Analyse des métriques de tokenisation et leur corrélation avec les performances de traduction automatique multilingue

Serge Molina[1, 2] Josiane Mothe[1, 3]
(1) Université Paul Sabatier, 118 Rte de Narbonne, 31400, Toulouse, France
(2) IRIT, UMR5505, CNRS
(3) Université Jean-Jaurès, UT2J, INSPE, 79 Av de l'URSS, 31400, Toulouse, France
Serge.Molina@irit.fr, Josiane.Mothe@irit.fr

RÉSUMÉ

Cet article étudie le résultat de la tokenisation obtenu par deux tokeniseurs basés sur des sous-mots dans le contexte de la traduction automatique multilingue en utilisant le corpus FLORES-200. Il s'intéresse à 9 langues issues de 3 familles, avec différents niveaux de présence numérique. Tout d'abord nous évaluons les performances de traductions entre paires de langues et de familles de langue. Ensuite nous explorons les liens potentiels entre des mesures de caractérisation des résultats de la tokenisation et les performances de traduction. Notre objectif est de confirmer ou réfuter l'hypothèse selon laquelle la capacité d'un modèle de traduction multilingue de type encodeur-decodeur à traduire un texte peut être prédite en examinant les caractéristiques des tokens générés. Nous nous appuyons sur des mesures calculées à partir de la phrase source et de sa version traduite par un expert (version de référence), représentées sous forme de chaînes de caractères et de listes de tokens générés par le tokeniseur du modèle considéré. Nous n'avons pas détecté de corrélation fortes entre entre les caractéristiques des tokens et l'efficacité de traduction. Cette étude préliminaire propose toutefois des pistes pour de futures recherches visant à intégrer des mesures de caractérisation plus qualitatives pour évaluer l'impact des représentations tokenisées sur les performances de traduction.

ABSTRACT

Analysis of Tokenization Metrics and their Correlation with Multilingual Machine Translation Performance

This article studies the tokenization outcomes produced by two subword-based tokenizers in the context of multilingual machine translation using the FLORES-200 corpus. It focuses on 9 languages from 3 families, each with varying levels of digital presence. First, we evaluate translation performance between language and family pairs. Then, we explore the potential relation between measures characterizing the tokenization results and translation performance. We aim to validate or invalidate the hypothesis that the difficulty a multilingual translation model would have in translating a text could be inferred from features related to the tokens produced by its tokenizer. The measures are calculated based on the source sentence and its expert-produced reference translated version. We could not detect strong correlation between the features related to tokens and the translation effectiveness. This preliminary study suggests however directions for future research aiming to include more qualitative features to evaluate the impact of tokenized representations on translation performance.

MOTS-CLÉS : Traitement Automatique des Langues, Traduction Automatique Multilingue, Traduction Automatique Neuronale, Tokeniseur, Tokenisation, SentencePiece, BPE, FLORES-200, Modèles de Langue Pré-entraînés, BLEU, chrF, BERTScore.

KEYWORDS: Natural Language Processing, Multilingual Machine Translation, Neural Machine Translation, Tokenizer, Tokenization, SentencePiece, Byte Pair Encoding, FLORES-200, Pretrained Language Models, BLEU, chrF, BERTScore.

Atelier EvalLLM2024 à TALN 2024, Toulouse

1 Introduction

Les tâches de traitement automatique des langues (TAL), comme la traduction automatique, nécessitent le pré-traitement des textes. Dans le contexte des modèles de langues pré-entraînés (PLM), les tokeniseurs sont utilisés en préalable aux traitements des textes afin de les segmenter en unités plus petites (tokens) identifiées de manière unique et auxquelles sont associées des représentations denses. Les tokeniseurs qui divisent le texte en sous-mots permettent de gérer les mots rares et les variations de mots et sont plus à même de traiter des textes en différentes langues en comparaison avec les tokeniseurs utilisant un vocabulaire constitués de mots entiers. Par exemple l'algorithme SentencePiece (Kudo & Richardson, 2018) avec Byte-Pair Encoding (BPE) (Sennrich *et al.*, 2015) apprend à fusionner de manière itérative les paires de caractères les plus fréquentes dans un corpus d'entraînement. Il constitue ainsi un vocabulaire qui représente le corpus d'entraînement et un modèle de langue basé sur ce vocabulaire.

Différentes études ont montré l'intérêt des tokeniseurs basés sur les sous-mots dans différentes tâches et différentes langues. Choo & Kim (2023) montrent que dans la tâche d'analyse d'opinion de revues en coréen, le tokeniseur non supervisé créant des sous-chaînes (SentencePiece) est meilleur que le tokeniseur Mecab-Ko, supervisé, basé sur les morphèmes et utilisé pour le coréen. Alyafeai *et al.* (2023) s'intéressent à l'arabe et comparent 6 tokeniseurs pour différentes tâches de classification.

L'étude préliminaire que nous présentons ici vise à étudier deux tokeniseurs basés sur des sous-mots. Nous étudions les liens possibles entre les caractéristiques des listes de tokens obtenus pour différentes langues et la performance d'un modèle pré-entraîné utilisant ce tokeniseur dans la tâche de traduction. Notre objectif est de vérifier l'hypothèse selon laquelle il serait possible de prédire l'efficacité de traduction par un modèle de langue pré-entrainé, à partir de caractéristiques obtenues sur les tokens produits par son tokeniseur. L'intérêt de cette approche est de mobiliser des représentations peu coûteuses à obtenir : les chaînes de caractères et liste de tokens, en amont de l'utilisation de l'architecture transformer, coûteuse en ressources de calcul.

