Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

La traduction automatique dialectale: état de l'art et étude préliminaire sur le continuum dialectal de l'occitan

Oriane Nédey
Inria Paris, France
oriane.nedey@inria.fr

RÉSUMÉ

Cet article dresse un état de l'art de la traduction automatique et de son évaluation pour les langues à variation dialectale, et en particulier pour les continuums dialectaux. Pour illustrer cet état de l'art, nous proposons une série d'expériences préliminaires sur le continuum occitan, afin de dresser un état des performances des systèmes existants pour la traduction depuis et vers plusieurs variétés d'occitan. Nos résultats indiquent d'une part des performances globalement satisfaisantes pour la traduction vers le français et l'anglais. D'autre part, des analyses mélangées à des outils d'identification de langues sur les prédictions vers l'occitan mettent en lumière la capacité de la plupart des systèmes évalués à générer des textes dans cette langue (y compris en *zero-shot*), mais révèlent aussi des limitations en termes d'évaluation de la diversité dialectale dans les traductions proposées.

ABSTRACT

Dialectal machine translation : survey and preliminary study on the Occitan dialect continuum

We present a state of the art of machine translation and its evaluation for languages with dialectal variation, and in particular for dialect continua. To accompany this overview, we propose a series of preliminary experiments working with the Occitan continuum, in order to assess the performance of existing systems with respect to translation from and into several varieties of Occitan. Our results indicate that translation into French and English is generally of good quality. Analyses combined with language identification tools applied to predictions into Occitan highlight the ability of most of the systems tested to generate texts in this language (even in *zero-shot settings*), but they also reveal limitations in terms of evaluation methods for the dialectal diversity in the proposed translations.

MOTS-CLÉS : traduction automatique, occitan, évaluation, langues peu dotées, dialectes.

KEYWORDS: machine translation, Occitan, evaluation, low-resource languages, dialects.

1 Introduction

L'amélioration des performances et la mise à disposition de modèles de langues pré-entraînés sur des jeux de données multilingues a permis des avancées importantes dans le domaine du traitement automatique pour les langues peu dotées. Dans le domaine de la traduction automatique (TA), les architectures multilingues montrent des résultats prometteurs en incluant des langues avec peu ou pas de données parallèles (Costa-jussà *et al.*, 2024). Cependant, pour augmenter davantage le nombre de langues outillées dans le monde, il est essentiel de se pencher sur le traitement des langues non standardisées et des continuum dialectaux. Ces langues présentent des phénomènes de variation importante, couplée à une faible quantité de données disponibles.

Actes des 18e Rencontres Jeunes Chercheurs en RI (RJCRI) et 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 190–238.

C’est le cas de l’occitan, dont l’aire géographique s’étend sur une grande partie du sud de la France ainsi que localement en Italie et en Espagne. Les contributions ouvertes pour l’occitan ont permis d’intégrer cette langue dans plusieurs systèmes de traduction disponibles en ligne, cependant la variation dialectale n’a pas été prise en compte explicitement dans leur développement.

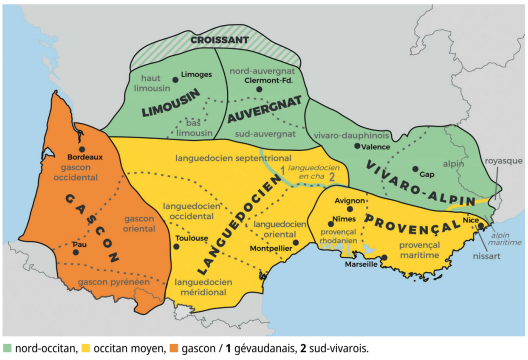

FIGURE 1 – Dialectes et sous-dialectes de l’occitan (Sibille, 2024).

Dans cet article, nous explorons les différences de performance de plusieurs modèles de traduction dans le contexte du continuum dialectal de l’occitan. Nos contributions sont :

— un état de l’art de la TA en situation de continuum dialectal
— des expériences préliminaires d’évaluation de modèles de traduction existants sur plusieurs variétés d’occitan, pour les direction occitan-français, français-occitan, et occitan-anglais.

2 État de l’art

2.1 La situation dialectale de l’occitan

En linguistique, la notion de *dialecte* est difficile à définir, notamment en raison de son opposition fréquente à celle de *langue*. Le terme *dialecte* est souvent employé comme un raccourci pour désigner une variété comportant des variations linguistiques systématiques par rapport à une ou plusieurs autres variétés apparentées. Chambers & Trudgill (1998) exposent plusieurs critères permettant de caractériser des situations dialectales, tout en insistant sur leur nature graduelle : chaque variété se situe quelque part sur un continuum pour chacun de ces critères. Nous en retenons trois : l’autonomie, l’intelligibilité mutuelle, ainsi que la continuité géographique et/ou sociale.

— Le critère d’**autonomie** indique un niveau d’indépendance d’une variété par rapport à d’autres. Ce critère repose principalement sur des facteurs politiques et culturels, comme la dénomination de la variété, la reconnaissance officielle, l’existence d’un organisme de standardisation ou l’usage de grammaires prescriptives. Par exemple, le suédois et le norvégien sont reconnues comme des langues distinctes, alors que le français québécois se réfère au français standard de France comme norme de référence.
— Le critère d’**intelligibilité mutuelle** dépend de la proximité linguistique entre différentes variétés, ainsi que des systèmes graphiques employés à l’écrit. Par exemple, bien que le serbe et le croate soient très proches d’un point de vue linguistique, l’usage du cyrillique pour le serbe limite leur intercompréhension à l’écrit.
— Le critère de **continuité dialectale** évalue si les distinctions entre variétés sont graduelles ou marquées par des frontières nettes, qu’elles soient d’ordre géographique ou social.

La langue occitane est implantée dans une aire linguistique qui s’étend sur la majeure partie du sud de la France ainsi que dans certaines régions frontalières d’Espagne et d’Italie. En France, il n’existe pas de standard officiel accepté par toute la communauté occitanophone ; pour autant, certains facteurs tendent vers une uniformisation de la langue. Ainsi, deux conventions graphiques - dites *classique*

et *mistralienne* - sont largement diffusées, dont la première vise à minimiser la variation dialectale à l'écrit (Lamuela, 2024). De plus, en dépit de la rupture de transmission familiale, de nouvelles générations de néo-locuteurs ont été formées grâce à des initiatives de revitalisation de la langue, comme l'enseignement de l'occitan à l'école, souvent dans une forme plus standardisée et supralocale (Brennan, 2024).

L'occitan est un continuum dialectal dont l'espace linguistique est le théâtre de variations tant sur le plan géographique que social (Brennan, 2024), et affectant plusieurs dimensions de la langue : phonétique, phonologie, morphologie, lexique et syntaxe (Esher & Sibille, 2024). Pour des raisons de simplification et de description linguistique, l'occitan est généralement divisé en six dialectes (Sibille, 2024) (cf. figure 1) : gascon, languedocien, limousin, auvergnat, vivaro-alpin, et provençal. Malgré l'absence de consensus sur cette classification, ce sont principalement ces étiquettes qui sont utilisées dans les corpus annotés en dialectes occitans (Miletic *et al.*, 2020; Séguier & Lo Congrès, 2023a,b; Lo Congrès, 2024). En raison du manque de données annotées de manière plus fine (avec des indications géographiques précises et des informations sur le profil des auteurs), le recours à une classification large reste nécessaire pour estimer les performances des systèmes de TAL sur le continuum occitan (Seguier, 2015; Miletić, 2023; Hopton & Aepli, 2024).

2.2 Traduction automatique dialectale

Les variétés dialectales (entendues au sens large) manquent souvent de ressources pour le TAL (Joshi *et al.*, 2020; Liu *et al.*, 2022), et ainsi les approches pour la TA des langues peu dotées (Haddow *et al.*, 2022; Ranathunga *et al.*, 2023) offrent des pistes prometteuses pour la TA dialectale. Sans traiter spécifiquement des variantes dialectales, certaines approches ont permis des avancées pour la traduction de langues similaires, en utilisant très peu voire aucune données parallèles selon les langues. En particulier, (Costa-jussà *et al.*, 2024; Bapna *et al.*, 2022; Kudugunta *et al.*, 2023) ont adapté des approches d'identification de langues, d'augmentation de données (fouille de données, traduction inverse [1]), de tokenisation, et de modélisation à un grand nombre de langues, qui comportent différents niveaux de standardisation et dont certaines sont généralement considérées comme des variantes dialectales (par exemple l'arabe tunisien et l'arabe marocain).

Les travaux de TA spécifiques aux contextes dialectaux ont suivi l'évolution des techniques pour la TA en général, avec des systèmes à base de règles (Armentano i Oller & Forcada, 2006; Scherrer, 2012), des modèles statistiques (Tiedemann, 2009; Sawaf, 2010; Salloum & Habash, 2011; Marujo *et al.*, 2011) ainsi que des réseaux de neurones (Costa-jussà *et al.*, 2018; Lakew *et al.*, 2018; Myint Oo *et al.*, 2019; Kumar *et al.*, 2021; Bafna *et al.*, 2025).

Approches de normalisation Une approche courante pour accroître la robustesse des modèles à la variation dialectale dans le cadre de la TA consiste à utiliser un processus de normalisation. L'objectif de la normalisation est de réduire les différences entre les textes dialectaux et une langue standard afin de faciliter leur traitement. Plusieurs méthodes ont été explorées dans ce cadre. Sawaf (2010) propose un système hybride combinant règles linguistiques et TA statistique (SMT) pour normaliser l'arabe dialectal. D'autres approches reposent sur la substitution lexicale, comme celles de Aminian *et al.* (2014) et Bafna *et al.* (2025), qui exploitent des ressources linguistiques (analyseurs morphologiques, corpus annotés en parties du discours) et des lexiques bilingues pour identifier et remplacer les mots

1. En anglais : *back-translation*

dialectaux par leur équivalent dans une langue standard proche au moment de l'inférence. Enfin, des méthodes basées sur la traduction statistique ou neuronale au niveau des caractères ont également été étudiées (Scherrer & Ljubešić, 2016; Honnet *et al.*, 2018) et offrent des gains de performance notamment sur les mots inconnus.

Approches par ajout de bruit Une autre approche gagne en popularité ces dernières années : l'ajout de bruit dans le corpus d'entraînement d'une langue standard. Cette approche est le plus souvent implémentée avec des opérations aléatoires d'échange, de remplacement, d'insertion et de suppression de caractères (Belinkov & Bisk, 2018; Heigold *et al.*, 2018; Aepli & Sennrich, 2022; Blaschke *et al.*, 2023). D'autres méthodes ont aussi été expérimentées pour que les transformations produites soient plus cohérentes avec les variétés dialectales visées. Anastasopoulos *et al.* (2019) utilisent le parsing syntaxique ainsi que des règles de transformation pour générer des erreurs grammaticales. Xia *et al.* (2019) et Jones *et al.* (2023) remplacent certains mots du corpus en langue standard par leur équivalent dans la langue cible s'il existe dans un lexique bilingue, créant ainsi du bruit de type *alternance codique* ou *mélange codique* (en anglais : *code-switching* et *code-mixing*). Brahma *et al.* (2023) sélectionnent les groupes de caractères à bruiter à partir des opérations de fusion de l'algorithme BPE sur les corpus de la langue originale et de la langue peu dotée. Bafna *et al.* (2024) et Bafna *et al.* (2025) définissent plusieurs bruiteurs pour plusieurs fonctions linguistiques : phonologique, morphologique et lexicale (mots porteurs de sens et mots fonctionnels).

Traduction vers des dialectes Ces techniques de normalisation et d'ajout de bruit sont particulièrement populaires et efficaces pour les modèles entraînés pour traduire vers des langues standardisées, mais il existe également des travaux pour la TA vers des dialectes (ou langues apparentées). Les langues sources utilisées dans cette direction sont généralement des langues proches (Armentano i Oller & Forcada, 2006; Fancellu *et al.*, 2014; Popović *et al.*, 2016; Wan *et al.*, 2020; Jerpelea *et al.*, 2024), mais parfois aussi des langues plus éloignées – le plus souvent l'anglais (Lakew *et al.*, 2018; Kumar *et al.*, 2021; Garcia & Firat, 2022). Concernant les variétés cibles, les recherches sont particulièrement rares dès lors qu'elles portent sur des dialectes non standardisés (Altintas & Cicekli, 2003; Haddow *et al.*, 2013; Jeblee *et al.*, 2014; Hassani, 2017; Myint Oo *et al.*, 2019; Kumar *et al.*, 2021; Her & Kruschwitz, 2024), voire des continuums dialectaux (Scherrer, 2012; Meftouh *et al.*, 2015; Abe *et al.*, 2018; Lambrecht *et al.*, 2022).

Parmi les approches spécifiques pour traduire vers des dialectes, Scherrer (2012) introduit des règles de transformation géoréférencées associées à des cartes de probabilités pour traduire vers le continuum suisse allemand. Plus récemment, les techniques de traduction inverse sont également plébiscités (Wan *et al.*, 2020; Kumar *et al.*, 2021; Lambrecht *et al.*, 2022; Her & Kruschwitz, 2024; Sánchez-Martínez *et al.*, 2024). Le développement de modèles multilingues et multi-tâches interrogés par des prompts (Raffel *et al.*, 2019; Brown *et al.*, 2020; Ouyang *et al.*, 2022) a permis à Garcia & Firat (2022) de comparer les performances du modèle mT5 (Xue *et al.*, 2021) en mode *zero-shot* lorsque des noms de dialectes sont mentionnés ou non dans le prompt (par exemple "Portuguese" vs. "Brazilian Portuguese"), et ouvrant la voie à la TA dialectale avec des grands modèles de langues (Sánchez-Martínez *et al.*, 2024).

Les approches d'apprentissage par transfert et d'apprentissage multilingue – très courantes pour la traduction des langues peu dotées – sont également des terrains d'expérimentation pour la traduction dialectale. Zbib *et al.* (2012) développent des modèles spécifiques à chaque dialecte pour traduire de deux dialectes de l'arabe vers l'arabe standard moderne. Salloum *et al.* (2014) introduisent un modèle

de classification de dialectes de l'arabe pour choisir le modèle (spécifique à un seul dialecte) le plus approprié en fonction du texte à traduire. Abe *et al.* (2018), Lakew *et al.* (2018) et Lambrecht *et al.* (2022) comparent des modèles spécifiques par dialecte à des modèles multi-dialectes, avec ou sans première phase d'entraînement sur l'ensemble des variétés apparentées. Abe *et al.* (2018) montrent que les modèles entraînés sur un seul dialecte avec peu de données sont moins performants que la *baseline* qui consiste à copier le texte source, pour la traduction des dialectes japonais vers le japonais standard. Lakew *et al.* (2018) expérimentent plusieurs approches de systèmes multi-dialectes, en intégrant notamment un token comme étiquette de variété au début du texte d'entrée, et en utilisant un modèle d'identification de variétés lorsqu'aucune étiquette de variété n'est disponible. L'implémentation multi-dialecte de Lambrecht *et al.* (2022) utilise quant à elle l'architecture Transformer, avec un encodeur unique pour les entrées en allemand standard ainsi qu'une série de décodeurs pour chacune des cinq variétés d'alémanique à générer en sortie du modèle.

2.3 Évaluation de la TA en contexte de variation dialectale

Évaluation humaine versus automatique Parmi les deux grands types d'évaluation des systèmes de TA, l'évaluation humaine, bien que coûteuse et complexe à organiser en raison de la nécessité de recruter des annotateurs qualifiés, est de plus en plus utilisée (Kocmi *et al.*, 2023, 2024). Elle permet non seulement de comparer des modèles avec une grande fiabilité, mais aussi d'obtenir des évaluations détaillées grâce à des annotations fines (Graham *et al.*, 2013; Lommel *et al.*, 2013; Freitag *et al.*, 2021). Toutefois, peu de travaux sur la TA dialectale intègrent une telle évaluation humaine (Jeblee *et al.*, 2014; Hassani, 2017; Costa-jussà, 2017; Costa-jussà *et al.*, 2018; Jerpelea *et al.*, 2024). La mise en œuvre d'une campagne d'évaluation humaine pose des défis supplémentaires pour les langues à faible nombre de locuteurs et pour les continuums dialectaux, notamment en ce qui concerne le recrutement de locuteurs-annotateurs, en particulier pour des annotations inter-dialectales.

Les techniques d'évaluation automatique de la TA ont beaucoup progressé ces dernières années (Freitag *et al.*, 2021), notamment grâce à l'arrivée de métriques basées sur des modèles de langue (ex. BERTscore, Zhang *et al.*, 2020) et entraînées sur des jugements humains (ex. COMET, Rei *et al.*, 2020 ; MetricX, Juraska *et al.*, 2024 ; Gemba-MQM, Kocmi & Federmann, 2023). Cependant, les langues peu dotées sont rarement incluses dans ces modèles, et ils n'ont pas été entraînés pour être robustes à la variation dialectale, les rendant ainsi peu appropriés pour évaluer la TA depuis et vers des dialectes peu dotés ou non standardisés. Pour la direction dialecte vers langue standard, Alam *et al.* (2024) montrent que les scores COMET sont plus faibles lorsque la phrase source est utilisée telle quelle que lorsqu'elle est remplacée par une chaîne vide. Pour la traduction vers des dialectes, Aepli *et al.* (2023) montrent que pour le suisse allemand, les métriques de surface (BLEU, Papineni *et al.*, 2001 ; chrF++, Popović, 2017) ont une corrélation très faible avec les jugements humains, au contraire de COMET (avec et sans référence) qui donne de meilleurs résultats.

Adaptation de métriques à la variation dialectale Sun *et al.* (2023) introduisent deux facteurs d'évaluation des métriques d'évaluation de la TA dialectale : la *robustesse dialectale* et la *conscience dialectale* (anglais : *dialectal awareness*). Une métrique *robuste* à la variation dialectale doit pénaliser le moins possible la variation dialectale entre un texte généré et un texte de référence, en particulier par rapport à la variation sémantique. En pratique, Sun *et al.* (2023) et Aepli *et al.* (2023) montrent que les métriques existantes pénalisent davantage la variation dialectale que des changements sémantiques introduits artificiellement. La notion de *conscience dialectale* implique quant à elle de développer des

métriques qui puissent systématiquement récompenser les textes générés qui comportent des traits du dialecte souhaité, et au contraire systématiquement pénaliser les textes dans d'autres dialectes.

Sun *et al.* (2023) proposent de continuer le pré-entraînement du modèle mT5 (Xue *et al.*, 2021) avec un objectif d'identification de variété avant de l'entraîner sur des jugements humains, permettant ainsi des améliorations en termes de robustesse et de conscience dialectales sur des variétés de portugais et de chinois. Aepli *et al.* (2023) s'inspirent de cette approche et l'adaptent pour le continuum dialectal suisse allemand. Leurs travaux incluent également une phase de continuation du pré-entraînement, suivie d'une phase de fine-tuning avec injection de bruit sur les données d'entraînement du modèle COMET. Cette approche permet d'accroître la robustesse dialectale pour les scores au niveau du système (jeu de test complet), mais la corrélation au niveau des segments individuels reste très faible.

En dehors des métriques classiques qui évaluent les modèles avec un score absolu, Faisal *et al.* (2024) proposent une série de métriques pour comparer l'*écart dialectal* sur 10 tâches de TAL et 281 variétés. Ces écarts peuvent être calculés pour différentes configurations de modèles (*zero-shot*, modèle affiné sur la variété standard, modèle affiné sur la variété à évaluer) et différentes paires de variétés, telles qu'un dialecte (ou cluster de dialectes) par rapport à une langue standard (apparentée ou anglais).

