Université de Strasbourg

Master Technologies des Langues
Option Traitement Automatique des Langues

Traduction automatique des dialectes alsaciens vers l'allemand standard

Franka Emilie WURPS

Sous la direction de
Delphine BERNHARD
Responsable pédagogique du Master Technologies des Langues

Table de matières

Table d'équitations

Table de figures

Table de tableaux

I. Liste des abréviations

(par ordre alphabétique)
Byte Pair Encoding (BPE)
Langue cible (LC)
Langue source (LS)
Large language model (LLM)
Groupe infinitif (GINF)
Groupe verbal (GV)
Traduction automatique (TA)
Traduction automatique neuronale (TAN)
Traduction automatique statistique (TAS)
Traitement automatique des langues / du langage (TAL)
Translation Edition Rate (TER)

II. Remerciements

Je tiens à exprimer ma gratitude à ma directrice de mémoire, Madame Delphine BERNHARD. Je la remercie de m'avoir encadrée, orientée, aidée et conseillée.

Je remercie également toute l'équipe pédagogique de l'université de Strasbourg ainsi que les professionnels et professionnelles en charge de ma formation, pour avoir assuré le suivi et le transfert de connaissances qui ont toujours été sur une base bienveillante.
Je remercie également mes camarades de classe pour les deux années de master passées ensemble. Pour les nouvelles connaissances, pour le soutien mutuel et l'écoute.

Je souhaite aussi dire merci à mes enseignants et enseignantes de ma licence, sans qui je n'aurais pas suivi cette voie des études germaniques, de la linguistique et finalement du traitement automatique des langues.

Enfin, je remercie toutes les autres personnes qui m'ont accompagné, écouté et soutenue dans la rédaction de ce mémoire.

1. Introduction

La traduction automatique est beaucoup utilisée et répandue de nos jours. Sur les réseaux sociaux, un simple clic de souris ou sur l'écran suffit et des paragraphes entiers sont traduits automatiquement en quelques secondes. Cette technique marche surtout bien pour des langues très répandues comme l'anglais, l'espagnol, l'allemand, le français ou autres. Mais la traduction automatique génère des phrases erronées quand il s'agit d'une langue moins connue. Ainsi, les dialectes et les langues régionales ne sont pas (entièrement) pris en compte non plus. Pourtant, les recherches dans ce domaine linguistique sont très actives. Il existe même déjà des systèmes de traduction automatique pour certaines langues régionales.

Par exemple, l'occitan est une langue régionale en France et *Revirada* est le premier traducteur automatique (pour l'occitan) développé par le Congrès[1]. Il est construit sur la base du moteur *Open source Apertium*. 8 dictionnaires français-occitan et occitan-français, la totalité des conjugaisons ainsi que les bases toponymiques et terminologiques font les ressources du traducteur en ligne. Il est possible de l'intégrer comme Plug-in pour traduire des articles ou pages Wordpress, le télécharger comme application ou faire traduire des sites web. L'occitan est avec un million de formes fléchies et plusieurs variétés de l'occitan un dialecte riche.

Dans ce travail, nous nous intéresserons plus particulièrement aux dialectes alsaciens et à leur traduction depuis et vers l'allemand standard. L'objectif sera donc de développer un premier outil de traduction automatique, selon le modèle de *Revirada*.

La traduction automatique (TA) est pour quelques langues déjà assez bien développée et recherchée, mais elle n'est pas évidente pour les langues peu dotées par exemple. La faible quantité de données disponibles, comme les corpus parallèles, rend la TA plus difficile (Haddow et al., 2022). Un autre facteur de complication majeur est le manque de normalisation à l'écrit de ces langues. Pour ce travail qui se base sur la TA des langues germaniques, les travaux récents sur les dialectes germaniques et alémaniques peuvent servir.

L'Alsace et l'Allemagne partagent une histoire intensive de guerres, d'annexions et de changements politiques. Les dialectes alsaciens et l'allemand standard sont également proches l'un de l'autre. Bien que les deux langues soient proches, les parlers dialectaux ont toujours subi un statut mineur par rapport à l'allemand. Travailler sur un système de traduction, focalisé sur l'alsacien, rend cette langue régionale plus visible, lui attribue une grande valeur et enfin rend l'accessibilité possible pour les germanophones ne comprenant pas ce dialecte. Et l'inverse, il rend également l'allemand accessible pour les dialectophones.

[1] https://revirada.eu/

2. Situation en Alsace

Cette première partie se concentre sur l'Alsace et son dialecte. Il ne s'agit pas de donner un historique complet de la région. Ces travaux ne visent pas non plus à analyser en profondeur la culture, la tradition, la société ou encore l'histoire de l'Alsace. Le but principal est la mise en contexte des changements linguistiques et les conséquences liées à cela. Les bases de ce que c'est le dialecte alsacien et dans quelle mesure il diffère de l'allemand sont également présentées.

2.1 Situation sociolinguistique et sociohistorique

Pour mieux comprendre les changements linguistiques et l'histoire langagière de l'Alsace, nous allons examiner la situation sociolinguistique et sociohistorique.
À partir de la création du Saint-Empire romain germanique en 962, l'Alsace en fait partie. Déjà avant, durant les royaumes mérovingien et carolingien, la population alsacienne a été germanophone (Huck, 2022).
C'est en 1648, après la guerre de Trente Ans, que l'Alsace est intégrée dans l'espace géopolitique de la France. Cette intégration conduit à une présence de la langue française plus grande qu'auparavant. C'est la langue du pouvoir politique et des classes dominantes. L'autre partie de la population continue cependant à parler les dialectes allemands (Huck, 2022). Bien que l'idéologie de la Révolution française soit que le français prend un rôle politique et idéologique central, en Alsace, l'allemand reste la langue majoritairement utilisé à l'écrit et à l'oral de la population durant la première moitié du 19ème siècle (Huck, 2022). Entre 1855 et 1870, les classes sociales dominantes et les intellectuels utilisent la langue française, même dans la sphère privée. En même temps une politique linguistique scolaire amène une connaissance passive du français aux jeunes enfants, en âge d'être scolarisés (Huck, 2022).
Avec la victoire de la Prusse en 1870/1871 après les guerres franco-allemandes, l'Alsace fait à nouveau partie d'un espace géopolitique allemand. Elle intègre avec une partie de la Lorraine l'Empire allemand, totalement sans autonomie politique et seulement comme *Reichsland* (Terre d'Empire) (Huck, 2022). À la suite de l'absence d'autonomie politique, une partie des intellectuels et des cercles culturels cherchent une mise à distance avec l'Empire en investissant dans les parlers dialectaux afin de marquer une identité alsacienne (Huck, 2022). Pendant ces années d'appartenance à l'Empire, l'allemand a été la seule langue officielle an Alsace (Huck, 2022).
En 1918, après la défaite de l'Allemagne et à la suite du traité de Versailles, l'Alsace réintègre la France. Une politique intense de diffusion du français est menée par l'état après avoir constaté que seulement 2% de la population utilisent le français de façon active et 8% possèdent des connaissances relatives (Huck, 2022).
Cette politique est surtout remarquable dans l'enseignement à l'école primaire. Il est intéressant d'observer les instructions pour l'enseignement donné aux jeunes enfants car c'est le moyen de déduire comment l'état français a voulu forger les nouvelles générations. Dans la circulaire du Recteur Charléty du 27 septembre 1919 et dans ses instructions du 15 janvier 1920, il est écrit clairement que l'allemand est vu « *comme instrument d'enseignement tant que le français sera*

insuffisamment connu des élèves »[2] et plus loin que « *le français doit être la langue essentielle […] l'enseignement de l'allemand doit être donné […] à la condition essentielle de ne pas porter préjudice à la diffusion de la langue française* »[3]. Dans l'ensemble des instructions du Recteur Charléty, le français est privilégié à l'allemand. La seule place pour l'allemand est l'enseignement religieux.

Malgré des conflits violents entre la majeure partie des élus alsaciens et les gouvernements français causés par des questions religieuses et sociales (Huck, 2022), la politique du français à la première place aboutit. Selon Huck (2022), 55,6% de la population déclare savoir parler le français en 1936. Environ 87% déclarent savoir parler le dialecte et environ 80% l'allemand (Huck, 2022).

Seulement peu d'années plus tard, la situation linguistique est à nouveau renversée. L'Alsace est annexée par le Troisième Reich en 1940. La législation ainsi que des structures nationales-socialistes sont introduites et l'allemand redevient la langue officielle (Huck, 2022). Dans l'idéologie hitlérienne, la présence et l'usage du français sont interdits sous peine de sanctions, l'usage du dialecte reste toléré à titre privé (Huck, 2022 et Denis, 2003).

Après la capitulation de l'Allemagne en mai 1945, l'Alsace redevient française. La politique française de 1918 est reprise mais encore plus renforcée. L'allemand, la langue de l'ennemi, est banni de la vie publique, le français est la seule langue présente. Des élus appartenant à la génération d'avant 1940, bataillent pour que l'allemand soit enseigné à l'école. C'est seulement en 1982, que l'allemand commence à être enseigné à nouveau à l'école (Huck, 2022). Cette politique en faveur absolue du français ainsi que la mise en relation entre la langue et la modernité et notamment les discours tenus sur les dialectes construisent des représentations stigmatisantes des dialectes et leurs locuteurs. Les parlers dialectaux ne sont pas vus comme des « vraies » langues mais comme handicap social, mental et cognitif pour l'acquisition du français (Huck, 2022).

Cette stricte politique aboutit, comme en 1936, à nouveau. Alors qu'en 1962, 80,7% de la population déclarent savoir parler le français, 84,7% le dialecte et 80,29% l'allemand, en 1999/2002 seulement 39% déclarent savoir parler le dialecte et 16,2% l'allemand (Huck, 2022). Ces données montrent une baisse des points de pourcentage en termes de la connaissance de 45,7 pour le dialecte et 64,09 pour l'allemand en presque 40 ans. Ce sont les effets de la politique linguistique qui mène à une accélération de la diffusion du français et un recul de l'usage des parlers dialectaux. Cela conclut finalement que pour la première fois depuis le Haut-Moyen Âge, les parlers dialectaux ne sont plus connus de la majorité de la population (Huck, 2022).

2.2 Les dialectes alémaniques

Les dialectes alémaniques sont un groupe de dialectes de langue germanique. Ils sont parlés par environ dix millions de personnes (Lambrecht et al., 2022). Le groupe dialectal se trouve notamment en Europe centrale, c'est-à-dire dans le sud-ouest de l'Allemagne, la Suisse germanophone, l'Autriche, le Lichtenstein et en Alsace. On peut répartir les dialectes en différentes

2 (Bulletin de l'enseignement, 1920, p. 1)
3 (Bulletin de l'enseignement, 1920, p. 1)

régions. Il existe par exemple le souabe, l'allemand de Bâle ou encore l'alsacien, qui englobe le haut-alémanique ainsi que le bas-alémanique. Tous ces dialectes se différencient de l'allemand standard par l'orthographe, la prononciation, le lexique et la grammaire. Mais même les dialectes se différencient entre eux (Lambrecht et al., 2022).

2.3 Est-ce que l'alsacien est de l'allemand ?

L'allemand est une langue ayant beaucoup des dialectes. Cela est en raison du fait que « l'Allemagne » n'était pas un pays uni dans son histoire. Le Saint-Empire romain germanique, qui existait jusqu'en 1789, était composé de nombreux royaumes individuels, de principautés (électorales), d'évêchés et de duchés. Par conséquent, il n'y avait pas de langue allemande standard et la variation diatopique était très marquée. Lors de la création de l'Empire allemand en 1871 dont l'Alsace en faisait partie, pour qu'une variante écrite de l'allemand soit introduite dans les institutions administratives et éducatives (Morin, 2020).
Il est assez souvent reproché aux parlers alsaciens que ces dialectes germaniques sont très proches de l'allemand standard ou encore la même langue que l'allemand. Mais ce n'est pas le cas. Tout de même en 1982, c'est l'allemand qui est qualifié comme langue régionale de l'Alsace et non l'alsacien. Aujourd'hui, c'est officiellement l'alsacien, classé sous langue germanique[4].
En 1982, lors de l'élection de François Mitterrand, les autorités académiques veulent installer l'enseignement des langues régionales en France. En raison du fait qu'en Alsace, l'enseignement de l'allemand est déjà proposé, le recteur Deyon décide de ne pas inclure les dialectes alsaciens mais l'allemand standard, car « *l'alsacien […] a pour expression écrite […] l'allemand* »[5](Denis, 2003).
On peut alors se questionner sur la distinction entre « langue » et « dialecte ». Est-ce qu'il s'agit d'une distinction linguistique, sociale ou encore autre ? Si on plonge dans l'analyse linguistique de l'alsacien, comme la grammaire, on constate vite qu'il ne s'agit pas de l'allemand standard. Dans cette partie, nous allons examiner la grammaire et syntaxe alsacienne pour montrer les différences avec l'allemand. Pour faire cela, nous allons traiter des différentes parties.

PHONETIQUE ET PHONOLOGIE

En ce qui concerne la phonétique et phonologie, les parlers dialectaux diffèrent beaucoup. Le bas alémanique est la variété majoritaire en Alsace. On trouve des inventaires phonémiques très vastes et différents d'une commune à l'autre (Huck , 2022). Nous n'allons pas approfondir cette thématique puisque la traduction automatique basée sur l'écrit et non pas sur l'oral est traitée dans ce travail.

[4] Source : https://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/Agir-pour-les-langues/Promouvoir-les-langues-de-France/Langues-regionales (consulté le 15/03/2023)
[5] Citation directe de Le dialecte alsacien : état des lieux par M. Denis, 2003

MORPHOSYNTAXE[6]

SYNTAXE POSITIONNELLE ET PROPOSITIONS SUBORDONNEES

En règle générale, la syntaxe positionnelle est assez similaire à celle de l'allemand.

La différence entre les deux syntaxes est la place des participes II des verbes de modalité et des verbes de perception ainsi que les infinitifs (Huck, 2022). On peut observer une disjonction du participe II du groupe infinitif (GINF) objet lorsque l'infinitif est postposé au participe II (1) (Huck, 2022).

(1) S Kind het zue sinere Màmme welle renne[7]

Pour les propositions subordonnées, le GINF objet est solidaire et précède le verbe de modalité au parfait (2). Le verbe de modalité est en forme conjuguée, appelé la forme personnelle à partir d'ici dans ce document, et précède la forme non personnelle, donc l'infinitif *(welle)*. La forme personnelle du verbe de modalité peut également précéder la forme non personnelle. Cependant, l'infinitif du GINF *(bsueche)* est disjoint du membre du GINF *(uns)* et postposé au verbe de modalité (3) (Huck, 2022).

(2) Ich weiss, àss r uns bsueche het welle[8]

(3) Ich weiss, àss r uns het welle bsueche[9]

(4) Ich weiss, àss r uns het bsueche welle[10]

Selon Huck, les formes (2) et (3) sont assez fréquentes. L'autre forme (4) est moins fréquente, elle pourrait avoir été influencée par l'allemand standard. Par ailleurs, c'est cette forme qui se rapproche le plus à la formulation en allemand standard.

Pour les exemples (1) à (3), l'allemand standard diffère un peu. Dans les structures à l'infinitif, c'est le verbe de modalité à l'infinitif qui se trouve en dernière position. S'il y a une forme personnelle du verbe de modalité comme en (2) et (3), il est postposé au GINF (*'Das Kind hat zu seiner Mutter rennen wollen'* (1) et *'Ich weiß, dass er uns hat besuchen wollen'* (2,3,4)). Il est à ajouter que les verbes à préfixes séparables peuvent être disjoints aussi. L'allemand standard ne fait pas de disjonction. Le verbe de modalité à l'infinitif reste comme aux autres exemples en dernière position (*'Er hat fortgehen wollen'*)

(5) Er het furtgehn welle / Er het furt welle gehn[11]

Quant au plus-que-parfait, l'ordre des mots est différent entre le nord et le sud de l'espace alémanique alsacien (Huck, 2022). C'est ainsi que l'ordre participe II du verbe de base *(furtgànge)*, forme personnelle de l'auxiliaire, participe II de l'auxiliaire (6) est employée majoritairement au nord. Au sud c'est l'ordre participe II du verbe, participe II de l'auxiliaire, forme personnelle de l'auxiliaire qui est employée (7). On peut également trouver l'ordre suivant (8) : forme personnelle de l'auxiliaire, participe II du verbe, participe II de l'auxiliaire (Huck, 2022).

---

[6] Tous les exemples utilisés sont tirés de *Les parlers dialectaux en Alsace* par D. Huck (2022)

[7] L'enfant a voulu courir chez sa mère

[8] Je sais qu'il a voulu nous rendre visite

[9] Voir 7

[10] Voir 7

[11] Il a voulu partir

(6) Wie er schun e Wil furtgànge isch gsin, isch m ebs inkomme[12]
(7) Wie er schun e Wil fortgànge gsi isch, isch m ebs inkomme[13]
(8) Wie er schun e Wil isch furtgànge gsin, isch m ebs inkomme[14]

Il semble que tous ces ordres de phrase sont acceptés. Alors qu'en allemand, ce n'est pas le cas. Contrairement à l'alsacien, il existe le prétérit en allemand standard. C'est le temps simple de la narration au passé. Dans notre cas des exemples (6) à (8), celui-ci serait employé. Si nous restons sur le temps du plus-que-parfait, il y a tout de même un ordre de mot à respecter : participe II du verbe, participe II de l'auxiliaire et finalement la forme personnelle de l'auxiliaire (*'[Als] er schon eine Weile fortgegangen gewesen war, ist [es] ihm [eingefallen]'*).
Bien évidemment, c'est l'ordre naturel dont on parle. Il est tout à fait possible de modifier l'ordre des phrases suites aux stratégies communicatives, stylistiques ou informatives. Cela s'applique pour l'allemand standard ainsi que pour les dialectes alsaciens (Huck, 2022).
Cependant, le français ne semble pas d'influencer la syntaxe positionnelle (Huck, 2022).

Morphologie nominale, Determinant & Pronoms

Comme l'allemand standard, l'alsacien est une langue ayant un système casuel. En allemand standard, il existe quatre cas : nominatif, accusatif, datif et génitif. Les dialectes alsaciens en possèdent trois, il n'existe pas de génitif. En règle générale, les marques de cas servent à affecter un genre (masculin, féminin ou neutre) et le nombre. Or, le système casuel en alsacien ne se réalise qu'au singulier (Huck, 2022).
Le manque du génitif est compensé de plusieurs manières, elles dépendent du type de génitif. En allemand standard, il y a un génitif qui exprime la possession, aussi appelé « génitif de possession » et le génitif résultant d'une rection prépositionnelle.
Le génitif de possession peut être remplacé par l'utilisation du datif avec la préposition *« von »* (de) (10) ou alors en passant par le pronom personnel (9) :

(9) (In) mim Sohn sin Auto isch knàllrot[15]
(10) S Auto von (in) mim Sohn isch knàllrot[16]

En allemand standard, l'exemple (9) n'est pas possible. On peut trouver cette version à l'oral, mais il ne s'agit pas de la règle grammaticale. L'exemple (10) avec la préposition évoquant le datif est également possible en allemand standard (*'Das Auto meines Sohnes ist knallrot'* au génitif classique et *'Das Auto von meinem Sohn ist knallrot'*).

Quant aux articles définis, le système casuel existe mais de manière plus réduite. Encore une fois, une distinction entre le nord et le sud de l'espace dialectal alsacien est faite (Huck, 2022). Le nord de cet espace retient la forme de l'accusatif, le sud celle du nominatif. Mais le nord ne distingue

[12] Après qu'il était déjà parti un certain temps, quelque chose lui est revenu
[13] Voir 10
[14] Voir 10
[15] La voiture de mon fils est rouge
[16] Voir 13

pas entre nominatif et accusatif masculin. Le datif est majoritairement utilisé avec la préposition *« in »* (à/en/dans) (Huck, 2022).
Les articles indéfinis ne présentent pas de distinction entre les cas de l'accusatif et du nominatif. La forme -e ou ses allophones est employée (Huck, 2022).
L'allemand standard au contraire fait une grande différence avec l'alsacien. Tous les articles, qu'ils soient définis ou indéfinis, sont marqués par leur cas. Il est ainsi possible de reconnaître les cas entre autres en s'intéressant aux articles.

Bien évidemment, le système casuel est également maintenu dans les pronoms, qu'ils soient possessifs ou personnels (Huck, 2022).
À nouveau, il n'y a pas de distinction entre le nominatif et l'accusatif dans l'intégralité de l'espace dialectal :

(11) Mi(n), di(n), [masc.] si(n), [fém.] ihr, [neutre] si(n) ; unser, ejer, ihr[17]

L'allemand standard fonctionne de la même manière, (*'mein, dein', sein/ihr/sein, unser, euer, ihr'*) en intégrant les quatre cas.
En ce qui concerne les pronoms personnels et les substantifs pronominaux, il n'y a pas une différence hors l'orthographe entre l'alsacien et l'allemand standard. Deux séries parallèles sont présentes dans les pronoms personnels et les substantifs pronominaux en alsacien (Huck, 2022). Une série est atone et l'autre tonique avec une possibilité de porter un accent contrastif (Huck, 2022).
En allemand standard, les pronoms démonstratifs sont toujours marqués par le système casuel. En alsacien, on constate encore une différence entre le nord et le sud de l'espace dialectal. Au nord, le système casuel est plus complet, c'est-à-dire qu'on distingue entre nominatif et accusatif (Huck, 2022).

[17] Les pronoms et déterminants possessifs

MORPHEMES DU PLURIEL

Dans les parlers dialectaux, il existe des morphèmes de pluriel. En allemand, il existe d'autres morphèmes du pluriel. Un tableau comparatif des morphèmes les plus courants suit :

Tableau 1 : Morphèmes courants des parlers dialectaux et d'allemand standard[18]

Parlers dialectaux	Allemand standard
- ø (morphème zéro)	- ø (morphème zéro)
- ¨ø : une palatalisation de la voyelle du radical du substantif, avec un morphème final zéro	- (¨)e : la palatalisation de la voyelle du radical du substantif n'est pas catégorique
- e	- en
- ¨er : une palatalisation de la voyelle du radical du substantif, avec le morphème final en -er	- n
	- (¨) er : la palatalisation de la voyelle du radical du substantif n'est pas catégorique
	- (¨) s : la palatalisation de la voyelle du radical du substantif n'est pas catégorique

Pour l'allemand standard, il est à noter que pour le morphème -e, un deuxième s peut être ajouté. Comme par exemple le substantif *der Bus (le bus)* devient *die Busse* en pluriel. Cette liste des morphèmes des pluriels n'inclut pas les nombreuses exceptions et irrégularités.
En comparant ces deux listes, on constate que les morphèmes ne sont pas tous pareil, mais ils possèdent souvent les mêmes caractéristiques, c'est-à-dire, la palatalisation de la voyelle.

