



HAL
open science

Ingénierie d'Invite pour la Traduction Automatique de Langues Vraiment peu Dotées - Etude Préliminaire du Cas de l'Occitan

Josiane Mothe, Serge Molina

► **To cite this version:**

Josiane Mothe, Serge Molina. Ingénierie d'Invite pour la Traduction Automatique de Langues Vraiment peu Dotées - Etude Préliminaire du Cas de l'Occitan. CORIA-RJCRI 2024, Apr 2024, La Rochelle, France. 10.24348/coria.2024.long_25 . hal-04727970

HAL Id: hal-04727970

<https://hal.science/hal-04727970v1>

Submitted on 9 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Ingénierie d’Invite pour la Traduction Automatique de Langues Vraiment peu Dotées - Etude Préliminaire du Cas de l’Occitan

Josiane Mothe^{1,2,*}, Serge Molina^{1,2}

¹IRIT, UMR5505, Toulouse, France

²Univ. de Toulouse, INSPE, Toulouse, France

¹IRIT, UMR5505, Toulouse, France

²Univ. de Toulouse, UPS, Toulouse, France

Abstract

Il y a plus de 7,100 langues vivantes dans le monde mais plus de 40% sont en danger; cela est le cas de l’occitan, langue du sud de la France et de parties de l’Espagne et d’Italie. Il existe peu de traducteurs automatiques pour l’occitan. Dans cet article, nous nous intéressons aux grands modèles de langue pour la traduction automatique en gascon, une variante de l’occitan. Cette méthode consiste à concevoir et à affiner des invites textuelles spécifiques pour guider les grands modèles de langage, afin de générer les sorties souhaitées. Nous nous sommes concentrés sur le modèle de langue pré-entraîné GPT-4 au travers de Chat-GPT. Nous avons testé différentes façons de rédiger des prompts et d’amener de la connaissance externe pour améliorer les traductions. Ici, nous discutons les différents résultats obtenus et présentons des pistes de réflexion relatives aux questions de recherche qui sont soulevées par ces résultats.

Keywords

Système d’Information, Recherche d’information, Langues en danger, Langues peu dotées, Traduction automatique, LLMs, Grands modèles de langues

1. Introduction

La langue n’est pas seulement un moyen de communication; elle permet également aux individus de partager des informations, des pensées, des idées et des émotions [1]. Le site Ethnologue ¹ recense plus de 7,100 langues vivantes dans le monde, parmi elles, plus de 40% sont en danger.

Selon www.ethnologue.com, “une langue en danger est une langue que la génération en âge de procréer ne transmet plus à ses enfants”. L’occitan est une langue en danger.

Les premiers textes écrits en occitan sont apparus au Xième siècle. Cette période est marquée par l’évolution de la langue à partir du latin et sa divergence par rapport aux autres langues romanes. L’occitan a connu une période florissante jusqu’au XIIIème siècle ; les troubadours² en sont un des témoignages. A partir du XIVème siècle, la langue a commencé à décliner au

CORIA 2024

✉ josiane.mothe@irit.fr (J. Mothe); serge.molina@irit.fr (S. Molina)

🌐 <https://www.irit.fr/~Josiane.Mothe> (J. Mothe)

🆔 0000-0001-9273-2193 (J. Mothe)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹<https://www.ethnologue.com/data-consulting>

²Les troubadours sont des poètes-musiciens médiévaux issues de la noblesse occitane qui composaient des poèmes lyriques sur les thèmes de la nature, de la chevalerie et de l’amour courtois.

L'accordéon diatonique

Comment ça marche ?

Le musicien tire (ouvre) ou pousse (ferme) le soufflet en appuyant sur les boutons à sa droite pour la mélodie et à sa gauche pour l'accompagnement. Il permet ainsi le passage de l'air dans les anches qui se situent à l'intérieur de l'instrument, entre les claviers et le soufflet. Chaque touche produit une note différente selon que le musicien tire ou pousse le soufflet (c'est ce qui différencie notamment l'accordéon diatonique du chromatique).

L'acordeon diatonic

Quin fonciona?

Lo musician que tira (qu'obreish) o que possa (que barra) l'acordeon en prèmer sus botons a man drete tà la melodia e a man esquèrra tà l'acompanhament. Que permet atau lo passatge de l'aire a las espiulas qui son plaçadas en dehens de l'instrument, enter los clavèrs e lo plec. Cada tòca que hè ua nòta desparièra segon que lo musician e tira o e posse lo plec (qu'ei çò qui diferèncià en especiau l'acordeon diatonic deu cromatic).

Figure 1: Extrait de fiches développées par le Centre occitan des musiques et danses traditionnelles (COMDT) sur les instruments de musique traditionnels, en français et gascon.

profil du français en lien avec un souci d'unification. Ce déclin s'est poursuivi avec l'école obligatoire, gratuite et en français post-révolution. Les efforts de revitalisation ont été tardifs. La première "calendreta"³ a ouvert en 1979. L'enseignement bilingue en langue régionale dans les établissements publics a été rendu possible plus tardivement⁴ avec un effort notable depuis 2021⁵.

L'occitan comprend plusieurs variantes et trois codes ISO639-3⁶: occitan, provençal (OCI), gascon (GSC) et occitan et provençal anciens (PRO). Un extrait de fiches développées par le Centre occitan des musiques et danses traditionnelles (COMDT)⁷ relatives à des instruments de musiques traditionnels est présenté en français et gascon dans la figure 1. Ces extraits serviront de base pour l'étude qui suit.

La préservation des langues en danger est cruciale. En effet, les langues reflètent de façon différente les dimensions linguistiques, culturelles, scientifiques et éthiques de la perception du monde et de la société humaine [2, 3].

