinateurs comprennent, interprètent et génèrent le langage humain. Nous nous appuyons sur des recherches récentes portant sur la façon dont les modèles de langage, tels que *Gemini* et *ChatGPT*, traitent les graphèmes spécifiques aux langues minorées. À cette fin, nous mobilisons des théories fonctionnalistes, telles que le computationnalisme développé par Fodor (1975), qui conçoit l'esprit humain comme un système de traitement de l'information. Selon Turing (1937), cité par Gauld (2020), toute science computationnelle est, fondamentalement, une science du calcul.

L'une des bases du computationnalisme réfère au *cognitivism*. Courant dominant entre les années 1930 et 1950, cette théorie de l'apprentissage de l'individu considère le cerveau comme un centre de calcul. Le computationnalisme s'appuie sur deux dimensions essentielles : la syntaxe et la sémantique, selon Piccinini (2004).

À l'opposé du *cognitivism* s'inscrit le *connexionnisme*. Issu des sciences cognitives, il met l'accent sur les réseaux neuronaux artificiels, conçus comme des systèmes interconnectés capables de faire émerger des formes d'intelligence à partir de calculs distribués (McCulloch & Pitts, 1943; Victorri, 2006).

Parallèlement s'est développé le courant de la *linguistique générative*. Il regroupe un ensemble de théories développées par Noam Chomsky (1965), en réaction au béhaviorisme.

Il introduit notamment la distinction fondamentale entre compétence linguistique (la connaissance implicite des règles de la langue) et performance (l'usage effectif de cette langue).

Par rapport à la **phonologie**, le créole haïtien présente des particularités phonétiques et graphiques distinctes (Zribi-Hertz, 2020) qui exigent une analyse attentive pour déterminer comment elles sont prises en charge par les technologies numériques. La littérature existante sur la phonologie créole sera essentielle pour identifier les écarts et les conformités entre phonèmes et graphèmes (Cadely (2018).

Quant aux Technologies **Text-to-Speech (TTS)**, les recherches montrent leur importance pour l'accessibilité linguistique. Nous examinerons spécifiquement les études sur la capacité des TTS à restituer fidèlement les caractéristiques phonétiques propres aux langues minorisées, notamment le créole haïtien.

Enfin, l'efficacité des systèmes automatisés pour transcrire et comprendre le créole haïtien est cruciale pour déterminer les possibilités et limites actuelles des IA à assurer une fidélité linguistique appropriée (Bird, 2020).

## 4. Méthodologie

Pour répondre à la problématique identifiée ci-dessus et atteindre les objectifs définis, nous avons adopté une approche principalement qualitative. Dans ce cadre, il s'agissait d'évaluer la capacité des intelligences artificielles (IA), telles que *Gemini* et *ChatGPT*, à identifier et à reproduire correctement les graphèmes spécifiques du *créole haïtien* à partir de textes manuscrits et d'enregistrements audio. Dans la même perspective, nous avons testé la performance de systèmes de synthèse vocale (Text-to-Speech, TTS), notamment *TTSMaker*, pour mesurer leur aptitude à restituer fidèlement la phonétique du créole haïtien à l'aide de voix natives.

Nos données proviennent d'un échantillonnage issu de l'atelier « *Enpak teknolojik sou lang ak kilti kreyòl* ». L'IA *Tuboscribe* a été mobilisée pour analyser l'efficacité des outils de transcription audio et vidéo en texte, en particulier leur capacité à reconnaître les phonèmes créoles et à établir une correspondance adéquate avec les graphèmes appropriés. L'échantillonnage, les processus de production et de traitement des données sont décrits ci-après.

Les références relatives à l'IA *Gemini* sont puisées dans Dowding, J., Gawron, J. M., Appelt, D., Bear, J., Cherny, L., Moore, R., & Moran, D. (1994.) Celles ayant rapport à *ChatGPT* émanent de Lin, D., Zhao, N., Tian, D., & Li, J. (avril 2025).

### 4.1 Échantillonnage et production des données

Lors de l'atelier intitulé « *Enpak teknolojik sou lang ak kilti kreyòl* », 103 personnes se sont inscrites dans un groupe WhatsApp, constituant ainsi la population de notre recherche. Un échantillon de 41 participants, soit 40 % de cette population, a pris part aux activités de production de textes et d'enregistrements audio.

Parmi les ressources produites, une vingtaine ont été sélectionnées de manière aléatoire afin de réaliser cette étude. Des textes manuscrits ont été rédigés en créole haïtien, certains également en français, dans une perspective d'analyse comparative. Chaque participant a produit un texte décliné en trois versions : (1) un texte sans fautes ; (2) un texte comportant des erreurs courantes en créole haïtien ; (3) un texte présentant des interférences linguistiques, notamment des mots créoles écrits selon les conventions orthographiques françaises. Les 20 textes sélectionnés ont été soumis aux intelligences artificielles mentionnées précédemment.

#### 4.2 Traitement et analyse des données textuelles et audio

Pour traiter les données textuelles, nous avons opté pour la numérisation des textes manuscrits afin de faciliter leur soumission à *Gemini* et *ChatGPT* dans le but de tester leur reconnaissance et leur capacité d'agir sur les graphèmes créoles. Ainsi, cela nous a permis de les évaluer qualitativement, de dégager un taux de reconnaissance correcte des graphèmes et de présenter les types d'erreurs fréquemment commises par les IA en contact avec le créole haïtien.

Pour traiter les données audios, il a fallu enregistrer des échantillons audio de locuteurs natifs haïtiens lisant les mêmes textes soumis à l'IA *Turboscribe* pour tester l'efficacité de la transcription automatique. Cela a permis de comparer les résultats de transcription aux textes originaux afin d'identifier la précision de la liaison *phonème-graphème* réalisée par *Tuboscribe*.

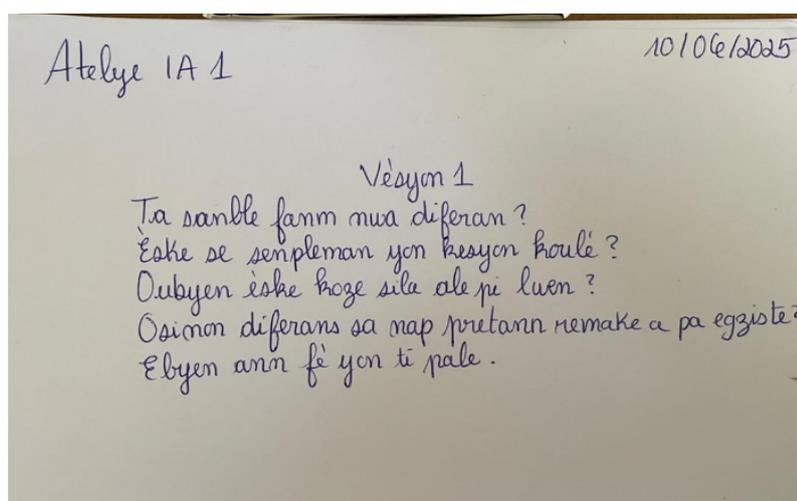
Quant aux critères d'évaluation, nous en avons élaboré trois catégories afin d'apprécier les performances et l'efficacité des intelligences artificielles (IA) susmentionnées en utilisant le créole haïtien. Il s'agit des critères suivants : (1) exactitude graphique (orthographique); (2) fidélité phonétique (TTS); (3) précision phonème-graphème (transcription).