Domingo *et al.* (2019) se sont également intéressés aux performances de traduction en fonction des tokeniseurs. Ils ont comparé 5 tokeniseurs sur 10 paires de langues et concluent que le tokeniseur peut amener un gain de 0.12 sur la mesure BLEU (cf Section 2 pour la description de cette mesure).

Pour notre étude, nous avons sélectionné un sous-ensemble de langues de trois familles. Pour chaque famille de langues, nous avons retenu trois langues ayant différent degrés de représentation sur Wikipédia. Ces 9 langues sont utilisées, soit en tant que langue source, soit en tant que langue destination de la traduction. Toutes les paires de langues sont considérées. Nous avons retenu deux tokeniseurs et le modèle de langue pré-entraîné associé. Nous avons choisi la collection FLORES-200 pour les évaluations. L'évaluation de la traduction s'appuie sur des mesures automatiques de similarité entre la traduction automatique et une référence. La caractérisation des tokeniseurs s'appuie sur des mesures de la littérature.

Le reste de cet article est structuré comme suit. Dans la section 2, nous présentons la collection de données, les modèles utilisés pour la traduction et leurs tokeniseurs ainsi que les mesures choisies pour évaluer la performance de traduction et celles pour caractériser les résultats des tokeniseurs. La section 3 précise le choix des langues et la méthodologie suivie pour l'analyse. La section 4 présente les résultats de cette étude préliminaire. Enfin, la section 5 conclu ce travail et indique les pistes pour nos travaux futurs.

2 Données et mesures

2.1 Flores-200 : un corpus multilingue

FLORES-200 est un ensemble de données multilingues créé par Meta pour évaluer les systèmes de traduction automatique dans 200 langues(Costa-jussà *et al.*, 2022) [1]. FLORES-200 étend FLORES-101 (Goyal *et al.*, 2022) et comprend des phrases qui ont été traduites manuellement à partir de phrases issues du web. Les 3 001 phrases sont réparties en ensembles de développement (dev) et test (devtest). Chaque phrase a été traduite par des experts depuis l'anglais vers les 200 langues du jeu de données. Cette collection de données est utilisée dans le domaine de la traduction automatique (Maillard *et al.*, 2023; Gu *et al.*, 2023) mais également pour d'autres tâches de traitement automatique des langues (Kargaran *et al.*, 2023).

2.2 Tokeniseurs et PLM pour la traduction

Les tokeniseurs NLLB-200 (NLLB Team *et al.*, 2022) et MADLAD-400 (Kudugunta *et al.*, 2023) utilisés dans notre étude sont basés sur l'algorithme Sentence Piece (Kudo & Richardson, 2018) avec une implantation de type Byte Pair Encoding (Gage, 1994) et ont été conçus respectivement pour les modèles de type encodeur-décodeur NLLB-200 (NLLB Team *et al.*, 2022) et MADLAD-400 (Kudugunta *et al.*, 2023). Ces deux modèles ont exclusivement été entraînés sur une tâche de traduction multilingue.

NLLB-200 : Le tokeniseur du modèle NLLB-200 a été entraîné sur un sous-ensemble de 100 million de textes issus du jeu de données également utilisé pour l'entraînement du modèle encodeur-décodeur. Deux approches ont été utilisées de manière à permettre au tokeniseur de représenter de manière équitable les langues supportées par le modèle avec un vocabulaire de 256 004 tokens. Tout d'abord un échantillonage de type *temperature sampling* (Chen *et al.*, 2021) est utilisé avec un paramètre de température de 5 (NLLB Team *et al.*, 2022) ; cela permet d'équilibrer la quantité de texte en langues peu dotées vue par le tokeniseur pendant son entraînement par rapport aux langues fortement dotées. Ce modèle n'utilise pas la technique du byte fallback, qui permet d'utiliser des tokens par défaut correspondant à chaque caractère UTF-8 pour représenter les expressions hors vocabulaire. Compte tenu du taux élevé de tokens inconnus pour les langues zho_Hans, zho_Hant et yue_Hant, les données dans ces langues ont été sur-échantillonnées lors de l'entraînement avec un facteur de 5 (NLLB Team *et al.*, 2022). Ce sur-échantillonage permet de maintenir la fréquence de tokens inconnus générés par le tokeniseur en dessous de 1% pour chacune des langues supportées par ce modèle.

MADLAD-400 : Le modèle MADLAD-400 utilise un vocabulaire de 256 000 tokens avec utilisation

1. https://github.com/facebookresearch/flores/blob/main/flores200/README.md

du byte fallback de manière à ne pas générer de tokens inconnus (Kudugunta *et al.*, 2023). Les détails de l'entraînement du tokeniseur ne sont pas présentés dans l'article introduisant MADLAD-400 (Kudugunta *et al.*, 2023).

Les deux tokeniseurs incluent des tokens spéciaux correspondant à la langue de destination pour la traduction, ces tokens doivent être insérés avant la séquence de tokens générés de manière à spécifier à l'encodeur-décodeur la langue vers laquelle le texte doit être traduit. Pour NLLB-200, il y a la possibilité optionnelle de spécifier la langue d'origine du texte à traduire, cette possibilité n'est pas disponible pour MADLAD-400.

2.3 Mesures de caractérisation des phrases

Nous avons choisi différentes mesures de la littérature afin de caractériser l'encodage des chaînes de caractères par les tokeniseurs et analyser leur lien avec les performance de traduction des PLM associés. Seules certaines de ces mesures ont été proposées dans un contexte de traduction multilingue par leurs auteurs.