Préserver une langue nécessite une approche multi-facettes, parmi lesquelles la création de ressources numériques textuelles occupe une place prépondérante pour l'éducation et la propagation linguistique. Cet article se concentre spécifiquement sur l'apport possible des grands modèles de langues pour la traduction automatique. La traduction automatique est souvent envisagée comme une solution pour enrichir et diversifier les ressources disponibles [4, 5, 6, 7]. Cette solution peut donc être naturellement envisagée pour faciliter l'accès et la diffusion de la langue dans un contexte numérique.

Il existe peu de traducteurs automatiques pour l'occitan. Lo Congrès⁸ propose un traducteur dans les deux sens entre les paires (français, gascon) et (français, languedocien). Il est à base de règles. La traduction qu'il propose du texte de la figure 1 est présentée dans la figure 2. Par ailleurs, la conférence WMT (Conference on Machine Translation) maintient un portail d'outils de traduction pour différentes langues. Pour l'occitan, cinq traducteurs sont recensés⁹ mais non

³Les établissements scolaires Calandreta (en occitan, petite alouette) sont des écoles maternelles et primaires, des collèges et actuellement un lycée immersifs en langue occitane. Les calandretas sont des établissements associatifs et laïques" Wikipédia, consulté le 2 février 2024.

⁴Circulaire n° 2001-166 du 5 septembre 2001

⁵Bulletin officiel n° 47 du 16 décembre 2021

⁶<https://documentation.abes.fr/guide/html/formats/CodesLanguesISO639-3.htm>

⁷<https://www.comdt.org/transmission/education-artistique-et-action-culturelle/ressources-en-ligne/>

⁸<https://locongres.org>

⁹<https://machinetranslate.org/occitan>

L'acordeon diatonic

Com aquò e marcha ?

Lo musician que tira (qu'obreish) o brot (fèrme) lo bohet en prement los botons a la soa dreita entà la melodia e a la soa esquèrra entà l'acompanhament. Que permet atau lo passatge de l'aire dens las espiulas qui's situan a l'interior de l'instrument, enter los clavèrs e lo bohet. Cada tòca que produseish ua nòta diferenta segon que lo musician tira o que possa lo bohet (qu'ei çò qui diferèncià sustot l'acordeon diatonic deu cromatic).

Figure 2: Traduction en gascon via le service La Revirada de Lo Congrès (à gauche), identifiée comme la traduction REV.

évalués: Apertium, NeuralSpace, ModernMT, Baidu Translate et Alibaba Translate. Apertium¹⁰ est également à base de règles. Il s'agit d'un outil ouvert et permet de rajouter des ressources linguistiques en particulier pour intégrer de nouvelles langues. Il se focalise sur des langues peu dotées. La version pour l'occitan a été réalisée en lien avec Lo Congrès.

Quelques techniques permettent de traduire automatiquement des textes dans des langues peu dotées, pour lesquelles des corpus parallèles de grande taille n'existent pas ou pour lesquels les outils linguistiques sont rares ou inexistant. Haddow et al. propose une étude sur la question [8] en se centrant sur les approches à base de réseaux de neurones. Durant la dernière édition de la conférence WMT consacrée à la traduction dans des langues peu dotées (ici 4 langues indiennes) [7], les participants se sont essentiellement focalisés sur l'utilisation de grands modèles de langues multi-lingues tels que IndicTrans2 [9] qui comprend 22 langues indiennes ou monolingues tels que T5 [10] et BART [11], et ont utilisé des méthodes de réglage fin des modèles [12], d'augmentation de données par rétro-traduction [13] et de débruitage des données [11].

Dans notre étude, nous nous sommes focalisés sur l'ingénierie des invites. Cette méthode consiste à concevoir et à affiner des invites textuelles spécifiques pour guider les grands modèles de langage, afin de générer les sorties souhaitées. L'avantage de cette méthode est qu'elle ne nécessite pas d'apprentissage du modèle contrairement aux méthodes d'apprentissage fin. Cette méthode s'appuie sur un meilleur usage des connaissances apprises par le modèle [14].

L'objectif de ce travail est une étude préalable autour de trois questions de recherche:

- Q1: Apprentissage par l'exemple
- Q2: Intégration des connaissances sur la langue
- Q3: Evaluation - Adéquation des mesures actuelles aux langues peu dotées

Dans le cas de langues très dotées en textes numériques, les grands modèles de langues sont capables de produire des textes de bonne qualité linguistique. Ils sont également efficaces pour la traduction entre langues très dotées mais sont moins efficaces sur les langues peu dotées [15, 16]. Brown et al. ont montré que l'ajout d'exemples pouvaient aider dans différentes tâches [17]. Nous testons cette méthode de contextualisation à partir d'exemples. Nous nous intéressons également à l'ajout de connaissances sur la langue dans le contexte de l'invite. En effet, les descriptions de la langue que développent les linguistes peuvent apporter des informations

¹⁰<https://github.com/apertium>

L'acordeon diatonic

Com aquò marcha ?

Lo musician tira (obre) o pòusa (tampa) lo soflèt en prement sus los botons a sa dreita per la melodia e a sa esquèrra per l'acompanhament. Aquò permet atau lo passatge de l'aire dens las ancas qu'ei situat a l'interior de l'instrument, entre los clavièrs e lo soflèt. Cada boton produseish ua nòta diferenta segon que lo musician tira o pòusa lo soflèt (aquò's cò que destria notablament l'acordeon diatonic deu cromatic).

L'acordeon diatonic

Com funciona?