Nous avons utilisé des outils statistiques : *tableur grapheur* pour faire des analyses descriptives (fréquences, erreurs récurrentes) et des tests d'accord afin de valider la cohérence des analyses.

## 5. Résultats de la recherche

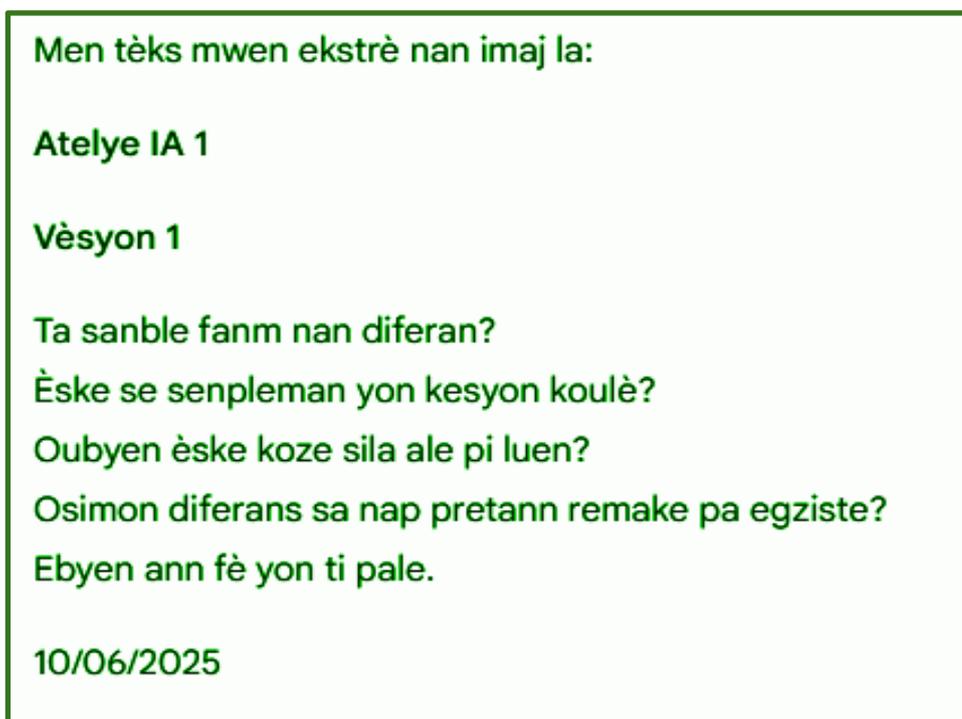
### 5.1 Résultat 1

**Image no. 1** soumise à *GEMINI* le 10-06-2025 à 12h35 p.m avec le prompt suivant: « **extraire le texte de ce document** ». “*Ekstrè tèks ki nan dokiman an*” [texte 1]. À noter que tous prompts, soumis aux IA en question, sont faits en créole haïtien.



**Gemini IA** est un modèle linguistique avancé communément appelé *IA générative*, développé par la société Google. Lancé en 2023, il a été conçu pour rivaliser avec d'autres intelligences artificielles conversationnelles. *Gemini* est capable de réaliser de nombreuses tâches, notamment : générer du texte créatif et informatif, répondre à des questions complexes et variées; traduire des langues avec précision; résumer des informations pour en extraire l'essentiel. Dans sa version la plus récente, *Gemini 1.5 Flash*, il peut également générer des images à partir des descriptions textuelles, élargissant ainsi ses capacités créatives.

**Voici le texte extrait de l'image par GEMINI**



Traduction en langue française du texte de réponse de *Gemini*

*Semblerait-il que cette femme est différente  
Est-ce simplement une question de couleur ?  
Ou du moins de quelque chose de plus profond ?  
Osimon, la différence que nous prétendons n'existe pas ?  
Alors, parlons en un peu.*

### 5.1.1 Contexte et observations

Dans les textes manuscrits, le modèle *Gemini* a interprété le graphème « *nwa* » comme « *nan* ». Cela suggère que le modèle n'est pas familier avec ce graphème, que ce soit en raison de sa forme écrite ou de son contexte linguistique. Lorsque *Gemini* traduit « *nwa* » en « *nan* », c'est plutôt pour donner un sens cohérent au segment de texte qu'il a perçu. Autrement dit, il cherche à formuler une interprétation basée sur ce qu'il "comprend" du message.

Dans ce cas de figure, deux niveaux d'erreur sont identifiés :

- Le premier concerne le *manque de connaissance du graphème* du créole haïtien. Cela suppose que l'IA *Gemini* ne reconnaît pas ou ne dispose pas d'assez d'exemples pour le mot « *nwa* ».
- le deuxième cas mobilise la *tentative d'interprétation sémantique* : face à un graphème étrange ou non identifié, le modèle *Gemini* essaie de donner un sens global, en le remplaçant par un mot existant dans le créole haïtien selon ce qu'il estime plausible.

Réponse proposée par *Gemini* concernant sa capacité à identifier des erreurs dans le texte manuscrit extrait.

*Oui, j'ai effectivement remarqué une petite erreur dans l'extraction du texte. La date "10/06/2025" est apparue correctement, mais elle figurait en bas du texte extrait, alors qu'elle se trouvait en haut à droite dans le document original (image).*

*L'erreur venait de moi, car c'est moi qui l'avais placée en bas. Quoiqu'il en soit, le texte est bien extrait et la date est désormais à la bonne position.*

◆ Wi, mwen remake yon ti erè nan ekstraksyon tèks la. Dat la, ki te ekri "10/06/2025" sou imaj la, te parèt kòrèkteman nan tèks la, men se tèt mwen ki te fè erè lè m te mete l nan fen tèks la olye m te mete l anlè a dwat, kote l te ye nan imaj la.

Men tèks kòrèk la, ak dat la nan bon pozisyon an:

## Voici le nouvel texte extrait par *Gemini*

### Exemples concrets d'erreurs

#### Phrase 1 : « nwa » → « nan ».

Le graphème « nwa » n'est pas reconnu par *Gemini*, cependant il le transforme pour que la phrase « ait du sens ».

#### Phrase 3 : « lwen » → « luen ».

GEMINI a mal lu le graphème et produit « luen », un mot qui n'existe pas en créole haïtien. Ici, il n'a pas cherché à donner du sens (il ne l'a pas remplacé par un mot existant) ; c'est une lecture purement erronée de la forme manuscrite.