Ces mesures sont de deux catégories : les mesures calculées sur un seul texte (dans notre contexte une phrase dans une langue) et les mesures calculées sur une paire de textes (ici, une paire de phrases parallèles dans deux langues).

Mesures calculées sur une phrase

Nom	Description	Formule
ctc (Zhang *et al.*, 2022)	Ratio entre le nombre de tokens générés et le nombre de caractères	$\frac{\lvert Tokens\rvert}{\lvert Caracteres\rvert}$
ur (Zhang *et al.*, 2022)	Proportion de tokens inconnus générés	$\frac{\lvert UnknownTokens\rvert}{\lvert Tokens\rvert}$
f (Rust *et al.*, 2021)	Ratio entre le nombre de tokens générés et le nombre de mots présents	$\frac{\lvert Tokens\rvert}{\lvert Mots\rvert}$
cp (Rust *et al.*, 2021)	Proportion de mots qui génèrent plusieurs tokens	$\frac{\sum_{Mots_i \in Mots} \lvert tokenize(Mots_i)\rvert > 1}{\lvert Mots\rvert}$
cpt (Limisiewicz *et al.*, 2023)	Ratio entre le nombre de caractères dans le texte et le nombre de tokens générés	$\frac{\lvert Caracteres\rvert}{\lvert Tokens\rvert}$
cm	Valeur médiane du nombre de tokens générés pour chaque mots dans le texte	$mediane(\{\lvert tokenize(Mots_1)\rvert; \ldots; \lvert tokenize(Mots_i)\rvert\})$

TABLE 1 – Mesures calculées sur un texte. $|X|$ est le nombre d'éléments de X. $tokenize(X)$ correspond à la fonction d'encodage d'une chaîne de caractères en liste de tokens.
Les noms originaux des mesures sont $ctcl$: Closeness to the Character Level, ur : UNK Rate, f : Fertility, cp : Continuation Proportion, cpt : Character Per Token, cm : Continuation Median

Ces mesures mobilisent un texte (ici une phrase) représenté sous forme de chaînes de caractères ainsi que les tokens produits par le tokeniseur pour ce texte. Le tableau 1 présente ces mesures.

Closeness to the Character Level et *UNK Rate* ont été définies pour mesurer l'impact de la sous-représentation des langues lors de l'entraînement d'un tokeniseur pour un modèle de langue (Zhang *et al.*, 2022). Les auteurs mettent en parallèle la performance spBLEU (une variante de la mesure $BLEU_N$ présentée plus loin) avec ces deux mesures, en fonction de la quantité de texte dans la langue considérée dans le jeu d'entraînement du tokeniseur. *Closeness to the Character Level* est présentée comme un indicateur d'une possible dégradation de performance de traduction lorsque sa valeur dépasse 0.87 ; cette valeur est de 3.7% pour *UNK Rate* (Zhang *et al.*, 2022).

L'étude qui introduit les mesures *Fertility* et *Continuation Proportion* montre que ces mesures sont corrélées négativement avec la performance sur différentes tâches de TALN (Rust *et al.*, 2021). Au contraire, la mesure *Character per Token* est corrélée positivement sur différentes tâches (Limisiewicz *et al.*, 2023).

Nous proposons une mesure supplémentaire, *Continuation Median* qui correspond au nombre de tokens médian généré par mots.

Mesures calculées sur une paire de phrases source et destination de traduction

Dans le cadre de la traduction automatique, ces mesures permettent de comparer des phrases tokenisées dans les langues source et destination. Leur calcul nécessite une connaissance à priori de la traduction de référence qui sera utilisée pour évaluer la traduction candidate.

p quantifie les inégalités d'accès aux grands modèles de langue, cette inégalité est due au sur-coût lié à la production en plus grand nombre de tokens lors du traitement de phrases en langues peu représentées dans le jeu d'entraînement du tokeniseur.

Nous proposons deux familles de mesures qui permettent de quantifier les mesures définies en section entre une phrase source et une phrase de destination : *mesure.ratio* et *mesure.diff*.

Le tableau 2 présente ces mesures.

Nom	Description	Formule
p (Petrov *et al.*, 2023)	Ratio entre le nombre de tokens générés pour une phrase de référence en langue de destination et cette phrase en langue source	$\frac{\lvert DestTokens\rvert}{\lvert SrceTokens\rvert}$
$mesure.ratio$	Ratio entre une mesure $Mesure$ calculée pour une phrase en langue de destination et cette phrase en langue source	$\frac{mesure(Dest)}{mesure(Srce)}$
$mesure.diff$	Différence entre une mesure $mesure$ calculée pour une phrase en langue de destination et cette phrase en langue source	$mesure(Dest) - mesure(Srce)$

TABLE 2 – Mesures calculées sur une paire de textes (phrases). Le nom original de la mesure p est Premium

2.4 Mesures d'évaluation de la traduction

Dans cet article, l'évaluation des résultats de la traduction utilise des mesures usuelles du domaine. Nous nous sommes restreints aux mesures qui peuvent être calculées de façon automatique afin de limiter les ressources humaines nécessaires pour l'évaluation : $BLEU_4$ (Papineni *et al.*, 2002), $chrF_6$ (Popović, 2015) et F_{BERT} (Tianyi *et al.*, 2020). Ces mesures sont utilisées dans le domaine de la traduction automatique (Pal *et al.*, 2023; Moghe *et al.*, 2024). Elles sont décrites ci-après.