Données de test multi-dialectes Un autre aspect de l'évaluation de la TA dialectale concerne les données de test. Alam *et al.* (2024) insistent sur la nécessité d'utiliser des exemples « contrastifs »[2] entre les variétés à évaluer pour obtenir des scores comparables. Plusieurs jeux de données multi-dialectes ont vu le jour ces dernières années. FLORES-200 (Costa-jussà *et al.*, 2024) contient des jeux de test multi-parallèles très utilisés pour l'évaluation de la TA multilingue ; une partie de ces données peut être utilisée comme exemples contrastifs pour certains groupes de variétés (notamment des dialectes de l'arabe). NTREX-128 (Federmann *et al.*, 2022) contient des traductions multi-parallèles dans 128 langues, et a été étendu à deux dialectes du suisse allemand par Aepli *et al.* (2023). TICO-19 (Anastasopoulos *et al.*, 2020) est un jeu de traductions multi-parallèles créé dans le cadre de la pandémie de COVID-19 avec 35 langues au total dont quelques paires de dialectes comme le malay et l'indonésien. MADAR (Bouamor *et al.*, 2019) comprend deux sous-jeux de données multi-dialectales et multi-parallèles avec une échelle d'annotation dialectale précise (25 villes). Enfin, Alam *et al.* (2024) ont publié de nouveaux jeux d'exemples contrastifs pour plusieurs groupes de variétés (basques, bangladaises, kurdes, etc.), créés à partir d'atlas linguistiques, de corpus pré-existants et de productions audiovisuelles, et impliquant de nouvelles campagnes de traduction.

Malgré ces avancées dans la création de ressources multi-dialectales, il arrive que certains jeux de test contrastifs ne comportent pas de traductions parallèles vers des langues distantes (notamment l'anglais). Alam *et al.* (2024) proposent alors un système de pseudo-références, qui consiste à prendre une paire d'exemples contrastifs dont une des variétés est standard et bien dotée (par exemple l'italien standard et le sicilien), puis d'utiliser la TA de l'exemple en langue standard vers la langue cible (en pratique l'anglais) comme pseudo-référence pour estimer la qualité de la prédiction candidate à partir du texte en dialecte.

Pour conclure cette revue de littérature, nous rappelons que la différence entre langue et dialecte est principalement politique et que les variétés désignées par le terme "dialecte" regroupent des situations diverses en termes de proximité linguistique, de standardisation et de contact. Ces caractéristiques en font des objets d'étude spécifiques dans le domaine de la traduction automatique, permettant d'une

2. Ils utilisent le terme « contrastif » plutôt que « parallèle » dans le cadre de l'évaluation impliquant plusieurs dialectes pour se référer à la terminologie de la dialectologie comparative.

part de profiter des avancées du TAL pour les langues peu dotées, mais nécessitant d'autre part des ressources et techniques adaptées, en particulier pour les variétés non standardisées et les continuums dialectaux. Enfin, la direction de traduction vers des langues standard et bien dotées constitue la vaste majorité des études existantes, alors que la direction inverse requiert encore beaucoup d'attention, tant sur le plan des techniques d'entraînement ou d'adaptation de modèles que de leur évaluation.

3 Évaluation de modèles de TA depuis et vers des dialectes occitans

Les expériences préliminaires [3] qui suivent ont pour but de dresser un aperçu global des performances quantitatives et qualitatives des modèles de traduction existants pour la traduction depuis et vers le continuum dialectal de l'occitan. Étant donnée la forte présence de la variété d'occitan languedocien sur le web ainsi que dans les jeux de données incluant de l'occitan (ex. WikiMatrix, Schwenk *et al.*, 2019 ; NLLB, Costa-jussà *et al.*, 2022 ; SoftwaresOccitanTranslations [sic], Séguier & Lo Congrès, 2023b), il en résulte inévitablement que cette variété domine les données d'entraînement des modèles prenant en charge l'occitan. De plus, les modèles multilingues récents ont des capacités *zero-shot* qu'il serait intéressant d'étudier dans le cadre de la TA dialectale.

Dans ce contexte, nos expériences préliminaires ont pour objectif de chercher des premiers éléments de réponse aux questions de recherche suivantes, en prenant l'occitan comme cas d'usage.

— Peut-on observer des écarts de performances importants entre la TA depuis la variété la plus représentée et depuis d'autres variétés ?
— Dans quelle mesure les grands modèles de langue (LLMs pour l'anglais *large language models*) possèdent-ils des capacités à générer des traductions en occitan, en particulier vers des variétés spécifiques ?
— Quel impact les techniques de prédiction en contexte [4] ont-elles sur l'adaptation des LLMs aux spécificités dialectales en TA ?

Alors que la plupart des travaux en TA sont centrés sur l'anglais, nous choisissons de nous concentrer sur la traduction de l'occitan **vers et depuis le français**, cette direction étant plus pertinente d'un point de vue culturel. Une partie de nos expériences comprend tout de même l'anglais comme langue cible afin de comparer les performances entre le français et l'occitan comme langues sources.

3.1 Évaluation des traductions vers le français

Pour l'évaluation de la traduction de l'occitan vers le français, nous utilisons les métriques de surface BLEU et chrF++, ainsi que des modèles COMET *avec* (COMET) ou *sans* (COMET-QE) référence [5]. Prenant compte des observations de Alam *et al.* (2024), nous implémentons également leur variante d'utilisation du modèle COMET (avec référence), qui consiste à remplacer le texte source par une chaîne vide ($\text{COMET}_{\text{nosrc}}$).

3. Code, configurations et résultats détaillés : https://github.com/DEFI-COLaF/odt-benchmark-recital2025
4. En anglais : *In-Context Learning*
5. Détails à l'annexe A.3. Les scores COMET rapportés sont systématiquement multipliés par 100 pour en faciliter la lecture.

3.2 Écarts dialectaux dans la traduction vers l'anglais

La traduction vers l'anglais permet de mesurer des écarts dialectaux tels que définis par Faisal *et al.* (2024), de manière à estimer le chemin restant à parcourir en termes de performances de traduction par rapport à des langues bien dotées comme le français. En particulier, nous comparons les écarts de performance entre le français et chaque dialecte séparément, avant de prendre la moyenne de ces mesures pour estimer l'écart de performance moyen entre le français et l'occitan de manière générale. Nous calculons les écarts normalisés de manière globale avec l'équation 1 [6], en sélectionnant d'une part le meilleur système pour la traduction depuis le français (meilleur score avec la métrique choisie), et d'autre part le meilleur système pour la traduction depuis chacun des dialectes.

$$\mathcal{G}(f, v) = \frac{\mathcal{S}_{best_f}(f) - \mathcal{S}_{best_v}(v)}{\mathcal{S}_{best_f}(f)} \tag{1}$$

Lorsque le jeu de test ne comporte pas de traductions en anglais, nous proposons l'alternative suivante :

1. utiliser les prédictions depuis le français comme pseudo-références pour estimer les performances de traduction occitan-anglais, selon la procédure décrite dans Alam *et al.* (2024), ce qui donne : $\mathcal{S}_{ocen} = comet(MT_{ocen}, MT_{fren})$
2. utiliser les scores COMET-QE de la direction français-anglais comme base de comparaison, afin de pouvoir calculer des écarts dialectaux avec les scores $\text{COMET}_{\text{nosrc}}$ de la direction occitan-anglais

La comparaison entre deux métriques différentes est proposée d'une part parce que le modèle COMET-QE est très fiable pour la paire de langues français-anglais, et d'autre part parce que la métrique $\text{COMET}_{\text{nosrc}}$ permet d'évaluer en ne tenant compte que du texte en anglais, pour éviter la comparaison avec un score probablement biaisé vers le bas (Alam *et al.*, 2024).

Malgré des réserves quant à la précision des scores produits, nous utilisons cette approche comme alternative face au manque de données contrastives et parallèles, permettant ainsi dans notre cas d'étudier les écarts entre six variétés différentes et non pas seulement deux.

3.3 Évaluation des traductions vers l'occitan

Pour cette direction, nos expériences préliminaires ont pour objectif d'évaluer la capacité des modèles à générer des traductions : (i) en occitan et (ii) dans des variétés d'occitan spécifiques.

Les métriques existantes utilisées pour la direction occitan-français ne sont pas fiables pour l'évaluation des prédictions en occitan. Nous choisissons tout de même de les calculer pour avoir une vue d'ensemble ainsi que pour repérer des différences notables de performances entre différents modèles et stratégies d'implémentation.

Un des écueils des métriques généralistes, et en particulier celles qui reposent sur des modèles de langue, est l'absence de vérification de la langue générée. Vérifier que les prédictions sont effectivement en occitan est d'autant plus important que tous les systèmes évalués (hormis Revirada)

6. $\mathcal{S}_{best_f}(f)$ correspond au score $\mathcal{S}$ sur les exemples f en français du meilleur système $best_f$ pour la traduction depuis le français ; $\mathcal{S}_{best_v}(v)$ correspond au score $\mathcal{S}$ sur les exemples v dans la variété d'occitan donnée avec le meilleur système $best_v$ pour la traduction depuis cette variété.

sont des modèles multilingues entraînés sur des langues proches de l'occitan. Appliquer des modèles d'identification de langue (LID) permet de proposer une validation supplémentaire des performances des différents modèles.

Nous choisissons d'observer les scores produits par trois modèles de LID sur les prédictions des systèmes de TA ainsi que les textes de référence : lid218e (Costa-jussà *et al.*, 2022), Idiomata Cognitor (Galiano-Jiménez *et al.*, 2024) et GlotLID (Kargaran *et al.*, 2023). L'analyse est faite tout d'abord sur les scores de l'étiquette 'occitan' (LID_{oc}), puis sur les étiquettes de quelques langues proches de l'occitan (français, espagnol, catalan, italien). En plus des scores LID moyens sur les jeux de tests, nous comparons les résultats pour chaque modèle avec les segments de référence en calculant l'erreur quadratique moyenne (MSE).

Les modèles LID utilisés ayant probablement été entraînés principalement sur la variété languedocienne pour l'étiquette occitan, il est probable que les scores de cette étiquette baissent lorsque les phrases de référence sont dans un autre dialecte. En partant de cette hypothèse, nous utilisons la métrique MSE_{oc} afin d'estimer la capacité des systèmes de TA à générer des traductions dans des dialectes proches de ceux des segments de référence. Ainsi, pour un modèle de TA qui ne traduirait que vers la variété languedocienne, on s'attend à ce que les scores MSE_{oc} soient très petits (c.-à-d. très bons) lorsque les segments de référence sont en languedocien, et au contraire plus élevés (c.-à-d. moins bons) pour d'autres variétés, en particulier pour le gascon et l'aranais qui sont des variétés nettement distinctes du languedocien.

3.4 Données de test

Les données de test utilisées proviennent de deux sources : *Flores+* [7] et *Occitan Corpus from Lo Congrès news* [8] (Séguier & Lo Congrès, 2023a).

Flores Ce dataset contient une version en occitan (languedocien) provenant de Flores-200 (Costa-jussà *et al.*, 2024), et a été récemment étendu avec une version en aranais [9] (Perez-Ortiz *et al.*, 2024). C'est à notre connaissance le seul jeu de données entièrement parallèle et contrastif pour des variétés d'occitan, permettant ainsi de pouvoir comparer des scores d'évaluation au niveau du corpus entre les deux variétés disponibles. Pour nos expériences, nous choisissons le jeu *devtest* dans ses versions en français, anglais, languedocien et aranais.

LoCongresNews Ce corpus a été créé par *Lo Congrès permanent de la lenga occitana* [10] à partir de la section d'actualités de leur site. Il est distribué sous forme de corpus parallèle, où chaque phrase, étiquetée avec sa variété d'occitan, est associée à sa traduction en français. Les étiquettes correspondent aux six variétés suivantes (dans l'ordre alphabétique) : auvergnat (AUV), gascon (GAS), limousin (LIM), languedocien (LAN), provençal (PRO) et vivaro-alpin (VIV). Pour autant, les articles n'ayant pas été traduits entre plusieurs variétés, il ne s'agit pas d'un corpus d'exemples contrastifs, et donc les scores d'évaluation ne seront pas comparables entre variétés pour ce corpus.

7. https://huggingface.co/datasets/openlanguagedata/flores_plus
8. https://zenodo.org/records/8411197
9. La version utilisée pour nos expérience est celle après correction par les auteurs d'un problème d'alignement
10. https://locongres.org/

Pour nos expériences, nous appliquons quelques étapes de prétraitement au corpus d'origine, décrites dans l'annexe A.1. Le tableau 1 donne quelques statistiques après prétraitement.

	AUV	GAS	LIM	LAN	PRO	VIV	Total
#exemples	15	1 803	77	2 097	130	33	4 155
#tokens	322	38 329	1 516	40 434	2 507	630	83 738

TABLE 1 – Nombre d'exemples et de tokens par dialecte dans *LoCongresNews* après prétraitement.

3.5 Modèles

Notre sélection de modèles à inclure dans le banc d'évaluation inclut des approches de TA diverses : un système à base de règles, des modèles neuronaux avec architecture encodeur-décodeur, et des grands modèles de langue avec décodeur seul (LLM). Nous avons choisi de nous concentrer sur des modèles relativement petits, laissant à des travaux futurs les plus gros modèles de langue tels que les vainqueurs de la campagne d'évaluation "General Machine Translation" à WMT24 (Kocmi *et al.*, 2024). Ci-dessous, une brève description des systèmes utilisés pour nos expériences :

Revirada [11] Un service de TA basé sur Apertium [12], permettant de traduire entre le français et l'occitan, en sélectionnant entre deux dialectes : languedocien et gascon. Pour nos expériences, nous avons comparé trois implémentations de ce système :

— LANGUEDOCIEN : toutes les requêtes sont faites en sélectionnant le dialecte languedocien
— GASCON : toutes les requêtes sont faites en sélectionnant le dialecte gascon
— ADAPT : lorsque l'exemple (dans LoCongresNews) est étiqueté 'gascon', ce dialecte est sélectionné pour la requête, sinon c'est le dialecte languedocien qui est sélectionné

NLLB (Costa-jussà *et al.*, 2024) Un modèle neuronal de type Transformer encodeur-décodeur permettant de traduire entre 200 langues différentes dont l'occitan, et dans n'importe quelle direction. Un modèle de 54 milliard de paramètres est d'abord entraîné avec une architecture *Sparsely Gated Mixture of Experts*, puis distillé sur des modèles denses plus petits. La direction de traduction est passée au modèle : le code de la langue source est donné à l'encodeur, et le code de la langue cible est donné au décodeur. Nous avons utilisé le modèle NLLB-200-distilled-600M [13] (NLLB-600) pour nos expériences.

Google Translate [14] Bien qu'il ne s'agisse pas d'un modèle open-source, nous avons choisi d'inclure ce service de traduction neuronale en ligne, car l'occitan y a été ajouté comme langue disponible [15] en juin 2024.

11. https://revirada.eu/. Requêtes par API.
12. https://github.com/apertium/apertium-oci-fra
13. https://huggingface.co/facebook/nllb-200-distilled-600M. Ce modèle a été choisi par Wikimedia Foundation dans son outil MinT (https://translate.wmcloud.org/) pour la traduction de et vers l'occitan.
14. Requêtes via l'API Google Cloud Translation.
15. https://support.google.com/translate/answer/15139004?hl=fr

Aya-23 (Aryabumi *et al.*, 2024) Un LLM spécialisé pour la génération multilingue, dont la TA. Bien que l'occitan ne fasse pas partie des 23 langues explicitement utilisées pour l'entraînement, nous avons choisi ce modèle pour l'exploration de la traduction de l'occitan en mode *zero-shot* et *few-shot*, dont les performances pourraient être favorisées par la forte présence de français, d'espagnol et d'italien dans les données d'entraînement. Nous avons utilisé la version Aya-23-8B pour nos expériences.

Llama3 (Dubey *et al.*, 2024) Un LLM généraliste de taille modeste, avec des capacités multilingues. Huit langues sont officiellement prises en charge, dont le français, l'espagnol et l'italien, mais pas l'occitan. Nous avons choisi d'intégrer le modèle Llama3.1-8B-Instruct à nos expériences pour avoir un point de comparaison avec un LLM généraliste de taille comparable avec les autres modèles sélectionnés, et également pour des facilités d'implémentation pour cette phase d'expériences préliminaires [16].

Pour les LLMs (Aya-23 et Llama3), nous avons comparé plusieurs types de prompts (détails d'implémentation à l'annexe A.2) :

— LANG-TGT : seul le nom de la langue cible est donné, mais pas celui de la langue source
— LANG-SRC : les noms des langues source et cibles sont donnés
— DIA-TGT [17] : seul le nom de la langue cible est donné, et s'il s'agit de l'occitan, le nom du dialecte est précisé (par exemple *Occitan 'limousin'*)
— DIA-SRC : les noms des langues source et cible sont donnés, et pour l'occitan, le nom du dialecte est précisé

De plus, nous avons testé trois approches avec ou sans exemples, pour chacun des types de prompt indiqués ci-dessus :

— 0-SHOT : pas d'exemples
— 3-SHOTS : trois exemples fixes dont la version occitane est en languedocien
— 3-SHOTS-ADAPT : trois exemples fixes dont la version occitane est dans le même dialecte que la traduction de référence de l'exemple à traduire

3.6 Résultats

Le nombre de métriques et de configurations de modèles étant trop grand pour que tous les scores puissent être présentés ici, les tableaux de résultats complets sont mis en annexes A.4. De par la proximité linguistique entre l'occitan et le français, ainsi que la présence d'éléments en français (notamment des titres d'ouvrages) dans le jeu LoCongresNews, nous reportons également les scores pour la baseline COPIE, qui consiste à générer des traductions en recopiant le texte source.

3.6.1 Résultats depuis l'occitan

Les résultats des évaluations de traduction de l'occitan vers le français laissent se dessiner quelques tendances. D'une part, le système Google Translate semble être le plus performant (voir table 2), avec les meilleurs scores sur Flores pour toutes les métriques calculées, et les meilleurs scores

16. Les modèles Llama3 de tailles supérieures requièrent plusieurs GPUs à l'inférence.
17. Pour la direction occitan-français, la stratégie DIA-TGT est identique à LANG-TGT

avec les métriques COMET sur LoCongresNews. D'autre part, les scores de tous les modèles sont systématiquement plus élevés sur LoCongresNews que sur Flores. Cela peut sembler surprenant sachant que le domaine *Wiki* de Flores est un domaine majeur des données d'entraînement de beaucoup de modèles de langues. Cependant, les exemples de Flores sont aussi particulièrement longs et techniques, alors que LoCongresNews contient plus de diversité de longueurs et plus de segments courts. Aussi, la présence d'éléments en français dans le texte source (occitan) de LoCongresNews se traduit par des scores déjà élevés pour la *baseline* COPIE.