MORPHOSYNTAXE NOMINALE

Pour introduire un relatif comme expansion à droite d'une base nominale, l'alsacien passe par le pivot *« wo »* (ou ses allophones *wie/wü*). Celui-ci est utilisé lorsque le relatif représente un sujet ou un objet à l'accusatif (Huck, 2022) :

(12) De Mànn, wo i getroffe hàb, kenn i guet[19]

Quand le relatif a une fonction dans le groupe verbal (GV), équivalant à un datif, une autre construction est appliquée (Huck, 2022) :

(13) S Kind, wo i im e Velo gschenkt hàb, het sich gfrait[20]

Ici, le relatif sert qu'à la mise en relation entre le GV et son antécédent. Le pronom personnel prend alors en charge la fonction au sein du GV (Huck, 2022).
Il existe deux possibilités pour les verbes du GV de la relative si elle présente une rection prépositionnelle (Huck, 2022). La première possibilité est analogue à l'exemple incluant le datif :

[18] Source : (Huck, 2022) et moi-même
[19] Je connais bien l'homme que j'ai rencontré
[20] L'enfant à qui j'ai offert un vélo s'en est réjoui

(14) S isch dr Frind, wo ich uf ne gwàrt hàb[21]

Ici, le rôle du « relatif » est rempli par le relatif invariable suivant l'antécédent. La préposition (uf) se combine cependant avec un pronom personnel anaphorisant l'antécédent (Huck, 2022). Dans la deuxième possibilité, la préposition se combine avec un démonstratif qui est repris par le pivot relatif. Cette possibilité fonctionne pour les inanimés ainsi que les animés :

(15) S isch dr Frind, uf denne, wo ich gewàrt hàb/hà[22]

(16) De Brief, wo i druf gewàrt hà, isch endli ànkomme[23]

En allemand standard, le relatif ne se fait pas par *'wo'*. Un pronom relatif (der, die ou das) est employé (*'Den Mann, den ich getroffen habe, kenne ich gut'* pour l'exemple (12) et *'Das Kind, dem ich ein Fahrrad geschenkt habe, hat sich gefreut'* pour l'exemple (13) avec le datif). Quand il s'agit d'une rection prépositionnelle, elle est combinée avec le pronom relatif (*'Das ist der Freund, auf den ich gewartet habe'* pour les exemples (14) et (15) et *'Der Brief, auf den ich gewartet habe, ist endlich angekommen'*). Les pronoms relatifs sont accordés au cas. À l'oral, on peut entendre des versions ayant *'wo'*, mais il ne s'agit pas de la règle grammaticalement correcte.

Morphologie verbale

Les parlers dialectaux possèdent deux formes non personnelles : l'infinitif et le participe II (Huck, 2022). En allemand standard, il s'ajoute à ces deux formes encore le participe I. Pour les formes personnelles, les parlers dialectaux sont plus restreints que l'allemand en ce qui concerne le système et le nombre de modes et de temps (Huck, 2022).

En allemand standard, il existe trois modes principaux : l'indicatif, le subjonctif et l'impératif. Le conditionnel s'ajoute comme mode supplémentaire. En alsacien cependant, seuls l'indicatif et le subjonctif sont présents (Huck, 2022).

Les temps grammaticaux sont le présent, le parfait et le plus-que-parfait. Le temps grammatical qui pourrait apparaître comme un futur I n'a pas une valeur temporelle mais modale (Huck, 2022).

Cependant, l'allemand standard possède également le présent, le parfait et le plus-que-parfait. Mais, comme mentionné auparavant, il existe également le prétérit. Le futur I, aussi appelé le futur simple et le futur II, l'équivalent allemand du futur antérieur, ont une valeur temporelle.

L'allemand standard et les parles dialectaux ne diffèrent pas en ce qui concerne les voix : il y a une voix active, une voix passive processuelle et une voix du passif-bilan (Huck, 2022).

Marques de personnes-nombre

Dans les parlers dialectaux, deux séries de marques de personne-nombre existent. La première série est employée pour les verbes faibles et forts au présent de l'indicatif. La deuxième série est employée pour tous les autres cas (Huck, 2022) :

[21] C'est l'ami que j'attendais
[22] Voir 14
[23] La lettre que j'attendais est enfin arrivée

Tableau 2 : Marques de personne-nombre dans les parlers dialectaux[24]

	1ère personne S	2ème personne S	3ème personne S	Pluriel
Série 1	- Ø	- sch	- t	- e/en
Série 2	- ø	- sch	- ø	- e

Des alternances vocaliques peuvent apparaître. Mais elles ne concernent que des verbes forts au singulier pour lesquels la voyelle du radical est un [ɛ] ou [a] (Huck, 2022). À l'exception des petites régions, l'ensemble de l'espace dialectal adopte la voyelle du radical du singulier au pluriel (Huck, 2022).

Voici une table pour illustrer les marques de personne-nombre en allemand standard à l'indicatif présent :

Tableau 3 : Marques de personne-nombre en allemand standard[25]

	1ère personne	2ème personne	3ème personne
Singulier	- e	- st	- t
Pluriel	- en	- t	- en

Cette table s'applique pour les verbes forts et les verbes faibles. Il est à ajouter que la voyelle du radical des verbes forts peut changer. Ce changement concerne notamment la deuxième et troisième personne du singulier.

LES TEMPS DE L'INDICATIF

Comme déjà évoqué auparavant, il existe plusieurs temps grammaticaux dans l'indicatif. Nous allons les examiner : le présent, le parfait, le plus-que-parfait, le subjonctif I et II ainsi que les formes non personnelles que sont l'infinitif et le participe II.

En ce qui concerne le présent, l'allemand standard et l'alsacien ne disposent pas de différences. Les verbes avoir et être sont des auxiliaires. En alsacien, ces auxiliaires présentent un paradigme irrégulier (Huck, 2022).

Le parfait joue non seulement le rôle du passé, mais également de temps du récit. Selon les contextes, il peut être temporel, aspectuel ou distanciel (Huck, 2022). Le parfait est formé par l'auxiliaire conjugué et le participe II du verbe (Huck, 2022). Cette même construction se retrouve en allemand standard.

Les parlers dialectaux et l'allemand se différencient dans la formation du plus-que-parfait. En alsacien, il se forme du parfait de l'auxiliaire et le participe II du verbe (Huck, 2022).

(17) D Kinder hàn gspielt ghet. / D Kinder han gspielt ghà[26]

[24] Source : (Huck, 2022) et moi-même
[25] Source : (Huck, 2022) et moi-même
[26] Les enfants avaient joué

En allemand standard, le plus-que-parfait est formé du prétérit de l'auxiliaire et le participe II du verbe. Cela donne pour l'exemple précédent *'Die Kinder hatten gespielt'*.
Selon Huck (2022), les locuteurs dialectaux tendent à utiliser le parfait à la place du plus-que-parfait. Cela fait que les fonctions du parfait sont encore plus élargies.
Dans la moitié de l'espace dialectal, le subjonctif I n'est plus utilisé. Il est autant plus utilisé dans la partie méridionale de l'espace. Mais cela uniquement pour les verbes *'si'* et *'hà'*, donc les auxiliaires, au présent et au parfait. Il est formé par l'auxiliaire au subjonctif I et le participe II (Huck, 2022). Ce temps sert essentiellement au discours rapporté (Huck, 2022).
Le subjonctif II connait un usage courant dans l'espace dialectal. Bien qu'il soit formé par les auxiliaires, le verbe si(n) prend la forme du *'war(t)/wär(t)'*. La construction du subjonctif II des verbes faibles est exclusivement périphrastique, c'est-à-dire *'tuen'* (faire) au subjonctif II et un verbe à l'infinitif (Huck, 2022). Pour les verbes forts, les formes synthétiques du subjonctif II tiré d'un verbe sont peu nombreuses (Huck, 2022). Les verbes de modalités, présentent une forme synthétique et non analytique comme les verbes faibles (Huck, 2022).
En allemand standard, le subjonctif I est formé en ajoutant un suffixe personnel à la base du verbe. Il est principalement employé dans les propositions subordonnées pour exprimer une action possible ou une situation hypothétique. Le subjonctif I peut également être employé pour exprimer, des souhaits ou des demandes polies. La construction du subjonctif II est différente. Elle consiste du prétérit de *'haben'*, *'sein'* ou *'werden'*, la particule conjuguée de *'würde'* et l'infinitif du verbe. Le prétérit prend assez souvent un changement de voyelle du radical. Ce subjonctif est employé pour exprimer des doutes, pour prendre de la distance du discours rapporté ou pour des actions peu probables ou hypothétiques.
L'alsacien possède l'infinitif I et l'infinitif II. Le premier est l'infinitif simple d'un verbe tandis que le deuxième est composé par l'auxiliaire à l'infinitif et participe II. Le morphème d'infinitif est majoritairement -e (ou un allophone) (Huck, 2022).
L'allemand standard ne présente pas un premier et deuxième infinitif mais un infinitif présent et un infinitif passé. Ces formes ressemblent à celles de l'alsacien. L'infinitif présent est formé par l'ajout du morphème -en à la racine du verbe. L'infinitif passé se construit par l'auxiliaire et le participe II. La préposition *'zu'* est souvent employée dans les structures d'infinitif.
En alsacien, il n'existe pas de participe I (Huck, 2022).
En allemand standard, le participe I est une forme verbale utilisée pour exprimer une action en cours de réalisation. Le suffixe *'-d'* est ajouté à la racine du verbe pour les verbes faibles et pour les verbes forts, le suffixe *'-end'* est ajouté. Il a le même fonctionnement que le gérondif en français. En conséquence de la fréquence de l'emploi du participe II, celui-ci joue un rôle important. Pour les verbes faibles et forts, le préfix *'g-/ge-'* est ajouté au radical. S'il s'agit d'un verbe faible, le suffixe *'-t'* est ajouté, dans le cas contraire, c'est le suffixe *'-e'* ou un allophone. Les verbes forts peuvent présenter des changements vocaliques ou même consonantiques (Huck, 2022). Cependant, les verbes présentant déjà un préfixe inséparable, ne prennent pas le préfixe du participe II. Les verbes de modalité quant à eux, possèdent deux formes du participe II. Une forme est égale à la forme des verbes faibles. L'autre est formée comme un infinitif, elle est uniquement utilisée quand le verbe de modalité a un objet à l'infinitif (Huck, 2022).

Tandis que le participe I exprime en allemand une action en cours de réalisation, le participe II exprime l'action achevée ou qui a eu lieu dans le passé. Pour les deux genres de verbe, le préfixe *'ge-'* est ajouté à la base du verbe. Pour les verbes faibles, le suffixe *'-t'* complète le participe. Pour les verbes forts, c'est une conjugaison spécifique au verbe en question. Le participe II est employé dans les phrases complexes avec les temps composés comme le plus-que-parfait ou le passé antérieur. Il peut également être utilisé comme adjectif pour décrire par exemple une action ou une personne.

SPHERE PREPOSITIONNELLE

Comme en allemand standard, les prépositions dans les parlers dialectaux présentent une rection casuelle (Huck, 2022). Le tableau suivant démontre une comparaison. Les prépositions en allemand standard sont ici non exhaustives, il s'agit seulement de montrer les différences des cas pour les prépositions en commun en dialecte et en allemand.

Tableau 4 : Prépositions en allemand standard et en dialecte[27]

	Accusatif	Datif	Mixtes	Génitif
Allemand	bis, durch, für, gegen, ohne, um (…)	aus, auβer, bei, mit, nach, seit, zu, zuwider	an, auf, in, über, unter, vor, zwischen	während, wegen (…)
Dialecte	dur(ich), fir, geje, ohne, um	bi, gejeniwer, mit, no(ch), sàmt, trotz, üs, üsser, vo(n), waje/wage, während, zitter/sitter, zue	àn, hinter, in, iwwer, newe/nawe, uf, unter, vor, zwische	

Il est à ajouter ici que les rections prépositionnelles ne sont pas les mêmes pour l'allemand standard dû au fait que quatrième cas, le génitif, s'ajoute. Le nominatif a été soumis ici.
Selon Huck, le français semble exercer une influence dans les choix des prépositions sur les locuteurs dont leur langue maternelle est majoritairement le français. Cette tendance peut aller jusqu'à la transposition pure et simple (Huck, 2022).

[27] Source : (Huck, 2022) et moi-même

Pour conclure, on peut dire que la syntaxe, la morphosyntaxe et la morphologie sont peu impactées par des phénomènes de langues issu du français. Cela est probablement lié au fait que les systèmes morphosyntaxiques de ces deux langues sont assez distincts. Or, on peut déduire que l'allemand standard influence l'alsacien à un certain point. Mais compte-tenu des différences linguistiques entre les parlers dialectaux et l'allemand standard, il est incontestable qu'il s'agit des langues différentes, mêmes si les origines sont partagées. Des corpus parallèles alsacien-allemand ainsi qu'alsacien-français contenant ces exemples ont été construits.

2.4 Différentes zones dialectales en Alsace

Les dialectes alsaciens sont parlés dans le nord-est de la France (Bernhard et al., 2021), donc l'Alsace. L'Alsace fait 1,5 % du territoire national français et qui est habité par 3% de la population française (Huck et al., 2007). À côté du français, les dialectes alsaciens sont parlés.
Bien que le bas alémanique soit la variété majoritaire en Alsace (Huck, 2022), les autres variétés ne sont pas à ignorer. Il existe trois familles dialectales de l'allemand présentes en Alsace : le francique, l'alémanique et le francique rhénan méridional (Huck et al., 2007). Toutes ces familles font partie du haut-allemand, qui se divise en allemand moyen et en allemand supérieur. Le francique fait partie de l'allemand moyen, cependant, l'alémanique de l'allemand supérieur (Huck et al., 2007 et Bernhard et al., 2021).
La carte suivante montre la répartition du francique mosellan, du francique rhénan et celui du sud, du bas-alémanique du nord et du sud ainsi que du haut-alémanique. On peut également constater les différences phonétiques entre les familles dialectales. Cette constatation évoque à nouveau le fait que les parlers alsaciens ne sont pas normalisés, comme déjà mentionné auparavant. Les différences phonétiques sont accompagnées par des différences au niveau du lexique et éventuellement de la grammaire.

Figure 1 : Zone dialectale en Alsace et en Moselle[28]

[28] (carte créée par A. Horrenberger en 2017, publiée dans l'ouvrage « Collecting and annotating corpora for three under-resourced languages of France: Methodological issues » par Bernhard et al., 2017)

3. Traduction automatique

Cette partie est dédiée à la traduction automatique. Nous allons avoir un aperçu de l'histoire de la TA et présenter les différentes approches existantes. En outre, nous expliquerons une étape importante, à savoir la tokénisation. Les premières applications de l'intelligence artificielle en liaison avec les parlers dialectaux alsaciens seront nommés et commentés. Enfin, une place importante sera faite aux travaux de Lambrecht et al. (2022) qui ont beaucoup contribué à la traduction automatique vers l'allemand standard.

3.1 Bref historique

L'histoire de la traduction automatique est divisée en quatre grandes périodes : d'abord les années 1950 à 1960, ensuite le rapport ALPAC[29] avec ses conséquences, puis les années 1970 à 1980 et enfin la dernière période débutant dans les années 1990 allant jusqu'à aujourd'hui.
Déjà pendant la Seconde Guerre Mondiale, des tentatives de déchiffrage de code de l'ennemi ont été faites (Kübler et al., 2007). Après la guerre, ces techniques de déchiffrage grâce aux premiers ordinateurs ont été appliqués à la traduction automatique. La première conférence sur la TA, organisée en 1952, donne le résultat suivant : une traduction automatique de bonne qualité sans intervention humaine, que ce soit par pré-édition ou post-édition, n'est pas réalisable (Kübler et al., 2007; Volkart, 2018). En effet les premiers systèmes de TA n'étaient pas très puissants pour plusieurs raisons : les ordinateurs possèdent des capacités de stockage limitées et sont peu puissants et les théories linguistiques et syntaxiques ne sont pas très poussées. Seulement quelques années plus tard avec les analyses syntaxiques de Noam Chomsky, les modèles syntaxiques de la langue ont été pris en considération dans la recherche sur la TA et en traitement automatique des langues en général (Kübler et al., 2007).
Mais avec la publication du rapport ALPAC en 1966, l'avancement de la recherche dans la TA a essuyé un revers. Le rapport a conclu que, en comparaison à la traduction humaine, la TA était plus lente, moins efficace et beaucoup plus chère. La recommandation du rapport est d'arrêter les financements de la recherche dans la TA et se consacrer plutôt aux outils d'aide à la traduction (Kübler et al., 2007; Volkart, 2018).

Pendant la décennie suivante, le développement de la TA s'est poursuit, notamment au Canada et en Europe selon Kübler et al. (2007). Le système Systran par exemple est mis en place en 1976 et est utilisé encore aujourd'hui.
Durant la troisième période, les années 1980, les gros systèmes commercialisables ont été mis au point et les aides à la traduction comme les mémoires de traduction ont été développées. En plus de ces mémoires de traduction, des dictionnaires et des concordanciers ont été construits (Kübler et al., 2007).

[29] Automatic Language Processing Advisory Committee ; rapport consultable sous : Front Matter | Language and Machines: Computers in Translation and Linguistics |The National Academies Press

Avec la montée en puissance des ordinateurs et le développement d'Internet dans les années 1990, la recherche de TA a trouvé un nouvel essor. Les systèmes sont devenus plus individuels et efficaces (Kübler et al., 2007; Volkart, 2018).
Aujourd'hui, la TA est quasi omniprésente, surtout sur internet. Les sites web ainsi que les réseaux sociaux proposent des traductions générées de façon automatique de leurs contenus. Malgré ces avancements, la TA est parfois encore mal perçue par le public car la qualité reste médiocre (Volkart, 2018). Cela concerne notamment les langues moins courantes.

Il existe plusieurs approches et systèmes dans la traduction automatique. Ceux-ci sont expliqués sous la section 3.2 (Différentes approches existantes en TA). Néanmoins, la TA est un élément inéluctable dans la recherche du traitement automatique des langues.

3.2 Différentes approches existantes en TA

Comme nous l'avons déjà constaté précédemment, la traduction automatique a un passé riche en facettes et en péripéties. Au fil du temps, différentes approches ont été développées, améliorées ou perfectionnées. Cette partie du travail a pour objectif de faire une synthèse des différentes approches de la traduction automatique. Ces approches ont été choisie car ce sont les approches les plus courantes.
En général, les approches se basent soient sur des données linguistiques, soit sur des corpus bilingues, donc sur des données empiriques. En premier temps, nous allons présenter les approches linguistiques. Ensuite, les approches empiriques voire statistiques sont présentées.

3.2.1 Approche par traduction directe

L'approche par traduction directe, aussi appelée traduction mot à mot, est une approche très basique et moins efficace dans la traduction automatique. Le nom de cette approche vient du fait que la traduction se fait directement, de la langue source à la langue cible. Aucune représentation intermédiaire a lieu (Volkart, 2018) .
Elle consiste en trois étapes pour arriver à une traduction finale. Durant la première étape, la phrase d'entrée est segmentée en mots. Pendant l'analyse morphologique, un dictionnaire unilingue est consulté par le système afin de classer les segments dans leurs catégories grammaticales. L'identification des formes fléchies se base sur des règles morphologiques (Volkart, 2018). Dans cette étape, la gestion des temps de conjugaison des mots est également poursuivie (Bouhrim & Zenkouar, 2017). La deuxième étape correspond à trouver des équivalents de la phrase d'entrée dans un dictionnaire dans la langue cible. La troisième étape génère le résultat final en faisant un réordonnancement dans la langue cible (Bouhrim & Zenkouar, 2017).
Les problèmes de cette approche sont qu'elle est très spécifique. L'approche par traduction directe marche pour une seule paire de langue et dans une seule direction. La structure syntaxique est ignorée. La construction d'un dictionnaire bilingue de très grande taille est également nécessaire

(Bouhrim & Zenkouar, 2017). L'ambiguïté pose également des problèmes. C'est pourquoi ces systèmes peuvent seulement traduire des langues qui sont assez proches (Volkart, 2018).

3.2.2 Approche par interlangue

Contrairement à l'approche par traduction directe, les systèmes par interlangue créent des représentations intermédiaires. Ils disposent également d'une grammaire détaillée et d'une grammaire comparative de chaque langue (Volkart, 2018).
La traduction elle-même se fait en deux étapes. Premièrement, la phrase d'entrée, donc la langue source, est analysée et transformée en l'interlangue, la représentation intermédiaire. Cette représentation contient toutes les informations principales concernant la syntaxe et la sémantique. Comme deuxième étape, la phrase de sortie, la langue cible, est générée à partir de la représentation interlangue (Volkart, 2018).
Vu que cette approche est indépendante des langues sources et cibles, elle ouvre la voie à la création de systèmes multilingues. Car pour ajouter une nouvelle langue, il faut plutôt ajouter les modules d'analyse. Avec ces modules, il s'agit d'un module qui est capable de générer la représentation après une analyse et un module complémentaire, qui génère à partir de la représentation la phrase de sortie (Volkart, 2018).
Il reste tout de même des difficultés dans cette approche. Pour une représentation totalement indépendante de la langue, toutes les informations syntaxiques, sémantiques, morphologiques, relationnelles ainsi que conceptuelles doivent être prise en compte. Surtout les concepts exprimés par le lexique, qui peuvent, selon les langues, être représentés de plusieurs façons (Volkart, 2018).