— $BLEU_N$ (BiLingual Evaluation Understudy) est la proportion de n-grammes au niveau des mots du texte résultant de la traduction (phrase candidate) dans le texte (phrase) de référence. Ce ratio est pondéré pour ne pas donner trop d'importance aux mots qui sont répétés et pour favoriser les traductions qui ont une longueur similaire au texte de référence. La longueur maximale des n-grammes (N) est un paramètre de la mesure (Papineni *et al.*, 2002).

— $chrF_N$ combine la précision et le rappel des n-grammes au niveau des caractères. $chrF_N$ intègre le paramètre β qui permet de donner plus de poids à la précision qu'au rappel, la longueur des n-grammes est définie par le paramètre N (Popović, 2015).

Nous utilisons N=6.

— F_{BERT} utilise le calcul de mesure F1 qui combine la précision et le rappel ; ceux-ci sont obtenus à partir d'un alignement de plongements lexicaux contextuels des tokens entre la phrase de référence et la phrase candidate. Ces plongements sont obtenus à l'aide d'un encodeur de type BERT. Cet alignement se fait par recherche de paires de tokens maximisant la similarité cosinus des plongements lexicaux, une fois l'alignement réalisé les scores de similarités sont utilisés pour pondérer les mesures de rappel et précision (Tianyi *et al.*, 2020).

3 Méthodologie de production des données

Dans cette étude, nous avons souhaité étudier en détail les résultats de traduction automatique. Plutôt que de sélectionner l'ensemble des 200 langues, nous nous sommes concentrées sur 9 langues issues de 3 familles de manière à permettre une évaluation croisée entre langues de différentes familles.

Pour sélectionner les langues étudiées, nous avons procédé comme suit :

1. Suppression des langues non supportées par MADLAD-400 et NLLB-200 ; 139 langues sont supportées par les deux outils ;
2. Tri des familles par somme du nombre de personnes ayant contribué au Wikipedia de chaque langue au sein de chaque famille ;
3. Regroupement des langues par famille ; 3 langues sont représentées pour chaque famille retenue à l'étape précédente ;
4. Tri au sein de chaque famille par nombre de personnes ayant contribué au Wikipedia de chaque langue ; la valeur maximum est de 122 498, la valeur minimum est de 75 ;
5. Sélection de 3 langues par famille en fonction de l'ordre de tri précédent : première, seconde et dernière langue.

Nous avons choisi d'utiliser le nombre de personnes ayant contribué aux articles Wikipédia comme critère de choix des langues, pour deux raisons. D'une part Wikipedia est une ressource utilisée dans les deux modèles que nous avons choisis. D'autre part, le nombre d'articles Wikipédia est parfois utilisé comme critère ; il nous semble que le nombre de contributeurs est une meilleure évaluation de

l'intérêt de la communauté numérique pour une langue, en s'affranchissant mieux des robots d'édition automatique qui peuvent amener une homogénéité dans les textes produits. Par exemple, la langue Cebuano est la deuxième en terme de nombre d'articles Wikipedia mais la majorité des articles sont générés automatiquement par un robot [2].

Les langues sélectionnées sont décrites dans le tableau 3.

iso-639-3	Langue	Famille	Script	#Contributeur Wikipedia	Support digital
eng	Anglais	Germanique	Latin	122 498	Prospère
nld	Hollandais	Germanique	Latin	3 685	Prospère
ltz	Luxembourgeois	Germanique	Latin	75	Vital
fra	French	Romane	Latin	17 122	Prospère
spa	Espagnol	Romane	Latin	14 871	Prospère
glg	Galicien	Romane	Latin	260	Vital
rus	Russe	Balto-Slave	Cyrillique	9 570	Prospère
pol	Polonais	Balto-Slave	Latin	4 371	Prospère
mkd	Macédonien	Balto-Slave	Cyrillique	241	Vital

TABLE 3 – Détails des langues sélectionnées. Le nombre de contributeurs Wikipedia pour chaque langue est issu de https://commons.wikimedia.org/wiki/Data:Wikipedia_statistics/data.tab. Le classement du niveau de support digital est issu de https://www.ethnologue.com/ (Simons *et al.*, 2022)

Nous avons encodé pour les 9 langues choisies l'ensemble des 2009 phrases des partitions dev et devtest du jeu de données FLORES-200 à l'aide des tokeniseurs de chaque modèle puis nous les avons traduites automatiquement à l'aide des modèles associés pour les 72 directions de traductions, produisant au total 144 648 paires de phrase.

2. https://en.wikipedia.org/wiki/Cebuano_Wikipedia consulté en Juin 2024

4 Résultats

4.1 Liens entre les mesures d'évaluation de la traduction

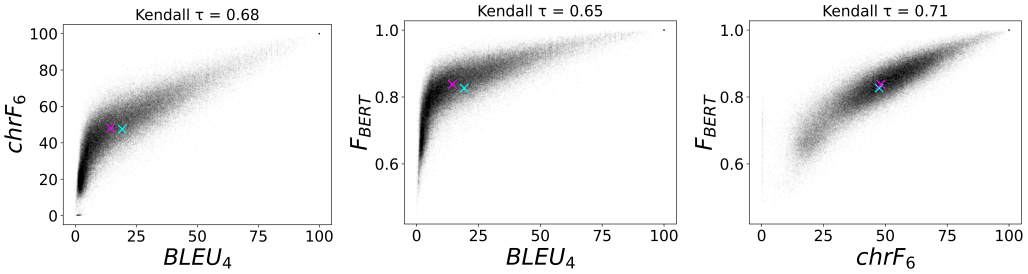
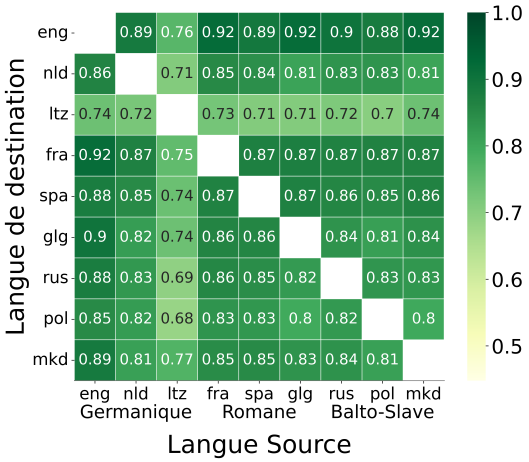