	Flores			LoCongresNews		
	BLEU	chrF++	$\text{COMET}_{\text{nosrc}}$	BLEU	chrF++	$\text{COMET}_{\text{nosrc}}$
Baseline COPIE	2,99	26,07	46,05	15,48	40,71	57,51
Revirada-ADAPT	6,25	48,57	40,10	**65,91**	**81,17**	84,20
NLLB-600	31,46	55,09	75,65	56,88	74,88	82,66
Google Translate	**41,33**	**63,16**	**83,20**	57,15	75,17	**86,10**
Aya23-8B						
0-SHOT LANG-SRC	24,47	49,40	68,80	47,05	66,23	74,05
3-SHOTS LANG-SRC	25,17	48,62	72,23	36,77	53,73	73,74
Llama3.1-8B Instruct						
0-SHOT LANG-SRC	29,69	54,43	76,37	59,80	75,80	82,17
3-SHOTS LANG-SRC	30,09	55,05	77,78	57,75	73,94	82,08

TABLE 2 – Extrait des scores des modèles pour la traduction de l'occitan vers le français (dialectes combinés). Détail complet des scores à l'annexe A.4.1 (table 10 et table 11).

Nos observations des prédictions du système Revirada révèlent de nombreux mots partiellement traduits sur Flores, ce qui a pour conséquence de grandes pénalités, surtout pour les scores BLEU. Cette particularité de Revirada est au contraire à son avantage pour les segments de LoCongresNews qui requièrent une traduction plus légère, plaçant cette fois-ci le système en premier pour les métriques BLEU et chrF++.

Concernant la traduction avec les LLMs, la comparaison entre les modes 0-SHOT et 3-SHOTS ne laisse pas apparaître de différence nettement visible sur les scores tous dialectes confondus, si ce n'est une potentielle légère amélioration pour les deux modèles Aya-23 et Llama3 sur le jeu Flores. Quelques améliorations plus marquées sont observables dans les scores $\text{COMET}_{\text{nosrc}}$ par dialectes (cf. table 15), notamment pour le modèle Aya-23 sur les dialectes aranais et auvergnat. Concernant les stratégies de prompts liées aux noms des langues, il apparaît nettement que l'indication du nom de la variété d'occitan dans le prompt (stratégie DIA-SRC) impacte négativement les résultats avec Aya-23.

En regardant le détail des scores en fonction des variétés d'occitan données en entrée aux modèles (cf. table 3), il n'apparaît pas d'écart important entre les différentes variétés à notre disposition, pour le modèle le plus performant (Google Translate). Certes, le score $\text{COMET}_{\text{nosrc}}$ pour le languedocien est supérieur à celui pour l'aranais dans l'évaluation sur Flores (de plus dans une configuration contrastive). Cependant, les scores avec la *baseline* COPIE sont eux aussi plus élevés pour le languedocien, et de plus, les traductions de la version aranaise ayant été produites à partir du catalan, l'alignement entre les versions aranais-français est probablement de moins bonne qualité que pour la paire languedocien-français.

Pour autant, les écarts de performances entre dialectes diffèrent d'un modèle à l'autre, laissant

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	43,13	48,96	54,95	54,21	58,81	60,07	61,15	58,75
Revirada-ADAPT	51,34	76,50	75,94	83,26	74,15	85,86	80,38	76,65
NLLB-600	70,73	80,50	83,36	80,34	79,03	84,76	82,57	83,50
Google Translate	**81,82**	**84,53**	**86,74**	**85,37**	**83,45**	**86,84**	**86,13**	**86,89**
Aya23-8B								
3-SHOTS LANG-SRC	68,31	76,14	77,83	69,13	74,34	77,56	74,83	75,31
Llama3.1-8B Instruct								
3-SHOTS LANG-SRC	75,08	80,66	82,90	80,51	80,35	83,52	80,83	84,94

TABLE 3 – Scores $COMET_{nosrc}$ par dialecte (Flores/LoCongresNews) pour la traduction de l'occitan vers le français. Les meilleurs scores par dialecte sont mis en gras.

apparaître certaines faiblesses pour certaines combinaisons de modèles et de dialectes de l'occitan. Avec Revirada, la différence de scores $COMET_{nosrc}$ sur Flores entre languedocien et aranais ($\Delta = 25{,}16$ avec la version LANGUEDOCIEN du système) est bien plus importante qu'avec la baseline ($\Delta = 5{,}83$), avec une nette amélioration lorsque la version GASCON du modèle est utilisée (ce qui est cohérent, le dialecte aranais pouvant être considéré comme une variante du gascon). Le modèle NLLB-600 présente également une grande différence de scores ($\Delta = 9{,}77$) entre ces deux dialectes. Concernant les sous-jeux par dialectes dans LoCongresNews, alors que les scores $COMET_{nosrc}$ de Google Translate sont très proches (autour de 86, sauf pour le limousin légèrement inférieur), Revirada montre une différence de performances entre les dialectes, avec des scores élevés pour le languedocien (85,86) et le gascon (83,26) d'une part, et des scores plus faibles pour les autres variétés (le plus faible étant à nouveau le limousin avec 74,15).

3.6.2 Écarts des performances vers l'anglais

	Score $COMET_{nosrc}$	Écart
Français	86,82	-
Aranais	83,31	4,04
Languedocien	86,94	-0,15
Moyenne	85,1	2,0

TABLE 4 – Écarts dialectaux dans les traductions de Flores de dialectes occitan vers l'anglais, par rapport à la traduction depuis le français. Système de TA : Google Translate.

	Score FR^{QE}	Score OC	Écart
AUV	84,21	83,64	0,68
GAS	82,96	83,14	-0,21
LIM	81,09	81,00	0,12
LAN	83,07	84,45	-1,66
PRO	82,91	83,71	-0,96
VIV	83,16	83,65	-0,59
Moyenne	82,90	83,26	-0,44

TABLE 5 – Écarts dialectaux dans les traductions des sous-jeux par dialecte dans LoCongresNews. Direction : français/occitan vers anglais. La métrique utilisée est COMET-QE pour le français, et $COMET_{nosrc}$ pour l'occitan. Système de TA : Google Translate.

Le modèle Google Translate est celui qui a obtenu les meilleurs scores $\text{COMET}_{\text{nosrc}}$ pour la traduction vers l'anglais depuis chaque variété d'occitan ainsi que depuis le français. Les résultats de notre analyse des écarts dialectaux (cf. table 4 pour Flores et table 5 pour LoCongresNews) montrent que les performances entre la traduction depuis le français et depuis l'occitan sont très proches, avec même parfois des scores plus élevés depuis l'occitan que depuis le français (-0,44 en moyenne pour LoCongresNews). Une évaluation humaine resterait cependant nécessaire pour tirer des conclusions certaines et pour comparer les types d'erreurs fréquentes entre ces deux directions de traduction.

3.6.3 Résultats vers l'occitan

Contrairement à la traduction vers les langues standard français et anglais, les résultats de la traduction du français vers l'occitan montrent des performances plus faibles et disparates en fonction des systèmes.

	Flores			LoCongresNews		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Baseline COPIE	2,98	26,75	64,65	15,47	41,65	72,35
Revirada-ADAPT	17,97	44,80	63,71	**60,94**	**79,22**	**79,53**
NLLB-600	16,32	42,49	64,49	41,46	65,35	75,47
Google Translate	**19,89**	**46,04**	**65,18**	37,05	62,49	74,73
Aya23-8B						
0-SHOT LANG-SRC	4,87	30,87	63,27	18,97	46,95	71,22
0-SHOT DIA-SRC	5,07	31,72	62,47	19,20	47,49	70,29
3-SHOTS DIA-SRC	5,78	31,79	61,83	15,07	39,41	67,41
3-SHOTS-ADAPT DIA-SRC	5,80	32,04	62,05	14,27	37,49	66,34
Llama3,1-8B Instruct						
0-SHOT LANG-SRC	12,74	39,84	63,85	37,67	62,20	73,90
0-SHOT DIA-SRC	12,72	39,92	63,59	25,22	56,67	67,35
3-SHOTS DIA-SRC	13,45	40,84	64,41	38,45	63,10	73,28
3-SHOTS-ADAPT DIA-SRC	13,49	40,82	64,46	38,77	63,36	73,40

TABLE 6 – Extrait des scores des modèles pour la traduction du français vers l'occitan (dialectes combinés). Détail complet des scores à l'annexe A.4.2 (tables 17 et 18)

À travers les scores globaux des métriques de traduction (cf. table 6), le système Revirada domine largement sur LoCongresNews, et Google Translate obtient la première place sur Flores en dépassant de peu Revirada et NLLB-600. De manière plus surprenante, Llama3 obtient des scores très légèrement inférieurs alors que le modèle ne supporte pas officiellement l'occitan.

Parmi les métriques, on observe que les scores BLEU sont globalement bas, au contraire de ceux basés sur COMET, qui sont plutôt élevés et numériquement rapprochés. Les scores donnés par la métrique COMET-QE semblent être inutilisables pour notre analyse, car ils sont les plus faibles pour les modèles a priori meilleurs, et les plus élevés pour la baseline COPIE, c'est-à-dire lorsque la traduction candidate ressemble le plus au texte source.

Les scores des métriques de traduction révèlent des performances plus élevées lorsque les segments de référence sont en dialecte languedocien, par rapport aux autres dialectes (cf. tables 19 et 21) – à

nuancer cependant car les sous-jeux de LoCongresNews ne sont pas contrastifs. Les différences entre dialectes sont particulièrement marquées avec l'aranais sur Flores (26 points BLEU de différence avec Google Translate), et avec le gascon sur LoCongresNews (20 points BLEU avec NLLB-600).

La stratégie DIA-SRC est intéressante à analyser pour Llama3, avec au départ une forte dégradation des scores en mode 0-SHOT, révélant que le nom du dialecte perturbe les prédictions du modèle. Cependant, cette perturbation semble être réparée successivement par les modes 3-SHOT puis 3-SHOT-ADAPT, ce qui se traduit par une évolution systématiquement positive des scores, sur les deux jeux de test et sur tous les sous-jeux par dialectes, toutefois sans atteindre les meilleurs scores parmi toutes les stratégies testées pour ce modèle.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Revirada-ADAPT	0,43	**0,15**	**<0,01**	**0,85**	**1,16**	**1,14**	**2,36**	**0,05**
NLLB-600	**0,40**	0,23	**<0,01**	2,10	1,84	1,72	2,63	2,86
Google Translate	0,44	0,19	**<0,01**	2,36	1,71	2,24	4,87	1,67
Aya23-8B								
0-SHOT DIA-SRC	73,26	80,83	69,93	91,76	49,55	55,72	51,06	50,93
3-SHOTS DIA-SRC	65,39	71,36	37,39	49,19	40,14	43,62	34,60	39,81
3-SHOTS-ADAPT DIA-SRC	67,93	71,36	18,52	44,81	38,06	43,62	32,97	33,27
Llama3.1-8B Instruct								
0-SHOT DIA-SRC	0,88	0,98	0,03	11,78	13,00	11,75	15,43	8,18
3-SHOTS DIA-SRC	0,45	0,22	**<0,01**	2,41	2,08	2,85	4,13	1,60
3-SHOTS-ADAPT DIA-SRC	0,44	0,22	**<0,01**	2,25	2,16	2,85	3,18	1,68

TABLE 7 – Scores MSE_{oc} (multipliés par 100) sur entre les traduction candidates (Flores/LoCongresNews) et leurs segments de référence, pour la direction de traduction français vers l'occitan. Résultats en fonction du dialecte du segment de référence. Modèle LID utilisé : lid218e. Détail complet des scores à l'annexe A.4.2, table 25.

Les résultats de nos analyses via LID donnent des scores très différents entre les trois modèles d'identification employés (cf. tables 22, 23 et 24). Avec les modèles lid218e et GlotLID, Revirada obtient les scores MSE_{oc} les plus faibles pour tous les dialectes de LoCongresNews, alors qu'Idiomata Cognitor donne cette première place seulement sur les sous-jeux gascon, languedocien et vivaro-alpin. Les trois modèles s'accordent cependant sur l'amélioration du score MSE_{oc} lorsque la stratégie de Revirada correspondant au dialecte (gascon ou languedocien) est utilisée. Pour les autres dialectes, on observe des scores MSE_{oc} particulièrement faibles pour l'auvergnat - sûrement lié aux scores LID de référence très élevés - et au contraire des scores MSE_{oc} particulièrement élevés pour le provençal, surtout avec Google Translate - là où les scores LID de référence sont bas.

Concernant les LLMs, Llama3 obtient également de très bons scores LID_{oc} et MSE_{oc}, surpassant parfois les modèles de TA qui supportent officiellement l'occitan, en fonction des dialectes et des métriques LID utilisées. Les analyses des différentes étiquettes LID permettent d'expliquer les mauvaises performances du modèle Aya-23 pour la direction français-occitan, en révélant que le modèle génère principalement des traductions dans des langues proches de l'occitan, en particulier le catalan et le français (cf. tables 28 et 29). Pour autant, ces analyses montrent aussi que les stratégies de prédiction en *few-shot* permettent de mieux contrôler la langue générée, pour les modèles Aya-23 comme pour Llama3.

Enfin, contrairement à notre hypothèse, les scores LID_{oc} ne sont pas meilleurs pour les sous-jeux en languedocien que pour les autres dialectes. Cela s'explique d'une part par la forte présence d'alternance codique avec le français dans le jeu LoCongresNews, et d'autre part par la présence probable d'exemples valides pour plusieurs variétés d'occitan, perturbant la distinction entre les sous-jeux par dialecte. En conséquence, les scores MSE_{oc} ne permettent pas d'extraire de grandes tendances pour estimer la capacité des modèles (notamment les LLMs) à générer différentes variétés d'occitan.

4 Conclusion

Après avoir défini la notion de dialecte et décrit la situation dialectale de l'occitan, nous avons établi un état de l'art des recherches pour la traduction et son évaluation en contexte dialectal. Les approches existantes couvrent à la fois des modèles à base de règles - qui malgré leur ancienneté restent parfois utilisés dans le contexte de langues similaires -, des modèles statistiques, et des modèles neuronaux. Ces derniers modèles sont particulièrement plébiscités pour leurs capacités multilingues favorables aux langues peu dotées, ce qui est le cas de la plupart des dialectes dans le monde. Alors que les techniques de traduction pour la direction *dialecte vers langue standard apparentée* ont été largement explorées (notamment la normalisation et l'ajout de bruit dans les données), la traduction vers des dialectes non standardisés et vers des continuum dialectaux est encore dans une phase exploratoire.

Nos expériences préliminaires de la traduction depuis et vers l'occitan montrent que les modèles existants offrent globalement de bonnes performances pour la traduction vers le français et l'anglais, même si des écarts existent en fonction du dialecte source, le languedocien figurant parmi les variétés les mieux traduites. En revanche, les résultats de la traduction vers l'occitan sont plus faibles et diffèrent en fonction des modèles. Bien que plusieurs modèles neuronaux dont un LLM atteignent de bons scores, seul Revirada – un modèle ancien basé sur des règles linguistiques – génère des traductions dont la langue est plus proche des références selon les scores LID, et ce, pour toutes les variétés étudiées.

Certaines des principales limitations des travaux présentés ici proviennent des données utilisées. Flores est peu représentatif de la diversité dialectale du continuum occitan, et présente des limites de contrastivité entre les versions aranaise et languedocienne. LoCongresNews est plus représentatif, avec ses 6 dialectes couverts, mais contient beaucoup de données en français, et son manque de contrastivité entre les sous-jeux par dialecte est un frein important pour comparer de manière fiable les performances d'un dialecte à l'autre. De plus, nos résultats ont tous été obtenus sur des jeux de tests utilisant la graphie classique, alors que d'autres graphies existent, notamment la graphie mistralienne. Les métriques d'évaluation posent également des difficultés, et celles que nous avons utilisées sont relativement élémentaires. Dans la direction *langue standard apparentée vers dialecte*, l'évaluation des contenus dialectaux est particulièrement délicate, car les métriques automatiques actuelles ne permettent pas de prendre en compte les traits dialectaux des contenus générés, que ce soit en termes de robustesse à la variation dialectale ou pour vérifier la cohérence avec le dialecte souhaité.

Pour terminer, nous souhaitons insister sur le fait que le travail sur des langues minoritaires ne peut se faire sans l'implication d'une communauté de locuteurs. Nous sommes en lien avec différents acteurs de la communauté occitaniste, et nous prévoyons déjà de les consulter pour suite de nos travaux, notamment pour la création de jeux de test et pour implémenter des processus d'évaluation humaine de modèles de traduction vers le continuum occitan.

Remerciements

Je tiens à remercier mes encadrants Benoît Sagot, Rachel Bawden et Thibault Clérice pour leur aide durant toute la préparation de cet article, ainsi qu'Armel Zebaze pour son aide pour l'utilisation de Llama3. Merci également à Tristan Gahús pour la traduction d'exemples pour l'inférence en few-shot, et à Lo Congrès permanent de la lenga occitana (en particulier Aure Séguier) pour la mise à disposition de leur corpus. Ce travail a été financé par Inria dans le cadre du DÉFI Inria COLaF, ainsi que par les chaires de Rachel Bawden et Benoît Sagot au sein de l'institut PRAIRIE financé par l'ANR dans le cadre du programme "Investissements d'avenir" sous la référence ANR-19-P3IA-0001.

Références

ABE K., MATSUBAYASHI Y., OKAZAKI N. & INUI K. (2018). Multi-dialect Neural Machine Translation and Dialectometry. *Information and Computation*.

AEPLI N., AMRHEIN C., SCHOTTMANN F. & SENNRICH R. (2023). A Benchmark for Evaluating Machine Translation Metrics on Dialects without Standard Orthography. In P. KOEHN, B. HADDOW, T. KOCMI & C. MONZ, Éds., *Proceedings of the Eighth Conference on Machine Translation*, p. 1045–1065, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wmt-1.99.

AEPLI N. & SENNRICH R. (2022). Improving Zero-Shot Cross-lingual Transfer Between Closely Related Languages by Injecting Character-Level Noise. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Findings of the Association for Computational Linguistics : ACL 2022*, p. 4074–4083, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.321.

ALAM M. M. I., AHMADI S. & ANASTASOPOULOS A. (2024). CODET : A Benchmark for Contrastive Dialectal Evaluation of Machine Translation. arXiv :2305.17267 [cs], DOI : 10.48550/arXiv.2305.17267.

ALTINTAS K. & CICEKLI I. (2003). A Machine Translation System Between a Pair of Closely Related Languages. *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*.

AMINIAN M., GHONEIM M. & DIAB M. (2014). Handling OOV Words in Dialectal Arabic to English Machine Translation. In P. NAKOV, P. OSENOVA & C. VERTAN, Éds., *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, p. 99–108, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/W14-4213.

ANASTASOPOULOS A., CATTELAN A., DOU Z.-Y., FEDERICO M., FEDERMANN C., GENZEL D., GUZMÁN F., HU J., HUGHES M., KOEHN P., LAZAR R., LEWIS W., NEUBIG G., NIU M., ÖKTEM A., PAQUIN E., TANG G. & TUR S. (2020). TICO-19 : the Translation Initiative for COvid-19. In K. VERSPOOR, K. B. COHEN, M. CONWAY, B. DE BRUIJN, M. DREDZE, R. MIHALCEA & B. WALLACE, Éds., *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.nlpcovid19-2.5.

ANASTASOPOULOS A., LUI A., NGUYEN T. Q. & CHIANG D. (2019). Neural Machine Translation of Text from Non-Native Speakers. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), p. 3070–3080, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1311.

ARMENTANO I OLLER C. & FORCADA M. L. (2006). Open source machine translation between small languages : Catalan and Aranese Occitan. In *Proceedings of the 5th Workshop on Strategies for developing machine translation for minority languages*, p. 51–54, Genoa, Italy.