Lo musician tira (dubris) o pren (clau) lo soflèt tot prement sus los botons a sa dreita per la melodia e a sa mancha per l'acompanhament. Atau, permet que l'aire passe per las lengüetas plaçadas a l'interior de l'instrument, entre los clavièrs e lo soflèt. Cada clau hè una nòta diferenta se lo musician tira o pren lo soflèt (aquò's cò que hè la diferèncià principau entre l'acordeon diatonic e lo cromatic).

Figure 3: Traduire X en Y du français au gascon; deux versions fournies par ChatGPT. Identifiées comme traductions GPT1 et GPT1b

précieuses dans le cas de langues peu présentes dans les collections d'entraînement des grands modèles de langues.

Nous nous sommes concentrés sur le modèle de langue pré-entraîné GPT-4 au travers de ChatGPT. Nous avons testé différentes façons de rédiger des prompts et d'amener de la connaissance externe pour améliorer les traductions. Il s'agit d'un travail préliminaire qui permet d'illustrer l'aide que peuvent apporter les grands modèles de langues dans la traduction dans des langues peu apprises par les modèles pré-entraînés. Min et al.[14] distinguent plusieurs approches pour utiliser les invites dans des tâches de traitement automatique des langues. Nous avons choisi d'étudier l'utilisation d'instructions, sans et avec des exemples ainsi que les instructions augmentées de connaissances externes.

Cet article comprend cinq sections. Après cette introduction, la section 2 présente les traductions automatiques obtenues par les différentes approches à base d'invites que nous avons testées, sur un texte court. La section 3 discute les différents résultats obtenus et présente des pistes de réflexion relatives aux questions de recherche qui sont soulevées par ces résultats. Nous discutons des travaux reliés dans la section 4. Finalement, la section 5 conclut cet article.

2. Apprendre via les instructions

2.1. Instruction traduire

L'instruction typique ici est "traduire X en Y" [10] qui ne fonctionne que si le modèle a vu suffisamment de données. Nous l'avons testée sur l'exemple proposé en figure 1, à partir de la version française. Li et al. a montré que la façon de rédiger l'invite pouvait avoir également une influence sur les résultats [18]; nous n'avons pas étudié cet aspect ici.

La figure 3 à gauche (que nous identifions comme GPT1) a été obtenue avec l'instruction: *Traduire le texte suivant qui est en français vers le gascon* . La figure 3 (GPT1b) à droite a été obtenue avec le texte *Peux-tu proposer une autre traduction?* .

2.2. Avec quelques exemples

Brown et al. [17] ont montré que les modèles génératifs étaient également efficaces lorsque des exemples (entrée, sortie) leur sont fournis. Nous avons donc proposé les exemples suivants, issus de d'autres fiches: *F1: La cornemuse languedocienne. Comment ça marche ?*

L'acordeon diatonic.

Com aquò marcha ?

Lo musician tira (obre) o pren (tampa) lo soflèt en premènt sus los botonets a sa dreita per la melodia e a sa esquèrra per l'acompanhament. Aquò permet atau lo passatge de l'aire dins las anshas qu'ei plaçat a l'interior de l'instrument, entre los clavièrs e lo soflèt. Cada tecla produsís ua nòta diferenta segon que lo musician tira o pren lo soflèt (aquò's cò que destria notablament l'acordeon diatonic deu cromatic).

L'acordeon diatonic.

Quin fonciona ?

Lo musician qu'envia aire en bohant en lo bufet e que prem la pòcha dab lo son braç tà har lo son. La melodia que's hè suu graile gràcias aus horats de jòc. Cada boton qu'ei ligat a ua nòta diferenta, que depen se lo musician tira o pòusa lo soflèt, cò qu'establis la diferèncià màger enter l'acordeon diatonic e lo cromatic.

Figure 4: Traduction du texte de la Figure 1 après quelques exemples. Identifiées comme traductions GPT2 et GPT2b dans la suite

Le musicien envoie de l'air en soufflant dans le bufet et presse la poche avec son bras pour produire le son. La mélodie se joue sur le graile grâce aux trous de jeu.

G1: La bodega. Quin fonciona ?

Lo musician qu'envia aire en bohant en lo bufet e que prem la pòcha dab lo son braç tà har lo son. La melodia que's hè suu graile gràcias aus horats de jòc.

F2: La cornemuse languedocienne est la plus grosse cornemuse de France, car elle est fabriquée grâce à une peau de chèvre entière : elle peut ainsi contenir entre 80 et 100 litres d'air !

G2: La bodega qu'ei la mei bèra boha de França, per'mor qu'ei fabricada gràcias a ua pèth de craba sancèra qui pòt tièner enter 80 e 100 litres d'aire !

F3: La vielle à roue. Comment ça marche ?

Le musicien tourne la roue grâce à la manivelle pendant qu'il joue de la main gauche la mélodie sur les touches du clavier. Ainsi, la roue frotte les cordes qui se mettent à vibrer.

G3: La sonsaina. Quin fonciona ?

Lo musician que vira l'arròda gràcias a la manivèla mentre que jòga dab la man esquèrra la melodia sus las tòcas deu clavèr. Atau, l'arròda que hrega las còrdas qui's hican a bronzinar.

F4: Le bourdon qui passe par le chien s'appelle la «cigale» ou la «trompette».

G4: Lo bordon qui passa peu can que s'apèra la « cigala » o la « trompeta ».

F5: Sautereaux : petites pièces en bois qui appuient sur les cordes lorsqu'une touche du clavier est actionnée, ce qui crée des notes comme le font les doigts sur le manche d'un violon.