3.2.3 Approche par système de transfert

Comme le système par interlangue, les systèmes de transfert fonctionnent également avec une représentation intermédiaire. Mais celle-ci est dépendante de la langue (Volkart, 2018).
Le système de transfert est beaucoup utilisé durant les années 1990. L'entrée dans la langue source est analysée. Un deuxième module contient des règles de transfert ayant des règles syntaxiques ainsi que des dictionnaires. En plus de seulement traduire les phrases mot par mot, des arbres syntaxiques complets sont traduits (Bouhrim & Zenkouar, 2017). Ensuite, une traduction dans la langue cible est générée. L'analyse de la langue source, le transfert et la génération de la langue cible sont les trois modules nécessaires pour ces systèmes (Volkart, 2018). La croissance des documents disponible sur support électronique, a enrichi les dictionnaires et a mené à la création des dictionnaires spécialisés (Kübler et al., 2007).
Cette approche évite les difficultés de création de représentation intermédiaire dû au fait qu'elle est moins abstraite. Par conséquent, il est plus difficile d'ajouter une nouvelle langue dans ces systèmes, car cela nécessite 3 nouveaux modules (Volkart, 2018).

3.2.4 Approche par TA basée sur des exemples

Dans cette approche, la traduction est basée sur des exemples déjà existants. Les étapes sont : trouver les correspondances, puis aligner les phrases et enfin les combiner (Bouhrim & Zenkouar, 2017). C'est par des analyses sur des corpus traduits et alignés que les traductions les plus fréquents sont détectées (Kübler et al., 2007).
Le désavantage de cette approche est que la chance de retrouver exactement la phrase correspondante est peu élevée. Le système cherche alors des petites parties, très courtes, ce qui rend la génération de la phrase cible plus compliquée (Volkart, 2018).
L'avantage de cette approche est qu'une formalisation et une construction des règles de transfert compliquées sont évitées.

3.2.5 Approches par des systèmes statistiques

L'approche statistiques se base sur des calculs de probabilité. Elle travaille pour cela avec des corpus bilingues. À partir de ces corpus, un modèle statistique est créé qui contient toutes les traductions possibles observées dans les corpus, il est donc similaire à un grand dictionnaire bilingue. En fonction de la fréquence d'occurrence dans le corpus, chaque traduction est associée à une probabilité. Des n-grammes, les probabilités des séquences de mots dans la langue cible, sont également calculés (Volkart, 2018).
Pour arriver à une traduction finale, le système cherche parmi toutes les traductions possibles, celle avec la probabilité la plus élevée selon les modèles de traduction et de langue (Volkart, 2018).
Dans la TA statistique (TAS), il existe deux principaux cadres de modélisation (le *noisy-channel* et le *log-linear*) permettant de trouver la traduction la plus probable. Peu importe quel cadre est utilisée, les deux demandent deux éléments principaux : un modèle de langue et un modèle de traduction (Volkart, 2018). Les systèmes utilisant le cadre de modélisation du *log-linear* ont la possibilité d'attribuer un poids plus ou moins importants au modèle de langue et au modèle de traduction. Cet ajustement est autrement appelé *tuning*. Le *log-linear* permet également d'intégrer des composantes supplémentaires. C'est ainsi que des probabilités de traduction bidirectionnelles ou des pénalités pour les phrases trop longues ou trop courtes peuvent être prise en compte (Volkart, 2018).
Une traduction automatique statistique consiste en deux étapes, l'entraînement avec des corpus et la traduction. Les deux modèles de langue et de traduction sont créés durant l'entraînement (Volkart, 2018).
Pour mieux expliquer ces deux étapes d'entraînement et de traduction, nous allons faire des sous-sections pour chacune de ces étapes.

ENTRAINEMENT

Comme mentionné auparavant, dans cet étape le modèle de langue et celui de traduction sont créés. Il s'agit des modèles indispensables pour un système de traduction automatique statistique. D'abord, nous allons voir le fonctionnement du modèle de langue.

Le modèle de langue se base sur un corpus monolingue dans la langue cible, il certifie la fluidité de la traduction. Le modèle contribue également au choix des mots ainsi que leur ordre (Volkart, 2018).
Il existe trois sous-modèles pour calculer la probabilité d'une phrase qui reposent sur des modèles de n-grammes. La probabilité se construit par des multiplications des n-grammes. Le modèle unigramme est le plus simple. La probabilité d'apparition par mot est calculée comme suit : le mot en question est repéré dans le corpus, ensuite le nombre d'occurrences du mot est divisé par le nombre totale des mots. Comme chaque mot a sa propre probabilité, les phrases mal formées sont difficiles à détecter (Volkart, 2018).
Le modèle bigramme prend en compte des séquences de deux mots dans le corpus. Il s'agit de la probabilité que le deuxième mot suit le premier (Volkart, 2018). Si nous avons par exemple « mot1 » et « mot2 », les occurrences de « mot1-mot2 » sont comptées et divisées par le nombre total des occurrences de « mot1 ».
Le modèle trigramme repose sur le même principe que le modèle bigramme, mais avec une séquence de trois mots. C'est-à-dire la probabilité que le troisième mot suit les deux précédents (Volkart, 2018). Le nombre d'occurrences de « mot1-mot2-mot3 » est divisé par le nombre d'occurrence de « mot1-mot2 ».
Plus une traduction basée sur des modèles de n-grammes longs, plus elle sera fluide (Volkart, 2018). La taille du corpus d'entrainement importe sur le choix de modèle de n-grammes. Parce que le plus que le corpus a une taille importante, plus il est possible de prendre en compte des séquences de mots longues (Volkart, 2018). Tout de même, il reste toujours une chance que des n-grammes ne soient pas présents dans le corpus d'entrainement, mais apparaissent dans les phrases à traduire (Volkart, 2018).
Pour éviter qu'une probabilité d'une phrase soit égale à 0 en raison de l'absence des n-grammes, il existe des mécanismes de *smoothing* (Volkart, 2018). Ces mécanismes, dont il existe différents types, modifient les probabilités de manière empirique pour obtenir « *une réserve de probabilités à attribuer aux n-grammes absents du corpus* »[30] (Volkart, 2018).
À côté de ces mécanismes, il existe d'autres moyens pour résoudre le problème de l'absence des n-grammes. L'*interpolation* ou le *back-off* peuvent être employés. Une combinaison des différents modèles de langues (modèle unigramme, bigramme ou trigramme) et donc une combinaison des différentes longueurs des n-grammes est effectuée lors de l'interpolation. Un poids plus important est attribué aux modèles ayant des n-grammes plus longs (Volkart, 2018).
Le back-off est un cas de priorité. D'abord, les modèles de langue de n-grammes longs sont utilisés et si des n-grammes ne se trouvent pas dans le corpus, des modèles de langue de n-grammes courts sont utilisés.

[30] Citation directe de *Traduction automatique statistique vs. neuronale : Comparaison de MTH et DeepL à La Poste Suisse* par L. Volkart, 2018

Nous venons de voir que le modèle de langue assure la fluidité. Le modèle de traduction assure cependant la fidélité. Contrairement au modèle de langue, il est basé sur un corpus bilingue aligné (Volkart, 2018).
Ici, la probabilité d'une traduction d'un mot ou d'une séquence de mots dans la langue cible est calculée. Toutes les traductions possibles ainsi que leurs probabilités sont représentées sous forme de tableau (Volkart, 2018).
En raison du fait que le modèle de langue peut être créé sur différentes longueurs des n-grammes, le modèle de traduction s'adapte au modèle de langue (Volkart, 2018). Il n'y a pas de chevauchement des probabilités des systèmes basés sur des mots isolés et des systèmes basés sur des séquences de mots.
Comme mentionné auparavant, ce modèle est basé sur un corpus bilingue et aligné. Or, les corpus ne sont pas alignés sur des mots mais à la base des phrases. Afin de pouvoir identifier les paires source-cible, l'outil *expectation maximization algorithm* est employé (Volkart, 2018).
Cet algorithme permet d'aligner le corpus sur des mots. Pour faire cela, il y a trois étapes : *initialisation*, *expectation* et *maximisation* (Volkart, 2018).
Dans la première étape, toutes les paires de mots possibles sont considérées et la même probabilité leur est assignée. Ce modèle est ensuite, dans la deuxième étape, appliqué aux données du corpus. Une probabilité à chaque paire en fonction de leur fréquence est donnée (Volkart, 2018). Enfin, lors de la dernière étape, les nouvelles probabilités sont ajoutées au modèle. Afin d'améliorer les estimations, la deuxième et troisième étapes sont répétées à plusieurs reprises. C'est ainsi que les paires les plus probables pour chaque traduction peuvent être déterminées (Volkart, 2018).
Cet algorithme permet alors d'établir l'alignement du corpus basé sur des mots et dans ce même principe sur des séquences de mots.

Traduction

La traduction, aussi appelé le décodage, est la deuxième étape dans la TA statistique. On pourrait dire qu'il s'agit d'une application de l'étape d'avant en pratique. La traduction est complexe car le nombre de traductions possibles est très grand (Volkart, 2018).
Les systèmes basés sur des séquences de mots sont davantage utilisés que ceux sur des mots à cause de leur performance plus élevée. Volkart (2018) explique cela par les problématiques des modèles unigrammes mentionnés plus haut : ces modèles ont du mal à traiter les cas où le mot source peut être traduit par deux mots cibles. Utiliser les modèles des segments résout les ambiguïtés.
Durant cette étape, les modèles décrits plus hauts sont utilisés pour trouver la traduction la plus probable. La phrase source est segmentée de toutes manières possibles, ensuite les traductions des segments sont cherchées par le système dans les tables de traduction du modèle de traduction (Volkart, 2018). Le modèle de traduction fournit alors à ce point une certaine probabilité pour chaque traduction. C'est le modèle de langue qui assigne par la suite également une probabilité à chaque traduction (Volkart, 2018).
Les deux probabilités sont après combinées pour obtenir le meilleur score de probabilité d'une phrase (Volkart, 2018).

Un modèle de réordonnancement est aussi appliqué pour les systèmes basés sur les segments (Volkart, 2018).

3.2.6 Approches par des réseaux neuronaux

L'approche par des réseaux neuronaux a visé l'amélioration de la performance des systèmes statistiques. Ensuite, des approches purement neuronales pour une traduction automatique ont été développés. Cette approche reste assez récente, c'est pourquoi la recherche dans ce domaine est très active (Volkart, 2018).
Les méthodes de la traduction automatique neuronale (TAN) peuvent être classées en trois catégories : supervisée, semi-supervisée et non supervisée. La TAN supervisée est l'architecture par défaut qui s'appuie sur des ensembles de données à grande échelle (Ranathunga et al., 2023). Quand un corpus parallèle à disposition est petit, il est possible de fusionner ces données avec autres données monolingues de la langue en question. Ici, on parle de la TAN semi-supervisée (Ranathunga et al., 2023). On parle alors des méthodes non supervisées lorsqu'il n'y a pas de données parallèles. Il s'agit du cas extrême (Ranathunga et al., 2023). Il est connu qu'il n'existe pas une grande quantité des données parallèles pour les langues peu dotées. Une solution consiste à générer synthétiquement des données, ce que l'on appelle l'augmentation des données. La traduction inversée en fait partie (Ranathunga et al., 2023).
Cette approche par réseaux neuronaux n'est cependant plus statistique. En théorie, il s'agit d'un système qui apprend à l'aide d'un corpus d'entrainement. Ces systèmes disposent d'un encodeur pour l'analyse de la langue source et d'un décodeur pour la génération de la langue cible. Les systèmes savent prendre en compte la totalité de la phrase en langue source. L'analyse se fait de manière totalement autonome, la génération s'appuie sur cette analyse (Volkart, 2018)
Ce sont l'encodeur et le décodeur qui possèdent les réseaux de neurones. Ils sont entrainés généralement sur un corpus bilingue, mais peuvent également être utilisés pour les corpus monolingues. Contrairement à la TA statistique, le système neuronal n'a besoin qu'un seul modèle (Volkart, 2018).
Le fonctionnement d'un système neuronal ressemble à celui par interlangues vu en section 3.2.2 : une représentation intermédiaire du sens de la phrase est utilisée pour créer la traduction. Or, le fonctionnement d'un système neuronal est beaucoup plus complexe (Volkart, 2018).
C'est l'encodeur qui donne la représentation de la phrase source et le décodeur s'occupe à utiliser cette représentation pour prédire la phrase cible (Volkart, 2018).
Cette représentation est appelée *word embedding* ou *plongement lexical* en français (Volkart, 2018). Le plongement lexical représente les sens du mot en se basant sur la supposition que des mots apparaissant dans le même contexte sont similaires (Volkart, 2018). Au plongement lexical, il s'agit d'un espace multidimensionnel. Ici, chaque mot est représenté par un vecteur, les mots qui ont des propriétés (sémantiques, syntaxiques ou morphologiques) communes, se retrouvent proches les uns aux autres dans une certaine dimension (Volkart, 2018).

C’est ainsi que le plongement lexical peut regrouper des mots en fonction de leur sens. Et par conséquence, la capacité de traiter des séquences de mots inconnus, qui n’apparaissent pas dans le corpus d’entrainement, est donnée.
Cela et la prise en compte des larges contextes sont des avantages de la TAN. Elle est à un certain point flexible si des énoncés inconnus sont rencontrés (Volkart, 2018). Si par exemple le système traite une séquence de cinq mots dont un est inconnu, la tokénisation suivi par le plongement lexical permettent tout de même de traiter la phrase. Cela est seulement possible si la phrase se retrouve dans le corpus d’entrainement mais à la place du mot inconnu, il y a un mot qui lui est proche sémantiquement (Volkart, 2018).
Construire des modèles de langue pour chaque paire de langue n’est pas pratique. Une solution pour cela sont les *multilingual-NTM models* (modèles multilingue-TAN, court multi-TAN). Ils facilitent la traduction entre les paires de langue en utilisant qu’un seul modèle. En général, ils sont basés sur la TAN supervisée (Ranathunga et al., 2023).
La multi-TAN peut être catégorisée en trois types. Premièrement la traduction d’une langue source (LS) vers plusieurs langues cibles (LCs) *(one-to-many)*. Il s’agit essentiellement d’un problème multitâche, où chaque cible devient une nouvelle tâche. Deuxièmement, la traduction de plusieurs LSs vers une LC *(many-to-one)*. Il s’agit alors d’un problème multi-source. Il est considéré comme plus facile que le problème multitâche. Dernièrement, la traduction de plusieurs langues vers plusieurs langues *(many-to-many)*, c’est un problème de multi-source, multi-cible et par conséquent le scénario le plus difficile (Ranathunga et al., 2023).
Les systèmes pour cette traduction supervisée possèdent également différents types d’encodeur ou de décodeur. Ici, il existe à nouveau trois catégories : 1) un seul encodeur-décodeur pour toutes les langues : toutes les phrases de la LS sont entrées dans l’encodeur, quelle que soit la langue. Le décodeur est capable de générer n’importe quelle LC ; 2) un encodeur-décodeur par langue : chaque LS a son propre encodeur et chaque LC a son propre décodeur ; 3) un seul encodeur-décodeur partagé d’un côté avec un décodeur/encodeur par langue de l’autre côté (Ranathunga et al., 2023). L’objectif principal de ces différents types d’architecture est de maximiser les informations partagées entre les langues en conservant les informations spécifiques à la langue afin de la distinguer (Ranathunga et al., 2023). Presque toutes les architectures de la multi-TAN sont basées sur les modèles à base de transformers.
Le modèle à base de transformers consiste d’un encodeur ainsi qu’un décodeur. Selon le modèle utilisé, il peut avoir un tas de six couches d’encodeur et de décodeur (Xu, 2021). Outre la matrice du plongement lexical *(Embedding)* et du plongement lexical positionnel *(Positional Embedding)* dans l’encodeur et le décodeur, le décodeur possède également un classifieur *(Classifier)* afin de produire des tokens traduits (Xu, 2021). Dans l’encodeur, chaque couche se compose d’un réseau d’auto-attention *(Self-Attention)* qui prend en compte l’ensemble de la séquence d’entrée pour construire des représentations contextuelles, ainsi d’un réseau neuronal à propagation avant *(Feed-Forward)* pour traiter l’information recueillie. Le décodeur se compose également de ces deux réseaux, mais un réseau d’attention-croisée *(Cross-Attention)* s’ajoute entre les deux réseaux d’auto-attention et d’anticipation. Il apporte des informations concernant la représentation de la source d’encodeur (Xu, 2021).

La figure suivante est une illustration simplifiée de ce modèle.

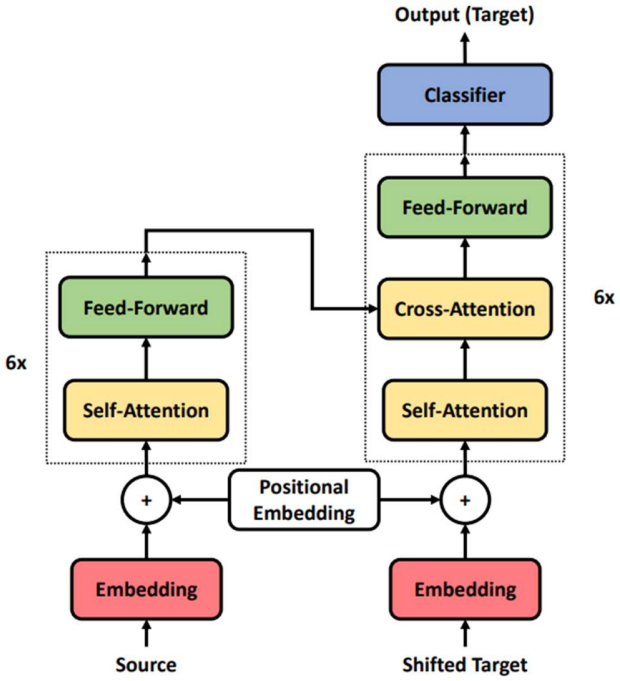

Figure 2 : Modèle de transformer[31]

La TA neuronale est, comme mentionné plus haut, encore un domaine de recherche très actif. Cela veut dire aussi qu'il existe des difficultés (Volkart, 2018) : bien qu'on puisse accéder à des systèmes neuronaux en ligne, ils ne peuvent éventuellement pas être spécialisés aux données de l'utilisateur. Une grande puissance computationnelle est également exigée (Volkart, 2018). Par conséquence, la taille du vocabulaire des systèmes est limitée. En outre, pour un déroulement sans problème et une traduction satisfaisante, l'entrainement nécessite une quantité de données très élevée (Lambrecht et al., 2022; Volkart, 2018). Enfin, la complexité des systèmes neuronaux rend l'identification et la correction des erreurs difficiles (Volkart, 2018). Les deux problèmes principaux de la TA des dialectes sont l'acquisition des données et l'absence d'une normalisation de l'orthographe (Lambrecht et al., 2022).
Ces deux problèmes mènent à la réflexion comment bien traduire des langues peu dotées. Dans la section suivante, des possibles solutions ou approches sont présentées.

3.3 Traduction des mots rares

Tandis que beaucoup de mots sont faciles à traduire, il n'existe pas toujours un équivalent 1-1 (Sennrich et al., 2016). La traduction des mots peut être transparente dans le cas où le mot en question est traduisible pour un traducteur compétent, même s'il ne connait pas le mot. Cela est possible par une traduction des unités de sous-mot connues (Sennrich et al., 2016). Les catégories

[31] Source : Transformer-Based NMT: Modeling, Training and Implementation (Xu, 2021)

dont la traduction semble transparente sont les noms entités, les mots apparentés et emprunts ainsi que des mots morphologiquement complexes (Sennrich et al., 2016).
La théorie de Sennrich et al. (2016) est que si on segmente des mots rares en unités de sous-mots, un système de traduction neuronal est capable d'apprendre des traductions transparentes et par la suite, pouvoir généraliser ce savoir pour générer des mots inconnus (Sennrich et al., 2016). Un algorithme de segmentation basé sur l'encodage en paire d'octet est employé afin d'obtenir un vocabulaire qui apporte un taux de compression du texte (Sennrich et al., 2016).
Les résultats montrent que les systèmes neuronaux sont capables de la traduction des mots rares et mots inconnus par une séquence de sous-mots (Sennrich et al., 2016). Cela pourrait être applicable pour la plupart de paires de langues et éliminer le besoin d'un grand vocabulaire pour la traduction automatique neuronale (Sennrich et al., 2016).

Si nous parlons de la segmentation des mots, il faut également expliquer ce que c'est concrètement. Cette segmentation est appelée la tokénisation. Il existe la tokénisation au terme classique et notamment au nouveau terme en rapport avec la TAN.

3.3.1 Tokénisation

Depuis l'introduction du terme de la tokénisation, le sens a légèrement changé. Dans le sens classique et historique, la tokénisation est une des premières étapes dans le traitement automatique des langues (TAL). Dans le processus d'analyse automatique des textes, elle se situe avant le traitement lexical et les analyses syntaxiques ainsi que sémantiques. C'est grâce à cette étape de la tokénisation, que finalement un texte peut être traité et par conséquent, être traduit.
La tokénisation consiste à découper un texte en phrases et les phrases sont à nouveau découpées en unités lexicales. Ces unités lexicales, ainsi dénommées tokens, peuvent être constitués de plusieurs choses : des unités mono-lexicales, des expressions idiomatiques, des mots composés, des entités nommées, des locutions ou des chiffres. Un token n'est pas découpé en unités encore plus petites (Bernhard et al., 2017).
Afin de marquer des frontières de phrase et par la suite des mots, la tokénisation se base sur des séparateurs ou délimiteurs. En général, il s'agit des signes de ponctuations et d'espace. Cependant, certains délimiteurs comme l'apostrophe, le tiret ou le point sont ambigus et nécessitent un traitement plus soigné.
Dans le contexte récent des modèles neuronaux, la tokénisation consiste en découpage en sous-mots. Ici, les tokens ne doivent pas forcément porter un sens comme les expressions idiomatiques. C'est un découpage en petites unités afin de traiter ces unités par la suite par un plongement lexical par exemple.
Bernhard et al. (2017) décrivent dans leur article sur les problèmes de tokénisation pour deux langues régionales de France, qu'un autre problème relève les textes de spécialité car ils sont non seulement adaptés aux besoins de leur domaine mais aussi bien liés aux phénomènes particuliers.
Quant à la tokénisation pour l'alsacien, d'autres problèmes existent. Bien que la morphosyntaxe se rapproche beaucoup à l'allemand standard et est comparable avec le dernier, les dialectes alsaciens

restent non-normés. Par conséquent, il n'existe pas des règles ou des normes sur la grammaire et l'orthographe. Cela complique la tokénisation.
Dans l'article mentionné plus haut, un tokéniseur pour l'alsacien a été développé. La tokénisation est faite à l'aide des expressions régulières qui permet de distinguer plusieurs types d'unités. C'est ainsi que les articles, les prépositions et conjonctions, les pronoms situés à droite d'un verbe, les nombres, les abréviations, les adresses mail, les URLs et des suites de caractères situés en milieu de mot comme unités indépendantes. Ce tokéniseur spécifique aux dialectes alsaciens est plus performant que le tokéniseur générique fourni avec le TreeTagger (Schmid, 1997). Il reste tout de même des erreurs. Ceux-ci se sont produits selon Bernhard et al. (2017) au niveau des caractères ambigus qui doivent être délimités ou non selon le contexte.