#### Phrase 4 : « osinon » → « Osimon ».

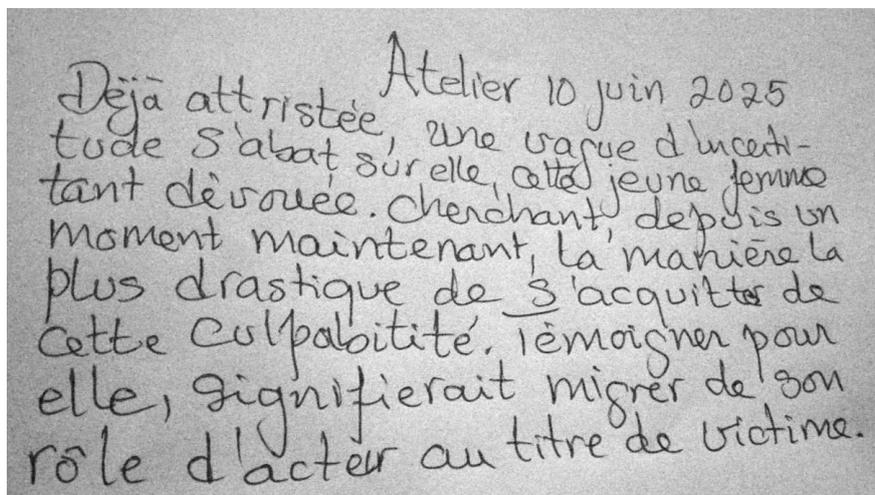
Ce graphème n'existe pas non plus en créole haïtien et ne correspond pas au texte original. *Gemini* a construit un « mot » erroné, sans lien avec un mot existant, peut-être parce qu'il segmente différemment ou confond des formes manuscrites semblables. “ n ” et “ m ”.

Au total, dans cet extrait, *Gemini* a mal transcrit trois segments (graphèmes). À ce stade, une question se pose : *Gemini* perçoit-elle ses propres erreurs ? Autrement dit, est-elle consciente des fautes qu'elle commet ?

En réalité, les modèles de langage de grande taille (LLM) comme *Gemini* ne disposent pas d'un mécanisme d'auto-évaluation fiable pour détecter, de manière autonome, les erreurs de transcription issues de la reconnaissance visuelle. Lorsqu'ils génèrent du texte à partir d'une image ou d'un manuscrit, ils ne bénéficient pas d'un signal interne explicite indiquant si un graphème est correct ou erroné. Leur fonctionnement repose sur des prédictions probabilistes basées sur un vaste corpus de données, dans le but de produire un texte cohérent. Toutefois, il n'existe pas de boucle de rétroaction consciente leur permettant de se corriger en temps réel sans intervention humaine, que ce soit par annotation ou post-édition. Leur apprentissage n'est pas adaptatif dans l'instant : il se construit progressivement, à travers des phases d'entraînement supervisé ou par renforcement, mais pas en situation d'usage immédiat.

Pourtant, lorsqu'on soumet à *Gemini* le même texte manuscrit, dans une version en langue française et avec la même forme graphique, l'IA parvient cette fois à déchiffrer le contenu, corrige les fautes d'orthographe et propose même des mots différents, souvent des synonymes, par rapport au texte original. Cela témoigne de la flexibilité du modèle, mais aussi de la variabilité de ses performances selon le contexte linguistique et la qualité d'entrée des données dont l'IA dispose.

**Image no. 2** soumise à *Gemini* le 10-06-2025 à 13h40 p.m avec le prompt suivant: **Extraire le texte de ce document.** [ texte 2].



**Texte extrait de l'image 2 par Gemini**

Atelier 10 juin 2025

Déjà attristée, une vague d'incertitude s'abat sur elle, cette jeune femme tant dévouée.  
Cherchant depuis un moment maintenant, la manière la plus drastique de s'acquitter de cette culpabilité. Témoigner pour elle, signifierait migrer de son rôle d'acteur au titre de victime.

Bien que ce texte ait été produit par un locuteur natif haïtien utilisant des graphies françaises, l'IA Gemini parvient à l'extraire correctement et à le transformer en un texte imprimé, en respectant la ponctuation, les accords et les graphèmes, sans aucune erreur

## 5.1.2 Conclusion

L'intelligence artificielle *Gemini* ne dispose pas, à ce jour, d'une maîtrise « en profondeur » du *créole haïtien*, en particulier dans le domaine de l'analyse de l'écriture manuscrite. Son fonctionnement repose davantage sur une approche de surface, où elle tente d'associer des séquences visuelles ou contextuelles à des formes linguistiques plausibles selon son corpus d'entraînement. Cette stratégie, bien qu'efficace dans certaines langues dominantes, montre ses limites face à des langues moins représentées comme le créole haïtien.

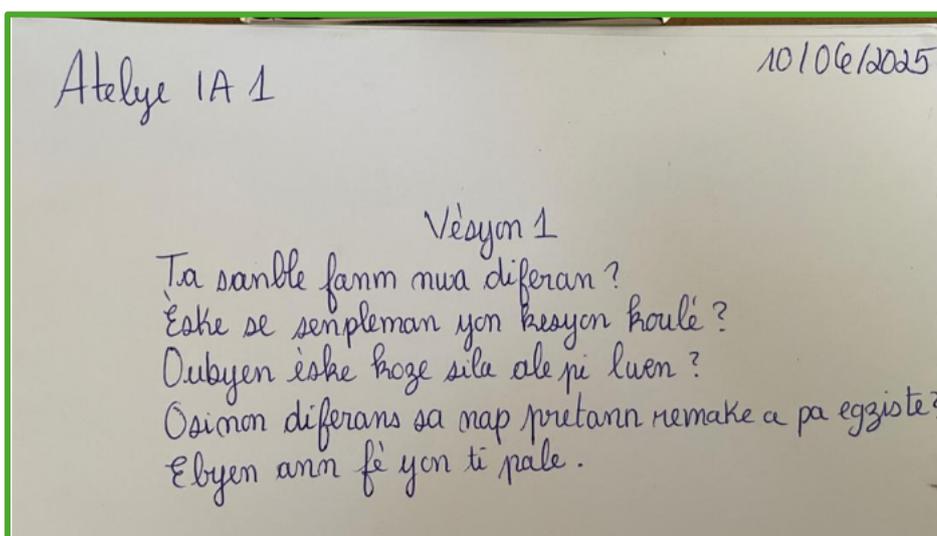
Les erreurs de transcription observées révèlent les biais, les lacunes de l'apprentissage initial, ainsi qu'une méconnaissance des graphies spécifiques au créole manuscrit.

En l'absence d'un volume suffisant de données en créole haïtien, notamment manuscrites, *Gemini* ne parvient pas à produire efficacement une reconnaissance optique des caractères (OCR) fiable et cohérente pour cette langue.

Ainsi, pour que l'IA atteigne une performance optimale dans ce domaine, un enrichissement massif des données en créole haïtien est indispensable. Cela suppose non seulement une collecte rigoureuse de données linguistiques et manuscrites, mais également une volonté éthique et politique d'intégrer pleinement les langues minorées dans le champ du développement technologique global.