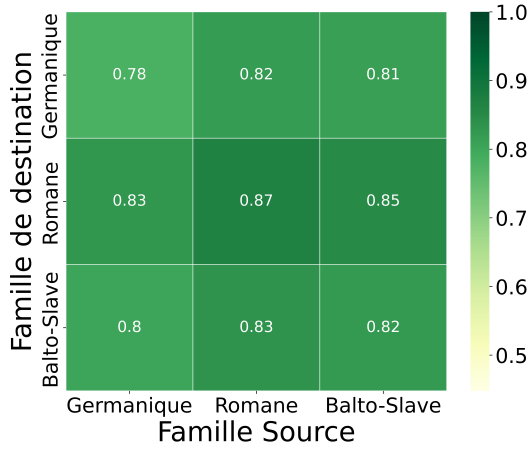
FIGURE 1 – Lien entre les paires de mesures d'évaluation de traduction et leur corrélation de Kendall ; les valeurs moyennes (resp. médianes) sont en cyan (resp. magenta)

La figure 1 présente la corrélation de Kendall entre les paires de mesures de performance de la traduction automatique en considérant l'ensemble des paires de phrases. La valeur de corrélation de Kendall varie entre 0.65 et 0.71 en fonction des paires de mesures considérées. Les mesures sont fortement corrélées (bien que non linéairement). L'observation de la tendance du score $BLEU_4$ semble indiquer un biais pour les valeurs faibles tandis que les scores $chrf_6$ et F_{BERT} semblent plus équilibrés avec des valeurs moyennes (cyan) et médianes (magenta) plus centrées sur les intervalles de définition des mesures. Dans la suite des analyses, nous avons choisi de garder la mesure F_{BERT}.

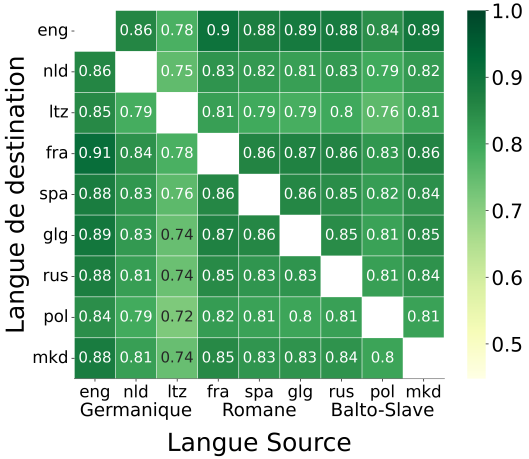
4.2 Performances de traduction



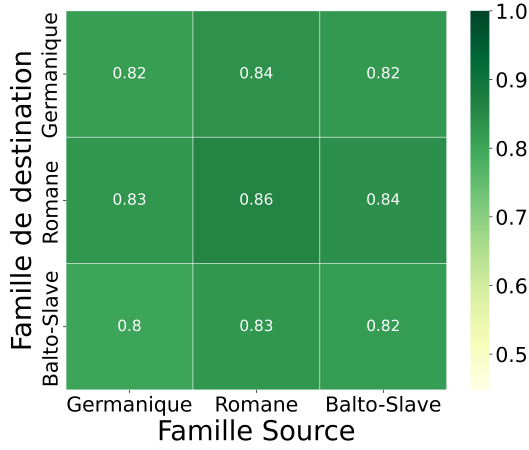
a) agrégé par langue pour MADLAD-400



b) agrégé par famille de langues pour MADLAD-400



c) agrégé par langue pour NLLB-200



d) agrégé par famille de langues pour NLLB-200

FIGURE 2 – Performance de la traduction avec MADLAD-400 (a) entre les paires de langues sélectionnées, (b) par famille de langues ; avec NLLB-200 (c) entre les paires de langues (d) par famille de langues, mesurées avec la mesure F_{BERT}. Pour (a) et (c), les données sont ordonnées par famille, puis par nombre de contributeurs sur wikipédia.

La figure 2 montre les performances obtenues pour la mesure F_{BERT} pour chaque modèle en fonction de la langue source et de destination ainsi que lorsque les langues sont regroupées par famille. Il n'y a pas de schéma qui se dégage clairement de la figure 2. On peut cependant constater que les deux modèles obtiennent des performances similaires pour chacune des directions.