ARYABUMI V., DANG J., TALUPURU D., DASH S., CAIRUZ D., LIN H., VENKITESH B., SMITH M., CAMPOS J. A., TAN Y. C., MARCHISIO K., BARTOLO M., RUDER S., LOCATELLI A., KREUTZER J., FROSST N., GOMEZ A. N., BLUNSOM P., FADAEE M., ÜSTÜN A. & HOOKER S. (2024). Aya 23 : Open weight releases to further multilingual progress. *CoRR*, **abs/2405.15032**.

BAFNA N., CHANG E., ROBINSON N. R., MORTENSEN D. R., MURRAY K., YAROWSKY D. & SIRIN H. (2025). DialUp ! Modeling the Language Continuum by Adapting Models to Dialects and Dialects to Models. arXiv :2501.16581 [cs], DOI : 10.48550/arXiv.2501.16581.

BAFNA N., MURRAY K. & YAROWSKY D. (2024). Evaluating Large Language Models along Dimensions of Language Variation : A Systematik Invesdigatiom uv Cross-lingual Generalization. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 18742–18762, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.1044.

BAPNA A., CASWELL I., KREUTZER J., FIRAT O., ESCH D. V., SIDDHANT A., NIU M., BALJEKAR P., GARCIA X., MACHEREY W., BREINER T., AXELROD V., RIESA J., CAO Y., CHEN M. X., MACHEREY K., KRIKUN M., WANG P., GUTKIN A., SHAH A., HUANG Y., CHEN Z., WU Y. & HUGHES M. (2022). Building Machine Translation Systems for the Next Thousand Languages. arXiv :2205.03983 [cs], DOI : 10.48550/arXiv.2205.03983.

BELINKOV Y. & BISK Y. (2018). Synthetic and Natural Noise Both Break Neural Machine Translation. arXiv :1711.02173 [cs], DOI : 10.48550/arXiv.1711.02173.

BLASCHKE V., SCHÜTZE H. & PLANK B. (2023). Does Manipulating Tokenization Aid Cross-Lingual Transfer ? A Study on POS Tagging for Non-Standardized Languages. In Y. SCHERRER, T. JAUHIAINEN, N. LJUBEŠIĆ, P. NAKOV, J. TIEDEMANN & M. ZAMPIERI, Éds., *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, p. 40–54, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.vardial-1.5.

BOUAMOR H., HASSAN S. & HABASH N. (2019). The MADAR shared task on Arabic fine-grained dialect identification. In W. EL-HAJJ, L. H. BELGUITH, F. BOUGARES, W. MAGDY, I. ZITOUNI, N. TOMEH, M. EL-HAJ & W. ZAGHOUANI, Éds., *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, p. 199–207, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4622.

BRAHMA M., MAURYA K. & DESARKAR M. (2023). SelectNoise : Unsupervised Noise Injection to Enable Zero-Shot Machine Translation for Extremely Low-resource Languages. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 1615–1629, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.109.

BRENNAN S. C. (2024). 21 La situation sociolinguistique de l'occitan du début du XXe siècle à nos jours. In L. ESHER & J. SIBILLE, Éds., *Manuel de linguistique occitane*, p. 593–621. De Gruyter. DOI : 10.1515/9783110733433.

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D. M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S.,

RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language Models are Few-Shot Learners. arXiv :2005.14165 [cs], DOI : 10.48550/arXiv.2005.14165.

CHAMBERS J. K. & TRUDGILL P. (1998). *Dialectology*. Cambridge University Press. Google-Books-ID : 9bYV43UhKssC.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J., SUN A. Y., WANG S., WENZEK G., YOUNGBLOOD A., AKULA B., BARRAULT L., GONZALEZ G. M., HANSANTI P., HOFFMAN J., JARRETT S., SADAGOPAN K. R., ROWE D., SPRUIT S., TRAN C., ANDREWS P., AYAN N. F., BHOSALE S., EDUNOV S., FAN A., GAO C., GOSWAMI V., GUZMÁN F., KOEHN P., MOURACHKO A., ROPERS C., SALEEM S., SCHWENK H. & WANG J. (2022). No language left behind : Scaling human-centered machine translation. *CoRR*, **abs/2207.04672**. DOI : 10.48550/ARXIV.2207.04672.

COSTA-JUSSÀ M. R. (2017). Why Catalan-Spanish Neural Machine Translation ? Analysis, comparison and combination with standard Rule and Phrase-based technologies. In P. NAKOV, M. ZAMPIERI, N. LJUBEŠIĆ, J. TIEDEMANN, S. MALMASI & A. ALI, Éds., *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, p. 55–62, Valencia, Spain : Association for Computational Linguistics. DOI : 10.18653/v1/W17-1207.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J., SUN A., WANG S., WENZEK G., YOUNGBLOOD A., AKULA B., BARRAULT L., GONZALEZ G. M., HANSANTI P., HOFFMAN J., JARRETT S., SADAGOPAN K. R., ROWE D., SPRUIT S., TRAN C., ANDREWS P., AYAN N. F., BHOSALE S., EDUNOV S., FAN A., GAO C., GOSWAMI V., GUZMÁN F., KOEHN P., MOURACHKO A., ROPERS C., SALEEM S., SCHWENK H., WANG J. & NLLB TEAM (2024). Scaling neural machine translation to 200 languages. *Nature*, **630**(8018), 841–846. Publisher : Nature Publishing Group, DOI : 10.1038/s41586-024-07335-x.

COSTA-JUSSÀ M. R., ZAMPIERI M. & PAL S. (2018). A Neural Approach to Language Variety Translation. In M. ZAMPIERI, P. NAKOV, N. LJUBEŠIĆ, J. TIEDEMANN, S. MALMASI & A. ALI, Éds., *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, p. 275–282, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., YANG A., FAN A., GOYAL A., HARTSHORN A., YANG A., MITRA A., SRAVANKUMAR A., KORENEV A., HINSVARK A., RAO A., ZHANG A., RODRIGUEZ A., GREGERSON A., SPATARU A., ROZIÈRE B., BIRON B., TANG B., CHERN B., CAUCHETEUX C., NAYAK C., BI C., MARRA C., MCCONNELL C., KELLER C., TOURET C., WU C., WONG C., FERRER C. C., NIKOLAIDIS C., ALLONSIUS D., SONG D., PINTZ D., LIVSHITS D., ESIOBU D., CHOUDHARY D., MAHAJAN D., GARCIA-OLANO D., PERINO D., HUPKES D., LAKOMKIN E., ALBADAWY E., LOBANOVA E., DINAN E., SMITH E. M., RADENOVIC F., ZHANG F., SYNNAEVE G., LEE G., ANDERSON G. L., NAIL G., MIALON G., PANG G., CUCURELL G., NGUYEN H., KOREVAAR H., XU H., TOUVRON H., ZAROV I., IBARRA I. A., KLOUMANN I. M., MISRA I., EVTIMOV I., COPET J., LEE J., GEFFERT J., VRANES J., PARK J., MAHADEOKAR J., SHAH J., VAN DER LINDE J., BILLOCK J., HONG J., LEE J., FU J., CHI J., HUANG J., LIU J., WANG J., YU J., BITTON J., SPISAK J., PARK J., ROCCA J., JOHNSTUN J., SAXE J., JIA J., ALWALA K. V., UPASANI K., PLAWIAK K., LI K., HEAFIELD K., STONE K. & ET AL. (2024). The llama 3 herd of models. *CoRR*, **abs/2407.21783**. DOI : 10.48550/ARXIV.2407.21783.

ESHER L. & SIBILLE J., Éds. (2024). *Manuel de linguistique occitane*. De Gruyter. DOI : 10.1515/9783110733433.

FAISAL F., AHIA O., SRIVASTAVA A., AHUJA K., CHIANG D., TSVETKOV Y. & ANASTASOPOULOS A. (2024). DIALECTBENCH : A NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. arXiv :2403.11009 [cs], DOI : 10.48550/arXiv.2403.11009.

FANCELLU F., WAY A. & O'BRIEN M. (2014). Standard language variety conversion for content localisation via SMT. In M. CETTOLO, M. FEDERICO, L. SPECIA & A. WAY, Éds., *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, p. 143–149, Dubrovnik, Croatia : European Association for Machine Translation.

FEDERMANN C., KOCMI T. & XIN Y. (2022). NTREX-128 – News Test References for MT Evaluation of 128 Languages. In K. AHUJA, A. ANASTASOPOULOS, B. PATRA, G. NEUBIG, M. CHOUDHURY, S. DANDAPAT, S. SITARAM & V. CHAUDHARY, Éds., *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, p. 21–24, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2022.sumeval-1.4.

FREITAG M., FOSTER G., GRANGIER D., RATNAKAR V., TAN Q. & MACHEREY W. (2021). Experts, Errors, and Context : A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, **9**, 1460–1474. DOI : 10.1162/tacl_a_00437.

GALIANO-JIMÉNEZ A., SÁNCHEZ-MARTÍNEZ F. & PÉREZ-ORTIZ J. A. (2024). Idiomata cognitor.

GARCIA X. & FIRAT O. (2022). Using natural language prompts for machine translation. arXiv :2202.11822 [cs], DOI : 10.48550/arXiv.2202.11822.

GRAHAM Y., BALDWIN T., MOFFAT A. & ZOBEL J. (2013). Continuous Measurement Scales in Human Evaluation of Machine Translation. In A. PAREJA-LORA, M. LIAKATA & S. DIPPER, Éds., *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 33–41, Sofia, Bulgaria : Association for Computational Linguistics.

HADDOW B., BAWDEN R., BARONE A. V. M., HELCL J. & BIRCH A. (2022). Survey of Low-Resource Machine Translation. *Computational Linguistics*, **48**(3), 673–732. DOI : 10.1162/coli_a_00446.

HADDOW B., HERNÁNDEZ A., NEUBARTH F. & TROST H. (2013). Corpus development for machine translation between standard and dialectal varieties. In C. VERTAN, M. SLAVCHEVA & P. OSENOVA, Éds., *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*, p. 7–14, Hissar, Bulgaria : INCOMA Ltd. Shoumen, BULGARIA.

HASSANI H. (2017). Kurdish Interdialect Machine Translation. In P. NAKOV, M. ZAMPIERI, N. LJUBEŠIĆ, J. TIEDEMANN, S. MALMASI & A. ALI, Éds., *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, p. 63–72, Valencia, Spain : Association for Computational Linguistics. DOI : 10.18653/v1/W17-1208.

HEIGOLD G., VARANASI S., NEUMANN G. & VAN GENABITH J. (2018). How Robust Are Character-Based Word Embeddings in Tagging and MT Against Wrod Scramlbing or Randdm Nouse ? In C. CHERRY & G. NEUBIG, Éds., *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1 : Research Track)*, p. 68–80, Boston, MA : Association for Machine Translation in the Americas.

HER W.-H. & KRUSCHWITZ U. (2024). Investigating Neural Machine Translation for Low-Resource Languages : Using Bavarian as a Case Study. In M. MELERO, S. SAKTI & C. SORIA, Éds., *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, p. 155–167, Torino, Italia : ELRA and ICCL.

HONNET P.-E., POPESCU-BELIS A., MUSAT C. & BAERISWYL M. (2018). Machine Translation of Low-Resource Spoken Dialects : Strategies for Normalizing Swiss German. arXiv :1710.11035 [cs], DOI : 10.48550/arXiv.1710.11035.

HOPTON Z. & AEPLI N. (2024). Modeling Orthographic Variation in Occitan's Dialects. In Y. SCHERRER, T. JAUHIAINEN, N. LJUBEŠIĆ, M. ZAMPIERI, P. NAKOV & J. TIEDEMANN, Éds., *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, p. 78–88, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.vardial-1.6.

JEBLEE S., FEELY W., BOUAMOR H., LAVIE A., HABASH N. & OFLAZER K. (2014). Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic. In N. HABASH & S. VOGEL, Éds., *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, p. 196–206, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/W14-3627.

JERPELEA A.-I., RĂDOI A.-C. & NISIOI S. (2024). Dialectal and Low Resource Machine Translation for Aromanian. arXiv :2410.17728, DOI : 10.48550/arXiv.2410.17728.

JONES A., CASWELL I., FIRAT O. & SAXENA I. (2023). GATITOS : Using a New Multilingual Lexicon for Low-resource Machine Translation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 371–405, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.26.

JOSHI P., SANTY S., BUDHIRAJA A., BALI K. & CHOUDHURY M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6282–6293, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.560.

JURASKA J., DEUTSCH D., FINKELSTEIN M. & FREITAG M. (2024). MetricX-24 : The Google submission to the WMT 2024 metrics shared task. In B. HADDOW, T. KOCMI, P. KOEHN & C. MONZ, Éds., *Proceedings of the Ninth Conference on Machine Translation*, p. 492–504, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.wmt-1.35.

KARGARAN A. H., IMANI A., YVON F. & SCHUETZE H. (2023). GlotLID : Language identification for low-resource languages. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 6155–6218, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.410.

KOCMI T., AVRAMIDIS E., BAWDEN R., BOJAR O., DVORKOVICH A., FEDERMANN C., FISHEL M., FREITAG M., GOWDA T., GRUNDKIEWICZ R., HADDOW B., KARPINSKA M., KOEHN P., MARIE B., MONZ C., MURRAY K., NAGATA M., POPEL M., POPOVIĆ M., SHMATOVA M., STEINGRÍMSSON S. & ZOUHAR V. (2024). Findings of the WMT24 General Machine Translation Shared Task : The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, p.̃46, Miami, USA.

KOCMI T., AVRAMIDIS E., BAWDEN R., BOJAR O., DVORKOVICH A., FEDERMANN C., FISHEL M., FREITAG M., GOWDA T., GRUNDKIEWICZ R., HADDOW B., KOEHN P., MARIE B., MONZ C., MORISHITA M., MURRAY K., NAGATA M., NAKAZAWA T., POPEL M., POPOVIĆ M. & SHMATOVA M. (2023). Findings of the 2023 Conference on Machine Translation (WMT23) : LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, p. 1–42, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wmt-1.1.

KOCMI T. & FEDERMANN C. (2023). GEMBA-MQM : Detecting translation quality error spans with GPT-4. In P. KOEHN, B. HADDOW, T. KOCMI & C. MONZ, Éds., *Proceedings of the*

Eighth Conference on Machine Translation, p. 768–775, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wmt-1.64.

KUDUGUNTA S., CASWELL I., ZHANG B., GARCIA X., CHOQUETTE-CHOO C. A., LEE K., XIN D., KUSUPATI A., STELLA R., BAPNA A. & FIRAT O. (2023). MADLAD-400 : A Multilingual And Document-Level Large Audited Dataset. arXiv :2309.04662 [cs], DOI : 10.48550/arXiv.2309.04662.

KUMAR S., ANASTASOPOULOS A., WINTNER S. & TSVETKOV Y. (2021). Machine Translation into Low-resource Language Varieties. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, p. 110–121, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-short.16.

LAKEW S. M., EROFEEVA A. & FEDERICO M. (2018). Neural Machine Translation into Language Varieties. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 156–164, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6316.

LAMBRECHT L., SCHNEIDER F. & WAIBEL A. (2022). Machine Translation from Standard German to Alemannic Dialects. In M. MELERO, S. SAKTI & C. SORIA, Éds., *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, p. 129–136, Marseille, France : European Language Resources Association.

LAMUELA X. (2024). 20 Codification et élaboration linguistiques. In L. ESHER & J. SIBILLE, Éds., *Manuel de linguistique occitane*, p. 563–589. De Gruyter. DOI : 10.1515/9783110733433.

LIU Z., RICHARDSON C., HATCHER R. & PRUD'HOMMEAUX E. (2022). Not always about you : Prioritizing community needs when developing endangered language technology. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3933–3944, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.272.

LO CONGRÈS (2024). ReVoc Corpus. DOI : 10.5281/zenodo.11566430.

LOMMEL A. R., BURCHARDT A. & USZKOREIT H. (2013). Multidimensional quality metrics : a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK : Aslib.

MARUJO L., GRAZINA N., LUIS T., LING W., COHEUR L. & TRANCOSO I. (2011). BP2EP - Adaptation of Brazilian Portuguese texts to European Portuguese. In M. L. FORCADA, H. DEPRAETERE & V. VANDEGHINSTE, Éds., *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium : European Association for Machine Translation.

MEFTOUH K., HARRAT S., JAMOUSSI S., ABBAS M. & SMAILI K. (2015). Machine Translation Experiments on PADIC : A Parallel Arabic DIalect Corpus. In H. ZHAO, Éd., *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, p. 26–34, Shanghai, China.

MILETIC A., BRAS M., VERGEZ-COURET M., ESHER L., POUJADE C. & SIBILLE J. (2020). A Four-Dialect Treebank for Occitan : Building Process and Parsing Experiments. In M. ZAMPIERI, P. NAKOV, N. LJUBEŠIĆ, J. TIEDEMANN & Y. SCHERRER, Éds., *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, p. 140–149, Barcelona, Spain (Online) : International Committee on Computational Linguistics (ICCL).

MILETIĆ A. (2023). Outiller l'occitan : nouvelles ressources et lemmatisation. In C. SERVAN & A. VILNAT, Éds., *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, p. 217–231, Paris, France : ATALA.

MYINT OO T., KYAW THU Y. & MAR SOE K. (2019). Neural Machine Translation between Myanmar (Burmese) and Rakhine (Arakanese). In M. ZAMPIERI, P. NAKOV, S. MALMASI, N. LJUBEŠIĆ, J. TIEDEMANN & A. ALI, Éds., *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, p. 80–88, Ann Arbor, Michigan : Association for Computational Linguistics. DOI : 10.18653/v1/W19-1408.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). Training language models to follow instructions with human feedback. arXiv :2203.02155 [cs], DOI : 10.48550/arXiv.2203.02155.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2001). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, p. 311, Philadelphia, Pennsylvania : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PEREZ-ORTIZ J. A., SÁNCHEZ-MARTÍNEZ F., SÁNCHEZ-CARTAGENA V. M., ESPLÀ-GOMIS M., GALIANO JIMENEZ A., OLIVER A., AVENTÍN-BOYA C., PARDOS A., VALDÉS C., SANS SOCASAU J. L. & MARTÍNEZ J. P. (2024). Expanding the FLORES+ multilingual benchmark with translations for Aragonese, aranese, Asturian, and Valencian. In B. HADDOW, T. KOCMI, P. KOEHN & C. MONZ, Éds., *Proceedings of the Ninth Conference on Machine Translation*, p. 547–555, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.wmt-1.41.

POPOVIĆ M. (2017). chrF++ : words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, p. 612–618, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4770.

POPOVIĆ M., ARČAN M. & KLUBIČKA F. (2016). Language Related Issues for Machine Translation between Closely Related South Slavic Languages. In P. NAKOV, M. ZAMPIERI, L. TAN, N. LJUBEŠIĆ, J. TIEDEMANN & S. MALMASI, Éds., *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, p. 43–52, Osaka, Japan : The COLING 2016 Organizing Committee.

POST M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6319.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv :1910.10683 [cs] version : 2, DOI : 10.48550/arXiv.1910.10683.

RAMÍREZ-SÁNCHEZ G., ZARAGOZA-BERNABEU J., BAÑÓN M. & ROJAS S. O. (2020). Bifixer and Bicleaner : two open-source tools to clean your parallel data. In A. MARTINS, H. MONIZ, S. FUMEGA, B. MARTINS, F. BATISTA, L. COHEUR, C. PARRA, I. TRANCOSO, M. TURCHI, A. BISAZZA, J. MOORKENS, A. GUERBEROF, M. NURMINEN, L. MARG & M. L. FORCADA, Éds., *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 291–298, Lisboa, Portugal : European Association for Machine Translation.