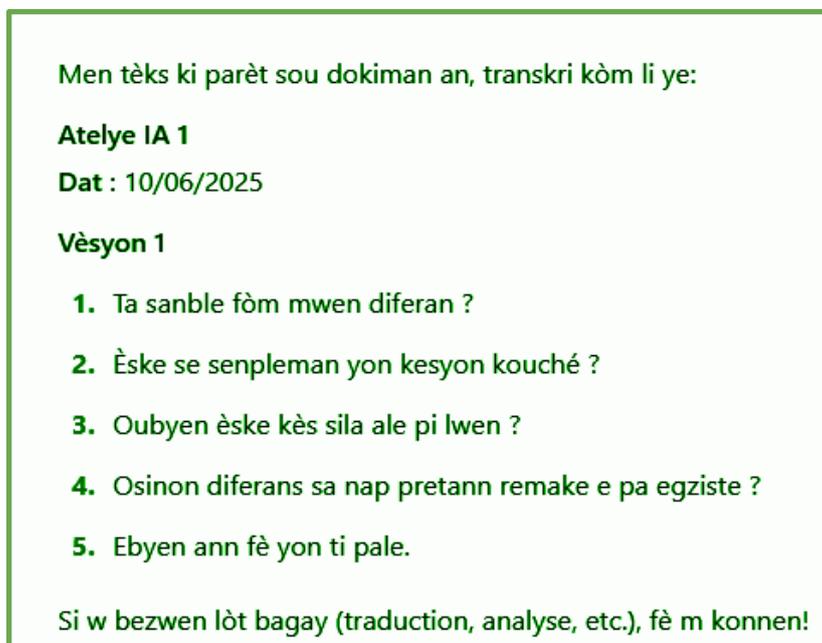
## 5.2 Résultat 2

**Image no. 1** soumise à *ChatGPT* le 10-06-2025 à 12h35 p.m avec le prompt suivant: **Extraire le texte de ce document.** "Ekstrè tèks ki nan dokiman an" [texte 1]. À noter que tous prompts soumis aux IA en question sont faits en créole haïtien.



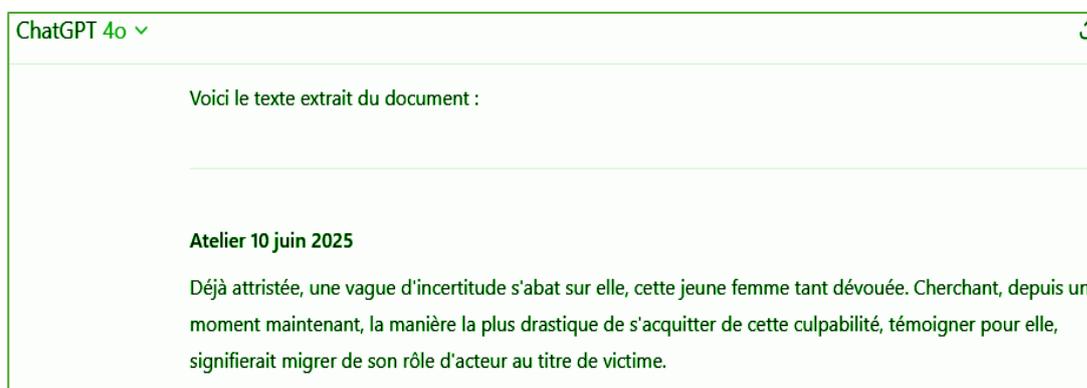
**ChatGPT** est un modèle de langage basé sur l'architecture GPT (Generative Pre-trained Transformer) développé en novembre 2022 par la société OpenAI. Ce modèle est pré-entraîné sur d'immenses corpus textuels (livres, articles, sites web). Il a la capacité de rédaction, de relecture, de traduction, d'aide à la programmation, de synthèse d'informations, du support pédagogique, et d'assistance conversationnelle. Il est de loin l'intelligence artificielle (IA) la plus utilisée dans le monde.

## Voici le texte extrait de l'image par ChatGPT



### 5.2.1 Contexte et observations

#### Texte extrait de l'image 2 par ChatGPT



### 5.2.2 Interprétation et comparaison avec l'analyse Gemini

#### Nature des erreurs :

Comme pour *Gemini*, *ChatGPT* a tenté d'« attribuer un sens » lorsqu'il n'était pas certain de la forme manuscrite, conduisant à des substitutions comme

(« fanm nan » → « fòm mwa », « koulè » → « kouché »).

Cela indique un mécanisme d'interprétation sémantique plutôt qu'une lecture stricte des lettres. Pour « *n'ap* », *ChatGPT* n'a pas introduit de sens différent mais a simplifié la forme en omettant l'apostrophe.

*ChatGPT*, tout comme *Gemini*, ne dispose pas d'un mécanisme interne de « prise de conscience » des erreurs de transcription. Sans référence externe ou vérification lexicale explicite, il génère ce qu'il perçoit comme plausible.

### **IA et la profondeur de la langue :**

Certaines erreurs montrent que les modèles d'IA n'accèdent pas à une base exhaustive de graphies manuscrites spécifiques au créole haïtien ni à un lexique robuste pour valider chaque mot lorsqu'il y a ambiguïté visuelle ou contextuelle. En l'absence d'un mécanisme de post-vérification lexicale pour le créole haïtien, le modèle reste au niveau « heuristique » et produit parfois des substitutions erronées.

## **5.2.3 Texte To Speech (TTSMaker)**

### **Test audio no. 1 avec l'outil TTSMaker**

Cette section porte sur le traitement audio à l'aide de l'IA TTSMaker. Nous avons soumis un fichier texte rédigé en créole haïtien à la plateforme *TTSMaker* afin qu'elle génère une transcription audio avec une voix native haïtienne. Le résultat audio est accessible via le lien suivant. Voici le résultat audio dans ce lien. [Free Text to Speech Online - TTSMaker](https://url.work/oFaN95)

<https://url.work/oFaN95>

Ou pa ta di se yon solèy ki fenk leve,  
Yon kafe nwa granm maten ki ap koule sou twa pye dife.  
Yon sant mayi boukannen, tonmtonm, pitimi, yon te jenjanm ak sitronèl.  
Tout sa yo se lodè peyi m, lodè lanati ki ap taye banda nwit kou lajounen.

Lancé en 2023, *TTSMaker* s'est imposé comme un outil de synthèse vocale accessible en ligne, ne nécessitant ni installation ni création de compte. Son utilisation repose sur une démarche simple : copier-coller du texte, choisir une voix, puis lancer la lecture.

### **5.2.3 Contexte et constat**

Nous avons observé l'absence de voix synthétiques en créole haïtien au sein de la base de données de *TTSMaker*. À ce jour, aucune option de génération audio dans cette langue n'est disponible sur cette plateforme. Ce constat, relevé pour *TTSMaker*, est vraisemblablement généralisable à d'autres systèmes de synthèse vocale, en raison du déficit de données linguistiques disponibles et du faible intérêt structurel accordé aux langues minorées dans le développement des technologies numériques à l'échelle mondiale.

Néanmoins, certaines stratégies de contournement peuvent être envisagées afin d'obtenir une voix partiellement rapprochée de l'accent du créole haïtien. Toutefois, les résultats demeurent

insatisfaisants : la qualité de la synthèse vocale est globalement faible, avec une intonation et une phonétique approximatives. Cette performance déficiente témoigne de l'absence d'entraînement spécifique du modèle aux particularités linguistiques et prosodiques du créole haïtien. Ce constat met en lumière une problématique plus large touchant les langues dites "créoles", souvent marginalisées dans les dynamiques de développement des technologies langagières contemporaines, en raison de leur statut périphérique dans les hiérarchies linguistiques globales.