Le luxembourgeois (ltz) obtient les valeurs les plus faibles pour les deux modèles et cela que ce soit en langue source ou destination, et quelque soit l'autre langue de la paire (couleurs plus claires

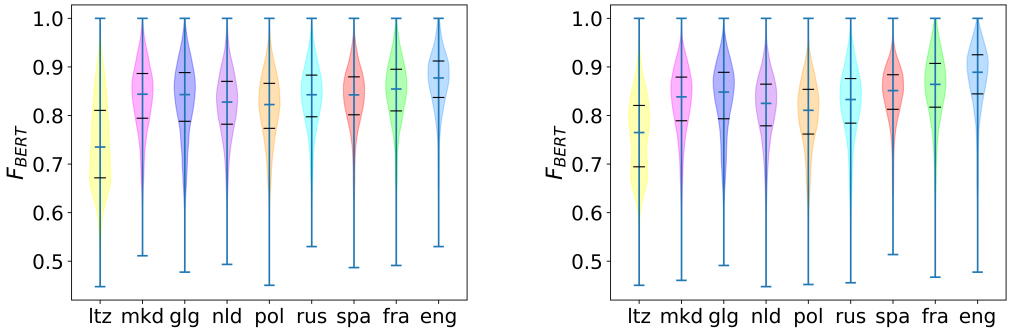
sur la 3ième ligne et 3ième colonne). Il s'agit de la langue la moins dotée en termes de personnes ayant contribué à Wikipedia. Il s'agit de la langue pour laquelle les deux modèles ont le plus de différences lorsqu'elle est langue cible ; NLLB-200 est légèrement plus performant. Les différences relatives varient de 0.06 à 0.11. Si l'on exclu le luxembourgeois, les différences relatives sont de 0.05 maximum.

Nous n'observons pas de performance plus forte au sein d'une même famille. Nous observons des performances légèrement plus fortes sur les langues ayant le plus de contributeurs (voir table 3), donc a priori les plus dotées également dans l'étape d'apprentissage des tokeniseurs (1ière, 4ième et 7ième lignes et colonnes).

Le script utilisé (latin ou cyrillique) ne semble pas avoir d'influence notable.

Les meilleures performances en moyenne sont obtenues pour l'anglais (1ière ligne / colonne, la plus fortement colorée).

En ce qui concerne les familles de langues (sous-figure (b)), nous pouvons observer une performance légèrement meilleure sur les paires qui font intervenir les langues romanes. Le croisement de script de la langue ne semble pas dégrader les résultats.



a) performances par langue source

b) performances par langue de destination

FIGURE 3 – Distributions des performances obtenues pour chacune des langues en tant que (a) source de traduction et (b) destination de traduction, mesurées avec F_{BERT} et ordonnées de gauche à droite en fonction du nombre de contributeur·ice·s Wikipedia

La figure 3 montre la distribution des performances F_{BERT} sur les 997 phrases de la partition dev du jeu de données, en considérant ensemble les deux modèles. Les langues sont ici ordonnées par ordre croissant du nombre de personnes contribuant à Wikipédia, cela pour chacun des modèles. Il n'y a pas de différence notable entre les deux configurations, sauf pour le luxembourgeois. Il n'y a pas non plus de tendance forte qui se dégage en fonction des langues. Pour deux familles de langues, le faible nombre de contributeurs est lié à des performances légèrement plus basses (ltz et spa).

4.3 Lien entre les mesures de caractérisation des tokeniseurs et les performances de traductions

Pour vérifier l'hypothèse selon laquelle des caractéristiques calculées à partir de la tokenisation pourraient prédire l'efficacité de la traduction par un PLM, nous avons calculé la corrélation de Kendall entre chaque mesure de caractéristique et la mesure F_{BERT}. Cette analyse a été effectuée pour chaque modèle et pour différents types de regroupements de directions de traduction.

Les types de regroupement effectué sont les suivants :

1. Pas de regroupement, n'importe quelle langue en destination et source
2. Regroupement par langue de destination, toutes langues en source, figure 5 (a)
3. Regroupement par langue source, toutes langues en destination, figure 5 (b)
4. Regroupement par famille de langue de destination, toutes langues en source
5. Regroupement par famille de langue source, toutes langues en destination
6. Regroupement par paire famille de langue source - destination, figure 4

Les corrélations ont été calculées entre les caractéristiques et les performances pour chaque combinaison au sein de chaque type de regroupement. Les résultats ont été triés selon la valeur de corrélation afin d'identifier la caractéristique la plus significative et la combinaison maximisant cette corrélation.

La corrélation de Kendall maximale entre une caractéristique et la mesure F_{BERT} est de -0.56. Les langues sources sont de famille romane et les langues de destination sont de famille germanique. La distribution des données pour cette combinaison est présentée dans la figure 4. Nous présentons également les resultats pour les groupements par langue de destination en figure 5 (a) et groupement par langue source en figure 5 (b).

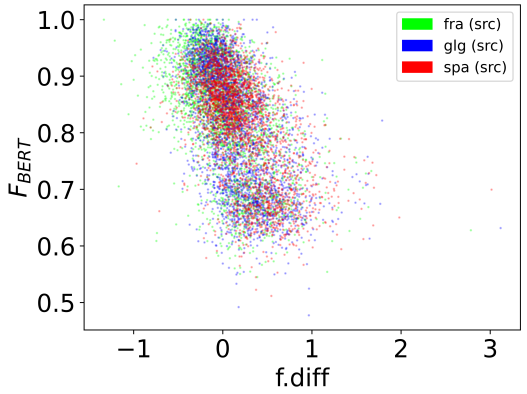
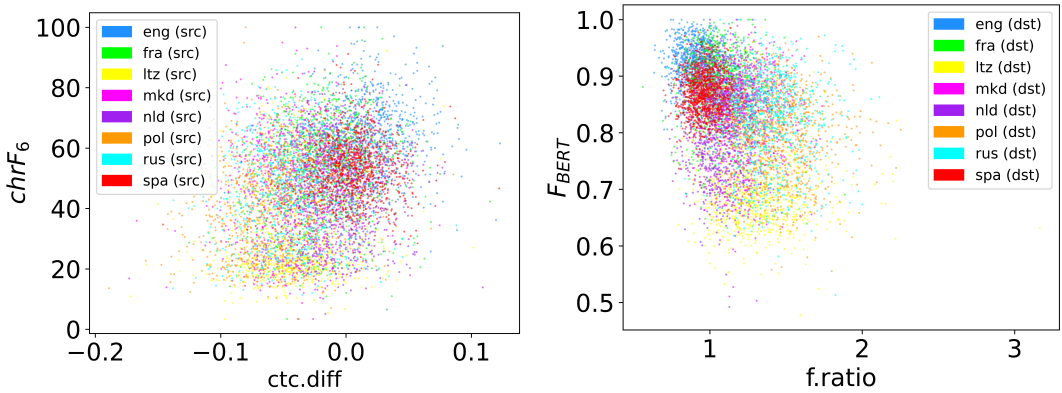