RANATHUNGA S., LEE E.-S. A., PRIFTI SKENDULI M., SHEKHAR R., ALAM M. & KAUR R. (2023). Neural Machine Translation for Low-resource Languages : A Survey. *ACM Comput. Surv.*, **55**(11), 229 :1–229 :37. DOI : 10.1145/3567592.

REI R., STEWART C., FARINHA A. C. & LAVIE A. (2020). COMET : A Neural Framework for MT Evaluation. arXiv :2009.09025 [cs].

SALLOUM W., ELFARDY H., ALAMIR-SALLOUM L., HABASH N. & DIAB M. (2014). Sentence Level Dialect Identification for Machine Translation System Selection. In K. TOUTANOVA & H. WU, Éds., *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 772–778, Baltimore, Maryland : Association for Computational Linguistics. DOI : 10.3115/v1/P14-2125.

SALLOUM W. & HABASH N. (2011). Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In J. JANCSARY, F. NEUBARTH & H. TROST, Éds., *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, p. 10–21, Edinburgh, Scotland : Association for Computational Linguistics.

SAWAF H. (2010). Arabic Dialect Handling in Hybrid Machine Translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas : Research Papers*, Denver, Colorado, USA : Association for Machine Translation in the Americas.

SCHERRER Y. (2012). *Generating Swiss German sentences from Standard German : a multi-dialectal approach*. Thèse de doctorat, Université de Genève. DOI : 10.13097/archive-ouverte/unige :26361.

SCHERRER Y. & LJUBEŠIĆ N. (2016). Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, p. 248–255, Bochum, Germany.

SCHWENK H., CHAUDHARY V., SUN S., GONG H. & GUZMÁN F. (2019). WikiMatrix : Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. arXiv :1907.05791 [cs], DOI : 10.48550/arXiv.1907.05791.

SEGUIER E. (2015). Reconnaissance automatique des dialectes occitans à l'écrit. Mémoire de master, Université Toulouse Jean Jaurès, Toulouse, France.

SIBILLE J. (2024). 16 Les dialectes occitans. In L. ESHER & J. SIBILLE, Éds., *Manuel de linguistique occitane*, p. 423–471. De Gruyter. DOI : 10.1515/9783110733433.

SUN J., SELLAM T., CLARK E., VU T., DOZAT T., GARRETTE D., SIDDHANT A., EISENSTEIN J. & GEHRMANN S. (2023). Dialect-robust Evaluation of Generated Text. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 6010–6028, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.331.

SÁNCHEZ-MARTÍNEZ F., PEREZ-ORTIZ J. A., JIMENEZ A. G. & OLIVER A. (2024). Findings of the WMT 2024 Shared Task Translation into Low-Resource Languages of Spain : Blending Rule-Based and Neural Systems. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, p.1̃5, Miami, USA.

SÉGUIER A. & LO CONGRÈS (2023a). Occitan Corpus from Lo Congrès news. DOI : 10.5281/zenodo.8411197.

SÉGUIER A. & LO CONGRÈS (2023b). SoftwaresOccitanTranslations corpus. DOI : 10.5281/zenodo.8411351.

TIEDEMANN J. (2009). Character-Based PSMT for Closely Related Languages. In L. MÀRQUEZ & H. SOMERS, Éds., *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain : European Association for Machine Translation.

WAN Y., YANG B., WONG D. F., CHAO L. S., DU H. & AO B. C. H. (2020). Unsupervised Neural Dialect Translation with Commonality and Diversity Modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(05), 9130–9137. Number : 05, DOI : 10.1609/aaai.v34i05.6448.

XIA M., KONG X., ANASTASOPOULOS A. & NEUBIG G. (2019). Generalized Data Augmentation for Low-Resource Translation. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 5786–5796, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1579.

XUE L., CONSTANT N., ROBERTS A., KALE M., AL-RFOU R., SIDDHANT A., BARUA A. & RAFFEL C. (2021). mT5 : A massively multilingual pre-trained text-to-text transformer. arXiv :2010.11934 [cs], DOI : 10.48550/arXiv.2010.11934.

ZBIB R., MALCHIODI E., DEVLIN J., STALLARD D., MATSOUKAS S., SCHWARTZ R., MAKHOUL J., ZAIDAN O. F. & CALLISON-BURCH C. (2012). Machine Translation of Arabic Dialects. In E. FOSLER-LUSSIER, E. RILOFF & S. BANGALORE, Éds., *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 49–59, Montréal, Canada : Association for Computational Linguistics.

ZEBAZE A., SAGOT B. & BAWDEN R. (2025). Compositional Translation : A Novel LLM-based Approach for Low-resource Machine Translation. arXiv :2503.04554 [cs], DOI : 10.48550/arXiv.2503.04554.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating Text Generation with BERT. arXiv :1904.09675 [cs], DOI : 10.48550/arXiv.1904.09675.

A Annexes

A.1 Prétraitement du corpus LoCongresNews

Les étapes de prétraitement du corpus d'origine sont :

1. Suppression des doublons exacts (texte occitan, texte français, étiquette)
2. Suppression des exemples très courts (moins de quatre tokens)
3. Annotation des exemples parallèles avec le package `bicleaner-hardrules` [18] (Ramírez-Sánchez *et al.*, 2020), puis analyses et décisions manuelles par type d'erreur pour les segments rejetés par l'outil.

Pour l'annotation avec `bicleaner-hardrules`, nous avons appliqué tous les filtres disponibles, à l'exception du modèle de langue (indisponible pour l'occitan), de l'option *porn_removal* et du filtre sur la longueur minimale des exemples (que nous avions déjà implémenté dans l'étape précédente).

Suite à l'analyse manuelle des segments rejetés par type d'erreur, nous avons décidé d'appliquer ou d'ignorer certains filtres comme indiqué dans la table 8. En particulier, les résultats pour l'identification de la langue contenant principalement du bruit, nous avons dû ignorer ce filtre.

18. https://github.com/bitextor/bicleaner-hardrules - version utilisée : 2.8.0

Filtres appliqués	*Filtres ignorés*
• segments identiques	• identification de la langue
• ratio de longueur	• mots répétés
• titres	• mots agglutinés
• parenthèses	• problèmes d'encodage
• trop long	• *breadcrumbs*
• nombres uniquement	
• *unicode_noise*	

TABLE 8 – Décisions par filtre de `bicleaner-hardrules`.

A.2 Détails d'implémentation des modèles

A.2.1 Prompts

Pour la traduction avec les LLMs, nous utilisons différents prompts en fonction de la stratégie choisie et du modèle utilisé.

Pour Llama3-8B-Instruct, nous utilisons la librairie vLLM [19]. Tous les prompts pour ce modèle sont entourés du template avec les tokens spéciaux requis :

```
<|begin_of_text|><|start_header_id|>user<|end_header_id|>
{prompt}
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Pour Aya-23-8B, nous utilisons la librairie `transformers` [20] qui applique automatiquement le template via la fonction `tokenizer.apply_chat_template()`.

Les prompts utilisés en 0-SHOT appliquent le format suivant pour Aya-23 et Llama3 :

```
Translate {lang_src} into {lang_tgt}: {text}
```

Pour les prédictions en *few-shot*, nous appliquons une stratégie différente pour Aya-23 et Llama3.

Pour Aya-23, nous utilisons le prompt 0-SHOT dans une série de messages avec alternance des rôles `user` et `assistant`.

Pour Llama3, nous réutilisons le prompt *few-shot* "vanilla" proposé par (Zebaze *et al.*, 2025), en remplaçant le nom de la langue source par le mot "source" pour les stratégies LANG-TGT et DIA-TGT :

19. https://docs.vllm.ai/en/latest/index.html. Version utilisée : 0.7.3
20. https://huggingface.co/docs/transformers/index. Version utilisée : 4.48.3

```
Given the following sentence-translation pairs written by a
professional translator:

<Demonstrations>
1. {lang_src} sentence
{shot_1_src}
{lang_tgt} translation
{shot_1_tgt}

2. {lang_src} sentence
{shot_2_src}
{lang_tgt} translation
{shot_2_tgt}

3. {lang_src} sentence
{shot_3_src}
French translation
{shot_3_tgt}

</Demonstrations>

Please write a high-quality {lang_tgt} translation of the
following {lang_src} sentence