G5: Sauterèus : peçòtas en husta qui premen sus las còrdas quan ua tòca deu clavèr ei tocada, cò qui crea nòtas com ac hèn los dits suu mange d'un vriolon.

La figure 4 de gauche (que nous identifions comme GPT2) est obtenue avec l'instruction *Traduire le texte*. Celle de droite (GPT2b) avec l'instruction *peux-tu faire une autre version en apprenant des exemples que je t'ai fournis?*

Ici (voir figure 4 partie droite), nous pouvons constater que le modèle a déjà sur-appris, mélangeant un peu les instruments puisqu'il indique que le joueur souffle dans l'accordéon. Il a en revanche bien appris la façon choisie de traduire "Comment ça marche?" issue des exemples fournis dans le contexte.

2.3. Avec des connaissances extérieures

Le gascon est une langue peu dotée et il n'est pas possible de savoir sur quels textes GPT-4 a été appris. Nous faisons l'hypothèse que des règles complémentaires sur la langue peuvent être

L'acordeon diatonic

Com funcionèra ?

Lo musician tira (descura) o empòrta (encèra) lo fuelhet en empètant sus los botonets a sa dreita per la melodia e a sa esquèrra per l'acompanhament. Que permet aishi lo passatge de l'aire dins las anshas que se situan a l'interior de l'instrument, entre los clavèrs e lo fuelhet. Cadun botòn prodús ua nòta diferenta segon se lo musician tira o empòrta lo fuelhet (qu'ei aquò que diferènciá notadament l'acordeon diatonic deu cromatic).

L'acordeon diatonic

Com funcionèra ?

Lo musician tira (descura) o empòrta (encèra) lo fuelhet en empètant sus los botonets a sa dreita entà la melodia e a sa esquèrra entà l'acompanhament. Que permet lo passatge de l'aire dins las anshas, que se situan a l'interior de l'instrument, entre los clavèrs e lo fuelhet. Cadun botòn prodús ua nòta diferenta segon se lo musician tira o empòrta lo fuelhet (aquò qu'ei çò que diferènciá notadament l'acordeon diatonic deu cromatic).

Figure 5: Traduction du texte de la figure 1 après l'apport du document sur l'utilisation de la particule Que en gascon. Identifiées comme traductions GPT3 et GPT3b

utiles à la traduction.

Nous lui avons donc fourni un document en français expliquant l'usage de la particule "que", une particularité du gascon¹¹, mais d'autres ressources pourraient être apportées comme les dictionnaires préférés, des conjugaisons ou des règles de grammaire.

La figure 5 de gauche est obtenue avec l'instruction *peux-tu utiliser cette connaissance pour traduire le texte suivant* : , après lui avoir fourni le document. Celle de droite avec l'instruction *peux-tu produire une autre version?* .

Dans ce cas particulier de la particule "que", les exemples ont été plus formateurs que le texte décrivant l'utilisation de la particule.

2.4. Traduction dans d'autres langues romanes

L'évaluation est normalement réalisée entre le texte de référence, générale produit par un humain, et le texte produit automatiquement, tous deux dans la langue cible de traduction. Nous avons constaté (cf. figure 7) que les scores entre français pris comme référence et gascon pris comme texte traduit étaient assez élevés. Nous avons donc voulu évaluer la ressemblance avec les autres langues romanes. Les langues romanes ont en effet des traits communs.

Nous nous sommes intéressés aux traductions dans les plus communes de ces langues. Bien que nous n'ayons pas les références dans ces langues, nous souhaitons utiliser ces traductions pour les comparer aux traductions en gascon. Selon Wikipédia, les langues romanes "les plus parlées dans le monde sont l'espagnol (480 millions de locuteurs natifs), le portugais (227), le français (93), l'italien (71), le roumain (24) et le catalan (4)¹²". Nous avons donc traduit le texte dans chacune de ces langues avec l'instruction *traduire le texte suivant qui est en français vers le L* où L est successivement remplacée par la langue mentionnée (cf. les traductions obtenues dans la figure 6.

¹¹<http://linguistica-oc.com/wp-content/uploads/2013/07/Linguistica-Occitana-4-Morin.pdf>