Les deux parties suivantes présentent des algorithmes de tokénisation ou de la gestion de la traduction de mots rares.

3.3.2 Byte Pair Encoding, encodage en paire d'octet

Byte Pair Encoding est une simple technique de compression des données (Sennrich et al., 2016). Il s'agit d'un algorithme qui remplace la paire d'octet la plus fréquente dans une séquence par un seul octet non-utilisé (Sennrich et al., 2016). Sennrich et al. (2016) proposent de ne pas fusionner des paires d'octet fréquentes, mais des caractères ou séquences de caractères. C'est ainsi que le symbole de vocabulaire et celui du caractère sont d'abord initialisés et chaque mot représenté comme une séquence de caractères. Afin de pouvoir garder la tokénisation initiale, un symbole spécifique est mis en place à la fin des mots ('·'). Ensuite, les paires de symboles sont repérées et chaque occurrence de la paire la plus fréquente ('A', 'B') est remplacée par un nouveau symbole fusionné ('AB') (Sennrich et al., 2016). Cela génère pour chaque opération de fusionnement un nouveau symbole représentant un n-gramme de caractères. Les n-grammes de caractères fréquents sont à nouveau fusionnés en un seul symbole. La taille du vocabulaire des symboles final correspond à la taille du vocabulaire initial, plus le nombre d'opérations de fusionnement (Sennrich et al., 2016).
C'est ainsi que l'algorithme peut être employé sur un dictionnaire extrait d'un texte. Chaque mot est pondéré par sa fréquence (Sennrich et al., 2016).
Sennrich et al. (2016) déclarent qu'avec cette méthode, les séquences de symboles sont toujours interprétables en tant que « *subword units* » (unités de sous-mot) (Sennrich et al., 2016). Mais aussi que le réseau peut généraliser pour les traduire et générer des nouveaux mots, mêmes inconnus ou jamais rencontrés lors de l'entraînement sur la base de ces unités de sous-mot (Sennrich et al., 2016).
Deux méthodes d'application du BPE peuvent être considérées selon Sennrich et al. (2016) : premièrement l'apprentissage de deux encodages indépendants, à savoir un pour la langue source et l'autre pour le vocabulaire cible. Deuxièmement l'apprentissage d'encodage sur la fusion de deux vocabulaires, aussi appelé « *joint BPE* » (raccord / assemblage d'encodage en paires d'octet) (Sennrich et al., 2016).

Le BPE n'est qu'un exemple d'algorithme de découpage en sous-mots parmi d'autres. Il s'est avéré pratique pour la tokénisation des mots inconnus ou rares. La grande variété orthographique dans les dialectes alsaciens et le fait que c'est une langue peu dotée, pourrait faire que les mots sont vus comme des mots rares.

3.4 Mesures d'évaluations

Quant à l'évaluation d'une traduction automatique, plusieurs méthodes existent. Il existe naturellement l'évaluation humaine, où un ou plusieurs humains évaluent la traduction générée. Ces évaluations sont très souvent basées sur la fluidité et la naturalité de la traduction, les humains ne doivent pas être des traducteurs eux-mêmes pour évaluer une traduction dans la langue maternelle. Avec l'avancée de la technologie, différentes méthodes d'évaluation automatique ont été développées. Ces méthodes se basent sur la comparaison par exemple d'une phrase complète ou les caractères de n-grammes ou le taux d'erreurs.

3.4.1 Évaluation par BLEU

En 2002, Papineni et al. présentent un système d'évaluation des traductions. Pour ce système appelé « *bilingual evaluation understudy* » (BLEU) deux éléments de base sont nécessaires : une métrique numérique de *proximité de traduction* et un corpus de traductions humaines de référence (Papineni et al., 2002). La métrique de proximité de traduction est façonnée après la métrique *word error rate* (taux d'erreur de mot) qui est aussi utilisée dans la reconnaissance vocale. L'idée principale du système d'évaluation est d'utiliser une moyenne pondérée des correspondances de segments de longueur variable par rapport aux traductions de référence (Papineni et al., 2002).
Pour évaluer une traduction automatique, il suffit de comparer les correspondances de n-grammes entre chaque traduction candidate (donc la TA) et les traductions de référence (Papineni et al., 2002). Les correspondances ne sont pas dépendantes de la position dans la phrase. On suppose que plus le nombre de correspondances est grand, meilleure est la traduction (Papineni et al., 2002).
Afin d'obtenir une précision modifiée de n-grammes, tous les n-grammes d'une traduction candidate apparaissant dans les traductions de référence sont comptés et ensuite divisés par le nombre total de n-grammes de la traduction candidate (Papineni et al., 2002). Une fois qu'une correspondance dans les traductions de référence a été trouvée, elle est considérée comme épuisée et ne sera pas reconnue une deuxième fois (Papineni et al., 2002). Cette précision de n-grammes assure deux aspects importants lors d'une évaluation de traduction : la fluidité et l'adéquation (Papineni et al., 2002).
Lorsqu'un mot est employé dans la traduction candidate alors qu'il n'apparaît pas dans une traduction de référence, la traduction candidate est pénalisée. Elle est également pénalisée si un mot est plus fréquemment utilisé que dans le maximum d'occurrence de référence (Papineni et al., 2002). Les traductions candidates ayant des scores élevés doivent correspondre aux traductions de référence en longueur, en choix de mot et en ordre de mots (Papineni et al., 2002). Si une traduction candidate a la même longueur qu'une traduction de référence, la *brevity penalty* (pénalité de

brièveté) est initialisée à 1. Cependant, la longueur de la phrase de référence la plus proche est considérée comme la « longueur de le meilleure correspondance » (Papineni et al., 2002).
La métrique BLEU varie entre 0 et 1. Peu de traductions auront un score parfait de 1 sauf si elles sont identiques avec la traduction de référence. Plus le nombre de traductions de référence par phrase est grand, plus haut sera le score (Papineni et al., 2002). Papineni et al. (2002) ont fait des tests d'évaluation avec un groupe de personnes bilingues et un groupe de personnes monolingues ainsi qu'avec BLEU. Les résultats montrent que le coefficient de corrélation est de 0,96 pour le groupe bilingue et 0,99 pour le groupe monolingue (Papineni et al., 2002).
L'utilisation des scores BLEU est devenue la méthode prédominante pour évaluer la qualité d'une traduction (Lambrecht et al., 2022). En comparaison à l'évaluation par l'humain, BLEU est moins cher, plus rapide et moins subjectif (Lambrecht et al., 2022; Papineni et al., 2002).
Il existe une variante d'implémentation de BLEU nommée sacreBLEU (Song et al., 2023). Elle a le même objectif d'évaluer la qualité de la traduction en mesurant la distance entre la phrase traduite et la phrase de référence (Song et al., 2023). Cette implémentation facilite l'utilisation de la métrique. L'implémentation supporte également d'autres métriques comme chrF(++) et TER. Des tokéniseurs différents sont également supportés comme ceux du chinois ou du japonais.
BLEU rencontre des problèmes avec des langues qui sont morphologiquement riches. En général, seulement une référence de traduction est disponible. Cela peut diminuer le score BLEU pour de telles langues (Lambrecht et al., 2022). Dans ces cas, une évaluation humaine pourrait être une option bénéfique (Lambrecht et al., 2022).
La formule BLEU est la suivante :

Équation 1 - BLEU

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Ici, N est le nombre maximal de n-grammes considérés. w_n est le poids attribué à chaque taille de n-gramme, p_n cependant correspond à la précision des n-grammes. Elle mesure la proportion de n-grammes dans la traduction automatique qui apparaissent également dans la référence. BP est la pénalité de brièveté (Papineni et al., 2002).

3.4.2 Évaluation par chrF++

Le score F des n-grammes de caractères compare les n-grammes de caractères dans les textes sachant le résultat de la TA et une ou plusieurs traductions de référence humaines. Il intègre des améliorations et des fonctionnalités supplémentaires pour fournir une évaluation plus précise et fiable. Certaines des principales caractéristiques et améliorations de chrF++ sont la flexibilité, le lissage et la normalisation. Le lissage est une technique qui permet de traiter les cas où les n-grammes de la traduction candidate ne sont pas présents dans la traduction de référence. La métrique chrF++ prend en compte les unigrammes et les bigrammes (Popović, 2017).

Le score est entre 0 et 100 représenté en pourcentage. De manière similaire au score BLEU, on essaie d'obtenir des scores se rapprochant le plus de 100.
Le formule pour cette métrique est la suivante :

Équation 2 - ChrF++

$$ngrF\beta = (1 + \beta^2)\frac{ngrP \times ngrR}{\beta^2 \times ngrP + ngrR}$$

Ici, ngrP représente le pourcentage de n-grammes dans l'hypothèse qui a une contrepartie dans la référence, ngrR représente le pourcentage de n-grammes dans la référence qui sont également présent dans l'hypothèse. Les deux représentent donc la précision et le rappel des n-grammes. β est un paramètre qui attribue β-fois plus de poids au rappel qu'à la précision (Popović, 2017).

3.4.3 Évaluation par Translation Edition Rate

Le *Translation Edit Rate*, ou TER (taux d'édition de la traduction) vise un score proche de 0. Il calcule le nombre d'éditions nécessaires pour corriger le résultat afin qu'il corresponde sémantiquement à une traduction de référence. Il compte donc le nombre d'éditions effectuées (par un humain) pour modifier l'hypothèse.
Si aucune référence de traduction existe, elle sera créée. Si une référence de traduction existe déjà, elle est comparée directement avec l'hypothèse, donc la traduction candidate. Ensuite, le nombre minimum de modifications est déterminé. Tout changement ou édition à un poids égal à 1. Cette méthode est très précise lorsque la référence est la plus proche possible de l'hypothèse.
Le score peut être calculé par la formule (Snover et al., 2006) suivante :

Équation 3 - TER

$$TER = \frac{nombre\ d'éditions}{nombre\ moyen\ de\ mots\ de\ référence}$$

3.5 Augmentation des données

Comme mentionné dans diverses sections dessus, le manque de données est en vrai problème dans la traduction automatique des langues peu dotées. Il existe plusieurs façons d'augmenter les données artificiellement. Dans la suite, quelques manières sont présentées.

3.5.1 Back-translation, la traduction inversée

Lors de l'entrainement, des modèles de langue source à langue cible et langue cible à langue source peuvent être employés. Il s'agit dans ce cas de la traduction inversée, en anglais *Back-translation* (Edunov et al., 2018). La traduction inversée est une alternative pour des données monolingues car elle est facile et simple à appliquer, car elle n'exige pas de modifier des algorithmes (Edunov et al., 2018). Des données synthétiques supplémentaires sont générées à partir des données monolingues de la langue cible (Edunov et al., 2018). Cette langue cible est généralement une langue bien dotée disposant des corpus monolingues en quantité suffisante. Quand les données parallèles sont rares, la traduction inversée a été jugée comme très bénéfique pour la TA par réseaux neuronaux (Edunov et al., 2018).

3.5.2 Pseudo-traduction

La pseudo-traduction inclut de simuler un processus de traduction de la langue source à une autre langue fictive. La sortie de cette langue fictive est à nouveau traduite vers la langue source. La langue fictive dépend des algorithmes et modèles utilisés pour générer le texte mais peut prendre les caractéristiques suivants : mélange des langues, ajout de caractère ASCII, incohérence grammaticales ou sémantiques (*Pré-traduction (TMS)*, s. d.).
Cette double traduction peut augmenter les données. Au lieu d'une langue artificielle, une autre langue disposant de ressources et proche à la langue cible, peut être utilisée car elle a une relation sémantique ou linguistique proche avec la langue à faibles ressources (Song et al., 2023).

3.5.3 Knowledge distillation, la distillation des connaissances

La distillation des connaissances consiste à migrer les connaissances obtenues d'un large modèle, aussi appelé *teacher model (modèle d'enseignant)*, ou des multiples modèles à un autre modèle plus léger, en anglais *student model (modèle d'étudiant)* (Song et al., 2023). De manière générale, ce *teacher model* est entraîné sur une langue bien dotée, le *student model* est souvent destiné à une langue peu dotée. C'est le modèle d'enseignant qui traduit les données d'entrainement, ce qui résulte en données synthétiques du côté cible (Haddow et al., 2022). Cette distillation des connaissances est basée sur des modèles de grande taille (Song et al., 2023). C'est justement avec cet entraînement sur une langue bien dotée et le transfert à une langue peu dotée que cette distillation des connaissances est intéressante. Cette méthode est moins utilisée que la traduction inversée (Haddow et al., 2022).

3.6 Travaux existants sur la TA vers l'allemand

Le travail de Lambrecht, Schneider et Waibel (2022) s'occupe de la traduction automatique de l'allemand standard vers des dialectes alémaniques. Lambrecht et al. (2022) ont collecté le corpus notamment à partir de Wikipédia en alémanique. Le travail traite le groupe dialectal alémanique et

pour avoir des résultats plus précis, des groupes de dialectes ont été constitués. Dans ces groupes se trouve l'alsacien (Lambrecht et al., 2022). Afin d'améliorer la traduction, plusieurs approches ont été testées et ensuite comparées. Un premier modèle de traduction de l'allemand vers le dialecte alémanique a été entrainé. Afin de traduire les textes monolingues en alémanique vers l'allemand standard, un second modèle de traduction inversée a été entrainé. Cela avait aussi pour but d'augmenter la quantité de données parallèles disponibles. À l'aide du corpus parallèle initial, ce modèle a été affiné pour améliorer les résultats (Lambrecht et al., 2022). Pour différents dialectes (l'alsacien non inclus), des modèles différents ont été entrainés. Enfin, la traduction multilingue a été exploitée (Lambrecht et al., 2022).
En raison de la petite quantité de données, le corpus a été divisé en 10% des données de test et le 90% restant également en 90% et 10% entre données d'entrainement et de validation. Pour les dialectes peu représentés, comme l'alsacien, dans le corpus parallèle, cela a fait un petit set de test (< 25 phrases) (Lambrecht et al., 2022).
Les données ont été normalisées, tokénisées, et un encodage byte paire a été appliqué (Lambrecht et al., 2022).
Les résultats ont été évalués avec BLEU (Lambrecht et al., 2022). Pour l'alsacien, la méthode de la traduction inversée a atteint le meilleur résultat avec un score de 45.8 (Lambrecht et al., 2022).

En 2022, Mohammadshahi et al. proposent SMaLL-100, un modèle de TA multilingue peu profond pour les langues à faibles ressources. 100 langues sont traitées dans le modèle. Les résultats démontrent que le modèle SMaLL-100 peut être une initialisation de modèle puissante et légère pour l'entrainement des différentes paires de langues (Mohammadshahi et al., 2022). Le modèle dépasse autres modèles multilingues en performance et en vitesse. (Mohammadshahi et al., 2022). Malheureusement, les langues traitées ne sont pas connues. Il est donc impossible à dire si ce modèle peut être bénéfique pour la traduction des dialectes alsaciens.

Dans le projet de Song et al. (2023), la question « Quelle est la solution la plus précise pour la TAN des langues à faibles ressources ? » est discutée. Le projet se concentre sur la traduction du luxembourgeois vers l'anglais. Le luxembourgeois et l'allemand standard partagent la racine langagière et quelques caractéristiques (Song et al., 2023).
Afin de trouver une réponse à la problématique de recherche, deux méthodes d'obtention des données parallèles ont été utilisées : la pseudo-traduction et la distillation des connaissances (Song et al., 2023).
Pour leur augmentation de données, la pseudo-traduction vers l'allemand et le français a été employée pour des raisons de proximité grammaire pour l'allemand et du lexique pour le français (Song et al., 2023).
Les résultats de la recherche prouvent que la meilleure technique de la traduction des langues à faibles ressources est l'utilisation de la distillation des connaissances. Elle est capable de produire des modèles bilingues de traduction de haute performance en utilisant des modèles méga-multilingues (Song et al., 2023).

4. Données et méthode

Cette partie des mémoires est la partie pratique. Les données sont introduites et expliquées. La méthodologie utilisée est également présentée, suivi par l'évaluation et ensuite les conclusions.

4.1 Corpus

Pour ce projet de recherche, deux corpus ont été utilisés. Ces corpus ont été exclusivement utilisés en tant que corpus d'évaluation.

Le premier corpus a été fourni par l'encadrante. Il s'agit d'un corpus parallèle contenant 300 phrases alsaciennes traduites en allemand et en français. Ces phrases ainsi que leurs traductions ont été récoltées de diverses sources, principalement des grammaires :

- Petit lexique français-alsacien-allemand de la METEO et de la QUALITE DE L'AIR,OLCA, https://www.olcalsace.org/fr/lexique_meteo
- Petit lexique français-alsacien-allemand de la nature, OLCA, https://www.olcalsace.org/fr/lexique_nature
- Adolf, P. (2006). Dictionnaire comparatif multilingue : Français-allemand-alsacien-anglais. Midgard.
- Jenny, A., & Richert, D. (1984). Précis pratique de grammaire alsacienne : En référence principalement au parler de Strasbourg. ISTRA.
- Zeidler, E., & Crévenat-Werner, D. (2008). Orthographe alsacienne : Bien écrire l'alsacien de Wissembourg à Ferrette. J. Do Bentzinger.

Les zones dialectales sont essentiellement le bas alémanique du nord, du sud et de Strasbourg. Quelques exemples en francique se trouvent également. Dans l'ensemble, on a constaté que des 300 phrases présentes dans les données de test, 145 ont été normalisées par ORTHAL (Zeidler & Crévenat-Werner, 2016). La normalisation de données touche notamment la graphie et l'écriture. La vaste variation dans l'orthographe dans les dialectes alsaciens, le fait que l'alsacien est une langue peu dotée et qu'une faible quantité de données est accessible, rendent la normalisation difficile.

La normalisation a été fait pour les graphies se trouvant dans les aires dialectales du bas alémanique du nord et celui du sud, du Bas-Rhin et du Haut-Rhin ainsi qu'un code géographique « AA ».

Le nom du corpus est *TestPairs* et c'est un fichier CSV.

Nous avons profité de la première partie de ce travail pour créer un deuxième corpus. Il s'agit des 21 phrases venant de la section 2.3 plus haut dans le document. Le corpus est aligné avec les phrases alsaciennes et allemandes. La source pour les phrases alsaciennes est Huck (2022), l'équivalent en allemand a été donné par moi-même. Ce corpus s'appelle *smallcorpus* et il s'agit d'un simple fichier texte.

4.2 Méthodologie

Avant de plonger dans la méthodologie, quelques aspects techniques doivent être revus. Des traductions ont été effectuées par des modèles multilingues adaptés. En ce qui concerne l'évaluation des traductions, des mesures classiques ont été utilisées. Dans la suite, les modèles ainsi que les mesures sont présentés.

4.2.1 Modèles multilingues

Pour ce projet de recherche, quatre types de modèles multilingues ont été utilisés. Ces modèles ont été appliqués directement, sans entrainement sur nos données de notre côté. Tous les modèles utilisés sont des modèles de transformers et pré-entraînés.

M2M100

M2M100-418M & M2M100-1.2B

Ce modèle est un modèle multilingue d'encodeur-décodeur. Son nom « M2M » suggère *many-to-many*, donc la traduction multilingue plusieurs (langues) à plusieurs. Il couvre 100 langues, dont l'allemand et le luxembourgeois, le danois, le norvégien, le néerlandais, l'islandais, l'afrikaans, le suédois ou l'anglais. Le modèle est capable de traduire entre 9900 directions de ces 100 langues. Ce modèle a pour objectif la génération du texte à texte (*Text2Text Generation*). Il est disponible sur la plateforme HuggingFace[32]. Les versions utilisées du modèle ont été M2M100-418M et M2M100-1.2B. (Fan et al., 2020)

SMALL-100

SMaLL-100 est la version « distillée », donc « réduite » de M2M-100. Il atteint des résultats comparables à M2M100 tout en étant plus petit et plus rapide. L'architecture et l'implémentation sont les mêmes que celui du modèle précédent, donc également un modèle transformer pré-entraîné. Les données d'entraînements, ainsi que le modèle, sont disponibles sur HuggingFace[33], il s'agit du dataset tatoeba. L'objectif principal de ce modèle est la traduction.
Ce modèle contient un tokéniseur spécifique, le SMaLL-100-tokenizer. (Mohammadshahi et al., 2022)

NLLB-200

NLLB-200-DISTILLED-600M & NLLB-200-DISTILLED-1.3B

Le modèle NLLB-200 est modèle de transformer pour la traduction. Les deux versions utilisées du modèle ont été nllb-200-distilled-600M et nllb-200-distilled-1.3B. Le modèle NLLB-200 est capable de gérer 200 langues de Flores-200 qui est un benchmark de traduction entre l'anglais et

[32] https://huggingface.co/facebook/m2m100_1.2B , https://huggingface.co/facebook/m2m100_418M
[33] https://huggingface.co/datasets/tatoeba

des langues à faibles ressources couvrant des langues germaniques. L'objectif du modèle est la traduction.
Dans ce modèle, il est possible de spécifier une langue source et une langue cible souhaitée. (Costa-jussà et al., 2022)

OPUS-MT

Toutes ces modèles proviennent du projet OPUS-MT[34]. Les modèles OPUS-MT reposent sur l'architecture de transformer. Ils sont pré-entraînés. (Tiedemann & Thottingal, 2020)

OPUS-MT-GEM-GEM

Cette version a pour but la traduction des langues germaniques aux langues germaniques. Il contient un pré-processus de normalisation et SentencePiece(spm32k, spm32k). SentencePiece est un tokéniseur de texte non-supervisée, qui implémente par ailleurs le BPE. Il est utilisé notamment pour les systèmes neuronaux de génération de texte[35]. L'intégralité du modèle est accessible sur HuggingFace[36].