{text}

Please make sure to consider the above information and provide
only the translation, nothing more.
```

Le tableau 9 indique les langues et dialectes indiqués dans les prompts en fonction de la stratégie employée.

	Langue source	Langue cible	Dialecte
LANG-TGT	non	oui	non
LANG-SRC	oui	oui	non
DIA-TGT	non	oui	oui
DIA-SRC	oui	oui	oui

TABLE 9 – Présence des noms des langues et dialectes en fonction de la stratégie de prompt utilisée.

Les noms des dialectes sont précisés en français avec le format `Occitan 'dialecte'`.

A.2.2 Exemples pour les prédictions few-shot

Inspirés du processus appliqué lors de la campagne d'évaluation WMT24 pour le benchmark des LLMs (Kocmi *et al.*, 2024), les prédictions en mode *few-shot* utilisent des exemples fixes choisis

manuellement dans d'autres sources de données pour éviter les intersections avec les exemples à traduire.

Afin d'éviter d'introduire des biais lors de la sélection des exemples pour le mode 3-SHOTS-ADAPT – où les exemples choisis sont dans le même dialecte que l'exemple à traduire – nous avons trouvé des exemples contrastifs pour l'ensemble des dialectes d'un même jeu de test.

Pour LoCongresNews, nous avons choisi trois exemples parallèles et contrastifs extraits du corpus *SoftwaresOccitanTranslations* [21].

Pour Flores, nous avons extrait manuellement trois phrases parallèles en anglais, français et aranais depuis un site web [22] du domaine du tourisme. Nous avons ensuite fait appel à un traducteur expérimenté pour traduire les phrases du français vers le languedocien.

A.2.3 Hyperparamètres

Lorsque cela était possible (c'est-à-dire pour NLLB-600, Aya-23 et Llama3), nous avons effectué les prédictions en mode *greedy*, c'est-à-dire sans *sampling* (*temperature* à 0) ni *beam search*.

A.3 Détails d'implémentation des métriques

Les scores BLEU [23] et chrF++ [24] sont calculés via la librairie `sacrebleu` [25] (Post, 2018).

Les scores COMET [26] *avec* référence (COMET, $\text{COMET}_{\text{nosrc}}$) sont obtenus via le modèle 'wmt22-comet-da'. Les scores COMET *sans* référence (COMET-QE) sont obtenus via le modèle 'wmt22-cometkiwi-da'. Les scores COMET rapportés sont systématiquement multipliés par 100 pour en faciliter la lecture.

Pour l'identification des langues dans les segments de référence et les prédictions des modèles, nous utilisons les modèles lid218e [27], Idiomata Cognitor [28] et GlotLID [29].

Avec les modèles lid218e et GlotLID, pour obtenir le score associé à une étiquette de langue particulière (occitan, français, espagnol, catalan), nous appelons le modèle via la librairie `fasttext` pour extraire le top 3 des étiquettes pour chaque exemple. Si l'étiquette voulue y est présente, nous utilisons le score LID qui lui est associé ; sinon nous attribuons un score LID de 0 pour l'exemple donné.

Avec le modèle Idiomata Cognitor – spécialisé pour l'identification des langues romanes –, le score que nous associons à l'étiquette "occitan" est la somme des probabilités prédites par le modèle pour les étiquettes "occitan" et "aranais".

21. https://zenodo.org/records/8411351
22. https://www.baqueira.es/oc-ar. Consulté en mars 2025.
23. Signature : `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1`
24. Signature : `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1`
25. https://github.com/mjpost/sacrebleu
26. https://github.com/Unbabel/COMET
27. facebook/fasttext-language-identification
28. https://github.com/transducens/idiomata_cognitor/tree/main, v1.0.1
29. https://huggingface.co/cis-lmu/glotlid, model_v3

A.4 Détails des résultats

A.4.1 Scores occitan vers français

Les scores de traduction occitan vers français pour les métriques utilisées ainsi que toutes les configurations de modèles sont présentés dans la table 10 pour Flores, et dans la table 11 pour LoCongresNews.

Les scores par dialecte sont donnés par métrique :

1. BLEU : table 12
2. chrF++ : table 13
3. COMET : table 14
4. $\text{COMET}_{\text{nosrc}}$: table 15
5. COMET-QE : table 16

	BLEU	chrF++	COMET	$\text{COMET}_{\text{nosrc}}$	COMET-QE
Baseline COPIE	2,99	26,07	41,76	46,05	38,02
Revirada					
ADAPT	22,09	48,57	62,61	63,92	59,53
LANGUEDOCIEN	22,09	48,57	62,61	63,92	59,53
GASCON	20,76	47,95	60,25	61,67	58,85
NLLB-600	31,03	55,09	75,12	75,62	67,83
Google Translate	**40,90**	**63,16**	**83,12**	**83,18**	**72,64**
Aya23-8B					
0-SHOT LANG-TGT	23,72	48,17	66,75	68,11	61,39
0-SHOT LANG-SRC	24,47	49,40	67,65	68,80	62,30
0-SHOT DIA-SRC	23,32	49,17	65,24	66,56	60,24
3-SHOTS LANG-TGT	25,73	49,59	71,93	72,64	63,77
3-SHOTS LANG-SRC	25,17	48,62	71,61	72,23	63,21
3-SHOTS DIA-SRC	20,89	42,09	67,08	67,52	59,07
Llama3.1-8B Instruct					
0-SHOT LANG-TGT	28,94	53,54	75,33	75,84	68,60
0-SHOT LANG-SRC	29,69	54,43	75,93	76,37	69,42
0-SHOT DIA-SRC	29,18	54,12	75,52	75,99	69,01
3-SHOTS LANG-TGT	30,09	55,01	77,44	77,78	70,11
3-SHOTS LANG-SRC	30,11	55,05	77,54	77,87	70,27
3-SHOTS DIA-SRC	30,29	55,16	77,56	77,88	70,24

TABLE 10 – Scores des modèles pour la traduction de l'occitan vers le français (dialectes combinés), pour le jeu de test Flores. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	BLEU	chrF++	COMET	$\text{COMET}_{\text{nosrc}}$	COMET-QE
Baseline COPIE	15,48	40,71	55,90	57,51	44,96
Revirada					
ADAPT	**65,91**	**81,17**	84,12	84,24	67,36
LANGUEDOCIEN	60,61	77,91	77,97	78,41	63,90
GASCON	63,14	79,54	81,79	82,03	65,72
NLLB-600	56,88	74,88	82,47	82,66	67,91
Google Translate	57,15	75,17	**86,14**	**86,12**	**70,43**
Aya23-8B					
0-SHOT LANG-TGT	43,76	63,52	71,59	72,52	60,39
0-SHOT LANG-SRC	47,05	66,23	73,23	74,05	61,49
0-SHOT DIA-SRC	42,90	64,22	68,55	69,54	58,21
3-SHOTS LANG-TGT	40,83	57,96	74,47	74,95	61,28
3-SHOTS LANG-SRC	36,77	53,73	73,26	73,74	59,78
3-SHOTS DIA-SRC	24,31	40,62	67,49	68,06	53,83
Llama3.1-8B Instruct					
0-SHOT LANG-TGT	57,15	73,32	79,78	80,03	66,13
0-SHOT LANG-SRC	59,80	75,80	82,04	82,17	67,48
0-SHOT DIA-SRC	57,55	73,27	76,99	77,48	63,51
3-SHOTS LANG-TGT	55,17	71,44	80,36	80,63	65,32
3-SHOTS LANG-SRC	57,75	73,94	81,90	82,08	66,87
3-SHOTS DIA-SRC	56,41	72,74	80,42	80,68	65,55

TABLE 11 – Scores des modèles pour la traduction de l'occitan vers le français (dialectes combinés), pour le jeu de test LoCongresNews. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	2,47	3,50	5,75	15,69	18,34	15,20	17,24	10,42
Revirada								
ADAPT	11,34	32,04	47,83	**62,60**	56,78	**69,92**	61,33	50,28
LANGUEDOCIEN	11,34	32,04	47,83	50,95	56,78	**69,92**	**61,33**	50,28
GASCON	12,61	28,37	45,57	**62,60**	55,98	64,48	57,61	50,99
NLLB-600	21,58	40,01	50,34	52,06	55,21	61,48	58,17	50,59
Google Translate	**32,98**	**48,35**	**58,21**	54,95	57,14	59,24	58,14	49,65
Aya23-8B								
0-SHOT LANG-TGT	15,44	31,68	36,58	35,82	45,82	50,81	49,60	38,89
0-SHOT LANG-SRC	16,05	32,83	38,19	39,04	50,43	54,13	52,39	40,76
0-SHOT DIA-SRC	16,60	29,52	35,86	36,42	43,29	48,94	48,17	40,04
3-SHOTS LANG-TGT	17,74	33,34	50,62	33,26	45,95	47,27	45,98	42,78
3-SHOTS LANG-SRC	17,41	32,56	48,13	29,66	43,86	42,90	39,71	37,08
3-SHOTS DIA-SRC	14,01	27,44	15,72	20,95	24,37	27,11	31,34	23,56
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	21,19	36,36	39,55	52,49	59,63	60,94	58,41	50,90
0-SHOT LANG-SRC	21,84	37,20	46,69	54,31	59,85	60,83	56,04	**52,71**
0-SHOT DIA-SRC	21,28	36,73	49,70	53,16	**60,73**	61,32	58,30	50,23
3-SHOTS LANG-TGT	22,21	37,53	48,75	51,90	58,08	58,42	52,16	49,56
3-SHOTS LANG-SRC	22,33	37,38	48,53	54,26	55,72	61,30	54,76	52,16
3-SHOTS DIA-SRC	22,31	37,74	48,56	52,45	58,81	60,52	50,75	49,06

TABLE 12 – Scores BLEU par dialecte (Flores/LoCongresNews) pour la traduction de l'occitan vers le français. Les meilleurs scores par dialecte sont mis en gras. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	24,32	27,82	34,05	40,64	43,73	40,63	42,45	38,55
Revirada								
ADAPT	39,16	57,99	72,64	**79,12**	74,14	**83,76**	**78,42**	71,44
LANGUEDOCIEN	39,16	57,99	72,64	72,05	74,14	**83,76**	**78,42**	71,44
GASCON	40,57	55,37	70,62	**79,12**	74,05	80,55	76,03	71,62
NLLB-600	48,03	62,18	71,56	71,71	72,97	78,03	75,08	71,39
Google Translate	**57,92**	**68,41**	**75,60**	73,87	73,71	76,50	75,57	70,53
Aya23-8B								
0-SHOT LANG-TGT	40,95	55,40	59,41	57,51	64,41	68,97	68,19	61,50
0-SHOT LANG-SRC	42,20	56,63	61,29	60,45	68,04	71,50	69,68	64,51
0-SHOT DIA-SRC	43,08	55,20	58,61	59,86	66,15	68,10	68,00	66,10
3-SHOTS LANG-TGT	42,87	56,34	71,44	51,29	61,80	63,53	63,99	62,00
3-SHOTS LANG-SRC	42,10	55,17	65,97	47,29	59,58	59,23	56,73	57,29
3-SHOTS DIA-SRC	35,80	48,38	31,63	37,49	41,70	43,18	47,51	39,11
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	47,10	59,99	66,60	70,47	74,87	76,00	74,32	71,21
0-SHOT LANG-SRC	48,12	60,75	72,89	72,98	**77,16**	78,47	76,60	71,16
0-SHOT DIA-SRC	47,64	60,62	71,80	70,42	75,86	75,92	73,83	69,82
3-SHOTS LANG-TGT	48,89	61,15	70,44	69,31	73,85	73,58	68,76	68,90
3-SHOTS LANG-SRC	49,06	61,07	69,88	71,56	73,26	76,42	71,62	**72,49**
3-SHOTS DIA-SRC	49,04	61,31	70,60	69,81	75,30	75,88	66,59	69,80

TABLE 13 – Scores chrF++ par dialecte (Flores/LoCongresNews) pour la traduction de l'occitan vers le français. Les meilleurs scores par dialecte sont mis en gras. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	37,85	45,67	52,02	51,59	57,42	59,28	60,58	57,36
Revirada								
ADAPT	48,69	76,54	75,44	82,79	73,15	86,10	80,29	76,07
LANGUEDOCIEN	48,69	76,54	75,44	68,61	73,15	86,10	80,29	76,07
GASCON	51,89	68,60	70,19	82,79	72,82	81,74	76,89	75,30
NLLB-600	69,54	80,69	83,57	79,66	78,40	84,99	82,66	83,89
Google Translate	**81,34**	**84,91**	**86,92**	**85,01**	**83,08**	**87,19**	**86,37**	**87,28**
Aya23-8B								
0-SHOT LANG-TGT	60,79	72,71	68,65	65,35	70,32	76,64	76,78	75,93
0-SHOT LANG-SRC	61,93	73,36	69,40	67,32	73,16	78,01	77,93	76,34
0-SHOT DIA-SRC	63,39	67,10	68,88	63,02	64,78	73,13	72,48	73,21
3-SHOTS LANG-TGT	67,64	76,23	77,93	69,01	73,50	78,83	78,89	79,16
3-SHOTS LANG-SRC	67,11	76,11	77,89	68,12	73,60	77,51	74,86	74,93
3-SHOTS DIA-SRC	62,66	71,49	54,74	64,12	65,78	70,35	71,36	64,52
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	71,35	79,30	79,13	77,01	78,18	82,09	81,29	82,31
0-SHOT LANG-SRC	72,52	79,33	82,44	79,34	81,41	84,25	84,19	81,84
0-SHOT DIA-SRC	71,30	79,75	80,56	74,28	77,25	79,19	78,72	76,39
3-SHOTS LANG-TGT	74,07	80,81	80,59	78,21	80,29	82,16	80,49	83,22
3-SHOTS LANG-SRC	74,19	80,89	82,98	79,90	80,04	83,69	80,86	85,35
3-SHOTS DIA-SRC	74,07	81,05	82,46	77,87	80,18	82,75	77,52	82,24

TABLE 14 – Scores COMET par dialecte (Flores/LoCongresNews) pour la traduction de l'occitan vers le français. Les meilleurs scores par dialecte sont mis en gras. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	43,13	48,96	54,95	54,21	58,81	60,07	61,15	58,75
Revirada								
ADAPT	51,34	76,50	75,94	83,26	74,15	85,86	80,38	76,65
LANGUEDOCIEN	51,34	76,50	75,94	69,83	74,15	85,86	80,38	76,65
GASCON	54,31	69,03	71,25	83,26	73,75	81,75	77,15	75,84
NLLB-600	70,73	80,50	83,36	80,34	79,03	84,76	82,57	83,50
Google Translate	**81,82**	**84,53**	**86,74**	**85,37**	**83,45**	**86,84**	**86,13**	**86,89**
Aya23-8B								
0-SHOT LANG-TGT	63,03	73,20	69,80	67,05	71,47	76,95	77,03	76,36
0-SHOT LANG-SRC	63,93	73,68	70,48	68,87	74,14	78,23	78,15	77,13
0-SHOT DIA-SRC	65,18	67,94	70,04	64,74	66,19	73,50	73,26	73,25
3-SHOTS LANG-TGT	68,96	76,33	77,96	70,04	74,22	78,85	78,97	79,47
3-SHOTS LANG-SRC	68,31	76,14	77,83	69,13	74,34	77,56	74,83	75,31
3-SHOTS DIA-SRC	63,59	71,45	55,21	65,07	66,42	70,61	71,57	64,96
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	72,49	79,19	79,35	77,76	78,71	81,93	81,20	82,24
0-SHOT LANG-SRC	73,57	79,18	82,50	79,92	81,58	84,02	84,01	81,87
0-SHOT DIA-SRC	72,41	79,57	80,93	75,13	77,53	79,42	78,79	76,68
3-SHOTS LANG-TGT	74,98	80,58	80,83	78,92	80,45	82,08	80,41	82,83
3-SHOTS LANG-SRC	75,08	80,66	82,90	80,51	80,35	83,52	80,83	84,94
3-SHOTS DIA-SRC	74,97	80,80	82,49	78,59	80,50	82,63	77,55	82,13

TABLE 15 – Scores $\text{COMET}_{\text{nosrc}}$ par dialecte (Flores/LoCongresNews) pour la traduction de l'occitan vers le français. Les meilleurs scores par dialecte sont mis en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	33,42	42,62	50,28	40,60	44,89	48,39	49,46	44,94
Revirada								
ADAPT	47,17	71,89	69,89	63,84	60,11	70,66	66,90	67,76
LANGUEDOCIEN	47,17	71,89	69,89	55,87	60,11	70,66	66,90	67,76
GASCON	50,97	66,74	65,05	63,84	59,91	67,63	64,19	66,33
NLLB-600	62,81	72,85	75,50	63,62	69,13	71,26	70,49	73,11
Google Translate	**69,97**	**75,31**	**76,65**	**66,73**	**69,79**	**73,28**	**73,74**	**76,84**
Aya23-8B								
0-SHOT LANG-TGT	55,15	67,64	65,63	53,46	60,57	65,77	67,27	67,48
0-SHOT LANG-SRC	56,53	68,08	65,71	54,96	61,84	66,62	67,34	66,86
0-SHOT DIA-SRC	57,33	63,15	68,79	52,24	54,98	63,02	62,39	64,99
3-SHOTS LANG-TGT	58,86	68,68	69,04	54,88	61,41	66,22	67,49	68,79
3-SHOTS LANG-SRC	58,19	68,24	68,50	53,49	61,38	64,70	64,25	65,24
3-SHOTS DIA-SRC	54,06	64,08	46,63	49,30	51,36	57,50	59,80	53,42
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	64,39	72,80	74,14	61,87	66,35	69,37	70,20	72,56
0-SHOT LANG-SRC	65,73	73,12	74,99	63,34	68,10	70,65	71,47	71,15
0-SHOT DIA-SRC	64,75	73,27	74,05	59,60	63,93	66,48	67,67	66,54
3-SHOTS LANG-TGT	66,44	73,78	72,39	61,25	65,74	68,44	68,50	71,54
3-SHOTS LANG-SRC	66,72	73,82	74,17	62,86	67,32	69,98	69,45	73,64
3-SHOTS DIA-SRC	66,48	73,99	74,21	61,08	65,52	69,21	66,01	71,54

TABLE 16 – Scores COMET-QE par dialecte (Flores/LoCongresNews) pour la traduction de l'occitan vers le français. Les meilleurs scores par dialecte sont mis en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

A.4.2 Scores français vers occitan

Le détail complet des scores de traduction (métriques BLEU, chrF++, COMET, $\text{COMET}_{\text{nosrc}}$ et COMET-QE) est présenté dans la table 17 pour Flores, et dans la table 18 pour LoCongresNews.

Les scores par dialecte sont donnés par métrique :

1. BLEU : table 19
2. chrF++ : table 20
3. COMET : table 21

Le détail complet des scores LID_{oc} est présenté dans les tables 22 (lid218e), 23 (Idiomata Cognitor) et 24 (GlotLID). Les tables MSE_{oc} correspondantes sont la table 25 (lid218e), la table 26 (Idiomata Cognitor) et la table 27 (GlotLID).

Pour les langues proches de l'occitan, seuls les scores obtenus avec le modèle Idiomata Cognitor sont présentés, à la table 28 pour les scores LID, et à la table 29 pour les scores MSE.

	BLEU	chrF++	COMET	$\text{COMET}_{\text{nosrc}}$	COMET-QE
Baseline COPIE	2,98	26,75	64,65	58,36	**84,99**
Revirada					
ADAPT	17,97	44,80	63,71	66,62	56,62
LANGUEDOCIEN	17,97	44,80	63,71	66,62	56,62
GASCON	10,16	39,94	58,03	62,31	45,59
NLLB-600	16,32	42,49	64,49	66,68	59,39
Google Translate	**19,89**	**46,04**	**65,18**	**67,84**	55,98
Aya23-8B					
0-SHOT LANG-TGT	4,73	30,85	63,13	61,68	74,49
0-SHOT LANG-SRC	4,87	30,87	63,27	61,90	74,12
0-SHOT DIA-TGT	4,97	31,63	62,65	62,27	71,57
0-SHOT DIA-SRC	5,07	31,72	62,47	62,34	70,44
3-SHOTS LANG-TGT	5,77	32,44	62,30	62,95	67,91
3-SHOTS LANG-SRC	5,79	32,25	62,17	62,88	67,24
3-SHOTS DIA-TGT	5,70	31,99	62,00	62,70	67,17
3-SHOTS DIA-SRC	5,78	31,79	61,83	62,58	66,89
3-SHOTS-ADAPT LANG-TGT	5,84	32,58	62,42	62,97	68,16
3-SHOTS-ADAPT LANG-SRC	5,82	32,37	62,33	63,00	67,34
3-SHOTS-ADAPT DIA-TGT	5,77	32,25	62,15	62,76	67,72