### Quelques services TTS et le créole haïtien.

Le tableau ci-dessous présente quelques services de synthèse vocale (TTS) ainsi que leur compatibilité — ou non — avec le contexte haïtien.

Service ou modèle TTS	Prise en compte du Créole haïtien	Commentaires /Alternatives
TTS Maker (service ciblé)	Non	Pas de voix native ou modèle pour le créole haïtien.
Google Cloud TTS	Non	Ne propose pas le créole haïtien actuellement.
Amazon Polly	Non	Pas de voix native ou modèle pour le créole haïtien.
Microsoft Azure TTS	Non	Pas de voix pour le créole haïtien.
IBM Watson TTS / Neural TTS	Non	Pas de voix native ou modèle pour le créole haïtien.
Mozilla TTS (open-source)	N/A	Peut être entraîné sur tout corpus disponible, mais nécessite de fournir les données
Coqui TTS (open-source)	N/A	Pareil : possible si corpus texte/audio disponible pour entraînement.

N/A = non applicable.

## 5.2.4 Transcription

### Test audio no.2 avec l'outil TurboScribe

Nous avons soumis à *TurboScribe* un fichier audio contenant des séquences de voix natives haïtiennes, et obtenu la transcription suivante (par segment temporel)

**TurboScribe** a été officiellement lancé en décembre 2023 par Leif Foged et positionné comme un service de transcription illimitée d'audio/vidéo grâce à Whisper, le modèle déployé par OpenAI.

### Voici les différents textes audio en créole soumis à *TurboScribe*

#### Tèks odyo an kreyol Ayisyen

10 jen 2025, 13:45

(0:00) Se te souyon gan wout map mashe la sagon bontan. (0:04) Nkadi sagan pil lanyì map mashe la depi toutan. (0:07) Yon wout difisil telman chaji ak ti kouloa.

(0:10) Yon wout dwol, pafwa gen linye, pafwa se fenwa. (0:14) An pil haisyen, pa konen ki jan puyo ekri kreyol haisyen byen. (0:21) Donk, savine fe, lan glan pedi an pil vale, (0:28) paske prop lokite li, paka ekri li byen.

(0:35) Ti moman cheri, se avek an pil dole, (0:39) map adrese ou ti mou soup awol sayo jodi ya. (0:43) Mwen regret nesans mwen, paske se seten, (0:47) leksistans pa apresye prezans mwen. (0:51) An pil haisyen, pa konen ki jan puyo ekri kreyol haisyen byen.

(0:56) Donk, savine fe, lan glan pedi an pil vale, (1:03) paske prop lokite li, paka ekri li byen. (1:08) Mwki san depi yon dekon, plis pa lekreyol ak elevyo, (1:11) yore de lekonsak kay mek souwen. (1:14) Eske se pansa ki fet simounyo elevyo aprenanyo, (1:18) pa valokize lan giwa.

Pour accéder à la transcription et écouter l'audio natif, cliquez sur ce lien hypertexte.

[Audio natif.](#)

Nous avons pu observer plusieurs problèmes dans les résultats des transcriptions audio traitées par le modèle *TurboScribe*, notamment au niveau de la phonétique et de l'orthographe, qui se révèlent souvent erronées. Des séquences telles que « *souyon gan* », « *sagon bontan* », « *Nkadi sagan pil lanyì* » ou encore « *ti kouloa* » semblent être des déformations approximatives issues de formes créoles plausibles, mais dont la sémantique reste difficile à interpréter.

Nous avons également constaté une certaine répétitivité et incohérence dans les segments générés. Certains passages apparaissent de manière récurrente, tandis que l'orthographe varie d'un segment à l'autre, sans stabilité ni cohérence linguistique. Le sens des mots demeure souvent flou, en particulier dans le cas du créole haïtien : de nombreuses phrases n'ont pas de signification claire ou correcte en créole haïtien standard. Cela suggère une correspondance phonétique approximative, résultant d'un traitement inadéquat des sons caractéristiques de la langue. Ces dysfonctionnements sont typiquement attribuables à une performance dégradée du modèle lorsqu'il ne reconnaît ni les phonèmes ni le lexique propres à la langue cible.

Ce déficit peut être accentué par plusieurs facteurs, tels que le bruit de fond, l'accent local, ou encore la diction du locuteur, autant d'éléments que le modèle semble mal interpréter en l'absence d'un entraînement spécifique sur des données en créole haïtien.

### 5.2.4 Créole haïtien et français : une comparaison entre une voix native et deux langues

Nous avons réalisé un comparatif en utilisant la même voix native haïtienne pour produire de l'audio en français, puis les avons soumis à *TurboScribe* pour transcription. Voici le texte transcrit, avec un taux d'erreur de 0 % :

#### Tèks odvo an fransè ( texte audio en français)

10 juin 2025 à 13 h 58

(0:00) Les relations sino-latino-américaines se sont fortement développées au 21<sup>e</sup> siècle, (0:08) marquant un tournant dans la diplomatie régionale. (0:13) Cette relation a des forces, enjeux et opportunités. (0:20) Si tu peux voir détruire l'ouvrage de ta vie et sans dire un seul mot te mettre à rebâtir,

Pour accéder à la transcription et écouter l'audio natif en français, cliquez sur ce lien

[Audio natif en français.](#)

## 6. Discussions

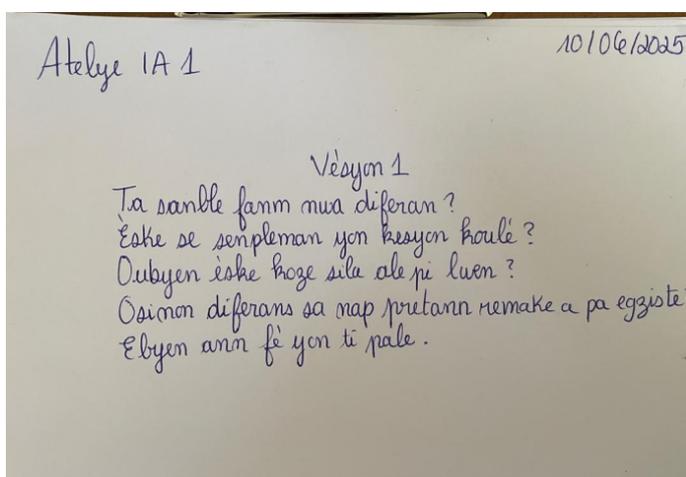
Cette partie "Discussion" porte sur les quatre outils et modèles présentés dans ce travail : *ChatGPT*, *Gemini*, *TTSMaker* et *TurboScribe*. Elle s'articule principalement autour des aspects suivants : l'application des modèles d'intelligence artificielle au créole haïtien, à travers la reconnaissance optique de caractères manuscrits, l'analyse contextuelle et linguistique, la complexité phonétique ainsi que la variation dialectale.