FIGURE 4 – Performances F_{BERT} (axe des Y) en fonction de la différence de la valeur Fertility ($f.diff$) pour la traduction des phrases en langues de familles romanes vers des langues de familles germaniques par le modèle MADLAD-400 (Kendall = -0.40)

La figure 4 présente la distribution des performances F_{BERT} en fonction de la valeur ($f.ratio$) calculées sur des traductions par le modèle MADLAD-400 avec les langues romanes en langue source et les langues germaniques en langue de destination. Cette combinaison étant celle maximisant la corrélation de Kendall, elle représente donc le meilleur cas où une caractéristique seule peut fournir

des indications sur l'efficacité de la traduction par un des deux PLM considérés. Cette corrélation (-0.56) n'étant pas significative, il semblerait donc que l'hypothèse structurant cette étude est invalidée.



(a) Galicien comme langue de destination et NLLB-200 comme modèle (Kendall = 0.20)

(b) Macédonien comme langue source et MADLAD-400 comme modèle (Kendall = -0.26)

FIGURE 5 – Performances F_{BERT} (axe des Y) en fonction :
(a) de la valeur de différence de *Closeness to the Character Level* ($ctc.diff$).
(b) de la valeur de ratio de *Fertility* ($f.ratio$).
Chaque paire de phrase est représentée par un point auquel est associée une couleur en fonction de la langue source (src) et destination (dst) pour la figure (a) et (b) (respectivement).

La figure 5 (a) présente la distribution des performances F_{BERT} en fonction de la différence de valeur *Closeness to the Character Level* ($ctc.diff$) calculées sur des traductions par le modèle NLLB-200 avec le Galicien comme langue de destination. On peut noter une distribution différente en fonction de la langue source avec les plus faibles performances pour le luxembourgeois (ltz). Par ailleurs, on note une tendance selon laquelle les langues bien dotées telles que l'anglais (eng) et l'espagnol (spa) ont une performance F_{BERT} plus élevée que pour le galicien, associée à une valeur $ctc.diff$ positive, correspondant à une production de tokens par caractère plus faible. Au contraire, pour les langues peu dotées telles que le luxembourgeois (ltz) la performance F_{BERT} est associée à des valeurs de $ctc.diff$ faible.

Lorsque la langue source est fixée, le maximum de corrélation est obtenu avec le macédonien, pour la mesure de performance F_{BERT} et pour la mesure de caractérisation calculant le ratio de la *Fertility* ($f.ratio$) pour des traductions faites avec le modèle MADLAD-400 (Voir figure 5 (b)). Le même type de commentaire s'applique ici où l'espace est occupé distinctement en fonction des langues de destination. Les valeurs faibles de $f.ratio$, indiquant une production de tokens par mots plus élevée pour la langue de destination, sont associées avec des performances élevées pour les langues bien dotées comme l'anglais (eng) et inversement ($f.ratio > 1$ et F_{BERT} faible) pour les langues moins dotées telles que le luxembourgeois (ltz) ou le polonais (pol).

5 Conclusion

Cette étude préliminaire nous a permis d'explorer la relation entre les performances de traduction de grands modèles de langue pré-entrainés et des caractéristiques sur le resultat de la tokenisation.

Malgré l'absence de corrélation notable entre les mesures et les performances de traduction, cette étude préliminaire nous a permis de comparer des mesures présentées séparément dans des études précédentes et dont le lien avec les performances de traduction multilingue par grands modèles de langues n'avait pas été exploré.

Pour poursuivre ces travaux, nous aimerions intégrer des mesures de comparaison des résultats des tokeniseurs plus qualitatives, afin de vérifier l'impact de la représentation des langues par les tokens. Nous pourrions utiliser des mesures telles que la perplexité (Jelinek *et al.*, 1977), la cross-entropie (Cover, 1999) ou une adaptation de la mesure de Fréchet qui mesure la similarité entre deux probabilités de distribution (Arabzadeh & Clarke, 2024). Nous pourrions également étudier les fréquences des tokens dans le vocabulaire de chaque langue et l'impact sur la qualité de traduction. L'étude de la similarité entre les tokens utilisés dans la phrase source et dans la phrase candidate nous parait également être une piste intéressante.

Remerciements

Ce travail a bénéficié d'une aide de l'État français gérée par l'Agence Nationale de la Recherche (project GUIDANCE, ANR-23-IAS1-0003 et projet D4R, ANR-21-CE38-0011).

Références

ALYAFEAI Z., AL-SHAIBANI M. S., GHALEB M. & AHMAD I. (2023). Evaluating various tokenizers for arabic text classification. *Neural Processing Letters*, **55**(3), 2911–2933.

ARABZADEH N. & CLARKE C. L. A. (2024). Fréchet distance for offline evaluation of information retrieval systems with sparse labels.