¹²https://fr.wikipedia.org/wiki/Langues_romanes, consulté le 12 février 2024

<p>ESPAGNOL</p> <p>El acordeón diatónico ¿Cómo funciona? El músico tira (abre) o empuja (cierra) el fuelle presionando los botones a su derecha para la melodía y a su izquierda para el acompañamiento. Esto permite el paso del aire a través de las lengüetas que se encuentran dentro del instrumento, entre los teclados y el fuelle. Cada tecla produce una nota diferente dependiendo de si el músico tira o empuja el fuelle (esto es lo que diferencia principalmente el acordeón diatónico del cromático).</p>	<p>PORTUGAIS</p> <p>O acordeão diatônico Como funciona? O músico puxa (abre) ou empurra (fecha) o fole pressionando os botões à sua direita para a melodia e à sua esquerda para o acompanhamento. Isso permite a passagem do ar pelas palhetas que estão localizadas no interior do instrumento, entre os teclados e o fole. Cada tecla produz uma nota diferente dependendo se o músico puxa ou empurra o fole (isso é o que diferencia especialmente o acordeão diatônico do cromático).</p>
<p>FRANCAIS</p> <p>L'accordéon diatonique Comment ça marche ? Le musicien tire (ouvre) ou pousse (ferme) le soufflet en appuyant sur les boutons à sa droite pour la mélodie et à sa gauche pour l'accompagnement. Il permet ainsi le passage de l'air dans les anches qui se situent à l'intérieur de l'instrument, entre les claviers et le soufflet. Chaque touche produit une note différente selon que le musicien tire ou pousse le soufflet (c'est ce qui différencie notamment l'accordéon diatonique du chromatique).</p>	<p>ITALIEN</p> <p>L'organetto diatonico Come funziona? Il musicista tira (apre) o spinge (chiude) il mantice premendo i pulsanti alla sua destra per la melodia e alla sua sinistra per l'accompagnamento. Questo permette il passaggio dell'aria nelle anche che si trovano all'interno dello strumento, tra le tastiere e il mantice. Ogni tasto produce una nota diversa a seconda che il musicista tiri o spinga il mantice (questo è ciò che differenzia in particolare l'organetto diatonico da quello cromatico).</p>
<p>ROUMAIN</p> <p>Acordeonul diatonic Cum funcționează? Muzicianul trage (deschide) sau împinge (inchide) folele apăsând pe butoanele de la dreapta pentru melodie și de la stânga pentru acompaniament. Acest lucru permite trecerea aerului prin anele care se află în interiorul instrumentului, între claviaturi și fole. Fiecare tastă produce o notă diferită în funcție de faptul că muzicianul trage sau împinge folele (acesta este aspectul care diferențiază în mod special acordeonul diatonic de cel cromatic).</p>	<p>CATALAN</p> <p>L'acordió diatònic Com funciona? El músic estira (obre) o empeny (tanca) el fuelle prement els botons a la seva dreta per a la melodia i a la seva esquerra per a l'acompanyament. Això permet el pas de l'aire a través de les llengüetes que es troben a l'interior de l'instrument, entre els teclats i el fuelle. Cada tecla produeix una nota diferent segons que el músic estiri o empenyi el fuelle (això és el que diferencia principalment l'acordió diatònic del cromàtic).</p>

Figure 6: Traduction du texte de la figure 1 dans les langues romanes les plus parlées - la version française est la version originale.

3. Evaluation et discussion

3.1. Méthodologie

Nous avons évalué les résultats selon deux approches:

- comparaison de la version de référence en gascon avec les versions traduites automatiquement,
- comparaison de la version de référence en français avec les versions en français après rétro-traduction en français des versions traduites automatiquement en gascon. La rétro-traduction est souvent utilisée dans les langues peu dotées pour augmenter les données [19], ici, nous l'utilisons comme moyen complémentaire d'évaluation.

3.2. Métriques

Nous avons évalué les résultats en utilisant des mesures calculées de façon automatique: BLEU [20], ROUGE [21], chrF [22] et TER [23]. Ces mesures sont utilisées dans le domaine de la

traduction automatique [7] et sont définies comme suit:

- BLEU (BiLingual Evaluation Understudy) se base sur le ratio entre les n-grammes présents dans le texte candidat et ceux présents dans le texte de référence. Ce ratio est pondéré de façon à ne pas surévaluer les répétitions de mots et à favoriser les traductions de longueur similaire entre le texte candidat et celui de référence. La longueur maximale des n-grammes considérés est un paramètre. Nous avons utilisé l'implémentation fournie par la librairie Python NLTK [24] qui correspond à l'équation 1.

$$BLEU_N(Ref, Cand) = bp * exp\left(\sum_{n=1}^N \frac{1}{N} * \log\left(\frac{\sum_{n_gram \in Cand} \min(occ_{Cand}(n_gram), occ_{Ref}(n_gram))}{\sum_{n_gram \in Cand} occ_{Cand}(n_gram)}\right)\right) \quad (1)$$

Où $bp = 1$ si le texte candidat traduit (Cand) est plus long que le texte de référence (Ref), sinon $bp = e^{1 - \frac{|Ref|}{|Cand|}}$. Occ signifie occurrence. N correspond à la longueur maximale des n-grammes considérés; nous avons choisi $N=4$, valeur suggérée par le papier original.

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) calcule le ratio entre le nombre de n-grammes communs au texte de référence et au texte candidat et le nombre de n-grammes dans le texte traduit (précision) multiplié par le ratio de n-grammes communs aux deux textes divisé par le nombre de n-grammes dans le texte original (rappel). Ce produit est ensuite divisé par la somme de la précision et du rappel. Il existe différentes variantes de ROUGE. Nous avons retenu ici ROUGE-L qui considère la longueur de la séquence commune la plus longue plutôt que le nombre de n-grammes communs et utilisé l'implémentation de la librairie Python Rouge [25] présentée dans l'équation 2.

$$ROUGE - L(Ref, Cand) = \frac{(1 + \beta^2) * R_{LCS}(Ref, Cand) * P_{LCS}(Ref, Cand)}{R_{LCS}(Ref, Cand) + \beta^2 * P_{LCS}(Ref, Cand)} \quad (2)$$

$$R_{LCS}(Ref, Cand) = \frac{LCS(Ref, Cand)}{|Ref|}$$

$$P_{LCS}(Ref, Cand) = \frac{LCS(Ref, Cand)}{|Cand|}$$

Où $LCS(Ref, Cand)$ calcule la longueur de la séquence commune à Ref et Cand la plus longue.

- chrF est une métrique qui agrège la précision et le rappel des n-grammes au niveau des caractères. Cette métrique intègre un paramètre β ajustable de manière à donner plus de poids à la précision par rapport au rappel des n-grammes. Dans ce papier, nous avons utilisé l'implémentation fournie par la librairie Python NLTK [24] présentée dans l'équation 3, avec $N=6$.