### a) Applications des modèles d'IA au créole haïtien

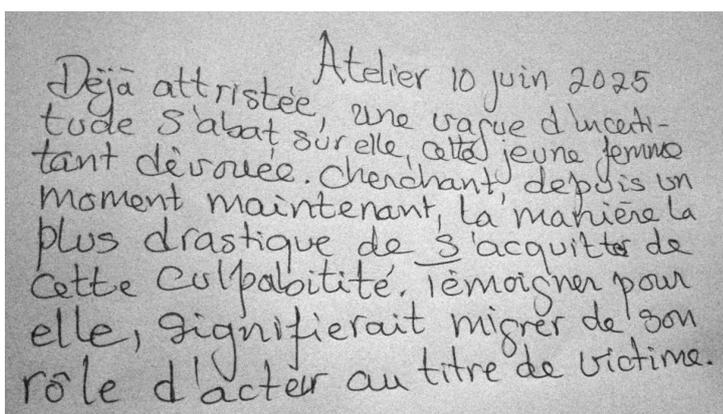
Le modèle *Gemini* est une intelligence artificielle avec des entrées multimodales capables de traiter du texte, de l'image, du code, des PDF, de la vidéo et de l'audio via des interfaces de programmation d'application (API) Google/Firebase. *Gemini*, couplé à ses capacités de vision, pourrait extraire du texte d'images de manuscrits en créole haïtien. Et c'est le cas pour *ChatGPT*.

#### Texte 1 et 2 proposés à l'extraction par Gemini et ChatGPT.

Texte 1 - en créole



Texte 2 - en français



Cependant, sans *fine-tuning* (**Prendre un modèle déjà entraîné sur beaucoup de données, et le ré-entraîner légèrement sur un petit jeu de données spécifiques, pour qu'il devienne meilleur sur une tâche particulière.**) spécifique sur la graphie cursive du créole haïtien, les erreurs de reconnaissance de formes (graphèmes) resteront élevées, nécessitant un post-traitement linguistique.

### Texte 1 (créole) extrait de l'image par Gemini

**Men tèks mwen ekstrè nan imaj la:**  
**Atelye IA 1**  
**Vèsyon 1**  
**Ta sanble fanm nan diferan?**  
**Èske se senpleman yon kesyon koulè?**  
**Oubyen èske koze sila ale pi lwen?**  
**Osimon diferans sa nap pretann remake pa egziste?**  
**Ebyen ann fè yon ti pale.**  
**10/06/2025**

### Texte 1 (créole) extrait de l'image par ChatGPT

**Atelye IA 1**  
**Dat : 10/06/2025**  
**Vèsyon 1**  
**1. Ta sanble fòm mwen diferan ?**  
**2. Èske se senpleman yon kesyon kouché ?**  
**3. Oubyen èske kès sila ale pi lwen ?**  
**4. Osinon diferans sa nap pretann remake e pa egziste ?**  
**5. Ebyen ann fè yon ti pale.**

### Texte 2 (français) extrait de l'image par GEMINI

Atelier 10 juin 2025

Déjà attristée, une vague d'incertitude s'abat sur elle, cette jeune femme tant dévouée. Cherchant depuis un moment maintenant, la manière la plus drastique de s'acquitter de cette culpabilité. Témoigner pour elle, signifierait migrer de son rôle d'acteur au titre de victime.

## Texte 2 (français) extrait de l'image par ChatGPT

ChatGPT 4o ▾



Voici le texte extrait du document :

### Atelier 10 juin 2025

Déjà attristée, une vague d'incertitude s'abat sur elle, cette jeune femme tant dévouée. Cherchant, depuis un moment maintenant, la manière la plus drastique de s'acquitter de cette culpabilité, témoigner pour elle, signifierait migrer de son rôle d'acteur au titre de victime.

### b) Reconnaissance optique de caractères (OCR) de manuscrits en créole haïtien

*Gemini et ChatGPT* peuvent combiner information visuelle (ex. image ou texte manuscrit, audio (ex. fichier vocal) et contexte textuel pour produire des analyses ou reformulations dans beaucoup de langues. Cependant, en contexte du créole haïtien, cette capacité se heurte à un problème d'entraînement et de disponibilité de données suffisantes.

Les technologies de reconnaissance optique de caractères (OCR) ouvrent des perspectives pour la valorisation et la diffusion du créole en milieu numérique. Elles permettent de transformer instantanément des documents imprimés ou manuscrits en des fichiers modifiables, l'OCR permet de constituer d'importants corpus textuels. Ces bases de données ainsi constituées, enrichies et structurées, alimentent les moteurs de recherche, facilitent l'étude linguistique et servent de matière première aux outils de traitement automatique de la langue (traduction, correction orthographique, synthèse vocale).

Pour tirer parti de l'OCR, il convient de relever plusieurs défis.

La variabilité orthographique du créole : ceci nécessite la création de modèles d'OCR spécifiquement entraînés sur ses différentes graphies. Les documents anciens ou de mauvaise qualité exigent un pré-traitement rigoureux (nettoyage d'image, amélioration du contraste). Enfin, l'implication de la communauté, écoles, centres culturels, utilisateurs, est essentielle pour numériser en masse et annoter des écrits variés, garantissant ainsi la représentativité et la richesse du corpus.

L'intégration de l'OCR dans la chaîne de production et de diffusion de textes en créole contribue à préserver le patrimoine linguistique haïtien, mais surtout à l'inscrire durablement dans l'écosystème numérique : ressources pédagogiques en ligne, interfaces mobiles et même assistants conversationnels capables d'échanger en créole.

### c) Analyse contextuelle et linguistique approfondie

*Gemini et ChatGPT* excellent dans la manipulation de textes — qu'il s'agisse de correction, reformulation, analyse stylistique, suggestion post-correction, génération de lexiques ou de règles grammaticales et orthographiques — dans une multitude de langues humaines. Concernant le créole haïtien, ces modèles peuvent contribuer à la vérification et à la correction orthographique ainsi qu'à l'amélioration des tournures de phrases, en se référant aux normes linguistiques du créole haïtien (lorsque le texte est disponible). Ils sont également capables de produire des traductions entre le français et le créole haïtien (FR→CR ou CR→FR), à titre indicatif, en s'appuyant sur leur corpus d'entraînement. Toutefois, la qualité de ces traductions peut varier, en raison de la sous-représentation du créole haïtien dans les données d'apprentissage. Par ailleurs, ces outils peuvent générer des suggestions de chaînes de traitement linguistique (pipelines), incluant le post-traitement de la reconnaissance automatique de la parole (ASR), la création de lexiques, ou encore la normalisation linguistique. De manière générale, les modèles d'intelligence artificielle générative surpassent les capacités humaines dans l'exécution de nombreuses tâches intellectuelles et cognitives (Taddei & Lucksh, 2025). Dans le domaine de la production et de la correction de textes dans des langues fortement représentées dans leur corpus d'entraînement, telles que l'anglais ou le français, le taux d'erreurs orthographiques ou syntaxiques est quasiment nul.