OPUS-MT-GMW-GMW

Cette version a comme groupe de langue source et groupe de langue cible les langues germaniques de l'ouest. Il a un pré-processus de normalisation et SentencePiece(spm32k, spm32k). La liste complète de données de test est accessible sur HuggingFace[37].

OPUS-MT-TC-BASE-GMW-GMW

Cette version a pour but la traduction des langues germaniques de l'ouest aux langues germaniques de l'ouest. Il est disponible sur HuggingFace[38]. Les sets de test tatoeba-test-v2021-08-07, flores101-devtest, multi30k_test_2016_flickr, multi30k_test_2017_flickr, multi30k_test_2017_mscoco, multi30k_test_2018_flickr, newssyscomb2009, news-test2008, news-test2009, news-test2010, news-test2011, news-test2012, news-test2023, newstest2014-deen, newstest2015-deen, newstest2016-deen, newstest2017-deen, newstest2018-deen, newstest2019-deen newstestB2020-deen ont été utilisés pour l'évaluation du modèle après qu'il a été entraîné.

OPUS-MT-TC-BIG-GMW-GMW

Comme la version précédente, cette version a pour but la traduction des langues germaniques de l'ouest aux langues germaniques de l'ouest. Cette version précise l'architecture du transformer-big. Pour cette version, les sets de tests pour l'évaluation du modèle ont été tatoeba-test-v2020-07-28-v2021-08-07, newstest2013, newstest2015-ende. Il est disponible sur HuggingFace.[39]

[34] https://github.com/Helsinki-NLP/Opus-MT
[35] https://github.com/google/sentencepiece
[36] https://huggingface.co/Helsinki-NLP/opus-mt-gem-gem
[37] https://huggingface.co/Helsinki-NLP/opus-mt-gmw-gmw
[38] https://huggingface.co/Helsinki-NLP/opus-mt-tc-base-gmw-gmw
[39] https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-gmw-gmw

4.2.2 Métriques d'évaluation utilisées

Lors de l'évaluation, des mesures classiques calculées à l'aide de l'outil SacreBLEU (Post, 2018) ont été utilisées. Plusieurs paramètres ou signatures des mesures d'évaluation ont été appliqués. Dans la suite, ces différentes mesures sont démontrées à l'aide des tableaux explicatifs :

```
BLEU :nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1
```

Tableau 5 : Paramétrage BLEU

nrefs:1	Nombre de traduction de référence (ici une traduction de référence)
case:mixed	Minuscules et majuscules mélangées
eff:no	Optimisations d'efficacité non appliquées (prétraitement, comptage des n-grammes, gestion des données éparses, parallélisation etc.)
tok:13a	Modèle de tokéniseur (tokéniseur par défaut)
smooth:exp	Méthode de lissage : lissage exponentiel
version:2.3.1	Version de BLEU

```
BLEU flores
200 :nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2
.3.1
```

Ici, seulement le tokéniseur par défaut est remplacé par le tokéniseur flores200. Flores200 est un set de données avec plusieurs langues différentes telles que l'allemand, le luxembourgeois, le limbourgeois, le danois, le norvégien nynorsk, le norvégien bokmål, le néerlandais, le féroïen, l'islandais, l'afrikaans, le suédois et l'anglais. Ces langues sont également présentes dans le modèle de traduction multilingue NLLB-200 (Costa-jussà et al., 2022).

```
chrF++ :nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.3
.1
```

Tableau 6 : Paramétrage chrF++

nrefs:1	Nombre de traduction de référence (ici une traduction de référence)
case:mixed	Minuscules et majuscules mélangées
eff:yes	Optimisations d'efficacité appliquées
nc:6	Nombre de n-grammes de caractères utilisés dans l'évaluation (ici des 6-grammes)
nw:2	Nombre de mots dans la fenêtre contextuelle
space:no	Indique si les espaces sont traités comme des caractères dans l'évaluation (ici ce n'est pas le cas)
version:2.3.1	Version de chrF++

Ces mesures sont généralement rapportées sous forme de pourcentages (donc entre 0 et 100), représentant la proportion d'éléments correspondants (n-grammes ou n-grammes de caractères) entre la traduction candidate et la (les) traduction(s) de référence. Le pourcentage est transformé en score final représenté sous 0 et 1. Lorsqu'une traduction est considérée comme réussie, elle aura un score le plus rapprochant à 1 que possible. Le TER au contraire vise un score proche de 0.

```
TER :nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|versio
n:2.3.1
```

Tableau 7 : Paramétrage TER

nrefs:1	Nombre de traduction de référence (ici une traduction de référence)
case:lc	Minuscules (TA et traduction de référence ; évite les non-concordances dues à des différences de capitalisation
tok:tercom	Tokéniseur : tercom
norm:no	Aucune normalisation n'est appliquée (conversion des caractères, des symboles ou des représentations avant l'évaluation)
punct:yes	Signes de ponctuation sont pris en compte
asian:no	Evaluation n'est pas spécifiquement optimisée pour les langues asiatique
version:2.3.1	Version d'outil TERCOM

Il est à mentionner que lors de l'évaluation, des scores de TER très haut ont été constatés. Cela peut être éventuellement expliqué par le fait que le nombre de traduction de référence était de 1. Le score est meilleur s'il existe plusieurs traductions de référence car la mesure d'évaluation peut considérer plus qu'une traduction comme « correcte ». Le nombre de traductions de référence n'est pas plus élevé à cause d'un manque de données général.

4.2.3 Expériences

Afin de pouvoir définir une méthodologie, il a fallu se familiariser avec les données du sujet.
Dans un premier temps des traductions produites par les modèles multilingues sans adaptation aux dialectes alsaciens a été faite. Ensuite, la qualité de ces traductions a été évaluée. Ici, le corpus TestPairs a été utilisé. La traduction a été faite sans adaptation aux dialectes mais en précisant tout de même une langue source proche. Comme langue source, des langues proches et/ou germaniques ont été choisies. C'est-à-dire que la phrase alsacienne est donnée telle que en prétendant qu'il s'agit par exemple du luxembourgeois, du néerlandais, de l'afrikaans ou de l'allemand pour la faire traduire vers l'allemand et vers le français.
Les résultats de cette première évaluation ont montré que le modèle NLLB-200-distilled-1.3B, en précisant le luxembourgeois comme langue source, a été le modèle le plus performant. Les résultats sont également assez convaincants en utilisant l'allemand comme langue source. En analysant les

résultats plus en détail nous avons tenté de trouver les erreurs fréquentes ou des traductions bien réussies. Le score BLEU était d'environ 13.

Les données normalisées en ORTHAL n'obtiennent pas de meilleurs résultats que les données non-normalisées. Quelques aires dialectales où la graphie est non-normalisée ont moins de six phrases d'exemple. Un petit nombre de phrases d'exemple peut éventuellement falsifier les résultats car le hasard de générer une traduction réussie qui compterait pour l'intégralité de cette aire est trop grand.
Pour l'ensemble du modèle NLLB-200 on peut relever les erreurs fréquentes suivantes (voir Tableau 8). Chaque partie du tableau est expliqué de façon plus détaillée en dessous. Dans la colonne de « Langue », la langue source (alsacien), la langue cible (allemand standard) ainsi que deux codes de langues (Ltz_Latn et Deu_Latn) sont précisés. La phrase en alsacien correspond à une phrase de notre corpus alsacien. La phrase en allemand standard, également en italique, est la traduction attendue. Les deux codes de langues correspondent aux langues sources qui ont été précisées en utilisant le modèle multilingue NLLB-200. Ici, Ltz_Latn correspond donc au luxembourgeois et Deu_Latn à l'allemand. La langue cible indiquée au modèle est, indépendamment de la langue source, toujours l'allemand.

Tableau 8 - Exemples d'erreur du modèle NLLB-200-distilled-1.3B sans adaptation

Langue / Numérotage	**Example**
Example 1	
Source : Alsacien	‘s ìsch kenna ku.
Cible : Allemand standard	Niemand ist gekommen.
Ltz_Latn (luxembourgeois)	Ich weiß, was ich tun kann.
Deu_Latn	- Ich weiß es.
Example 2	
Source : Alsacien	Dàs ìsch a Gschank fer dr Seppi.
Cible : Allemand standard	Das ist ein Geschenk für Seppi.
Ltz_Latn (luxembourgeois)	Das ist ein Schank für Dr. Seppi.
Deu_Latn	Das ist ein Schrei von Dr. Seppi.
Example 3	
Source : Alsacien	Dès Lokàl ísch bekànnt.
Cible : Allemand standard	Dieses Lokal ist bekannt.
Ltz_Latn (luxembourgeois)	Das Lokal ist ein Eisfest.
Deu_Latn	Der Lokal ist bekannt.
Example 4	
Source : Alsacien	Ìhr gadde besser nix sàge !
Cible : Allemand standard	Ihr würdet besser nichts sagen !
Ltz_Latn (luxembourgeois)	Sie sollten nichts sagen!
Deu_Latn	Sie sollten besser nichts sagen!

Une approche morphologique partielle est souvent observée si le mot est inconnu. Dans l’exemple 1, le mot « *kenna* » (personne) devrait être traduit par « *keine(r/n/m)/niemand* » mais les verbes « *kennen/können* » (connaître/pouvoir) sont retenus ce qui fait qu’aucune phrase est correcte.
L’exemple 2 montre que l’article « *Dr* » (le) n'est pas toujours reconnu comme article masculin mais comme abréviation docteur ou monsieur. Quand la phrase source dit « Dàs ìsch a Gschank fer dr Seppi » *(C’est/Ceci est un cadeau pour Seppi)*, le modèle ne comprend pas l’article masculin comme un article défini. De manière générale, et comme l’indique la traduction attendue, on ne retrouve pas l’article devant les noms propres en allemand standard. Certes, il s’agit d’une spécificité de l’alsacien mais dans ce contexte, ce « dr » ne fait en aucun cas référence à une abréviation.
Les pronoms « *diese, dieses, dieser, diesen* » (celui, celle) sont également problématiques car ils sont confondus avec l’article neutre, comme on peut constater avec l’exemple 3.
Le dernier exemple montre qu’il existe des problèmes avec les pronoms personnels, notamment « *euch/ihr* » (vous pluriel) et « *Sie/Ihnen* » (vous formel) et la déclinaison des cas. En allemand standard, cette distinction entre le vous formel et le vous pluriel ou aussi souvent informel, est importante car elle peut causer un glissement sémantique et par conséquence changer le contexte.

Ces erreurs sont exemplaires pour la totalité du corpus *TestPairs*. Naturellement, on peut relever davantage de fautes dans le Tableau 8 - Exemples d'erreur du modèle NLLB-200-distilled-1.3B sans adaptation. Ici, il s'agit des erreurs fréquentes qui pourtant ne se manifestent pas de manière que l'on puisse les corriger directement.
Hors des erreurs, on a pu repérer une chose qui a très bien fonctionné. Il s'agit du changement de verbe lorsque « *tuen* » (faire) est présent dans l'exemple alsacien. Comme expliqué dans la section 2.3 plus haut dans le document, ce verbe n'est pas utilisé comme auxiliaire pour dire faire en allemand standard. Le modèle réussit bien à supprimer ce verbe et à utiliser l'autre verbe en forme conjuguée.

En nous basant sur ces résultats d'analyse, une méthodologie a été développée afin d'obtenir des scores plus élevés. Puisque l'alsacien en prétendant qu'il s'agit du luxembourgeois et de l'allemand a pu être traduit vers l'allemand standard, nous avons choisi de transformer les corpus afin de le rapprocher de ces deux langues sources. L'hypothèse est que si un corpus (alsacien) peut être rapproché vers les deux langues sources (luxembourgeois et allemand) pré-entraînées dans le modèle multilingue NLLB-200-distilled-1.3B, la traduction sera plus réussie. Cela aurait comme avantage de pouvoir utiliser le modèle multilingue directement sans devoir l'affiner (*fine tuning*) avec des données d'entraînement et de validation, ce qui est problématique à cause du manque de données et ressources informatiques.
Pour transformer les corpus, cinq méthodes de transformation ont été développées. Dans la suite, ces méthodes sont présentées.

Transformation du corpus

Transformation basee sur vocabulaire

Une des transformations appliquées est la transformation basée sur vocabulaire, en anglais ***Vocabulary based transformation***.

Le script pour cette transformation a été développé par moi-même avec le soutien de l'encadrante. Le code pour cette transformation est disponible sur un dépôt en ligne.[40] Il prend en compte des lexiques bilingues (alsacien-français) afin de traduire les mots se trouvant dans les lexiques dans la langue cible. Le reste de la phrase reste en alsacien. Les lexiques ont des sources diverses comme le LOD (*LOD - Lëtzebuerger Online Dictionnaire - LOD*, s. d.), le Dictionnaire comparatif multilingue, des lexiques spécifique à des domaines de l'OLCA (Office pour la Langue et la Culture d'Alsace)[41], de-ElsassischWeb_diktionnair, le corpus annoté RESTAURE, les dictionnaires ACPA, Freelang, Wiktionary, Alsa Immer et des lexiques de mots de classes fermées (Bernhard & Ligozat, 2013).

Le lexique construit avec toutes les ressources citées plus haut contient dans l'ensemble 5799 entrées. Le tableau suivant démontre la répartition en parties du discours.

Tableau 9 - Répartition partie du discours

Partie du discours	*Nombre*
Adjectif	777
Adverbe	108
Article	16
Conjonction	17
Interjection	3
Non connu	1
Nombre	50
Nom propre	56
Particule	6
Préposition	32
Pronom	94
Substantif	3774
Verbe	865
Total	5799

Si plusieurs traductions existent, elles sont comparées et la plus proche est choisie. Cela a été fait à l'aide de la librairie python difflib et sa classe SequenceMatcher[42].

[40] https://git.unistra.fr/wurps/ta_alsacien_allemand_m2
[41] https://www.olcalsace.org/
[42] https://docs.python.org/3/library/difflib.html

TRANSFORMATION BASEE SUR DES REGLES

La deuxième transformation se base sur des règles de transformation, également appelée ***Rule based transformation***.

Un script qui gère la transformation *Rule based transformation* a été développé avec le soutien de l'encadrante. Le code pour cette transformation est disponible sur un dépôt en ligne.[43] Il prend en compte deux fichiers de règles de transformation de l'alsacien vers l'allemand et de l'alsacien vers le luxembourgeois. La base pour ces règles de transformation provient du lexique utilisé dans la transformation *Vocabulary based transformation*. Un exemple de cette transformation est par exemple que si un mot en alsacien dans le lexique bilingue commence par *Àl*, le mot correspondant en allemand commence par *Al* dans un certain nombre de cas. Cela fait alors que le mot *Àllergie* en alsacien devient *Allergie* en allemand (en français allergie). Même si la chance de générer une version fautive est moindre, elle existe car cette transformation ne prend pas en compte une vérification linguistique mais est basée sur la répétition et sa probabilité.

Pour un rapprochement vers l'allemand, 7537 règles ont été extraites. Pour le rapprochement vers le luxembourgeois, 10386 règles ont été extraites.

TRANSFORMATION BASEE SUR LES CARACTERES DIACRITIQUES

Une autre transformation est celle basée sur les caractères diacritiques ou ***Unaccented transformation***.

Cette transformation consiste à supprimer tous les accents qui n'existent pas en allemand ou en luxembourgeois. Les caractères diacritiques en allemand sont *ä*, *ö* et *ü* en majuscule et minuscule. Des éventuels trémas sur les voyelles *e* et *i* ainsi que les accents aigu et grave sur n'importe quelle voyelle seront enlevés.

Pour le luxembourgeois, la transformation suit le même principe mais adapté pour cette langue. Cela inclut donc le tréma sur le *ë* puisqu'il existe en luxembourgeois.

Cette transformation a été introduite par l'encadrante et affiné par moi-même.

TRANSFORMATION BASEE SUR LES MOTS DE CLASSE OUVERTE

Deux transformations ont été faite en impliquant les classes grammaticales.

La première, la transformation basée sur les mots de classe ouverte, aussi appelée ***Open class transposition*** implique la classe ouverte.

Les classes ouvertes accueillent facilement des nouveaux mots. Elles comprennent les noms, les verbes, les adjectifs et les adverbes. Elles se remarquent par l'acceptation des néologismes.

Pour cette transformation, un lexique de conversion de l'alsacien vers l'allemand de formes appartenant à ces classes a été utilisé. Ce lexique est produit par (Bernhard, 2014, 2021). Dans le cas où un mot alsacien a plusieurs traductions possibles, seule la plus fréquente est conservée (fréquence dans le corpus deu_news_2022_1M 1 (Goldhahn et al., 2012)). Ce lexique final comporte 6699 paires de mots alsacien-allemand.

Cette transformation a été développée par l'encadrante.

[43] https://git.unistra.fr/wurps/ta_alsacien_allemand_m2

L'idée avec cette transformation ainsi que la suivante était, contrairement aux *Rule based transformation* et *Vocabulary based transformation,* d'étudier plus précisément si le type de la classe de mots avait un impact ou non.

TRANSFORMATION BASEE SUR LES MOTS DE CLASSE FERMEE

La ***Closed class transposition*** en anglais ou la transformation basée sur les mots de classe fermée a été développé par l'encadrante.

Il s'agit d'une transformation à l'aide d'un lexique de conversion de l'alsacien vers l'allemand de formes appartenant aux classes fermées. Les pronoms, les prépositions, les conjonctions, les articles ainsi que les auxiliaires forment cette classe. Il s'agit des éléments de discours qui ont un nombre fixe de mots et n'acceptent pas de nouvelles formes.

Le lexique a été créé par (Bernhard & Ligozat, 2013) sans modification. Il contient 133 entrées et a été constitué par étude d'un petit corpus de cinq textes.

Dans le tableau suivant, toutes les transformations pour un rapprochement vers l'allemand sont démontrées. Les phrases en italique sont en allemand standard et donc la traduction attendue. Les modifications apportées par les transformations sont en gras.

Tableau 10 - Exemples des transformations

Phrase/transformation	Exemple 1	Exemple 2
Alsacien	Àm Dunnerschdàà ìsch Märrikdàà.	D'r Àppetit kommt mìt 'em Ësse.
Allemand standard	*Am Donnerstag ist Markttag. (Donnerstags ist Markttag.)*	*Der Appetit kommt mit dem Essen.*
Rule based	**um** Dunnerschdàà ìsch Märrikdàà.	D'r **A**ppetit kommt mìt 'em Ëssen.
Vocabulary based	**Am** Dunnerschdàà ìsche Märrikdàà.	Der Àppetit kommt **mit** 'em **essen**.
Unaccented	**A**m Dunnerschd**aa i**sch Märrikd**aa**.	D'r **A**ppetit kommt m**i**t 'em **E**sse.
Open class	Àm **Donnerstag** ìsch Märrikdàà.	D'r Àppetit kommt mìt 'em Essen.
Closed class	**Am** Dunnerschdàà **ist** Märrikdàà.	D'r Àppetit kommt mit 'em Ësse.

Ces cinq méthodes de transformation ont été appliquées au corpus *TestPairs*. Sur le *smallcorpus*, seules les transformations *Vocabulary based transformation* et *Rule based transformation* ont été appliquées. Les corpus modifiés ont ensuite été traduits à nouveau. Le modèle multilingue le plus performant ayant été déterminé auparavant, le modèle NLLB, a été utilisé pour la traduction des corpus modifiés. Nous avons également ajouté un nouveau modèle pour ce cycle de traduction : le modèle ChatGPT 3.5, version : 25 septembre 2023.

L'inclusion de ce modèle dans notre recherche a été inspirée par les travaux de (Robinson et al., 2023).
ChatGPT a été utilisé pour les deux corpus mais seulement pour traduire les corpus modifiés par les transformations *Rule based transformation* et *Vocabulary based transformation*. L'accès au modèle ChatGPT s'est fait par l'interface utilisateur. Pour chaque corpus transformé, un nouveau chat a été initialisé afin de maintenir les mêmes conditions préalables. Comme le petit corpus ne contient que 21 phrases, nous voulions conserver les mêmes conditions pour le corpus plus important. Par conséquent, ce corpus a été traduit par lots de 30 phrases. Les lots ont été saisis dans la même conversation. Les différences entre les conversations n'ont été faites que pour la langue et le corpus. Le prompt utilisé était le suivant, formulé en anglais : *[This is an Alsatian to German translation, please provide the German translation for these sentences: Please provide the translation for the following sentence. Do not provide any explanations or text apart from the translation.]*[44].Ce prompt a été choisi pour plusieurs raisons. Il s'agit du même prompt utilisé dans les travaux de Robinson et al. (2023). En plus, l'anglais permet de ne pas biaiser le modèle. Le français aurait pu être également possible, l'anglais a été choisi pour son caractère international.
Après la traduction, la qualité a été évaluée de la même manière qu'auparavant en utilisant les mêmes mesures.
Les résultats de la deuxième évaluation sont expliqués plus en détail dans la section 5.
Pour atteindre cet objectif, la prochaine étape de notre méthode a consisté à combiner les méthodes de transformation proposées. Les transformations *Unaccented transformation* et *Vocabulary based transformation*, par exemple, ont obtenu de bons résultats. Comme la méthode de transformation *Unaccented transformation* est facile à inclure dans nos algorithmes développés, nous l'avons incluse de façon constante dans les méthodes de transformation. Il est à mentionner qu'il y a des différences entre l'accentuation en allemand et en luxembourgeois. Notamment, la lettre e peut comporter un tréma (ë) en luxembourgeois. Les trémas en allemand existent seulement sur les lettres a, o et u. Ces différences ont été prises en compte dans la transformation combinée.
Dans les algorithmes, la transformation *Unaccented transformation* s'appliquait seulement aux tokens ne pouvant pas être transformés ou traduits.
Ces nouvelles transformations combinées ont été appliquées aux deux corpus alsaciens. Ensuite, la même procédure que précédemment a été suivie : traduction par NLLB et ChatGPT suivi par l'évaluation par les métriques pour une nouvelle analyse des résultats.

44 Traduction libre : *Il s'agit d'une traduction de l'alsacien vers l'allemand, fournis la traduction allemande des phrases suivantes. Ne fournis pas d'explications ou de texte en dehors de la traduction*

5. Evaluation

Dans cette partie, nous allons examiner les résultats de manière plus approfondie.