		BLEU_4		CHR6_6		ROUGE-L		TER	
		ORFR	ORGSC	ORFR	ORGSC	ORFR	ORGSC	ORFR	ORGSC
ORFR		1.00	0.45	1.00	0.36	1.00	0.13	0.00	89.02
ORGSC		0.45	1.00	0.34	1.00	0.13	1.00	92.41	0.00
REV		0.48	0.73	0.35	0.64	0.12	0.60	96.20	50.00
GPT1		0.48	0.64	0.35	0.51	0.14	0.47	87.34	58.54
G	GPT1B	0.42	0.60	0.31	0.46	0.13	0.42	94.94	64.63
a	GPT2	0.46	0.64	0.33	0.51	0.14	0.45	87.34	60.98
s	GPT2B	0.28	0.47	0.23	0.37	0.11	0.38	94.94	79.27
c	GPT3	0.46	0.64	0.34	0.52	0.15	0.46	87.34	59.76
o	GPT3B	0.46	0.65	0.34	0.52	0.15	0.47	88.61	59.76
n	MISTRAL	0.43	0.49	0.32	0.35	0.14	0.23	89.87	76.83
ORGSC-RETRO		0.79	0.42	0.77	0.35	0.68	0.12	37.97	93.90
F	REV-RETRO	0.84	0.41	0.83	0.34	0.75	0.12	20.25	90.24
r	GPT1-RETRO	0.86	0.43	0.84	0.35	0.75	0.12	20.25	90.24
a	GPT1B-RETRO	0.79	0.38	0.77	0.33	0.67	0.12	37.97	95.12
n	GPT2-RETRO	0.80	0.40	0.82	0.35	0.72	0.12	26.58	93.90
ç	GPT2B-RETRO	0.59	0.35	0.47	0.27	0.38	0.09	77.22	95.12
a	GPT3-RETRO	0.80	0.40	0.78	0.34	0.68	0.12	30.38	92.68
l	GPT3B-RETRO	0.81	0.40	0.79	0.34	0.68	0.12	30.38	91.46
s	MISTRAL-RETRO	0.75	0.41	0.66	0.34	0.55	0.11	45.57	91.46

Figure 7: Évaluation des traductions comparées aux textes originaux en français (ORFR) et en gascon (ORGSC). Le bloc de lignes "Gascon" correspond à la traduction en gascon de la version française, le troisième bloc correspond à la rétro-traduction du gascon en français. REV correspond à la traduction avec Revirada, les autres avec une invite sur chat-GPT4 et Mistral. Les blocs en colonnes correspondent aux différentes métriques. En gras, nous avons indiqué les valeurs les plus grandes par colonne. Les couleurs indiquent l'ordre de similarité des scores de rouge (peu similaire) à vert (très similaire).

catalan, la langue romane la plus proche du gascon dans l'ensemble des langues considérées, le français puis le portugais et l'espagnol. Nous observons également que les textes produits en gascon (cf la figure 7) sont bien mieux notés que n'importe lequel des textes dans les autres langues romanes (BLEU-4 jusqu'à 0.73 contre 0.46 ici).

		BLEU_4		CHR6_6		ROUGE-L		TER	
		ORFR	ORGSC	ORFR	ORGSC	ORFR	ORGSC	ORFR	ORGSC
ORFR		1	0.4527	1	0.3622	1	0.1345	0	89.0244
ORGSC		0.4523	1	0.3445	1	0.1345	1	92.4051	0
CATALAN		0.382	0.4588	0.2868	0.3295	0.1043	0.1724	96.2025	89.0244
ESPAGNOL		0.3566	0.4224	0.2802	0.3145	0.0877	0.1565	100	89.0244
ROUMAIN		0.3314	0.3508	0.2668	0.3076	0.0517	0.0855	98.7342	95.122
ITALIEN		0.3594	0.386	0.293	0.3254	0.0504	0.1	98.7342	92.6829
PORTUGAIS		0.3349	0.4267	0.2638	0.3269	0.0893	0.1062	96.2025	93.9024

Figure 8: Évaluation des traductions dans les langues romanes les plus utilisées comparées au texte original en français (ORFR) et à la traduction humaine en gascon (ORGSC). Les deux premières lignes correspondent aux versions originales en français et gascon; le bloc de lignes suivant correspond à la traduction en langues romanes de la version française réalisée à l'aide de chat-GPT4. Les autres conventions sont les mêmes que pour la figure 7.

Ces premiers résultats nous encouragent à étudier plus en détail la traduction du français vers l'occitan. Un traducteur plus performant serait intéressant pour produire par exemple des ressources pédagogiques dans l'enseignement public bilingue ou pour aider la diffusion de la langue à travers l'écrit.

Ces résultats posent cependant plusieurs questions.

Les résultats produits ici ne sont pas reproductibles; en effet le texte généré par une même invite ne sera pas le même à chaque essai. Il est alors difficile d'évaluer de façon sûre les résultats, même si l'amélioration du contexte semble être une piste prometteuse. La différence dans les mesures sur des variantes produites avec le même contexte pose également question quant à la généralisation des résultats.

Comme souvent dans les langues peu dotées, l'écriture n'est pas normée; nous ne savons pas quelle influence cela peut avoir sur l'évaluation par les mesures automatiques (et manuelles), ni sur les résultats produits.

L'intégration de connaissances dans le contexte pourrait combiner des exemples et des ressources linguistiques. L'étude d'une intégration plus systématique via des mécanismes de "retrieval augmented generation" [6] est une piste que nous souhaiterions poursuivre.

La création d'un grand modèle de langue propre aux langues romanes ou aux langues les plus proches de l'occitan est une piste intéressante. Différents travaux se sont intéressés à spécialiser des modèles sur des domaines ou des langues [28, 29, 30, 31]; d'autres ont montré que l'ajout de plusieurs langues semble améliorer les résultats de chacune [18].