### d) Limites des modèles d'intelligence artificielle (IA)

Les modèles d'intelligence artificielle sont fondamentalement dépendants des données sur lesquelles ils ont été entraînés, ce qui peut limiter leur performance dans l'exécution de certains prompts. Par exemple, en l'absence d'accès direct à un fichier audio ou à une image, *ChatGPT* ne peut pas évaluer la qualité brute d'une transcription : il dépend uniquement du texte fourni en entrée. Sa connaissance du créole haïtien repose sur les données textuelles disponibles lors de son entraînement, lesquelles demeurent souvent insuffisantes pour couvrir les variantes dialectales ou les formes linguistiques très spécifiques. L'un des enjeux majeurs pour les langues minorées, comme les langues créoles, est la capacité à produire des données (data) en quantité et en qualité suffisantes dans tous les domaines — éducatif, universitaire, scientifique, démographique, sanitaire, culturel, etc. — afin de permettre l'entraînement efficace des IA dans ces langues. Il s'agit d'un enjeu linguistique, technologique et sociétal déterminant pour l'avenir.

### e) Pourquoi la plupart des TTS (text-to-Speech) ne supportent pas le créole haïtien ?

La majorité des systèmes TTS commerciaux ne prennent pas en charge le créole haïtien, et cela s'explique par un ensemble de facteurs linguistiques, techniques et économiques. Cette limitation constitue un frein important à l'inclusion linguistique et à la diversité numérique.

Tout d'abord, l'absence de prise en charge du créole haïtien peut être attribuée à *la rareté des corpus linguistiques*. Le développement d'un système TTS de qualité nécessite un corpus conséquent, composé de paires texte/audio, enregistrées par des locuteurs natifs dans des conditions acoustiques maîtrisées.

Or, bien que le créole haïtien soit une langue très vivante et largement parlée, il souffre d'un manque de ressources linguistiques standardisées, comparativement à des langues dominantes comme l'anglais ou le français.

Ensuite, *la complexité phonétique et les variations dialectales* du créole haïtien posent des défis supplémentaires. Il existe une pluralité de prononciations, de rythmes et d'intonations selon les régions et les locuteurs, ce qui complique l'élaboration d'un modèle de synthèse vocale universel et représentatif.

Par ailleurs, *la diversité graphique et lexicale* du créole, de manière générale, complique l'uniformisation des ressources textuelles. Bien que l'orthographe officielle du créole haïtien ait été normalisée, son application demeure inégale dans les usages quotidiens ainsi que dans les documents disponibles. Or, pour entraîner un système TTS fiable, il est indispensable de disposer de textes transcrits de manière cohérente et standardisée, permettant un alignement précis entre l'écrit et l'oral. Ce processus repose en grande partie sur le traitement de la prosodie — notamment l'accentuation, l'intonation et les pauses — à travers des annotations et des analyses réalisées par des locuteurs natifs. Un tel travail permettrait également de normaliser les textes, en gérant la diversité des graphies, les abréviations, ainsi que les emprunts au français, qui sont fréquents dans l'usage courant.

Enfin, il faut souligner *le manque d'investissement commercial* dans les langues minorées. Il semblerait que le créole haïtien ne soit pas perçu comme un marché lucratif par les grandes entreprises technologiques, ce qui limite le développement de solutions adaptées. Pourtant, cette langue représente des millions de locuteurs dont les besoins en accessibilité numérique sont réels. Cela dit, la flexibilité des systèmes TTS reste un atout considérable pour diagnostiquer les erreurs de transcription, notamment en comparant l'image d'un manuscrit à sa transcription via un système de reconnaissance vocale (ASR). Cette approche peut aussi contribuer à l'enrichissement des corpus en créole haïtien, et donc, indirectement, à l'amélioration des modèles TTS à moyen terme.

Le non-support du créole haïtien par la plupart des systèmes TTS résulte d'une combinaison de facteurs structurels, linguistiques et économiques. Cette situation illustre de manière flagrante les inégalités d'accès à la technologie pour les langues minorées. Il est urgent de promouvoir des politiques linguistiques et technologiques plus inclusives, soutenant la création de corpus ouverts, la standardisation linguistique et l'investissement dans la diversité numérique.

#### **f) Pourquoi le modèle TurboScribe ne supporte-t-il pas le créole haïtien ?**

Le modèle *TurboScribe*, basé sur *Whisper* d'OpenAI, présente des difficultés à transcrire correctement le créole haïtien. Cette limitation s'explique par plusieurs facteurs liés à la nature du modèle, à la qualité des données disponibles et aux spécificités linguistiques du créole.

Tout d'abord, bien que *Whisper* ait été entraîné sur un large corpus multilingue, la quantité de données en créole haïtien reste très faible. Cette sous-représentation dans les données d'entraînement limite la capacité du modèle à généraliser sur cette langue, en particulier dans des contextes "*zero-shot*", où aucun ajustement spécifique (*fine-tuning*) n'a été effectué pour cette langue.

Ensuite, le corpus audio multilingue utilisé par *Whisper* contient une couverture inégale selon les langues. Pour des langues peu dotées comme le créole haïtien, cette couverture est insuffisante, ce qui se traduit par un taux d'erreur élevé (WER – *Word Error Rate*). De plus, les caractéristiques audio (qualité d'enregistrement, bruit de fond, accent, débit de parole, utilisation d'expressions idiomatiques ou de registres variés) peuvent également aggraver les performances, surtout en l'absence de données normalisées.

Parmi les autres facteurs limitants, on peut également citer :

- L'absence de fine-tuning spécifique sur le créole haïtien, ce qui empêche le modèle d'adapter ses prédictions aux spécificités phonétiques et syntaxiques de cette langue.
- Le manque de post-traitement linguistique adapté, notamment pour gérer la variation lexicale, les graphies divergentes et les emprunts au français.
- L'ambiguïté linguistique et la variation dialectale, qui compliquent davantage la tâche de transcription automatique sans ressources annotées finement.

Le non-support du créole haïtien par *TurboScribe* est symptomatique d'un problème plus large de sous-représentation des langues minorées dans les grands modèles de traitement automatique du langage. Pour y remédier, il est indispensable de construire des corpus audio et textuels en créole haïtien, d'enrichir les modèles par du fine-tuning ciblé, et d'intégrer des stratégies de post-traitement linguistique adaptées.

### **Perspectives :**

Parmi les initiatives technologiques récentes visant à promouvoir l'inclusion linguistique dans le domaine de l'intelligence artificielle, le projet haïtien *Palefò.ai*, développé par le groupe *Nerval AI*, retient particulièrement notre attention. Ce projet vise à concevoir des modèles de Reconnaissance Vocale Automatique (ASR) et de Synthèse Vocale (TTS) de haute qualité spécifiquement adaptés au créole haïtien. En ce sens, *Palefò.ai* se présente comme une alternative prometteuse aux solutions actuelles, encore largement inadaptées aux langues créoles, et pourrait répondre à plusieurs des limites identifiées dans cette étude. L'évolution de ce projet est suivie avec rigueur, et une étude comparative sera entreprise dès que l'agent conversationnel développé sera pleinement fonctionnel, afin d'évaluer son efficacité en contexte éducatif et sociolinguistique

## 7. Conclusion générale

Trois facteurs majeurs expliquent pourquoi, dans son état actuel, l'intelligence artificielle (IA) ne parvient pas à faire progresser efficacement la digitalisation des langues créoles, en particulier le créole haïtien :

1. *La non-intégration du créole haïtien dans les grandes bases de données linguistiques utilisées par les systèmes d'IA*

Le créole haïtien est quasiment absent des référentiels linguistiques standards sur lesquels s'appuient les modèles d'intelligence artificielle. Cette exclusion empêche l'IA de prendre en compte les structures syntaxiques, les règles grammaticales et les spécificités phonologiques propres à la langue. En conséquence, les modèles peinent à générer ou interpréter du contenu en créole de manière fiable.