Le premier tableau (Tableau 11 - Evaluation de modèle sans adaptation dialectale) démontre la toute première évaluation des modèles de traduction sans adaptation dialectale pour le corpus *TestPairs*. Nous retrouvons les quatre métriques d'évaluation ainsi qu'une colonne « Langues ». Cette colonne précise la langue source ayant été donné aux modèles multilingues, si cela était possible.
Cette évaluation a été faite avec plusieurs langues sources germaniques, le tableau montre seulement une sélection pour les langues allemande et luxembourgeoise. Le tableau complet se trouve dans la partie annexes (III) du document.
Cette évaluation se base sur les traductions de *TestPairs*. Le meilleur résultat pour chaque colonne pour le luxembourgeois est en gras, celui pour l'allemand souligné. Les résultats du modèle ChatGPT ont été ajouté dans l'après afin d'avoir une baseline avec ce modèle. La baseline pour le corpus *smallcorpus* est également ajouté.

Tableau 11 - Evaluation de modèle sans adaptation dialectale

Modèle	*Langues*	*BLEU*	*spBLEU*	*chrF++*	*TER*
small100	als_de	1,50	1,67	17,18	230,93
m2m100-1.2B	de	5,63	6,90	27,94	88,47
m2m100-418M	de	3,75	4,36	26,05	89,49
NLLB-200-distilled-1.3B	ltz_Latn	22,21	24,73	42,47	71,05
NLLB-200-distilled-1.3B	deu_Latn	15,37	18,40	36,66	76,28
NLLB-200-distilled-1.3B/small	ltz_Latn	13,05	12,14	33,02	76,92
NLLB-200-distilled-1.3B/small	deu_Latn	5,57	5,67	25,85	85,90
NLLB-200-distilled-600M	ltz_Latn	13,36	15,56	34,22	81,41
NLLB-200-distilled-600M	deu_Latn	10,99	13,69	31,74	86,39
NLLB-200-distilled-600M/small	ltz_Latn	6,10	4,82	27,76	80,77
NLLB-200-distilled-600M/small	deu_Latn	7,81	5,95	25,29	82,05
opus-mt-gem-gem	als_de	4,47	7,14	28,31	86,08
opus-mt-gmw-gmw	als_de	1,53	3,30	23,48	115,54
opus-mt-tc-base-gmw-gmw	als_de	2,00	4,28	24,35	104,88
opus-mt-tc-big-gmw-gmw	als_de	2,09	5,01	24,06	91,82
ChatGPT/TestPairs		65,04	65,13	76,16	25,61
ChatGPT/smallcorpus		42,73	47,50	70,05	40,74

Comme on peut voir, c'est le modèle NLLB-200-distilled avec ses deux versions qui atteint les meilleurs résultats, en précisant le luxembourgeois (ltz_Latn) et l'allemand (deu_Latn) en tant que

langue source. La figure suivante illustre ce tableau également en prenant en compte les deux métriques BLEU.

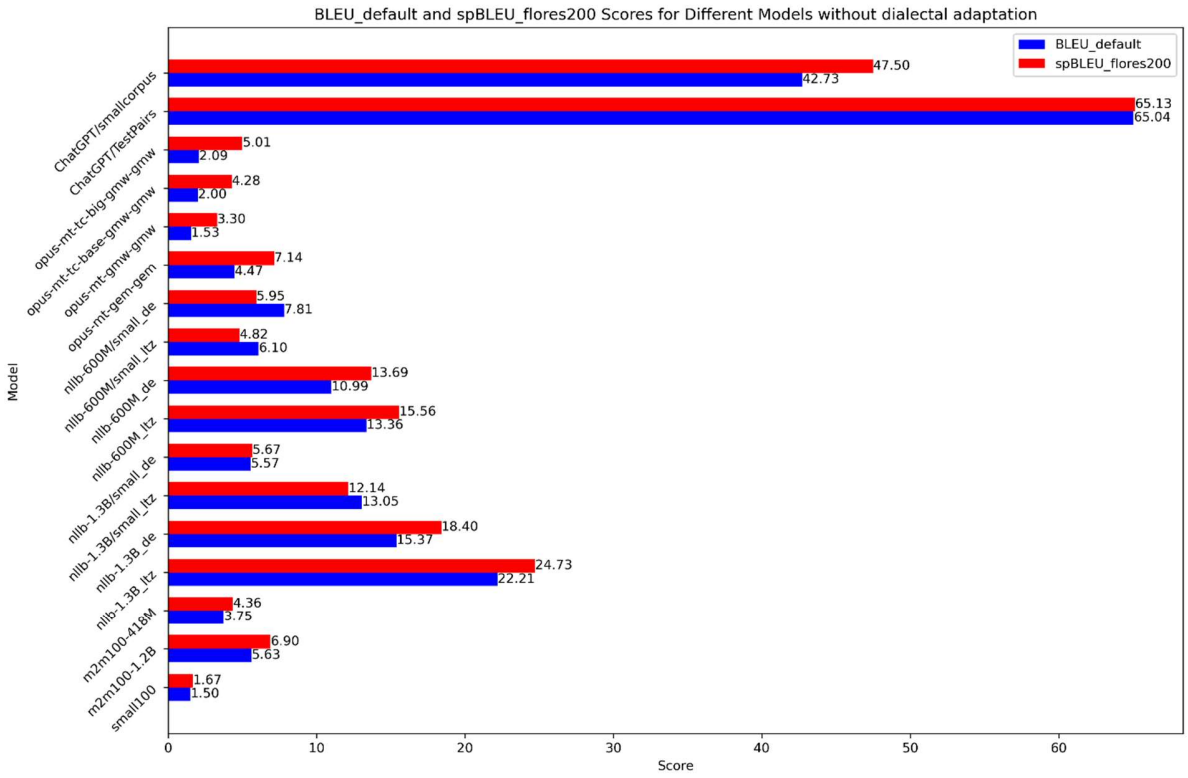

Figure 3 - Evaluation des modèles sans adaptation dialectale

Sur la base de cette performance des modèles, le modèle NLLB-200-distilled(-1.3B et -600M) a été choisi pour faire traduire les corpus transformés. Encore une fois, le modèle ChatGPT n'était pas encore introduit à ce moment de la prise de décision. Cette illustration sert aujourd'hui à démontrer les baselines.

La suite (Tableau 12 - NLLB-200-distilled-1.3B transformation simple vers l'allemand, Tableau 13 - NLLB-200-distilled-1.3B transformation simple vers le luxembourgeois Tableau 14 - NLLB-200-distilled-600M transformation simple vers l'allemand Tableau 15 - NLLB-200-distilled-600M transformation simple vers le luxembourgeois et Tableau 16 - ChatGPT transformation simple) montre les résultats de l'évaluation de la traduction après une transformation de corpus.
Afin de gagner de la place dans les tableaux, les transformations sont légèrement abréviées. Le même compte pour le *smallcorpus* qui s'appelle dans la suite seulement *small*. La colonne « BLEU » représente la valeur de la métrique BLEU avec le tokéniseur par défaut. « spBLEU » est la colonne de la métrique BLEU avec le tokéniseur flores200. Le meilleur résultat pour le corpus

TestPairs est mis en gras, celui pour le corpus *smallcorpus* est souligné. Les métriques montrent les résultats globaux. Le code de langue *ltz* représente un rapprochement vers le luxembourgeois. Respectivement, le code de langue *deu* représente le rapprochement vers l'allemand.
Les tableaux suivants démontrent les résultats globaux selon une transformation, un corpus et une métrique. Le modèle ayant été utilisé pour ces résultats est précisé dans la référence. Afin de mieux comparer les résultats obtenus sans transformation avec ceux suivi d'une transformation, la ligne « Baseline » est ajoutée.
Les tableaux suivants montrent les résultats de la simple transformation avec le modèle NLLB-200-distilled-1.3B pour un rapprochement vers l'allemand (Tableau 12) et un rapprochement vers le luxembourgeois (Tableau 13 - NLLB-200-distilled-1.3B transformation simple vers le luxembourgeois).

Tableau 12 - NLLB-200-distilled-1.3B transformation simple vers l'allemand

Métrique//transformation/corpus	*BLEU*	*spBLEU*	*chrF++*	*TER*
Baseline/TestPairs (deu_Latn)	15,37	18,40	36,66	76,28
Baseline/small (deu_Latn)	5,57	5,67	25,85	85,90
closed_class/TestPairs (deu_Latn)	23,47	25,45	44,48	66,12
closed_class/small (deu_Latn)	19,72	18,28	38,69	67,95
vocab_based_deu/TestPairs	19,87	22,42	40,97	73,24
vocab_based_deu/small	4,30	6,29	27,26	87,18
open_class/TestPairs (deu_Latn)	20,32	23,20	41,45	70,85
open_class/small (deu_Latn)	7,97	8,74	27,81	80,13
rule_based_deu/TestPairs	13,38	16,30	34,47	79,74
rule_based_deu/small	12,35	12,63	30,16	78,85
unaccented/TestPairs (deu_Latn)	24,64	26,57	44,88	66,07
unaccented/small (deu_Latn)	8,87	10,58	30,24	78,85

On peut constater avec ce premier tableau que la méthodologie du rapprochement des dialectes alsaciens vers une autre langue proche a été couronné de succès. Pour le corpus *TestPairs*, toutes les transformations sauf *Rule based transformation* ont pu atteindre un meilleur score. La transformation qui a obtenu le meilleur score est *Unaccented transformation*.
Le *smallcorpus* quant à lui a également obtenu des scores plus élevés. Notamment la transformation *Closed class transposition* a beaucoup amélioré la traduction. La transformation *Vocabulary based transformation* n'a pas atteint un meilleur score de la métrique BLEU avec le tokéniseur par défaut mais avec le tokéniseur flores200.
On peut remarquer que le corpus plus large *TestPairs* atteint des meilleurs résultats en global. Cela pourrait être expliqué avec la différence de taille entre les corpus. Un corpus plus large aura plus de possibilités d'obtenir des bonnes traductions et sera par conséquent moins puni pour des mauvaises traductions qu'un corpus de petite taille.
Le graphique suivant illustre le tableau de la simple transformation vers l'allemand.

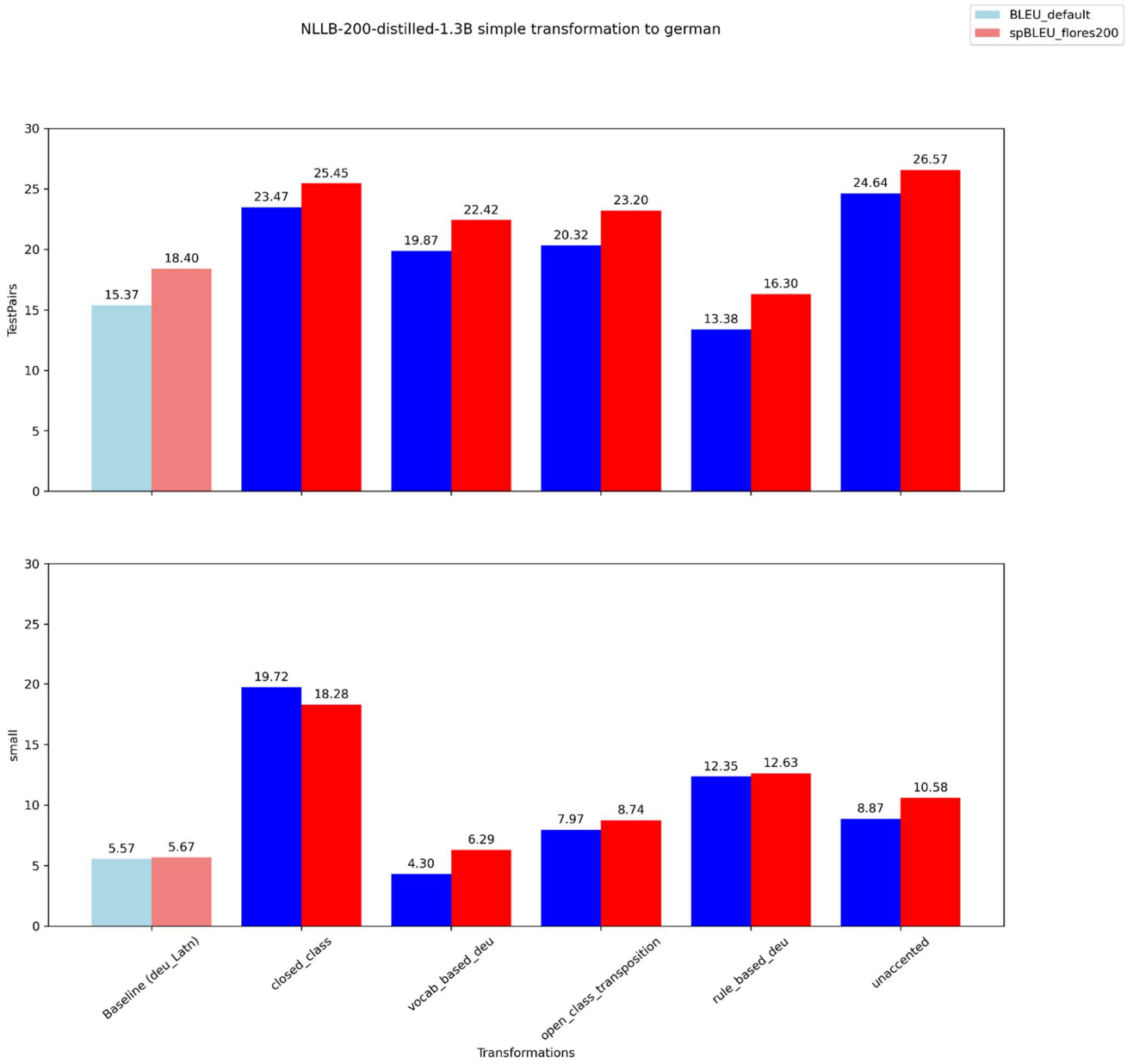

Figure 4 - NLLB-200-distilled-1.3B transformation simple vers l'allemand

Le modèle NLLB-200-distilled-1.3B n'a pas seulement traduit les corpus après une simple transformation vers l'allemand.
Le tableau avec les résultats du rapprochement vers le luxembourgeois suit.

Tableau 13 - NLLB-200-distilled-1.3B transformation simple vers le luxembourgeois

Métrique/transformation/corpus	*BLEU*	*spBLEU*	*chrF++*	*TER*
Baseline/TestPairs (ltz_Latn)	22,21	24,73	42,47	71,05
Baseline/small(ltz_Latn)	13,05	12,14	33,02	76,92
closed_class/TestPairs (ltz_Latn)	30,03	31,48	49,99	60,89
closed_class/small (ltz_Latn)	30,75	27,67	46,07	58,33
vocab_based_ltz/TestPairs	23,79	26,78	46,08	67,65
vocab_based_ltz/small	14,75	15,68	34,87	82,05
open_class/TestPairs (ltz_Latn)	25,23	28,32	46,58	65,21
open_class/small (ltz_Latn)	9,01	8,52	30,35	81,41
rule_based_ltz/TestPairs	17,35	20,28	38,93	75,42
rule_based_ltz/small	17,49	18,92	36,30	80,77
unaccented/TestPairs (ltz_Latn)	28,81	31,00	49,50	60,59
unaccented/small (ltz_Latn)	24,98	24,07	40,51	69,87

Aussi les transformations vers le luxembourgeois ont également permis d'obtenir de meilleurs scores d'évaluation par rapport à la baseline.
Ici, la transformation *Closed class transposition* atteint les meilleurs scores pour les deux corpus.
La figure suivante montre les résultats de ce tableau sous forme graphique. Elle montre les score BLEU, avec le tokéniseur par défaut et le tokéniseur flores200.

Même si les traductions obtenues avec ce modèle ont permis d'obtenir des meilleurs scores dans les deux langues de rapprochement testées, il faut aussi rappeler qu'un score BLEU de 30 n'est pas élevé pour une traduction automatique.

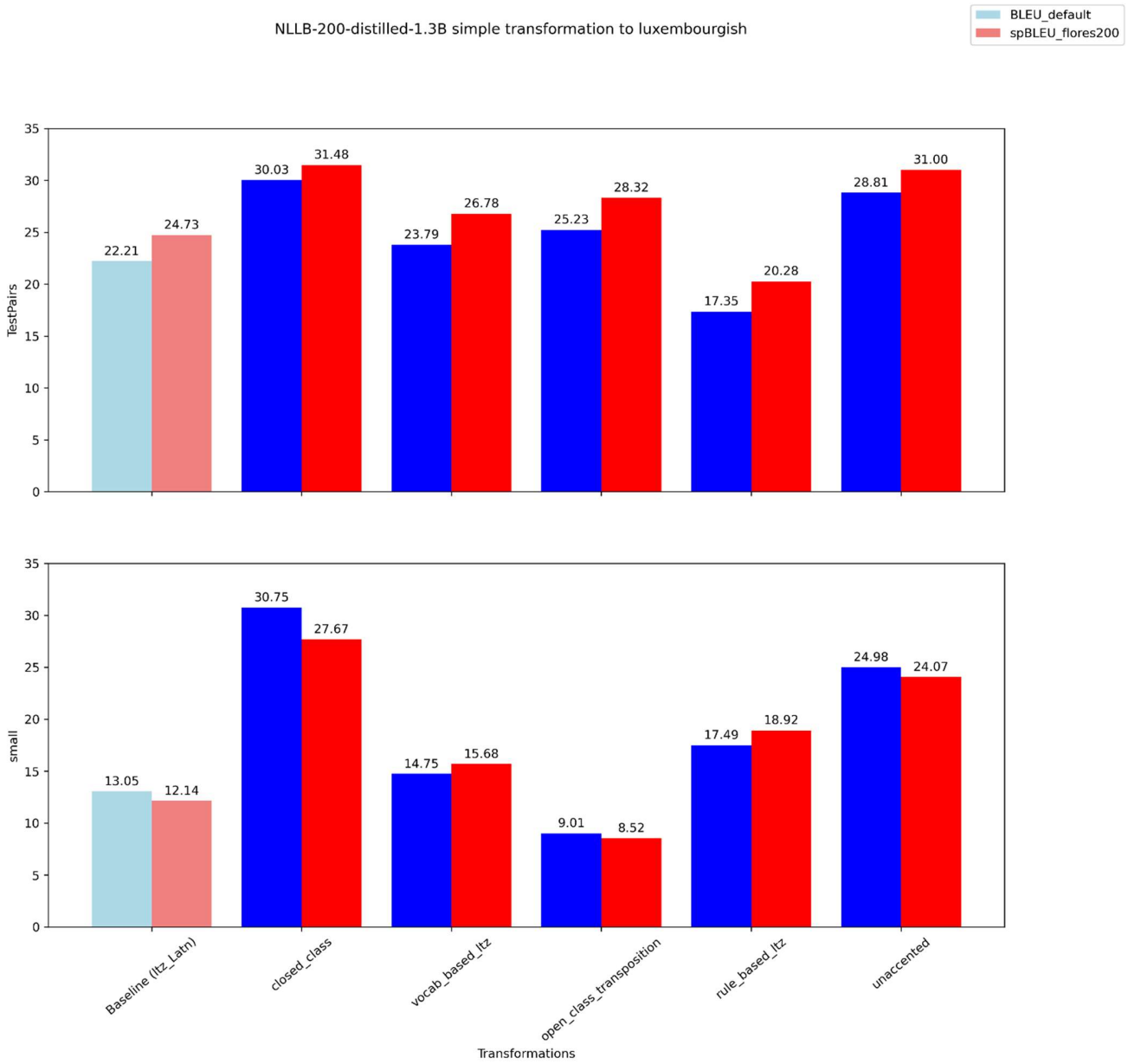

Figure 5 – NLLB-200-distilled-1.3B transformation simple vers le luxembourgeois

Sous les mêmes conditions, la version 600M du modèle NLLB-200-distilled a généré des traductions qui ont été évaluées. Les résultats sont montrés sous forme de tableau et forme graphique dans la suite.

Tableau 14 - NLLB-200-distilled-600M transformation simple vers l'allemand

Métrique/transformation/corpus	*BLEU*	*spBLEU*	*chrF++*	*TER*
Baseline/TestPairs (deu_Latn)	10,99	13,69	31,74	86,39
Baseline/small (deu_Latn)	7,81	5,95	25,29	82,05
closed_class/TestPairs (deu_Latn)	18,26	19,70	39,66	71,20
closed_class/small (deu_Latn)	15,49	15,26	32,79	71,79
vocab_based_deu/TestPairs	11,20	12,23	33,10	93,30
vocab_based_deu/small	3,79	6,07	26,88	98,08
open_class/TestPairs (deu_Latn)	14,72	17,03	36,18	77,50
open_class/small (deu_Latn)	7,33	5,62	26,21	81,41
rule_based_deu/TestPairs	8,89	9,68	30,01	88,01
rule_based_deu/small	8,58	9,29	25,87	98,08
unaccented/TestPairs (deu_Latn)	18,47	19,85	40,86	72,07
unaccented/small (deu_Latn)	12,29	12,90	30,11	76,92

Comme le tableau avant, celui-ci montre les résultats obtenus par le modèle multilingue NLLB-200-distilled, de version 600M sur les deux corpus après une transformation.
Cette version du modèle a une plus grande variation entre les scores obtenus que la version précédente, où les scores après la simple transformation sont relativement proches l'un à l'autre.

Néanmoins, on peut également constater ici, que la transformation vers l'allemand aide à obtenir des meilleurs scores. C'est encore une fois la méthode *Unaccented transformation* qui atteint les meilleurs scores pour le grand corpus *TestPairs*. Pour le corpus plus petit *smallcorpus* c'est également, comme avec la version plus large du modèle, la *Closed class transposition* qui obtient les meilleurs scores.