CHEN M., TWOREK J., JUN H., YUAN Q., DE OLIVEIRA PINTO H. P., KAPLAN J., EDWARDS H., BURDA Y., JOSEPH N., BROCKMAN G., RAY A., PURI R., KRUEGER G., PETROV M., KHLAAF H., SASTRY G., MISHKIN P., CHAN B., GRAY S., RYDER N., PAVLOV M., POWER A., KAISER L., BAVARIAN M., WINTER C., TILLET P., SUCH F. P., CUMMINGS D., PLAPPERT M., CHANTZIS F., BARNES E., HERBERT-VOSS A., GUSS W. H., NICHOL A., PAINO A., TEZAK N., TANG J., BABUSCHKIN I., BALAJI S., JAIN S., SAUNDERS W., HESSE C., CARR A. N., LEIKE J., ACHIAM J., MISRA V., MORIKAWA E., RADFORD A., KNIGHT M., BRUNDAGE M., MURATI M., MAYER K., WELINDER P., MCGREW B., AMODEI D., MCCANDLISH S., SUTSKEVER I. & ZAREMBA W. (2021). Evaluating large language models trained on code.

CHOO S. & KIM W. (2023). A study on the evaluation of tokenizer performance in natural language processing. *Applied Artificial Intelligence*, **37**(1), 2175112.

COSTA-JUSSÀ M. R., CROSS J., ÇELEBI O., ELBAYAD M., HEAFIELD K., HEFFERNAN K., KALBASSI E., LAM J., LICHT D., MAILLARD J. *et al.* (2022). No language left behind : Scaling human-centered machine translation. *arXiv preprint arXiv :2207.04672*.

COVER T. M. (1999). *Elements of information theory*. John Wiley & Sons.

DOMINGO M., GARCÍA-MARTÍNEZ M., HELLE A., CASACUBERTA F. & HERRANZ M. (2019). How much does tokenization affect neural machine translation ? In *International Conference on Computational Linguistics and Intelligent Text Processing*, p. 545–554 : Springer.

GAGE P. (1994). A new algorithm for data compression. *The C Users Journal archive*.

GOYAL N., GAO C., CHAUDHARY V., CHEN P.-J., WENZEK G., JU D., KRISHNAN S., RANZATO M., GUZMÁN F. & FAN A. (2022). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, **10**, 522–538.

GU T., CHEN K., OUYANG S. & LI L. (2023). Playground low resource machine translation system for the 2023 americasnlp shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, p. 173–176.

JELINEK F., MERCER R. L., BAHL L. R. & BAKER J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, **62**(S1), S63–S63.

KARGARAN A. H., IMANI A., YVON F. & SCHÜTZE H. (2023). Glotlid : Language identification for low-resource languages. *arXiv preprint arXiv :2310.16248*.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

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. DOI : 10.48550/ARXIV.2309.04662.

LIMISIEWICZ T., BALHAR J. & MAREČEK D. (2023). : [object Object]. DOI : 10.48550/ARXIV.2305.17179.

MAILLARD J., GAO C., KALBASSI E., SADAGOPAN K. R., GOSWAMI V., KOEHN P., FAN A. & GUZMÁN F. (2023). Small data, big impact : Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2740–2756.

MOGHE N., FAZLA A., AMRHEIN C., KOCMI T., STEEDMAN M., BIRCH A., SENNRICH R. & GUILLOU L. (2024). Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets. DOI : 10.48550/ARXIV.2401.16313.

NLLB TEAM, 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. (2022). No Language Left Behind : Scaling Human-Centered Machine Translation. DOI : 10.48550/arXiv.2207.04672.

PAL S., PAKRAY P., LASKAR S. R., LAITONJAM L., KHENGLAWT V., WARJRI S., DADURE P. K. & DASH S. K. (2023). Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, p. 682–694.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : A Method for Automatic Evaluation of Machine Translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of*

the 40th Annual Meeting of the Association for Computational Linguistics, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PETROV A., MALFA E. L., TORR P. H. S. & BIBI A. (2023). Language model tokenizers introduce unfairness between languages.

POPOVIĆ M. (2015). chrF : Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392–395, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/W15-3049.

RUST P., PFEIFFER J., VULIĆ I., RUDER S. & GUREVYCH I. (2021). How Good is Your Tokenizer ? On the Monolingual Performance of Multilingual Language Models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 3118–3135. DOI : 10.18653/v1/2021.acl-long.243.

SENNRICH R., HADDOW B. & BIRCH A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv :1508.07909*.

SIMONS G. F., THOMAS A. L. L. & WHITE C. K. K. (2022). Assessing digital language support on a global scale. In N. CALZOLARI, C.-R. HUANG, H. KIM, J. PUSTEJOVSKY, L. WANNER, K.-S. CHOI, P.-M. RYU, H.-H. CHEN, L. DONATELLI, H. JI, S. KUROHASHI, P. PAGGIO, N. XUE, S. KIM, Y. HAHM, Z. HE, T. K. LEE, E. SANTUS, F. BOND & S.-H. NA, Éds., *Proceedings of the 29th International Conference on Computational Linguistics*, p. 4299–4305, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

TIANYI Z., VARSHA K., FELIX W., KILIAN Q. W. & YOAV A. (2020). BERTScore : Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

ZHANG S., CHAUDHARY V., GOYAL N., CROSS J., WENZEK G., BANSAL M. & GUZMAN F. (2022). How Robust is Neural Machine Translation to Language Imbalance in Multilingual Tokenizer Training ? DOI : 10.48550/ARXIV.2204.14268.