2. *L'absence d'un corpus audio/textuel structuré et validé en créole haïtien chez les fournisseurs de données*

L'entraînement efficace de modèles linguistiques performants exige un corpus conséquent, standardisé et de haute qualité. Or, à ce jour, il n'existe pas de ressources en créole haïtien suffisamment riches, accessibles et validées pour entraîner de manière efficace une intelligence artificielle sur le créole haïtien.

Ce manque constitue un obstacle majeur à la modélisation des formes orales et écrites (syntaxe et phonologie) du créole par les technologies de l'IA.

3. *Le faible niveau de contribution des locuteurs natifs* La participation active des communautés linguistiques est cruciale pour garantir la représentativité et l'authenticité des corpus. Le nombre limité d'enregistrements produits par des locuteurs natifs haïtiens freine la capacité de l'IA à reproduire fidèlement les intonations, les accents régionaux et les usages contextuels du créole haïtien.

Ce constat appelle une mobilisation urgente. Sans une volonté affirmée de documenter, numériser et intégrer les langues créoles dans les projets technologiques, ces langues continueront à être marginalisées. Il est impératif de penser une inclusion linguistique qui dépasse la symbolique pour devenir pleinement opérationnelle dans les technologies émergentes. Par ailleurs, il est important de souligner que la dimension culturelle et anthropologique est encore trop peu prise en compte dans les modèles technologiques actuels, ouvrant la voie à des enjeux éthiques majeurs : biais algorithmiques, exclusion numérique et sous-représentation des langues minorées. Vers un écosystème IA créole durable, nous proposons des pistes concrètes :

- la création de corpus spécialisés et accessibles;
- l'implication active et continue des communautés locales;
- le renforcement de partenariats institutionnels entre chercheurs, gouvernements, organismes linguistiques et acteurs technologiques.

Enfin, les chercheurs, les décideurs politiques et les communautés linguistiques doivent œuvrer ensemble pour promouvoir un usage éthique, inclusif et efficace de l'intelligence artificielle, au service de la préservation et de la valorisation du créole haïtien.

## 8. Références bibliographiques

- Zribi-Hertz, A. (2020). *Phonologie du créole haïtien : présentation succincte pour la didactique du FLS*, CNRS
- Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2020). *Automatic Speech Recognition for Under-resourced Languages: A Survey*. *Speech Communication*, 56(1), 85-100.
- Blasi, D. E., Anastasopoulos, A., & Neubig, G. (2022). *Systematic Inequalities in Language Technology Performance across the World's Languages*. *ACL Anthology*, 2022.
- Dowding, J., Gawron, J. M., Appelt, D., Bear, J., Cherny, L., Moore, R., & Moran, D. (1994). *GEMINI: A Natural Language System for Spoken-Language Understanding*. *Computational Linguistics*
- DeGraff, M. (2019). *Against Apartheid in Education and in Linguistics: The Case of Haitian Creole in Neo-Colonial Haiti*. *Language*, 95(4), e344–e372.
- Gasser, M. (2018). *Computational Morphology and Low-resource Languages: Where to Go from Here?* Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages, *ACL Anthology*, 1–11.
- Hebblethwaite, B. (2012). *French and Underdevelopment, Haitian Creole and Development: Language Policy Problems and Solutions in Haiti*. *Journal of Pidgin and Creole Languages*, 27(2), 255-302.
- Jean Robert Cadely, *Aspects de la Phonologie du Créole haïtien*, Éditions universitaires européennes (2018)
- Jean-Baptiste, R., & Tardieu, C. (2018). *Le créole haïtien: phonologie, morphologie, syntaxe, lexicque*. Port-au-Prince : Éditions de l'Université d'État d'Haïti.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*. *ACL Anthology, Proceedings of the 58th Annual Meeting of the ACL*, 6282–6293.
- Valdman, A. (2015). *Haitian Creole: Structure, Variation, Status, Origin*. Sheffield: Equinox Publishing Ltd.
- Wang, C., Pan, X., & Nie, L. (2021). *Low-resource Languages: Theories, Methods, and Challenges*. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(4), 494-506.

DeGraff, M.; Frager, W. S; Miller, H. (2022). *Language Policy in Haitian Education: A History of Conflict over the Use of Kreyòl as Language of Instruction*. MIT & UNESCO IBE. [lingphil.scripts.mit.edu](http://lingphil.scripts.mit.edu).

ACL Anthology. *Challenges and Innovations in NLP for Haitian Creole* (2025). [aclanthology.org](http://aclanthology.org).

CreoleVal: Multilingual Multitask Benchmarks for Creoles (MIT Press, 2024). [direct.mit.edu](http://direct.mit.edu).

UNESCO. « Embracing Creole in Haitian Schools » (2023). [ibe.unesco.org](http://ibe.unesco.org).

Haitian Times. « Haitian Creole in the Digital Age: Embracing Technology » (mai 2024). [haitiantimes.com](http://haitiantimes.com).

ProAI (IDB). *Program to create digital opportunities in Haiti through AI*. [iadb.org](http://iadb.org).

Creole Solutions Blog. “How AI Can Revolutionize Education in Haitian Creole” (2023). [blog.creolesolutions.com](http://blog.creolesolutions.com).

Articles sur les enjeux éthiques et biais en IA pour créoles (p. ex. discussions LinkedIn et blogs spécialisés). [linkedin.com](http://linkedin.com) [aclanthology.org](http://aclanthology.org).

Akademi Kreyòl Ayisyen. Informations sur la régulation de la langue et initiatives de normalisation. [en.wikipedia.org](http://en.wikipedia.org).

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Haegeman, L. (1994). *Introduction to Government and Binding Theory*. Blackwell.

Boulogne, J. (2024) *Du nouveau pour les créoles et l'intelligence artificielle*.

Julien-Souverain, M. (2024). *Le créole haïtien à l'ère du numérique*. In *The Haitian Times*. Consulté en ligne le : [ 26 juin 2025]

Lin, D., Zhao, N., Tian, D., & Li, J. (2025, avril). *ChatGPT as Linguistic Equalizer? Quantifying LLM-Driven Lexical Shifts in Academic Writing*. ArXiv.

Gauld, Ch. (2020). *Une brève histoire des sciences computationnelles*. In revue *Médecine et philosophie* (4). DOI : 10.51328/100

Résultats de l'enquête sur les connaissances technologiques et l'IA (33 répondants, majoritairement étudiants). Lien de consultation: <https://lc.cx/ZWPvmU>

**Pour citer l'article :**

MABIEN, Bruny et PIERRE, Ralphson, (2025). *L'intelligence artificielle (IA) comme levier de digitalisation des langues créoles : impact culturel et technologique sur le créole haïtien*. In FED - Fondation Education & Développement, Haiti.