La figure suivante traduit le tableau en graphique :

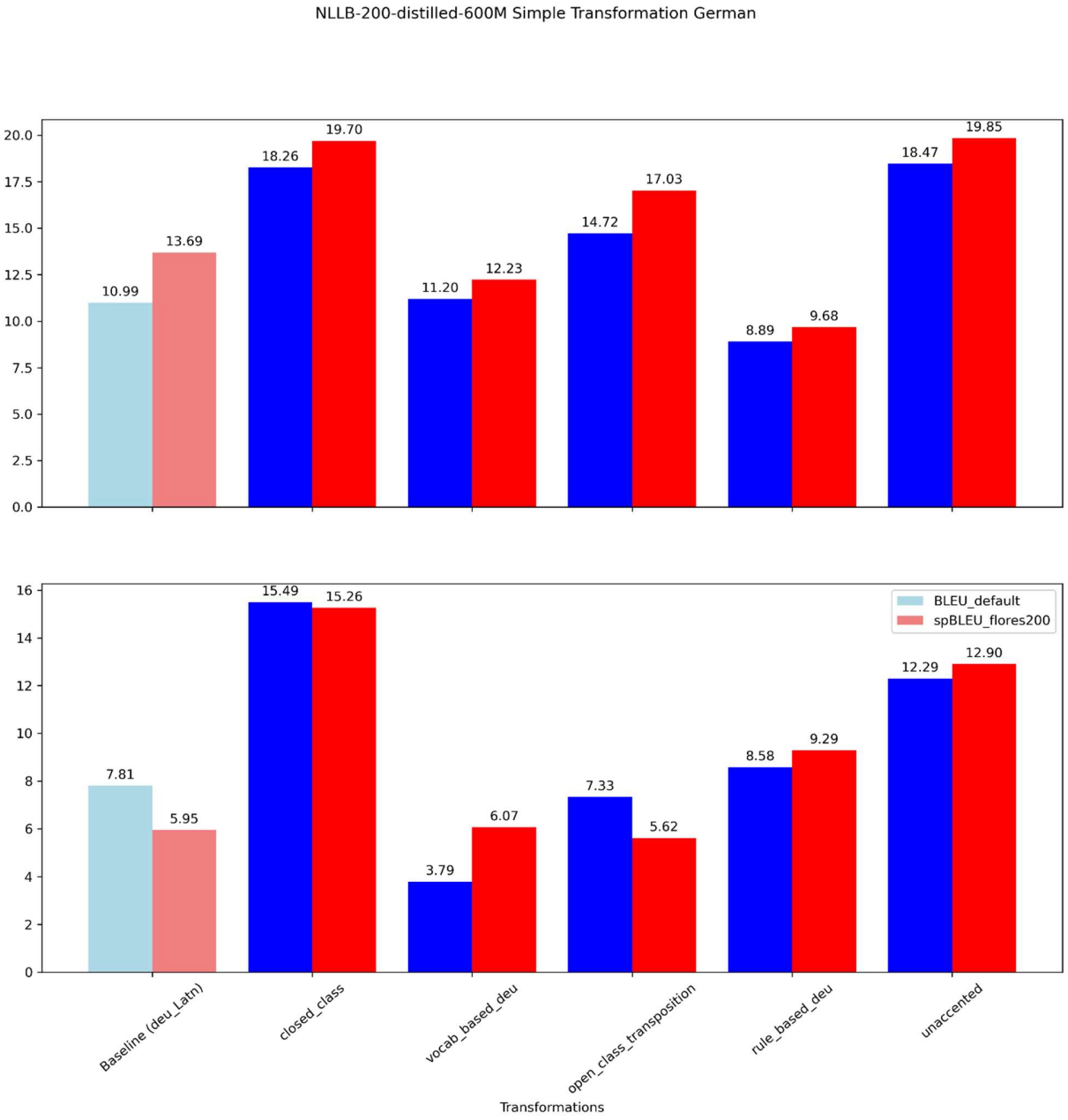

Figure 6 - NLLB-200-distilled-600M transformation simple vers l'allemand

Comme avec l'autre modèle, cette version-ci a également traduit les corpus après un rapprochement vers le luxembourgeois. Les résultats sont démontrés dans la suite.

Tableau 15 - NLLB-200-distilled-600M transformation simple vers le luxembourgeois

Métrique/transformation/corpus	*BLEU*	*spBLEU*	*chrF++*	*TER*
Baseline/TestPairs (ltz_Latn)	13,36	15,56	34,22	81,41
Baseline/small(ltz_Latn)	6,10	4,82	27,76	80,77
closed_class/TestPairs (ltz_Latn)	22,33	23,47	43,44	66,33
closed_class/small (ltz_Latn)	22,01	20,09	39,72	62,18
vocab_based_ltz/TestPairs	15,00	17,03	37,00	77,55
vocab_based_ltz/small	4,93	8,10	27,69	96,79
open_class/TestPairs (ltz_Latn)	16,72	19,38	38,71	75,88
open_class/small (ltz_Latn)	8,25	7,97	26,12	87,82
rule_based_ltz/TestPairs	7,68	10,70	30,37	85,88
rule_based_ltz/small	8,68	7,81	28,27	83,97
unaccented/TestPairs (ltz_Latn)	24,36	25,55	45,29	65,26
unaccented/small (ltz_Latn)	14,10	13,46	32,04	78,21

Pour un rapprochement vers le luxembourgeois, la transformation *Unaccented transformation* atteint le meilleur score pour le corpus *TestPairs*. Pour le corpus *smallcorpus* c'est la transformation *Closed class transposition*.
Avec ce modèle, la transformation *Rule based transformation* n'atteint pas de meilleurs scores que la baseline pour le *TestPairs* corpus. La *Vocabulary based transformation* dépasse la baseline à peine.
Pour le *smallcorpus*, la transformation *Closed class transposition* dépasse la baseline, les autres transformations obtiennent des scores plus élevés mais pas aussi élevés.
En général, on peut dire que les scores atteints ne sont pas élevés.
Ces résultats se retrouvent également dans la figure suivante de façon illustrée.

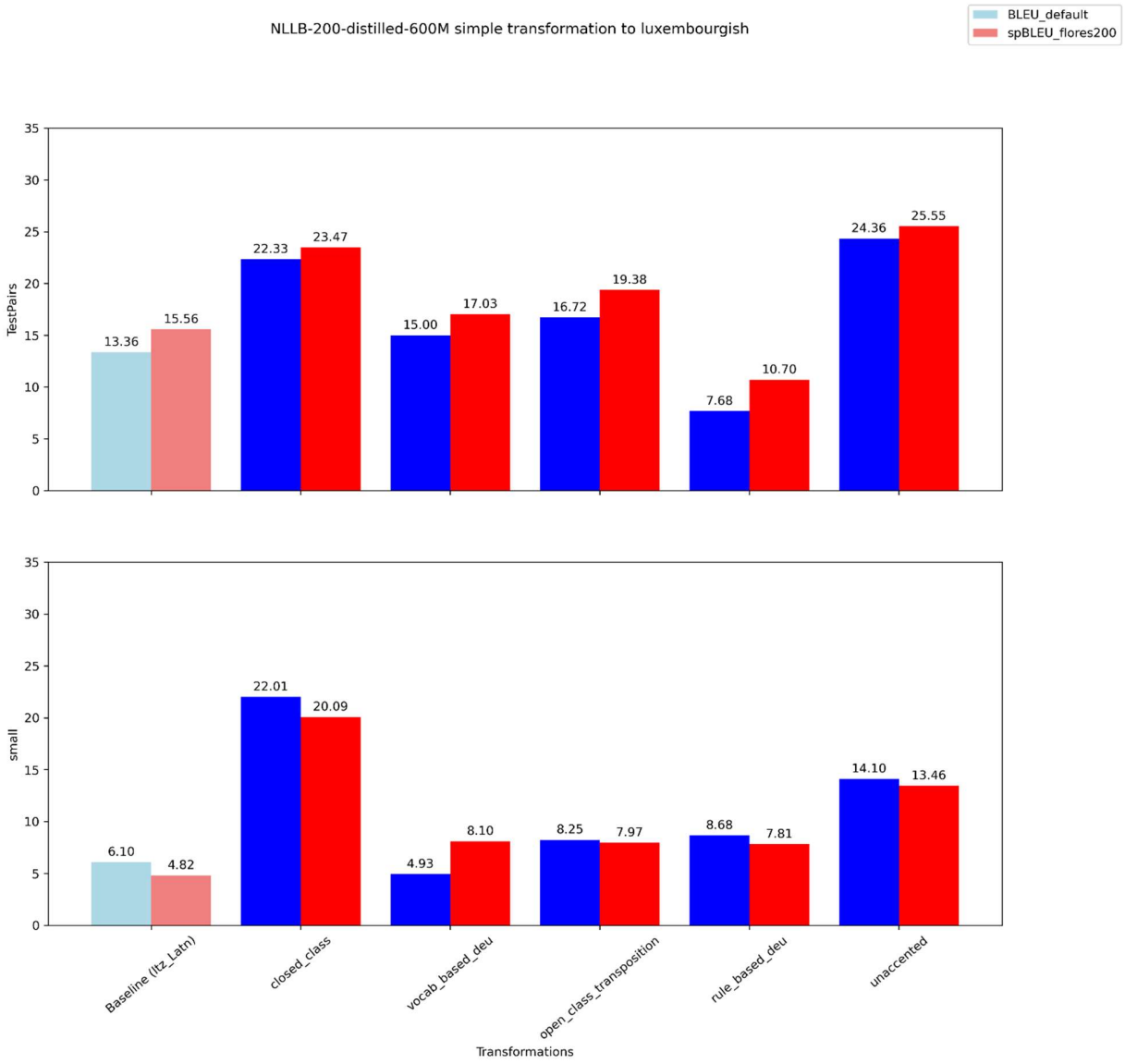

Figure 7 - NLLB-200-distilled-600M transformation simple vers le luxembourgeois

Dernièrement, le modèle ChatGPT a également produit des traductions qui ont été évaluées. Contrairement au modèle NLLB, celui-ci a seulement traduit des corpus où les transformations *Vocabulary based transformation* et *Rule based transformation* ont été appliquées. Une baseline pour les deux corpus sans adaptation dialectale a été créée.

Tableau 16 - ChatGPT transformation simple

Métrique/transformation/corpus	*BLEU*	*spBLEU*	*chrF++*	*TER*
Baseline/TestPairs	65,04	65,13	76,16	25,61
Baseline/smallcorpus	42,73	47,50	70,05	40,74
vocab_based_deu/TestPairs	50,75	53,97	71,75	39,82
vocab_based_deu/small	46,53	49,29	63,75	44,51
vocab_based_ltz/TestPairs	50,50	51,80	71,45	40,43
vocab_based_ltz/small	36,54	33,92	52,69	63,41
rule_based_deu/TestPairs	53,09	53,80	72,13	37,33
rule_based_deu/small	52,88	57,83	70,55	35,37
rule_based_ltz/TestPairs	54,03	55,48	71,86	36,47
rule_based_ltz/small	40,30	42,79	57,74	50,61

Ce tableau démontre les résultats moyens obtenu par le *large language model* ChatGPT pour la transformation simple, c'est-à-dire qu'une seule transformation appliquée. On peut constater que ces résultats dépassent largement ceux du modèle multilingue. Pour les deux corpus, la transformation *Rule based transformation* obtient les meilleurs scores. Cependant, il ne s'agit pas de la même langue de rapprochement. Pour le corpus *TestPairs*, un rapprochement vers le luxembourgeois et pour le corpus *small* un rapprochement vers l'allemand fonctionne mieux.

Il est à noter ici qu'aucune transformation a un impact positif sur le score par rapport à la baseline de *TestPairs*. Pour le *smallcorpus* le rapprochement vers l'allemand fonctionne mieux. Avec les deux transformations vers l'allemand la baseline est dépassée alors que cela n'est pas le cas avec un rapprochement vers le luxembourgeois.
La figure suivante montre également ces résultats.

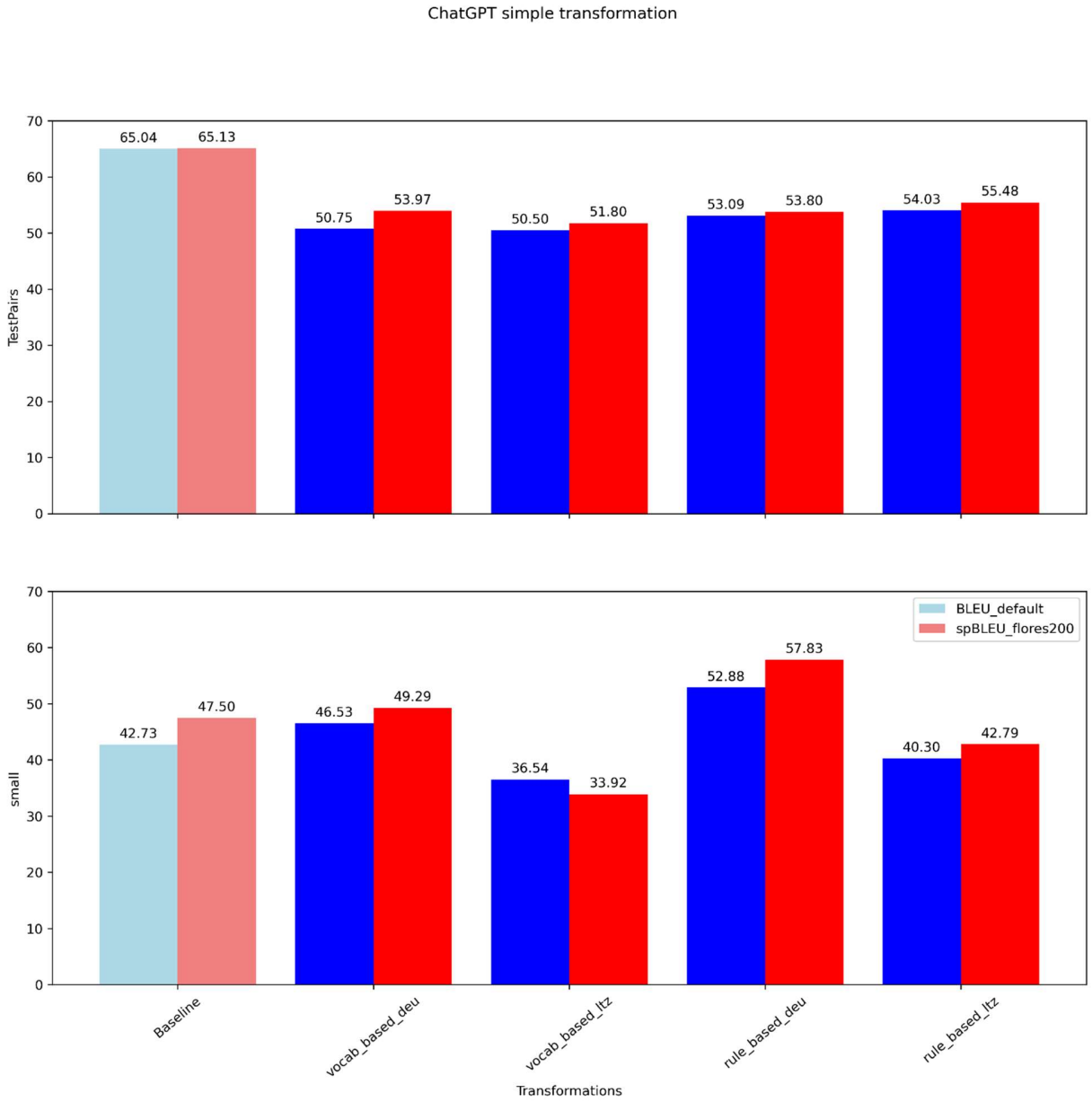

Figure 8 - ChatGPT transformation simple

Ces premiers résultats montrent clairement que le modèle ChatGPT atteint des scores plus élevés et qu'il s'agit donc d'un instrument plus élaboré. Par rapport à la version NLLB-200-distilled-1.3B, la version NLLB-200-distilled-600M atteint des scores moins bons, indépendamment de la transformation ou du corpus.
Cependant, nous souhaitons encore améliorer notre méthode en raison des limites de ChatGPT, expliquées ci-dessous.

Dans la suite, les résultats de la deuxième évaluation avec les corpus transformés par transformations combinées sont démontrés. La structure des tableaux est la même que celle dessus. Pour rappel, la transformation combinée consiste à combiner la transformation *Unaccented transformation* avec les transformations *Vocabulary based transformation* et *Rule based transformation*. Les même métriques et modèles sont gardés.

Tableau 17 – NLLB-200-distilled-1.3B transformation combinée

Métrique/transformation/corpus	*BLEU*	*spBLEU*	*chrF++*	*TER*
Baseline/TestPairs (deu_Latn)	15,37	18,40	36,66	76,28
Baseline/small (deu_Latn)	5,57	5,67	25,85	85,90
Baseline/TestPairs (ltz_Latn)	22,21	24,73	42,47	71,05
Baseline/small (ltz_Latn)	13,05	12,14	33,02	76,92
vocab_based_deu/TestPairs (deu_Latn)	20,09	23,33	42,87	71,61
vocab_based_deu/small (deu_Latn)	15,03	17,25	35,95	79,88
vocab_based_ltz/TestPairs (ltz_Latn)	26,82	29,71	49,10	62,06
vocab_based_ltz/small (ltz_Latn)	22,44	24,00	38,65	73,78
rule_based_deu/TestPairs (deu_Latn)	20,11	22,33	42,66	73,13
rule_based_deu/small (deu_Latn)	16,09	18,54	34,02	80,49
rule_based_ltz/TestPairs (ltz_Latn)	20,58	22,94	43,23	69,88
rule_based_ltz/small (ltz_Latn)	18,01	20,62	36,37	78,66

Ce tableau montre les résultats de la transformation combinée pour le modèle NLLB-200-distilled de version 1.3B. Il atteint des scores moins bons qu'après une simple transformation. Comme avant, le corpus *TestPairs* obtient de meilleurs résultats que le *smallcorpus.* Pour les deux corpus un rapprochement vers le luxembourgeois fonctionne le mieux avec la *Vocabulary based transformation*.
En général le modèle perd en termes de performance par rapport à la simple transformation.

La figure suivante illustre ce tableau.

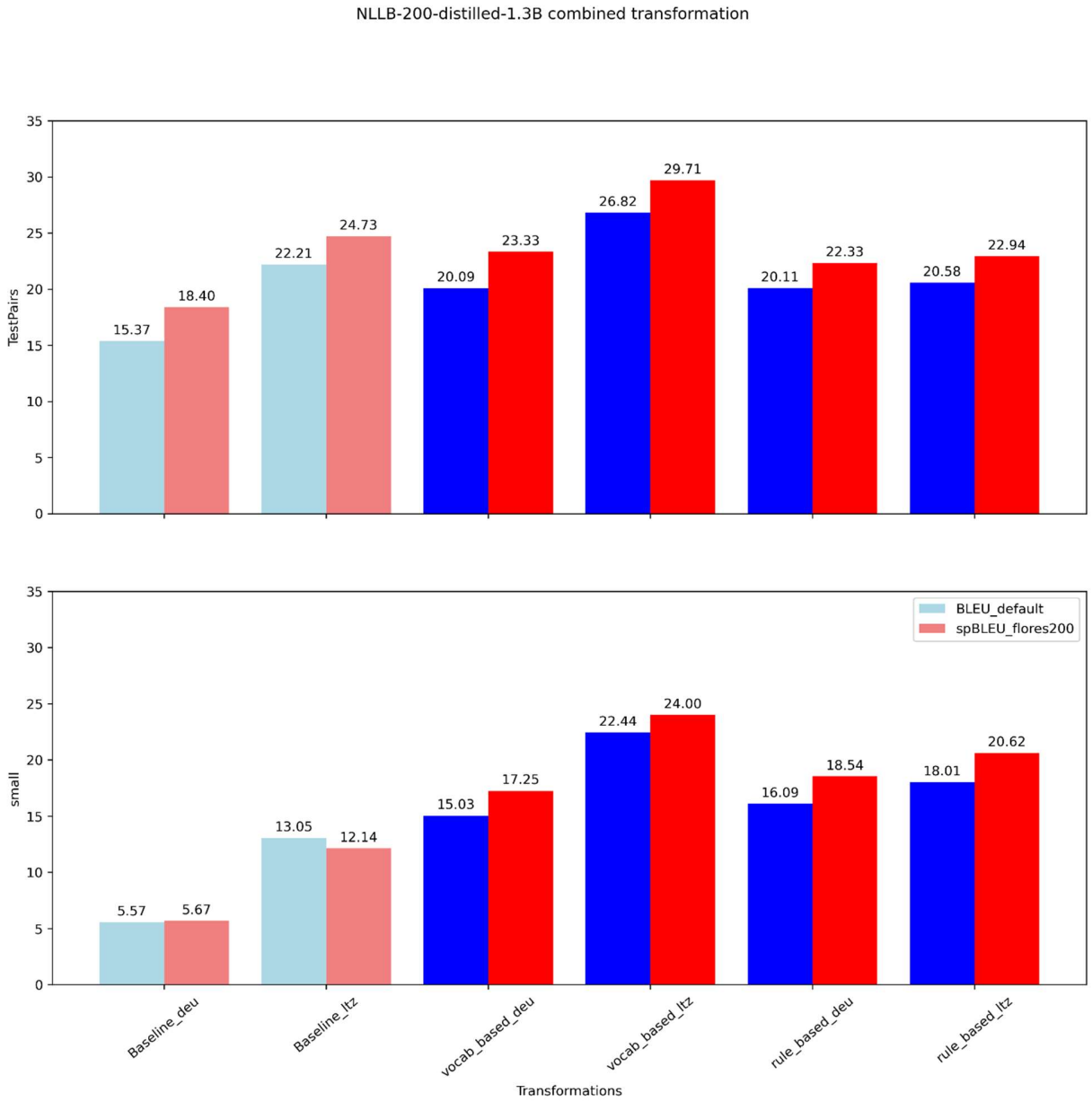

Figure 9 – NLLB-200-distilled-1.3B transformation combinée

Sous les mêmes conditions, la version 600M du modèle NLLB-200-distilled a généré des traductions qui ont été évaluées. Les résultats sont montrés sous forme de tableau et forme graphique dans la suite.

Tableau 18 - NLLB-200-distilled-600M transformation combinée

Métrique/transformation/corpus	*BLEU*	*spBLEU*	*chrF++*	*TER*
Baseline/TestPairs (deu_Latn)	10,99	13,69	31,74	86,39
Baseline/small (deu_Latn)	7,81	5,95	25,29	82,05
Baseline/TestPairs (ltz_Latn)	13,36	15,56	34,22	81,41
Baseline/small(ltz_Latn)	6,10	4,82	27,76	80,77
vocab_based_deu/TestPairs	16,54	18,79	40,27	73,74
vocab_based_deu/small	8,89	11,88	32,42	89,63
vocab_based_ltz/TestPairs	18,23	21,10	42,43	70,19
Vocab_based_ltz/small	23,30	23,22	40,23	76,83
rule_based_deu/TestPairs	15,00	16,66	37,83	77,96
rule_based_deu/small	9,43	8,94	25,43	101,83
rule_based_ltz/TestPairs	14,06	16,30	37,95	75,57
rule_based_ltz/small	10,55	10,01	29,07	84,76

En comparant les résultats obtenus par ce modèle multilingue NLLB-200-distilled-600M après la transformation combinée avec les résultats obtenus lors de la transformation simple, on peut constater que le modèle continue à améliorer sa performance. Encore une fois, comme avec la version plus large du modèle, le rapprochement vers le luxembourgeois obtient de meilleurs scores.