Finalement, se pose également le problème des capacités de calcul et de l'énergie nécessaires pour réaliser ce type de traduction en comparaison de modèles à base de règles et d'exemples.

4. Etude reliées

Différents travaux s'intéressent aux cas de langues peu dotées. L'état de l'art proposé par Haddow et al. en 2022 catégorisent ainsi les approches selon qu'elles s'intéressent à des corpus mono-lingue ou multi-lingue ou si elles intègrent d'autres ressources [8]. Handy et al. étudient la qualité des modèles GPT de type InstructGPT [32] qui utilisent un apprentissage par renforcement sur la base de préférences humaine par rapport à des systèmes commerciaux sur 18 paires de langues, certaines très dotées, d'autres peu dotées[16]. Leurs résultats indiquent que les modèles GPT sont compétitifs dans les langues à ressources élevées mais présentent des limitations dans les langues à faibles ressources. Selon eux, des approches hybrides combinant des modèles GPT avec des systèmes de traduction traditionnels peuvent améliorer la qualité de la traduction. La supériorité des modèles à base de grands modèles de langue par rapport à des modèles plus standard d'apprentissage n'est pas encore montrée. L'étude de WeLocalize semble indiquer l'inverse¹⁴. Pour améliorer la précision de la traduction, Liu et al. combinent la rétro-traduction avec le remplacement de mots peu fréquents et la correction d'erreurs grammaticales pour enrichir le corpus d'entraînement [6]. AlignInstruct s'appuie sur un discriminateur multilingue basé sur des alignements statistiques de mots pour fournir une supervision multilingue explicite lors de la mise au point de la traduction automatique [33].

5. Conclusion

Dans cet article, nous nous sommes intéressés à l'utilisation d'invites pour la traduction en gascon, une langue peu dotée mais pour laquelle quelques ressources existent toutefois [34, 35].

¹⁴<https://www.welocalize.com/do-llms-or-mt-engines-perform-translation-better/>

Notre objectif était d'illustrer le potentiel de l'utilisation des grands modèles de langues et leurs limites. Les tests préliminaires ont été réalisés sur un document unique. Nos travaux futurs visent à étudier à plus grande échelle ces phénomènes. Nous étudierons également dans quelle mesure les résultats sur les grands modèles de langues peuvent influencer sur d'autres types de modèles, moins voraces en données et moins énergivores.

Il existe des modèles pré-entraînés pour une langue particulière comme par exemple CamemBERT [28] ou FlauBERT [29] pour le français, GPT-SW3 [36] pour le suédois ou entraînés pour un domaine particulier comme DrBert [30] ou aliBert [31] entraîné sur des textes bio-médicaux en français. Cette piste de développement est intéressante car elle nécessite moins de données que les modèles multi-lingues. D'autres modèles sont plutôt volontairement multi-lingues comme BLOOM [37], entraîné sur 46 langues et 13 langages de programmation. La présence de nombreuses langues semblent permettre de diminuer les erreurs pour les différentes langues [18] et de nombreux travaux s'intéressent aux modèles multilingues comme mBert [38], un BERT multilingue pré-entraîné sur 104 langues les plus dotées sur Wikipedia ou mT5 [39]. Un niveau intermédiaire pourrait être celui de modèles multi-lingues mais pour des langues proches où pour lesquelles il existe un lien particulier. Le grand modèle de langue yugoGPT en est un exemple; HBS LLM intègre les langues croate, bosniaque, serbe et monténégrine¹⁵.

Nos travaux futurs dans le domaine s'intéresseront à différents types de questions: Serait-il pertinent de créer un grand modèle de langue qui s'appuierait sur la classification des langues romanes et en engloberait un plus ou moins grand nombre (cf. figure 9¹⁶)? Est-il plus pertinent de conserver des modèles de traduction à base de règles et d'apprentissage par des exemples? Comment intégrer mieux les connaissances linguistiques dans les différents types de modèles? Comment les modèles rendent-ils compte des variantes et comment s'accommodent-ils de l'absence de standardisation d'une langue?

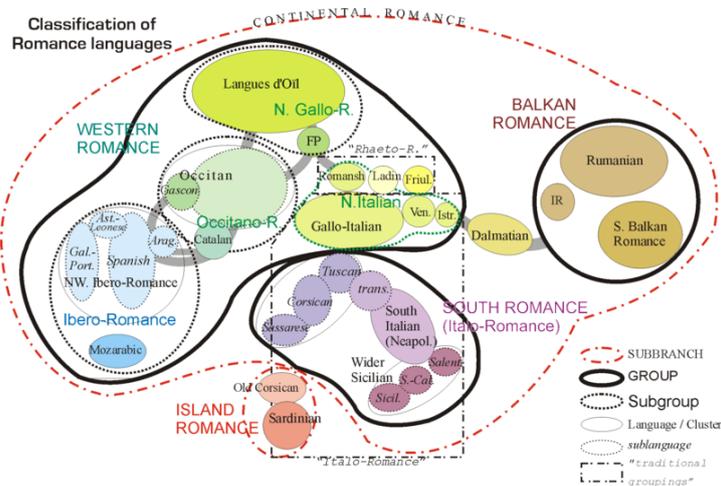


Figure 9: Classification des langues romanes - Wikipedia (février 2024). Image de Yuri B. Koryakov

¹⁵<https://multiplatform.ai/yugogpt-transforming-ai-for-south-slavic-languages/>

¹⁶https://fr.wikipedia.org/wiki/Langues_occitano-romanes Yuri B. Koryakov, Atlas of Romance languages, Moscou 2001, https://www.academia.edu/51180219/Koryakov_2001_Atlas_of_Romance_languages

Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche (project GUIDANCE, ANR-23-IAS1-0003).