3-SHOTS-ADAPT DIA-SRC	5,80	32,04	62,05	62,73	67,11
Llama3.1-8B Instruct					
0-SHOT LANG-TGT	12,90	40,00	64,36	66,12	63,46
0-SHOT LANG-SRC	12,74	39,84	63,85	65,98	61,43
0-SHOT DIA-TGT	13,02	40,15	63,84	65,96	61,51
0-SHOT DIA-SRC	12,72	39,92	63,59	65,80	60,75
3-SHOTS LANG-TGT	13,58	40,89	64,48	66,64	61,83
3-SHOTS LANG-SRC	13,46	40,89	64,43	66,63	61,78
3-SHOTS DIA-TGT	13,54	40,86	64,37	66,57	61,81
3-SHOTS DIA-SRC	13,45	40,84	64,41	66,61	61,78
3-SHOTS-ADAPT LANG-TGT	13,62	40,87	64,52	66,68	61,99
3-SHOTS-ADAPT LANG-SRC	13,50	40,91	64,46	66,65	61,90
3-SHOTS-ADAPT DIA-TGT	13,60	40,87	64,48	66,65	61,86
3-SHOTS-ADAPT DIA-SRC	13,49	40,82	64,46	66,65	61,78

TABLE 17 – Scores des modèles pour la traduction du français vers l'occitan (dialectes combinés), pour le jeu de test Flores. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	BLEU	chrF++	COMET	$\text{COMET}_{\text{nosrc}}$	COMET-QE
Baseline COPIE	15,47	41,65	72,35	67,10	**81,14**
Revirada					
ADAPT	**60,94**	**79,22**	**79,53**	**81,24**	54,62
LANGUEDOCIEN	52,81	73,95	78,12	79,56	58,62
GASCON	47,21	71,69	72,84	75,59	49,12
NLLB-600	41,46	65,35	75,47	76,62	61,70
Google Translate	37,05	62,49	74,73	76,29	59,51
Aya23-8B					
0-SHOT LANG-TGT	18,74	46,67	71,01	70,12	72,31
0-SHOT LANG-SRC	18,97	46,95	71,22	70,34	71,97
0-SHOT DIA-TGT	18,82	46,97	70,45	70,09	70,32
0-SHOT DIA-SRC	19,20	47,49	70,29	70,20	69,32
3-SHOTS LANG-TGT	17,82	44,44	69,42	69,88	65,83
3-SHOTS LANG-SRC	17,65	43,81	69,03	69,60	64,96
3-SHOTS DIA-TGT	15,82	40,87	68,11	68,54	64,36
3-SHOTS DIA-SRC	15,07	39,41	67,41	67,93	63,40
3-SHOTS-ADAPT LANG-TGT	17,64	43,64	69,01	69,61	64,55
3-SHOTS-ADAPT LANG-SRC	17,48	42,95	68,63	69,29	63,84
3-SHOTS-ADAPT DIA-TGT	15,17	39,29	67,27	67,82	62,80
3-SHOTS-ADAPT DIA-SRC	14,27	37,49	66,34	66,96	61,78
Llama3.1-8B Instruct					
0-SHOT LANG-TGT	37,60	62,20	74,12	75,01	63,10
0-SHOT LANG-SRC	37,67	62,20	73,90	75,02	62,05
0-SHOT DIA-TGT	31,76	60,16	70,60	72,22	58,78
0-SHOT DIA-SRC	25,22	56,67	67,35	69,34	56,26
3-SHOTS LANG-TGT	38,60	63,39	73,49	74,85	61,04
3-SHOTS LANG-SRC	38,98	63,69	73,69	75,06	61,18
3-SHOTS DIA-TGT	38,42	63,05	73,16	74,62	60,62
3-SHOTS DIA-SRC	38,45	63,10	73,28	74,72	60,65
3-SHOTS-ADAPT LANG-TGT	38,92	63,62	73,58	74,99	60,87
3-SHOTS-ADAPT LANG-SRC	39,20	63,87	73,73	75,16	61,02
3-SHOTS-ADAPT DIA-TGT	38,78	63,45	73,39	74,88	60,40
3-SHOTS-ADAPT DIA-SRC	38,77	63,36	73,40	74,88	60,49

TABLE 18 – Scores des modèles pour la traduction du français vers l'occitan (dialectes combinés), pour le jeu de test LoCongresNews. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	2,46	3,50	5,76	15,68	18,32	15,18	17,22	10,43
Revirada								
ADAPT	5,62	29,14	**45,75**	**57,59**	**46,83**	**65,77**	**48,89**	**40,99**
LANGUEDOCIEN	5,62	29,14	**45,75**	39,08	**46,83**	**65,77**	**48,89**	**40,99**
GASCON	5,88	14,18	32,39	**57,59**	37,34	38,47	35,34	31,65
NLLB-600	5,82	25,81	34,41	31,35	37,18	51,14	41,26	36,37
Google Translate	**6,19**	**32,14**	31,44	28,48	35,92	45,16	37,79	32,01
Aya23-8B								
0-SHOT LANG-TGT	3,44	5,99	11,16	17,35	21,07	19,79	19,94	14,71
0-SHOT LANG-SRC	3,48	6,22	10,86	17,53	21,52	20,07	19,39	15,53
0-SHOT DIA-TGT	3,64	6,23	9,90	17,40	19,45	20,16	20,67	12,97
0-SHOT DIA-SRC	3,74	6,34	12,05	17,38	21,94	20,83	21,05	13,85
3-SHOTS LANG-TGT	3,80	7,67	13,59	14,88	23,50	20,35	19,29	15,69
3-SHOTS LANG-SRC	3,83	7,68	14,81	14,64	21,84	20,37	18,22	14,77
3-SHOTS DIA-TGT	3,77	7,57	14,83	13,56	21,20	17,69	17,53	14,63
3-SHOTS DIA-SRC	3,87	7,62	17,74	12,94	18,25	16,80	18,09	13,91
3-SHOTS-ADAPT LANG-TGT	3,95	7,67	18,66	14,51	22,56	20,35	18,69	15,71
3-SHOTS-ADAPT LANG-SRC	3,89	7,68	16,72	14,19	23,48	20,37	17,73	14,49
3-SHOTS-ADAPT DIA-TGT	3,92	7,57	18,58	12,14	19,15	17,69	17,50	15,23
3-SHOTS-ADAPT DIA-SRC	3,91	7,62	19,01	11,21	18,32	16,80	16,36	15,86
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	5,23	19,95	30,48	30,66	37,46	44,23	36,06	34,65
0-SHOT LANG-SRC	5,07	19,78	30,11	30,94	38,17	43,93	36,49	34,51
0-SHOT DIA-TGT	5,19	20,21	27,41	27,21	31,59	35,70	33,28	25,22
0-SHOT DIA-SRC	5,08	19,68	30,71	21,80	26,12	28,14	24,16	26,38
3-SHOTS LANG-TGT	5,20	21,25	33,86	31,23	38,37	45,65	37,22	35,93
3-SHOTS LANG-SRC	5,10	21,10	33,39	31,34	38,56	46,27	38,24	35,49
3-SHOTS DIA-TGT	5,15	21,23	32,06	31,22	36,93	45,37	37,07	36,09
3-SHOTS DIA-SRC	5,08	21,07	33,05	31,05	36,90	45,55	37,48	35,80
3-SHOTS-ADAPT LANG-TGT	5,25	21,23	33,19	31,82	38,84	45,64	38,85	36,27
3-SHOTS-ADAPT LANG-SRC	5,14	21,13	33,06	31,73	38,80	46,27	39,07	36,99
3-SHOTS-ADAPT DIA-TGT	5,17	21,22	32,70	31,91	37,45	45,37	38,19	37,11
3-SHOTS-ADAPT DIA-SRC	5,17	21,01	33,69	31,67	37,32	45,55	38,55	36,71

TABLE 19 – Scores BLEU des modèles pour la traduction du français vers l'occitan, en fonction du dialecte du segment de référence. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	24,91	28,59	34,50	41,46	44,92	41,71	43,37	38,59
Revirada								
ADAPT	33,45	56,17	**69,90**	**77,34**	**70,15**	**82,11**	**71,95**	**64,99**
LANGUEDOCIEN	33,45	56,17	**69,90**	65,91	**70,15**	**82,11**	**71,95**	**64,99**
GASCON	**34,47**	45,43	60,00	**77,34**	64,94	67,27	64,01	59,86
NLLB-600	32,71	52,30	62,04	59,15	62,56	71,45	66,06	61,00
Google Translate	34,26	**57,86**	61,30	56,66	62,06	68,17	62,87	58,03
Aya23-8B								
0-SHOT LANG-TGT	28,42	33,29	43,16	44,88	49,68	48,30	47,68	43,34
0-SHOT LANG-SRC	28,28	33,47	42,50	45,23	49,52	48,55	47,50	44,32
0-SHOT DIA-TGT	29,05	34,22	42,00	45,11	49,17	48,71	48,52	40,30
0-SHOT DIA-SRC	29,31	34,13	43,70	45,54	50,16	49,28	49,20	42,25
3-SHOTS LANG-TGT	29,06	35,84	44,67	40,78	51,33	47,50	47,32	43,62
3-SHOTS LANG-SRC	28,92	35,59	46,04	40,08	47,25	47,14	45,20	42,83
3-SHOTS DIA-TGT	28,57	35,42	45,63	37,94	46,44	43,21	43,74	43,00
3-SHOTS DIA-SRC	28,50	35,08	47,56	36,71	43,35	41,44	44,11	41,77
3-SHOTS-ADAPT LANG-TGT	29,33	35,84	47,89	39,17	48,22	47,50	45,87	44,04
3-SHOTS-ADAPT LANG-SRC	29,15	35,59	47,01	38,20	48,97	47,14	43,52	42,82
3-SHOTS-ADAPT DIA-TGT	29,09	35,42	48,00	34,53	43,68	43,21	43,34	44,06
3-SHOTS-ADAPT DIA-SRC	29,01	35,08	48,24	32,72	42,17	41,44	39,88	44,35
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	31,79	48,22	59,20	57,46	62,94	66,82	61,79	59,74
0-SHOT LANG-SRC	31,61	48,09	58,85	57,58	63,37	66,64	62,36	59,31
0-SHOT DIA-TGT	31,75	48,56	59,58	55,98	61,31	64,05	61,78	55,24
0-SHOT DIA-SRC	31,70	48,15	59,78	52,95	58,31	60,08	57,12	55,95
3-SHOTS LANG-TGT	32,12	49,68	62,53	58,51	64,07	68,13	62,57	60,68
3-SHOTS LANG-SRC	32,10	49,71	61,97	58,55	64,46	68,68	63,36	60,39
3-SHOTS DIA-TGT	32,10	49,65	61,31	58,07	63,18	67,90	62,48	60,70
3-SHOTS DIA-SRC	32,06	49,63	61,71	58,05	63,41	67,99	62,80	60,37
3-SHOTS-ADAPT LANG-TGT	32,08	49,68	61,82	58,96	64,54	68,13	63,49	60,88
3-SHOTS-ADAPT LANG-SRC	32,09	49,75	61,58	58,88	64,59	68,68	64,01	61,20
3-SHOTS-ADAPT DIA-TGT	32,08	49,65	61,64	58,84	63,73	67,90	63,30	61,53
3-SHOTS-ADAPT DIA-SRC	32,04	49,61	61,68	58,60	63,57	67,99	63,16	60,00

TABLE 20 – Scores chrF++ des modèles pour la traduction du français vers l'occitan, en fonction du dialecte du segment de référence. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Baseline COPIE	61,44	67,86	71,40	69,42	72,21	74,71	74,85	72,92
Revirada								
ADAPT	59,20	68,22	74,11	**76,79**	**76,33**	**82,11**	**79,08**	**76,94**
LANGUEDOCIEN	59,20	68,22	74,11	73,54	**76,33**	**82,11**	**79,08**	**76,94**
GASCON	57,02	59,05	63,38	**76,79**	68,74	69,78	71,76	69,64
NLLB-600	60,27	68,72	**75,42**	72,06	72,89	78,39	76,99	76,34
Google Translate	60,50	**69,86**	74,86	71,43	75,02	77,48	76,01	74,21
Aya23-8B								
0-SHOT LANG-TGT	60,01	66,25	71,22	68,35	71,41	73,17	72,78	71,81
0-SHOT LANG-SRC	60,24	66,30	69,79	68,60	71,16	73,34	73,33	72,53
0-SHOT DIA-TGT	59,76	65,54	67,12	68,04	69,25	72,47	72,94	67,39
0-SHOT DIA-SRC	59,54	65,41	68,46	67,85	71,50	72,19	73,54	68,10
3-SHOTS LANG-TGT	59,35	65,25	70,22	66,67	70,32	71,60	70,76	72,81
3-SHOTS LANG-SRC	59,20	65,13	69,60	66,26	69,09	71,28	70,41	71,55
3-SHOTS DIA-TGT	58,95	65,04	70,85	65,75	68,25	69,93	69,84	72,92
3-SHOTS DIA-SRC	58,91	64,74	70,90	65,25	66,61	69,01	70,68	70,82
3-SHOTS-ADAPT LANG-TGT	59,58	65,25	71,10	65,77	69,35	71,60	70,57	72,81
3-SHOTS-ADAPT LANG-SRC	59,53	65,13	70,66	65,41	69,84	71,28	69,09	70,61
3-SHOTS-ADAPT DIA-TGT	59,26	65,04	69,75	63,94	66,89	69,93	69,31	72,21
3-SHOTS-ADAPT DIA-SRC	59,35	64,74	70,31	63,00	66,24	69,01	67,79	71,90
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	60,65	68,07	72,27	71,33	74,02	76,48	75,05	74,36
0-SHOT LANG-SRC	60,27	67,43	72,20	71,23	74,09	76,12	74,64	76,31
0-SHOT DIA-TGT	60,28	67,40	71,60	68,73	71,55	72,01	73,55	69,26
0-SHOT DIA-SRC	60,12	67,06	70,43	65,37	69,12	68,76	69,88	70,64
3-SHOTS LANG-TGT	60,74	68,21	73,55	70,93	72,72	75,64	74,23	75,38
3-SHOTS LANG-SRC	60,68	68,18	72,80	70,95	72,93	75,99	74,87	74,60
3-SHOTS DIA-TGT	60,65	68,09	72,66	70,45	72,70	75,45	73,98	74,72
3-SHOTS DIA-SRC	60,70	68,12	72,08	70,52	72,56	75,59	74,47	74,64
3-SHOTS-ADAPT LANG-TGT	60,82	68,22	72,81	71,12	72,88	75,63	74,87	74,97
3-SHOTS-ADAPT LANG-SRC	60,73	68,20	72,97	71,02	72,85	75,99	75,22	75,00
3-SHOTS-ADAPT DIA-TGT	**60,88**	68,08	72,49	70,87	73,24	75,45	74,84	75,51
3-SHOTS-ADAPT DIA-SRC	60,79	68,13	71,79	70,77	72,82	75,59	74,89	73,68

TABLE 21 – Scores COMET des modèles pour la traduction du français vers l'occitan, en fonction du dialecte du segment de référence. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
References	98.75	99.68	99.88	95.32	94.93	94.92	92.14	93.31
Revirada								
ADAPT	99.54	99.54	99.74	**95.94**	95.00	95.08	93.24	92.70
LANGUEDOCIEN	99.54	99.54	99.74	95.55	95.00	95.08	93.24	92.70
GASCON	99.57	99.57	99.84	**95.94**	93.98	**95.64**	**93.27**	92.85
NLLB-600	99.37	99.37	99.90	94.50	92.07	94.02	91.50	94.00
Google Translate	**99.65**	**99.65**	**99.93**	94.23	93.94	93.95	89.95	93.40
Aya23-8B								
0-SHOT LANG-TGT	8.06	8.06	12.08	20.00	26.39	21.08	21.69	27.36
0-SHOT LANG-SRC	13.23	13.23	10.30	23.55	29.05	25.65	21.81	32.21
0-SHOT DIA-TGT	12.63	9.94	14.72	22.30	34.09	27.77	30.17	32.95
0-SHOT DIA-SRC	18.88	13.74	22.73	27.55	37.52	32.43	32.98	38.53
3-SHOTS LANG-TGT	21.91	21.91	47.23	39.25	50.03	41.54	37.17	46.47
3-SHOTS LANG-SRC	23.78	23.78	42.20	41.53	49.77	44.42	43.03	48.24
3-SHOTS DIA-TGT	26.81	20.98	48.71	36.98	45.87	42.27	43.15	46.73
3-SHOTS DIA-SRC	25.39	21.48	51.41	39.43	47.12	44.19	48.95	50.34
3-SHOTS-ADAPT LANG-TGT	18.20	21.91	68,87	48,61	49,95	41,54	45,97	48,62
3-SHOTS-ADAPT LANG-SRC	23.45	23.78	61,58	50,25	53,66	44,42	49,32	49,61
3-SHOTS-ADAPT DIA-TGT	20.99	20.98	64,89	43,95	49,60	42,27	53,19	55,30
3-SHOTS-ADAPT DIA-SRC	23.62	21.48	73,92	44,72	49,67	44,19	51,99	56,14
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	94.82	94.82	98.92	89.37	90.09	88.65	87.19	91.25
0-SHOT LANG-SRC	97.87	97.87	99.40	91.59	92.38	90.63	89.78	91.49
0-SHOT DIA-TGT	98.17	98.25	99.41	91.32	87.84	92.17	87.46	88.39
0-SHOT DIA-SRC	98.53	98.23	99.41	84.52	80.56	85.39	78.86	86.21
3-SHOTS LANG-TGT	99,31	99,31	99,71	93,86	93,98	93,34	89,51	95,99
3-SHOTS LANG-SRC	99,17	99,17	99,72	94,06	95,34	93,20	88,79	91,62
3-SHOTS DIA-TGT	99,18	99,41	99,73	94,22	92,96	93,51	90,44	96,07
3-SHOTS DIA-SRC	99,18	99,28	99,72	94,31	93,07	93,46	89,55	96,05
3-SHOTS-ADAPT LANG-TGT	99,23	99,35	99,74	94,34	94,09	93,34	89,41	95,09
3-SHOTS-ADAPT LANG-SRC	99,13	99,15	99,87	94,29	**95,98**	93,20	91,50	94,11
3-SHOTS-ADAPT DIA-TGT	98,99	99,41	99,83	94,85	93,04	93,51	90,91	**96,18**
3-SHOTS-ADAPT DIA-SRC	99,05	99,28	99,88	94,49	93,19	93,46	91,07	96,11

TABLE 22 – Scores LID_{oc} moyens sur les traductions français-occitan, par dialecte. Modèle LID : lid218e (x100). Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
References	98,83	98,69	96,85	86,67	80,53	87,93	82,90	84,03
Revirada								
ADAPT	98,22	98,22	**98,20**	87,74	**86,11**	**88,87**	**85,44**	88,29
LANGUEDOCIEN	98,22	98,22	98,20	**89,19**	**86,11**	**88,87**	**85,44**	88,29
GASCON	97,22	97,22	97,18	87,74	84,23	87,32	83,38	85,55
NLLB-600	98,03	98,03	96,78	88,13	85,62	87,55	83,90	87,88
Google Translate	**98,70**	**98,70**	96,85	88,50	83,92	88,51	83,45	87,68
Aya23-8B								
0-SHOT LANG-TGT	15,62	15,62	11,99	21,86	26,52	24,52	20,26	26,18
0-SHOT LANG-SRC	20,31	20,31	12,66	24,53	27,99	27,36	19,82	32,59
0-SHOT DIA-TGT	22,62	19,13	15,15	24,74	31,98	30,96	25,81	31,16
0-SHOT DIA-SRC	29,95	23,04	23,08	29,82	36,06	35,23	28,95	35,31
3-SHOTS LANG-TGT	35,34	35,34	48,03	41,32	46,53	44,52	40,63	44,02
3-SHOTS LANG-SRC	37,13	37,13	39,63	43,21	46,69	46,45	43,74	44,95
3-SHOTS DIA-TGT	39,51	34,39	44,40	39,22	43,01	44,79	43,51	46,50
3-SHOTS DIA-SRC	38,93	35,47	50,35	42,02	44,67	46,91	47,29	49,53
3-SHOTS-ADAPT LANG-TGT	31,86	35,34	64,33	48,68	48,18	44,52	46,03	47,23
3-SHOTS-ADAPT LANG-SRC	36,86	37,13	59,02	50,17	51,00	46,45	47,70	49,63
3-SHOTS-ADAPT DIA-TGT	35,08	34,39	60,22	44,70	49,85	44,79	50,18	53,07
3-SHOTS-ADAPT DIA-SRC	37,82	35,47	68,40	45,26	52,08	46,91	50,56	53,05
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	92,99	92,99	91,46	82,17	80,39	81,46	78,15	85,59
0-SHOT LANG-SRC	95,94	95,94	92,57	83,89	82,82	83,26	79,57	85,31
0-SHOT DIA-TGT	96,41	96,37	94,79	85,26	81,27	85,43	80,27	83,02
0-SHOT DIA-SRC	96,76	96,31	95,08	81,28	78,02	81,65	76,70	78,70
3-SHOTS LANG-TGT	97,52	97,52	94,14	87,50	84,85	87,04	82,55	89,08
3-SHOTS LANG-SRC	97,39	97,42	94,18	87,96	86,10	87,31	82,03	87,34
3-SHOTS DIA-TGT	97,37	97,54	94,26	87,95	83,75	87,50	82,79	**89,20**
3-SHOTS DIA-SRC	97,35	97,54	94,24	88,16	83,88	87,55	82,91	89,04
3-SHOTS-ADAPT LANG-TGT	97,41	97,54	94,16	87,70	84,86	87,04	81,79	88,71
3-SHOTS-ADAPT LANG-SRC	97,39	97,44	94,30	87,82	85,73	87,31	83,52	88,39
3-SHOTS-ADAPT DIA-TGT	97,08	97,55	93,86	88,00	85,01	87,50	83,34	88,77
3-SHOTS-ADAPT DIA-SRC	97,14	97,54	94,40	87,80	84,34	87,55	83,33	88,29

TABLE 23 – Scores LID_{oc} moyens sur les traductions français-occitan, par dialecte. Modèle LID : Idiomata Cognitor (x100). Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
References	98,38	99,76	99,85	92,04	92,60	91,25	85,86	90,16
Revirada								
ADAPT	99,64	99,64	99,72	**92,93**	**94,35**	92,17	**87,68**	90,45
LANGUEDOCIEN	99,64	99,64	99,72	92,55	**94,35**	92,17	**87,68**	90,45
GASCON	**99,77**	**99,77**	99,79	**92,93**	92,65	**92,32**	86,29	86,08
NLLB-600	99,37	99,37	99,34	91,50	92,36	90,75	84,77	92,47
Google Translate	99,74	99,74	**99,96**	91,62	91,30	90,98	85,39	91,43
Aya23-8B								
0-SHOT LANG-TGT	8,05	8,05	10,35	16,31	25,61	17,79	12,03	23,53
0-SHOT LANG-SRC	13,08	13,08	8,35	20,18	27,55	21,85	12,03	24,08
0-SHOT DIA-TGT	13,56	10,00	14,95	18,91	33,07	24,83	21,18	23,91
0-SHOT DIA-SRC	21,47	14,37	24,90	24,82	36,48	28,90	23,87	30,22
3-SHOTS LANG-TGT	24,43	24,43	55,67	38,00	52,06	40,32	31,94	43,01
3-SHOTS LANG-SRC	26,35	26,35	45,70	40,52	51,79	43,85	38,45	45,20
3-SHOTS DIA-TGT	29,73	22,97	53,26	35,32	47,56	41,15	38,67	43,97
3-SHOTS DIA-SRC	29,47	24,17	61,35	38,75	49,16	43,62	44,67	48,83
3-SHOTS-ADAPT LANG-TGT	19,94	24,43	71,18	48,74	51,87	40,32	40,63	46,36
3-SHOTS-ADAPT LANG-SRC	26,21	26,35	67,43	50,37	59,22	43,85	44,18	46,79
3-SHOTS-ADAPT DIA-TGT	23,40	22,97	70,16	43,55	53,41	41,15	51,33	54,86
3-SHOTS-ADAPT DIA-SRC	26,45	24,17	80,83	44,46	55,61	43,62	47,32	55,50
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	95,34	95,34	98,34	85,81	88,66	84,92	80,05	87,82
0-SHOT LANG-SRC	98,15	98,15	98,94	87,86	91,26	86,65	81,92	88,28
0-SHOT DIA-TGT	98,50	98,41	98,95	88,97	87,44	89,51	82,77	85,77
0-SHOT DIA-SRC	98,94	98,37	98,98	82,54	80,87	83,22	76,83	84,65
3-SHOTS LANG-TGT	99,42	99,42	99,90	91,34	92,37	90,75	83,87	94,16
3-SHOTS LANG-SRC	99,30	99,30	99,90	91,86	93,86	90,95	83,40	89,30
3-SHOTS DIA-TGT	99,30	99,48	99,91	91,78	92,40	91,09	84,51	**94,44**
3-SHOTS DIA-SRC	99,28	99,43	99,91	92,08	93,29	91,25	85,26	94,02
3-SHOTS-ADAPT LANG-TGT	99,46	99,43	99,91	91,74	92,73	90,75	83,80	92,60
3-SHOTS-ADAPT LANG-SRC	99,21	99,31	99,91	91,93	93,65	90,95	85,48	91,50
3-SHOTS-ADAPT DIA-TGT	99,25	99,48	99,90	92,57	92,47	91,09	85,77	93,23
3-SHOTS-ADAPT DIA-SRC	99,23	99,43	99,92	92,42	92,42	91,25	85,66	93,21

TABLE 24 – Scores LID_{oc} moyens sur les traductions français-occitan, par dialecte. Modèle LID : GlotLID (x100). Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Revirada								
ADAPT	0,43	0,15	**<0,01**	**0,85**	**1,16**	**1,14**	2,36	**0,05**
LANGUEDOCIEN	0.43	0.15	0.01	1.16	**1.16**	**1.14**	2.36	**0.05**
GASCON	0.37	**0.05**	**<0.01**	**0.85**	2.41	1.21	**2.06**	0.23
NLLB-600	0,40	0,23	**<0,01**	2,10	1,84	1,72	2,63	2,86
Google Translate	0,44	0,19	**<0,01**	2,36	1,71	2,24	4,87	1,67
Aya23-8B								
0-SHOT LANG-TGT	86.61	88.16	82.65	70.22	60.88	68.02	63.55	64.58
0-SHOT LANG-SRC	80,14	81,64	86,05	65,89	57,31	62,78	63,18	58,18
0-SHOT DIA-TGT	80.73	85.53	80.37	67.49	52.97	60.91	55.59	55.75
0-SHOT DIA-SRC	73,26	80,83	69,93	91,76	49,55	55,72	51,06	50,93
3-SHOTS LANG-TGT	69.79	71.11	44.25	49.22	37.92	46.29	46.28	43.36
3-SHOTS LANG-SRC	67,53	68,83	51,20	46,57	37,64	43,06	40,00	41,27
3-SHOTS DIA-TGT	64.17	72.06	45.26	51.72	40.98	45.64	41.49	44.20
3-SHOTS DIA-SRC	65,39	71,36	37,39	49,19	40,14	43,62	34,60	39,81
3-SHOTS-ADAPT LANG-TGT	74.04	71.11	22,94	39,99	37,48	46,29	38,83	40,96
3-SHOTS-ADAPT LANG-SRC	68.30	68.83	28,50	38,52	32,03	43,06	35,00	38,98
3-SHOTS-ADAPT DIA-TGT	71.02	72.06	26,91	44,98	37,85	45,64	31,04	34,05
3-SHOTS-ADAPT DIA-SRC	67,93	71,36	18,52	44,81	38,06	43,62	32,97	33,27
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	3,88	3,84	0,07	5,49	4,54	5,65	6,43	4,17
0-SHOT LANG-SRC	1,42	1,28	0,03	3,86	3,01	4,64	5,28	3,53
0-SHOT DIA-TGT	1,12	0,88	0,03	5,22	7,70	6,11	6,49	6,39
0-SHOT DIA-SRC	0,88	0,98	0,03	11,78	13,00	11,75	15,43	8,18
3-SHOTS LANG-TGT	**0,32**	0,19	**<0,01**	2,32	2,63	2,82	3,91	1,60
3-SHOTS LANG-SRC	0,46	0,32	**<0,01**	2,35	1,40	2,92	4,43	4,80
3-SHOTS DIA-TGT	0,37	0,15	**<0,01**	2,59	2,15	2,87	3,46	1,61
3-SHOTS DIA-SRC	0,45	0,22	**<0,01**	2,41	2,08	2,85	4,13	1,60
3-SHOTS-ADAPT LANG-TGT	0,36	0,17	**<0,01**	2,38	2,55	2,82	4,45	1,97
3-SHOTS-ADAPT LANG-SRC	0,50	0,33	**<0,01**	2,48	1,31	2,92	2,96	2,93
3-SHOTS-ADAPT DIA-TGT	0,48	0,15	**<0,01**	2,08	2,35	2,87	3,09	1,66
3-SHOTS-ADAPT DIA-SRC	0,44	0,22	**<0,01**	2,25	2,16	2,85	3,18	1,68

TABLE 25 – Scores MSE_{oc} (multipliés par 100) entre les traduction candidates (Flores/LoCongresNews) et leurs segments de référence, pour la direction de traduction français vers l'occitan. Résultats en fonction du dialecte du segment de référence. Modèle LID utilisé : lid218e. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Revirada								
ADAPT	0,27	0,18	0,52	**1,06**	3,66	**1,16**	2,41	1,48
LANGUEDOCIEN	0,27	0,18	0,52	1,41	3,66	**1,16**	2,41	1,48
GASCON	0,36	0,28	0,52	**1,06**	3,88	1,47	2,51	**0,95**
NLLB-600	0,34	0,24	**0,09**	1,87	4,65	1,69	**2,40**	2,21
Google Translate	**0,19**	**0,11**	0,60	2,14	**2,54**	1,57	3,85	1,22
Aya23-8B								
0-SHOT LANG-TGT	74,55	74,28	74,44	53,38	42,13	51,80	52,58	48,03
0-SHOT LANG-SRC	68,89	68,66	76,65	50,29	41,13	48,49	52,39	43,70
0-SHOT DIA-TGT	65,03	68,97	72,95	50,12	38,58	44,85	46,23	43,11
0-SHOT DIA-SRC	56,03	64,44	62,32	44,30	33,65	40,04	41,64	40,43
3-SHOTS LANG-TGT	49,27	49,21	34,77	32,90	27,79	31,04	30,38	30,16
3-SHOTS LANG-SRC	47,42	47,30	45,65	30,97	24,60	28,89	27,33	31,37
3-SHOTS DIA-TGT	44,72	50,13	36,61	35,22	29,72	30,87	28,47	29,30
3-SHOTS DIA-SRC	45,28	48,89	28,88	32,55	26,64	29,09	25,13	26,60
3-SHOTS-ADAPT LANG-TGT	53,44	49,21	20,08	26,96	22,95	31,04	25,81	25,65
3-SHOTS-ADAPT LANG-SRC	48,19	47,30	25,75	25,63	20,67	28,89	24,04	25,25
3-SHOTS-ADAPT DIA-TGT	49,81	50,13	20,42	30,70	22,40	30,87	22,75	21,02
3-SHOTS-ADAPT DIA-SRC	46,90	48,89	12,50	30,97	19,01	29,09	22,22	21,57
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	3,65	3,53	2,38	4,70	6,31	4,77	5,07	2,43
0-SHOT LANG-SRC	1,29	1,19	2,01	3,74	4,71	4,05	5,03	0,97
0-SHOT DIA-TGT	0,86	0,79	0,88	3,51	5,41	3,96	4,34	2,25
0-SHOT DIA-SRC	0,68	0,92	0,87	5,23	7,43	5,71	6,07	4,82
3-SHOTS LANG-TGT	0,36	0,26	1,57	2,15	4,53	2,55	4,32	1,82
3-SHOTS LANG-SRC	0,41	0,30	1,57	2,27	4,44	2,53	4,83	1,71
3-SHOTS DIA-TGT	0,39	0,24	1,56	2,36	5,96	2,58	4,08	1,97
3-SHOTS DIA-SRC	0,40	0,25	1,57	2,34	5,65	2,53	4,57	1,83
3-SHOTS-ADAPT LANG-TGT	0,41	0,26	1,57	2,21	4,98	2,55	5,13	2,19
3-SHOTS-ADAPT LANG-SRC	0,43	0,29	1,57	2,16	5,15	2,53	4,10	1,97
3-SHOTS-ADAPT DIA-TGT	0,45	0,24	1,41	2,22	5,53	2,58	4,18	2,16
3-SHOTS-ADAPT DIA-SRC	0,47	0,25	1,56	2,14	6,37	2,53	3,96	2,22

TABLE 26 – Scores MSE_{oc} (multipliés par 100) entre les traduction candidates (Flores/LoCongresNews) et leurs segments de référence, pour la direction de traduction français vers l'occitan. Résultats en fonction du dialecte du segment de référence. Modèle LID utilisé : Idiomata Cognitor. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores		LoCongresNews					
	ARA	LAN	AUV	GAS	LIM	LAN	PRO	VIV
Revirada								
ADAPT	0,68	0,09	0,01	**1,07**	**1,40**	**1,63**	**3,47**	**0,04**
LANGUEDOCIEN	0,68	0,09	0,01	1,82	**1,40**	**1,63**	**3,47**	**0,04**
GASCON	0,73	**0,08**	**<0,01**	**1,07**	3,84	2,35	4,00	3,21
NLLB-600	0,80	0,26	0,03	2,69	2,19	2,73	3,88	1,27
Google Translate	0,73	0,14	**<0,01**	3,51	2,38	2,96	7,16	1,24
Aya23-8B								
0-SHOT LANG-TGT	87,43	89,16	86,85	71,76	61,25	69,05	69,71	66,23
0-SHOT LANG-SRC	81,54	83,20	90,02	67,22	59,66	64,61	68,58	62,18
0-SHOT DIA-TGT	80,44	86,26	82,01	68,84	53,96	61,55	59,63	60,96
0-SHOT DIA-SRC	71,53	81,06	70,09	62,42	49,11	57,17	54,94	59,61
3-SHOTS LANG-TGT	68,60	70,04	36,31	48,17	35,10	45,76	46,62	45,31
3-SHOTS LANG-SRC	66,00	67,55	51,33	46,17	35,14	42,23	40,62	44,84
3-SHOTS DIA-TGT	62,70	71,52	40,00	51,21	39,67	45,17	42,22	44,62
3-SHOTS DIA-SRC	62,69	69,95	32,45	47,86	38,53	42,79	36,16	39,98
3-SHOTS-ADAPT LANG-TGT	73,19	70,04	25,36	38,66	34,60	45,76	38,63	40,18
3-SHOTS-ADAPT LANG-SRC	66,48	67,55	29,05	37,26	27,71	42,23	34,64	41,09
3-SHOTS-ADAPT DIA-TGT	69,31	71,52	24,10	43,81	33,74	45,17	29,54	33,61
3-SHOTS-ADAPT DIA-SRC	65,91	69,95	14,16	43,64	32,07	42,79	33,62	33,15
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	3,85	3,56	0,20	6,85	4,20	6,67	8,07	6,68
0-SHOT LANG-SRC	1,65	1,16	0,13	5,00	2,44	5,84	6,70	3,77
0-SHOT DIA-TGT	1,29	0,96	0,13	5,97	6,79	7,07	4,75	5,48
0-SHOT DIA-SRC	1,06	1,01	0,13	11,42	13,33	11,52	10,68	7,60
3-SHOTS LANG-TGT	**0,61**	0,14	**<0,01**	3,04	2,72	3,56	4,79	2,14
3-SHOTS LANG-SRC	0,79	0,25	**<0,01**	3,17	1,43	3,83	5,51	5,48
3-SHOTS DIA-TGT	0,66	0,14	**<0,01**	3,23	3,32	3,94	3,71	2,53
3-SHOTS DIA-SRC	0,72	0,16	**<0,01**	3,27	2,97	3,92	4,45	2,09
3-SHOTS-ADAPT LANG-TGT	0,64	0,14	**<0,01**	3,00	3,40	3,56	5,46	7,28
3-SHOTS-ADAPT LANG-SRC	0,89	0,25	**<0,01**	3,30	2,76	3,83	4,00	6,31
3-SHOTS-ADAPT DIA-TGT	0,78	0,14	**<0,01**	3,16	3,81	3,94	3,65	5,92
3-SHOTS-ADAPT DIA-SRC	0,82	0,16	**<0,01**	2,88	3,17	3,92	3,68	5,97

TABLE 27 – Scores MSE_{oc} (multipliés par 100) entre les traduction candidates (Flores/LoCongresNews) et leurs segments de référence, pour la direction de traduction français vers l'occitan. Résultats en fonction du dialecte du segment de référence. Modèle LID utilisé : GlotLID. Les meilleurs scores sont en gras. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores				LoCongresNews			
	OC	FR	ES	CAT	OC	FR	ES	CAT
References	98,76	0,09	0,12	0,47	86,48	3,75	1,09	1,70
Revirada								
ADAPT	98,22	0,19	0,18	0,69	89,11	1,73	1,06	1,88
LANGUEDOCIEN	98,22	0,19	0,18	0,69	**89,35**	**1,63**	1,06	1,93
GASCON	97,22	0,45	0,23	0,82	87,57	2,28	1,15	**1,79**
NLLB-600	98,03	0,23	0,25	0,67	88,31	1,79	1,15	1,91
Google Translate	**98,70**	**0,10**	**0,13**	**0,49**	88,15	2,40	**1,05**	1,80
Aya23-8B								
0-SHOT LANG-TGT	15,62	27,14	1,85	50,94	21,89	22,40	2,31	43,25
0-SHOT LANG-SRC	20,31	31,04	1,78	42,69	24,16	22,60	2,60	39,68
0-SHOT DIA-TGT	20,88	13,04	2,25	58,88	26,63	15,86	2,48	44,24
0-SHOT DIA-SRC	26,49	12,78	2,35	52,74	31,41	18,04	2,69	36,63
3-SHOTS LANG-TGT	35,34	3,13	1,51	55,53	44,17	7,87	2,60	34,27
3-SHOTS LANG-SRC	37,13	3,13	1,61	53,65	44,11	8,98	2,79	32,15
3-SHOTS DIA-TGT	36,95	3,69	1,56	53,25	43,57	10,15	2,64	31,92
3-SHOTS DIA-SRC	37,20	3,33	1,63	53,15	46,80	10,18	2,79	27,97
3-SHOTS-ADAPT LANG-TGT	33,60	2,91	1,65	57,29	49,83	8,91	2,62	27,02
3-SHOTS-ADAPT LANG-SRC	37,00	2,89	1,76	53,65	50,66	10,94	2,59	24,27
3-SHOTS-ADAPT DIA-TGT	34,74	3,07	1,71	55,69	50,47	11,69	2,67	23,01
3-SHOTS-ADAPT DIA-SRC	36,64	2,95	1,77	53,83	52,71	10,79	2,89	20,60
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	92,99	3,07	0,48	1,90	83,20	4,45	1,47	2,89
0-SHOT LANG-SRC	95,94	1,17	0,39	1,04	84,57	4,34	1,35	2,30
0-SHOT DIA-TGT	96,39	0,65	0,38	1,18	85,01	3,12	1,26	2,09
0-SHOT DIA-SRC	96,53	0,69	0,38	1,04	81,91	5,04	1,30	2,08
3-SHOTS LANG-TGT	97,52	0,39	0,26	0,89	87,53	2,16	1,25	2,52
3-SHOTS LANG-SRC	97,41	0,40	0,27	0,90	87,49	2,19	1,22	2,57
3-SHOTS DIA-TGT	97,46	0,38	0,26	0,91	87,57	2,07	1,21	2,37
3-SHOTS DIA-SRC	97,44	0,39	0,26	0,89	87,63	2,07	1,19	2,35
3-SHOTS-ADAPT LANG-TGT	97,47	0,41	0,26	0,89	87,38	2,39	1,24	2,27
3-SHOTS-ADAPT LANG-SRC	97,42	0,41	0,26	0,89	87,84	2,15	1,19	2,25
3-SHOTS-ADAPT DIA-TGT	97,31	0,41	0,28	0,94	87,75	2,16	1,20	2,26
3-SHOTS-ADAPT DIA-SRC	97,34	0,39	0,28	0,93	87,62	2,11	1,21	2,22

TABLE 28 – Scores LID moyens sur les traductions français-occitan, par étiquette LID de langue proche. Scores obtenus à partir des moyennes par dialecte. Modèle LID : Idiomata Cognitor (x100). Les meilleurs scores sont mis en gras (les plus élevés pour l'occitan, et les plus bas pour les autres langues). Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

	Flores				LoCongresNews			
	OC	FR	ES	CAT	OC	FR	ES	CAT
Revirada								
ADAPT	0,22	0,03	**<0,01**	0,04	**1,72**	1,37	**0,01**	0,08
LANGUEDOCIEN	0,22	0,03	**<0,01**	0,04	1,77	1,38	0,02	0,09
GASCON	0,32	0,04	**<0,01**	0,05	1,73	**1,34**	0,02	**0,07**
NLLB-600	0,29	0,03	0,01	0,05	2,15	1,49	0,02	0,09
Google Translate	**0,15**	**0,01**	**<0,01**	**0,03**	1,99	1,57	0,03	**0,07**
Aya23-8B								
0-SHOT LANG-TGT	74,41	25,81	0,46	41,87	53,72	17,04	0,20	33,31
0-SHOT LANG-SRC	68,78	29,22	0,30	33,36	52,11	17,61	0,21	29,86
0-SHOT DIA-TGT	67,00	11,41	0,54	47,57	49,31	10,68	0,19	33,60
0-SHOT DIA-SRC	60,23	10,87	0,42	40,78	43,73	12,22	0,21	26,07
3-SHOTS LANG-TGT	49,24	1,57	0,10	41,36	31,18	3,79	0,16	22,60
3-SHOTS LANG-SRC	47,36	1,64	0,12	39,43	31,47	4,51	0,20	21,57
3-SHOTS DIA-TGT	47,43	2,13	0,10	39,07	31,70	5,79	0,14	20,75
3-SHOTS DIA-SRC	47,08	1,81	0,11	38,89	28,15	6,24	0,17	17,00
3-SHOTS-ADAPT LANG-TGT	51,33	1,67	0,12	43,31	25,42	4,62	0,15	16,18
3-SHOTS-ADAPT LANG-SRC	47,74	1,64	0,14	39,71	25,04	6,67	0,14	13,84
3-SHOTS-ADAPT DIA-TGT	49,97	1,77	0,13	41,62	24,69	7,56	0,13	12,74
3-SHOTS-ADAPT DIA-SRC	47,90	1,72	0,14	39,67	22,56	6,58	0,17	10,60
Llama3.1-8B Instruct								
0-SHOT LANG-TGT	3,59	2,47	0,06	0,52	4,28	2,89	0,05	0,31
0-SHOT LANG-SRC	1,24	0,68	0,02	0,13	3,42	2,86	0,04	0,19
0-SHOT DIA-TGT	0,82	0,25	0,02	0,14	3,39	1,92	0,04	0,12
0-SHOT DIA-SRC	0,80	0,29	0,02	0,10	5,02	3,26	0,05	0,15
3-SHOTS LANG-TGT	0,31	0,07	0,01	0,06	2,83	1,46	0,06	0,18
3-SHOTS LANG-SRC	0,35	0,05	0,01	0,06	2,89	1,54	0,04	0,23
3-SHOTS DIA-TGT	0,32	0,05	0,01	0,06	3,08	1,65	0,05	0,17
3-SHOTS DIA-SRC	0,32	0,06	0,01	0,05	3,08	1,71	0,04	0,17
3-SHOTS-ADAPT LANG-TGT	0,33	0,09	0,01	0,06	3,10	1,57	0,06	0,15
3-SHOTS-ADAPT LANG-SRC	0,36	0,07	0,01	0,06	2,91	1,44	0,04	0,20
3-SHOTS-ADAPT DIA-TGT	0,35	0,07	0,01	0,06	3,01	1,66	0,06	0,14
3-SHOTS-ADAPT DIA-SRC	0,36	0,07	0,01	0,06	3,13	1,65	0,04	0,14

TABLE 29 – Scores MSE (multipliés par 100) entre les traduction candidates (Flores/LoCongresNews) et leurs segments de référence, pour la direction de traduction français vers occitan. Résultats par étiquette LID de langue proche. Scores obtenus à partir des moyennes par dialecte. Modèle LID utilisé : Idiomata Cognitor. Les meilleurs scores pour Aya-23 et Llama3 sont soulignés.

A.4.3 Exemples de traductions pour la direction occitan-français

La table 30 présente un exemple en aranais de Flores, et les traductions (occitan-français) proposées par chacun des modèles.

Source	«Ara, auem arrats de quate mesi que solien èster diabetics e que dejà non ac son», agreguèc.
Référence (français)	« Nous avons à présent des souris de 4 mois qui ne sont pas diabétiques alors qu'elles l'étaient auparavant », a-t-il ajouté.
Revirada-GASCON	«Aride, nous avons des rats de quatre mesi que solien être diabétiques et que déjà ils ne le sont», agrégea.
NLLB-600	Maintenant, nous avons environ quatre mois que nous avons des diabétiques et que nous ne sommes plus là, je crois.
Google Translate	« Nous avons maintenant des rats de quatre mois qui étaient diabétiques et qui ne le sont plus », a-t-il ajouté.
Aya-23 0-SHOT LANG-SRC	«A partir d'aujourd'hui, nous avons des rats qui sont diabétiques et qui ne sont pas encore soignés», a-t-il ajouté.
Llama3 0-SHOT LANG-SRC	«Aujourd'hui, il y a des rats de quatre mois qui sont diabétiques et qui ne sont plus là», ajoutai-je.

TABLE 30 – Exemple en dialecte aranais, tiré du corpus Flores, ainsi que quelques traductions candidates par les modèles de notre banc d'évaluation.

A.4.4 Exemples de traductions pour la direction français-occitan

La table 31 présente un exemple du sous-jeu en dialecte 'limousin' de LoCongresNews et les traductions (français-occitan) proposées par chacun des modèles.

Source	Enfin, il est accessible dans l'application dicod'Òc, c'est-à-dire également depuis votre smartphone, alors… 'chabatz d'entrar !
Référence (occitan limousin)	Enfin, es accessible dins l'aplicacion dicod'Òc, quò vòu dire dempuei vòstre esmartfòne, 'laidonc... 'chabatz d'entrar !
Revirada-ADAPT	Enfin, es accessible dins l'aplicacion dicod'Òc, valent a dire egalament dempuèi vòstre telefòn intelligent, alara… 'chabatz d'entrar !
NLLB-600	Enfin, es accessible dins l'aplicacion dicodÒc, es a dire tanben dempuèi lo vòstre smartphone, doncas... 'causatz d'entrar !
Google Translate	Fin finala, es accessible dins l'aplicacion dicod'Òc, es a dire tanben dempuèi vòstre smartphone, alara… 'chabatz d'entrar !
Aya-23 0-SHOT LANG-SRC	Enfin, il est accessible dans l'application dicod'Òc, c'est-à-dire également depuis votre smartphone, alors… 'chabatz d'entrar !
Aya-23 3-SHOTS-ADAPT DIA-SRC	Finalament, es accessible dins l'aplicacion dicod'Òc, c'est-a-dire tamben dins vòstre smartphone, encara… 'chabatz d'entrar !
Llama3 0-SHOT LANG-SRC	Enfin, es es accessible dins l'aplicacion dicod'Òc, c'est a dire tanben dempuèi vòstra smartphone, alara... 'chabatz d'entrar !
Llama3 3-SHOTS-ADAPT DIA-SRC	Enfin, es es accessible dins l'aplicacion dicod'Òc, c'est a dire tanben dempuèi vòstre smartphone, alora... chabatz d'entrar !

TABLE 31 – Exemple en dialecte limousin, tiré du corpus LoCongresNews, ainsi que quelques traductions candidates par les modèles de notre banc d'évaluation.