References

- [1] J. M. Wilce, *Language and emotion*, 25, Cambridge University Press, 2009.
- [2] L. A. Grenoble, L. J. Whaley, *Saving languages: An introduction to language revitalization*, Cambridge University Press, 2005.
- [3] R. Smeets, Language as a vehicle of the intangible cultural heritage, *Museum international* 56 (2004) 156–165.
- [4] A. Ghafoor, A. S. Imran, S. M. Daudpota, Z. Kastrati, R. Batra, M. A. Wani, et al., The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing, *IEEE Access* 9 (2021) 124478–124490.
- [5] Q. Liu, M. Kusner, P. Blunsom, Counterfactual data augmentation for neural machine translation, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021*, pp. 187–197.
- [6] X. Liu, J. He, M. Liu, Z. Yin, L. Yin, W. Zheng, A scenario-generic neural machine translation data augmentation method, *Electronics* 12 (2023) 2320.
- [7] S. Pal, P. Pakray, S. R. Laskar, L. Laitonjam, V. Khenglawt, S. Warjri, P. K. Dadure, S. K. Dash, Findings of the WMT 2023 shared task on low-resource indic language translation, in: *Proceedings of the Eighth Conference on Machine Translation, 2023*, pp. 682–694.
- [8] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl, A. Birch, Survey of low-resource machine translation, *Computational Linguistics* 48 (2022) 673–732.
- [9] J. Gala, P. A. Chitale, R. AK, S. Doddapaneni, V. Gumma, A. Kumar, J. Nawale, A. Sujatha, R. Puduppully, V. Raghavan, et al., Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages, *arXiv preprint arXiv:2305.16307* (2023).
- [10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *The Journal of Machine Learning Research* 21 (2020) 5485–5551.
- [11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, *arXiv preprint arXiv:1910.13461* (2019).
- [12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020*, pp. 38–45.
- [13] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, *arXiv preprint arXiv:1511.06709* (2015).
- [14] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz,

- D. Roth, Recent advances in natural language processing via large pre-trained language models: A survey, *ACM Computing Surveys* 56 (2023) 1–40.
- [15] W. Jiao, W. Wang, J. Huang, X. Wang, Z. Tu, Is chatgpt a good translator? yes with gpt-4 as the engine, *arXiv preprint arXiv:2301.08745* (2023).
- [16] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, H. H. Awadalla, How good are gpt models at machine translation? a comprehensive evaluation, *arXiv preprint arXiv:2302.09210* (2023).
- [17] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [18] J. Li, H. Zhou, S. Huang, S. Cheng, J. Chen, Eliciting the translation ability of large language models via multilingual finetuning with translation instructions, 2023. *arXiv:2305.15083*.
- [19] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale, *arXiv preprint arXiv:1808.09381* (2018).
- [20] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: <https://aclanthology.org/P02-1040>. doi:10.3115/1073083.1073135.
- [21] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: <https://aclanthology.org/W04-1013>.
- [22] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: <https://aclanthology.org/W15-3049>. doi:10.18653/v1/W15-3049.
- [23] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, 2006, pp. 223–231. URL: <https://aclanthology.org/2006.amta-papers.25>.
- [24] S. Bird, E. Klein, E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*, ” O’Reilly Media, Inc.”, 2009. URL: <https://www.nltk.org/>.
- [25] P. Tardy, Rouge, a full python librarie for the rouge metric, GitHub repository (2021). URL: <https://github.com/pltrdy/rouge>.
- [26] M. Post, A call for clarity in reporting BLEU scores, in: *Proceedings of the Third Conference on Machine Translation: Research Papers*, Association for Computational Linguistics, Belgium, Brussels, 2018, pp. 186–191. URL: <https://www.aclweb.org/anthology/W18-6319>.
- [27] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: <https://aclanthology.org/W05-0909>.

- [28] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, arXiv preprint arXiv:1911.03894 (2019).
- [29] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, arXiv preprint arXiv:1912.05372 (2019).
- [30] Y. Labrak, A. Bazoge, R. Dufour, M. Rouvier, E. Morin, B. Daille, P.-A. Gourraud, Drbert: A robust pre-trained model in french for biomedical and clinical domains, bioRxiv (2023) 2023–04.
- [31] A. Berhe, G. Draznieks, V. Martenot, V. Masdeu, L. Davy, J.-D. Zucker, Alibert: A pre-trained language model for french biomedical text, in: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, 2023, pp. 223–236.
- [32] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems* 35 (2022) 27730–27744.
- [33] Z. Mao, Y. Yu, Tuning llms with contrastive alignment instructions for machine translation in unseen, low-resource languages, arXiv preprint arXiv:2401.05811 (2024).
- [34] B. Dazéas, Dicod’òc: du dictionnaire en ligne à la multi-application lexicographique, *Dicod’òc: du dictionnaire en ligne à la multi-application lexicographique* (2022) 157–174.
- [35] M. Bras, Nouvelles perspectives pour la linguistique occitane à partir de la base textuelle batelòc, *Les langues régionales de France: Nouvelles approches, nouvelles méthodologies, revitalisation* (2023) 121.
- [36] A. Ekgren, A. C. Gyllensten, E. Gogoulou, A. Heiman, S. Verlinden, J. Öhman, F. Carlsson, M. Sahlgren, Lessons learned from gpt-sw3: Building the first large-scale generative language model for swedish, in: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3509–3518.
- [37] B. Workshop, T. L. Scao, al., Bloom: A 176b-parameter open-access multilingual language model, 2023. arXiv:2211.05100.
- [38] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [39] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).