La transformation combinée *Vocabulary based transformation* vers le luxembourgeois se détache du tableau pour le *smallcorpus*. Les autres transformations sur les deux corpus sont relativement pareilles.

Ce comportement est illustré par le graphique suivant.

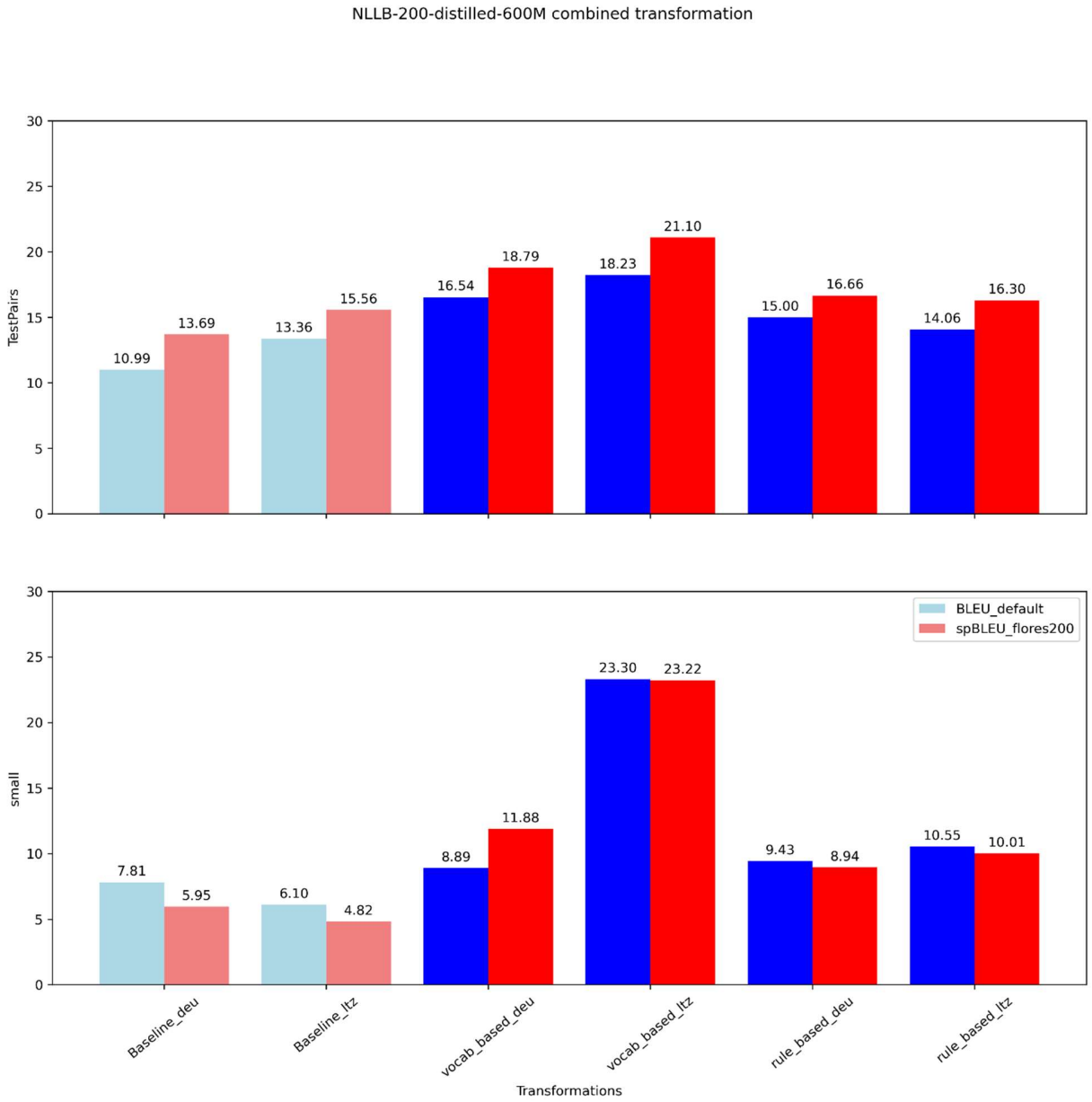

Figure 10 - NLLB-200-distilled-600M transformation combinée

Finalement, le tableau et la figure suivante démontrent les résultats du modèle ChatGPT avec l'évaluation des traductions à la suite des transformations combinées.

Tableau 19 - ChatGPT transformation combinée

Métrique/transformation/corpus	*BLEU*	*spBLEU*	*chrF++*	*TER*
Baseline/TestPairs	65,04	65,13	76,16	25,61
Baseline/smallcorpus	42,73	47,50	70,05	40,74
rule_based_deu/TestPairs	53,64	55,45	72,16	38,55
rule_based_deu/small	62,28	62,87	73,48	32,32
rule_based_ltz/TestPairs	54,85	56,56	72,70	36,47
rule_based_ltz/small	43,82	45,12	60,26	47,56
vocab_based_deu/TestPairs	56,84	58,74	74,73	35,35
vocab_based_deu/small	45,84	50,66	64,93	43,90
vocab_based_ltz/TestPairs	58,94	59,79	76,21	32,50
Vocab_based_ltz/small	54,98	53,30	64,58	34,15

Le modèle ChatGPT peut garder une aussi bonne performance de manière générale, comme l'indique le tableau ci-dessus montrant ses résultats après la transformation combinée. Ici, la transformation combinée *Rule based transformation* vers l'allemand atteint les meilleurs résultats pour le *smallcorpus.* Pour le corpus *TestPairs*, c'est la transformation combinée *Vocabulary based transformation* pour le luxembourgeois qui atteint les meilleurs résultats.
Il est intéressant de noter que le modèle ChatGPT garde une meilleure baseline, donc une traduction sans adaptation dialectale, pour le grand corpus *TestPairs* pendant qu'une amélioration de résultats pour le petit corpus *smallcorpus* est à constater. Ici, la baseline est à 42,73, le meilleur résultat de la transformation simple à 52,88 et pour la transformation combinée il est à 62,28.

Un début d'explication serait de dire que ChatGPT s'améliore avec son prompt. Comme il y avait plusieurs entrées de 30 phrases chacune, un contexte plus large a été proposé au modèle. On peut penser que les derniers batches ont été traduits avec beaucoup de succès, alors que le *smallcorpus* n'avait qu'un batch de pas tout à fait 30 phrases et que le modèle n'a donc pas eu assez de contexte. Ici, contexte ne porte pas une notion sémantique car sémantiquement, la plupart des phrases sont indépendantes les unes aux autres dans les deux corpus.
Cela n'explique cependant pas pourquoi les transformations n'obtiennent pas des meilleurs scores que la baseline avec *TestPairs* alors que la méthode de prompting a été la même.

Le graphique suivant démontre ces résultats.

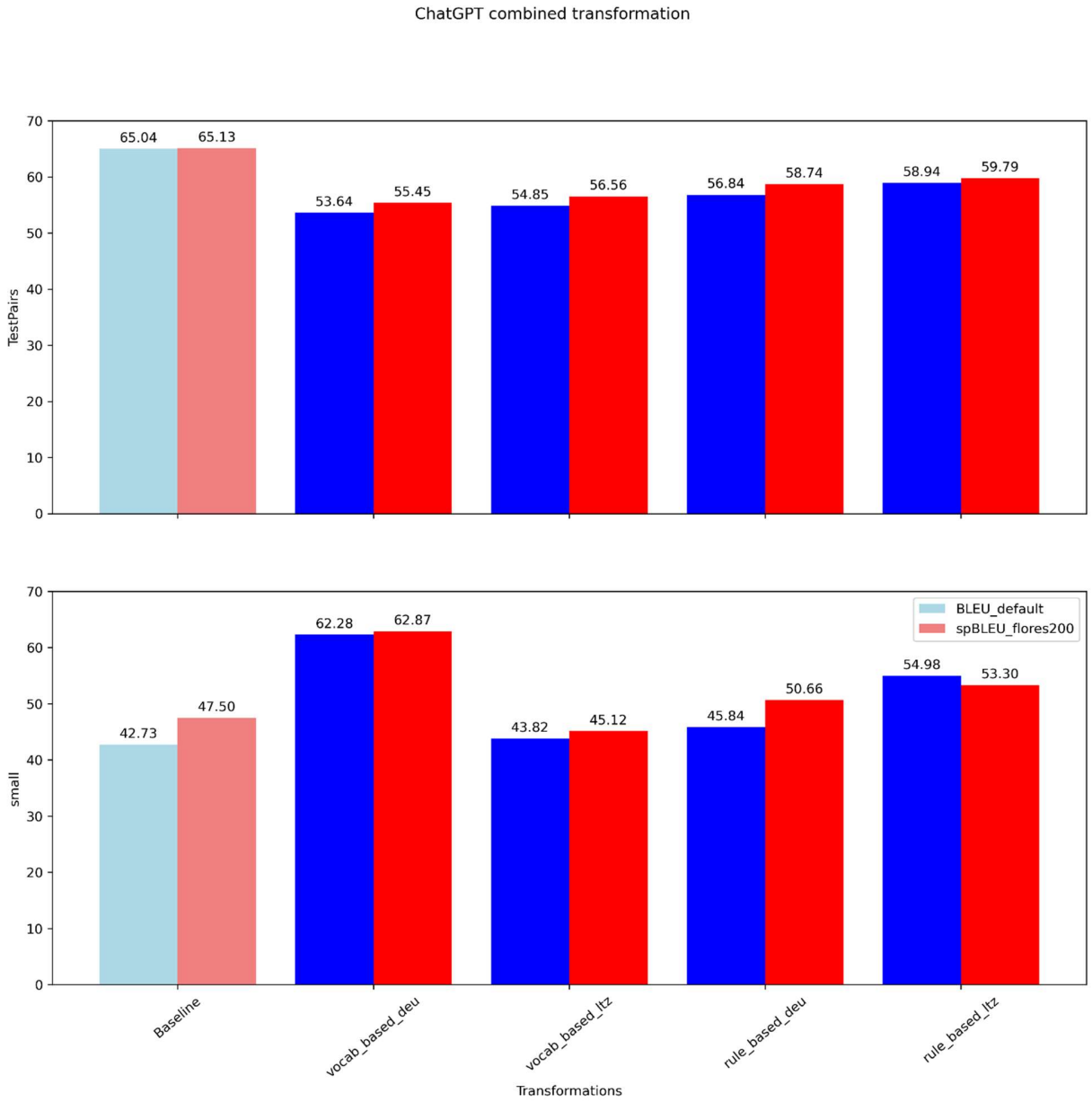

Figure 11 - ChatGPT transformation combinée

De manière générale on peut dire que le modèle ChatGPT obtient de meilleurs résultats globaux que le modèle multilingue NLLB-200-distilled. Alors que ChatGPT n'atteint pas non plus des scores parfaits, ils sont bien plus élevés que ceux de NLLB.
Le modèle multilingue gère mieux les corpus ayant été transformés qu'une seule fois. L'idée de combiner des manières de transformer un corpus pour le rapprocher vers une langue ne s'est pas révélée fructueuse.

En termes de scores obtenus, il est indéniable que le modèle ChatGPT dépasse les modèles multilingues. Un rapprochement vers une autre langue ne semble pas nécessaire quand les traductions peuvent se faire en plusieurs prompts. Cette hypothèse est cependant à vérifier. Nos expériences ont montré que la transformation combinée a été bénéfique pour le *smallcorpus*.
Cependant, nous avons pu constater une meilleure amélioration en termes de performances chez les modèles multilingues après une transformation. Le fait que les transformations combinées n'ont pas amélioré davantage la performance ou même conduit à une perte de performance n'a pas été attendu. Il serait envisageable de faire une étude plus précise pour comprendre à quoi cela est dû.

De manière générale, on peut constater que même si les transformations ont obtenu des scores plus élevés que l'utilisation des modèles multilingues sans adaptation dialectale, il est sans doute possible de les améliorer davantage.
Pour une telle amélioration la longueur des mots, une différenciation plus fine des régions ainsi que les verbes seront utiles à prendre en compte.

Notamment la longueur de mot se reflète dans les erreurs restantes qui sont démontrées dans la suite.

5.1 Erreurs restantes

Malgré les bonnes performances du modèle ChatGPT, on voit qu'il n'est pas parfait non plus. Nous allons essayer de trouver des erreurs restantes et éventuellement comment elles peuvent être améliorées.
Le tableau ci-dessous (Tableau 20 – Erreurs de traduction de ChatGPT) montre une partie des traductions faites par ChatGPT. Elles proviennent du *smallcorpus* ayant été rapproché avec la transformation combinée *Vocabulary based transformation* vers l'allemand. Il ne s'agit pas des seules erreurs produites par ce modèle, ce tableau sert à la démonstration des fautes du modèle. Afin de mieux comprendre l'évolution de la traduction, la phrase source en alsacien est fournie, la phrase après la transformation combinée, la traduction fournie par ChatGPT ainsi que la traduction attendue ou la phrase correcte en allemand standard. Dans la colonne « Transformation combinée » des mots en rouge soulignent des erreurs produites pendant les transformations, qui vont avoir un impact sur la traduction. Les mots en gras dans la colonne « Traduction ChatGPT » soulignent des erreurs venant du côté de ChatGPT. Une traduction en français est également fournie.

Dans la suite du tableau, les exemples sont commentés et analysés d'une façon plus détaillée. On peut constater des erreurs générales ainsi que spécifiques.
Il n'est pas évident pourquoi ChatGPT ajoute parfois des guillemets et parfois non. En plus, les guillemets ne correspondent pas au standard allemand. La forme est correcte mais les guillemets ouvrants se trouvent en bas et les guillemets fermants en haut alors que ChatGPT positionne les deux guillemets en haut.
Une autre erreur générale qui se produit lors des transformations est la gestion du verbe auxiliaire « het » (avoir), ici conjugué à la troisième personne du singulier. Ce verbe est faussement transformé en « heute » (aujourd'hui) ce qui change complètement le sens de la phrase. Ce problème arrive spécialement avec la transformation *Vocabulary based transformation* parce que le lexique pour cette transformation ne contient pas des verbes mais liste « het » comme adverbe avec la traduction *aujourd'hui* et son pendant allemand. Le lexique qui est utilisé pour la transformation *Closed class transposition* liste cette forme comme verbe.

Tableau 20 – Erreurs de traduction de ChatGPT

Exemple	Alsacien	Transformation combinée	Traduction ChatGPT	Allemand standard	Français[45]
1	S Kind het zue sinere Màmme welle renne.	sie Kind heute zu seiner Mamme wollen rennen.	Sie **möchten** heute zu **ihrer** Mutter rennen.	Das Kind hat zu seiner Mutter rennen wollen.	L'enfant a voulu courir chez sa mère.
2	Ich weiss, àss r uns het bsueche welle.	ich weiß , ass r uns heute bsueche wollen.	"Ich weiß, dass **sie** uns heute besuchen wollen."	Ich weiβ, dass er uns hat besuchen wollen	Je sais qu'il a voulu nous rendre visite.
3	Er het furtgehn welle.	ihr heute weggehen wollen.	Ihr **möchtet** heute weggehen.	Er hat fortgehen wollen.	Il a voulu partir.
4	Wie er schun e Wil fortgànge gsi isch, isch m ebs inkomme.	wohin ihr schun ein weil fortgange gsi ich , ich m ebs inkomme.	"**Wohin ihr** schon eine Weile fortgegangen seid, ich kann es **nicht** kommen."	Als er schon eine Weile fortgewesen war, ist es ihm eingefallen.	Après qu'il était déjà parti un certain temps, quelque chose lui est revenu
5	D Kinder hàn gspielt ghet.	du Kinder haben gspielt ghet.	Die Kinder haben gespielt **gegangen**.	Die Kinder hatten gespielt.	Les enfants avaient joué.

[45] Traduction libre par moi-même

Dans l'exemple 1, on peut constater tout d'abord que la lettre « S » n'a pas été transformée en article neutre « Das » comme il aurait dû mais en pronom personnel de la troisième personne du singulier (« sie », elle). ChatGPT a changé le verbe modal « wollen » (vouloir) en un autre verbe modal « möchten » (souhaiter) qui signifie la même chose mais qui est un peu plus soutenu et poli. « Möchten » exprime un souhait alors que « wollen » exprime ce souhait d'une façon un peu plus renforcée. Dans le contexte de la phrase, il serait préférable de maintenir « wollen » car c'est plus approprié pour un enfant. Le mot « Kind » est omis dans la traduction. ChatGPT change également la phrase initiale du singulier au pluriel.

Dans le deuxième exemple, la lettre « r » qui est le pronom personnel masculin de la troisième personne au singulier a été changé en pronom personnel féminin du pluriel. Ce changement n'est pas compréhensible considérant que le pronom personnel « er » (il) est plus proche de la simple lettre *r* que du pronom personnel du pluriel « sie » (elles).
La phrase du troisième exemple après la transformation combinée ressemble à un allemand approximatif. Le modèle arrive bien à traduire cet allemand approximatif en allemand standard mais qui n'est absolument pas en accord sémantiquement avec la phrase source qui veut dire *Il a voulu partir* alors que la traduction donne *Vous voulez sortir aujourd'hui.*
Il est tout de même à noter qu'ici les deux erreurs qui mènent à cette phrase fausse viennent du processus de la transformation où l'auxiliaire avoir a été transformé en « aujourd'hui » et le pronom personnel *il* en *vous* alors qu'il a déjà la bonne forme. Cela ne justifie pas pourquoi ChatGPT change le verbe modal comme dans le premier exemple. Dans le sens sémantique initial de la phrase alsacienne, il ne serait pas possible ou logique d'avoir la forme polie de « vouloir ».

Dans l'exemple 4 on peut constater la même erreur lors des transformations que dans l'exemple 3 : le pronom personnel masculin de la troisième personne du singulier au pronom personnel de la deuxième personne du pluriel. Le mot « Wie » (comme/comment) a été également changé à « wohin » (où). Dernièrement, le verbe *être* conjugué en troisième personne du singulier « isch » a été changé au pronom personnel de la première personne du singulier « ich ». Ce changement est spécifique à la transformation *Vocabulary based transformation*.
ChatGPT quant à lui commet aussi des erreurs. Dans la subordonnée, le modèle omet la simple lettre « m », qui signifie « ihm » (lui) dans la phrase source, mais ajoute une négation qui n'a pas lieu. ChatGPT ne respecte pas la syntaxe dans la proposition principale. En allemand, les mots interrogatifs sont généralement suivis immédiatement par un verbe. Il est possible que « wohin » ne fonctionne pas comme un interrogatif mais un adverbe relatif dans cette phrase. Dans ce cas, il serait possible d'avoir un pronom en deuxième position mais cette structure de phrase n'est pas très naturelle. Une fois de plus il existe une légère contradiction de sens entre « wohin » (où) et le verbe « fortgehen » (partir, quitter).

Dans le dernier exemple du tableau, on peut encore une fois constater une erreur lors de la transformation. La lettre « D » qui est ici l'article pluriel en alsacien est transformé en « du » (tu). ChatGPT retransforme ce pronom personnel avec succès en l'article pluriel.

Or le verbe n'est pas bien conjugué. Le plus-que-parfait se conjugue en allemand avec l'auxiliaire avoir ou être et un verbe en participe passé. Pour cette phrase, on aurait pu accepter une traduction de « Die Kinder haben gespielt gehabt » (littéralement « Les enfants ont avaient joué »), qui est acceptable en allemand dialectal, mais la bonne forme du plus-que-parfait aurait dû employer le prétérit de l'auxiliaire « avoir », donc « hatten ». La grande erreur que commet le modèle ici est que le verbe « aller » au participe passé est employé. Comme en français, aller se construit avec l'auxiliaire « être ». L'emploi de ce verbe ici est alors complètement faux.

De manière générale on peut dire que ChatGPT génère un allemand standard et moins parlé (wollen – möchten) mais qui n'arrive pas à toujours à saisir le sens ou contexte pour savoir quel verbe modal employer. Même si le modèle fournit des traductions avec des fautes, il est clair que la plupart des erreurs arrive pendant le processus de transformation et que ChatGPT continue par conséquence les fautes.

6. Conclusions

Dans ce travail, nous avons examiné la traduction automatique des dialectes alsaciens vers l'allemand standard. Tout d'abord, la situation socio-historique de l'Alsace a été expliquée et un inventaire de la langue alsacienne avec les variations dialectales a été présenté.
Ensuite, un bref historique de la traduction automatique a été dessiné. Des différentes approches existantes dans la TA ont été expliquées aussi. Enfin, les métriques d'évaluation ainsi que des défis dans l'intelligence artificielle en lien avec des dialectes, mais spécialement avec les dialectes alsaciens, ont été exposés.
Dans la deuxième partie de ce travail, les aspects techniques sont exposés. Les corpus de ce projet de recherche sont détaillés et les modèles utilisés pour la traduction sont exposés.
Ensuite, la méthodologie est présentée.
Comme expliqué dans la section sur la méthodologie (4.2), pour bien mener ce projet, une stratégie de travail a dû être adaptée à l'avancement et notamment aux résultats du projet. Ces résultats sont démontrés par la suite de façon exemplaire avec la métrique BLEU.
La stratégie initiale était de passer le corpus initial aux modèles multilingues sans aucune adaptation pour voir comment ces modèles multilingues arrivent à traduire une langue sur laquelle ils n'étaient pas entraînés. Les résultats n'ont pas été satisfaisant avec un score BLEU assez bas.
Au vu que les modèles multilingues ne sont pas tout à fait adaptés au corpus car ils traitent beaucoup des langues germaniques mais pas l'alsacien, la première évaluation « brut » a tout de même permis d'identifier le modèle multilingue le plus puissant pour ce projet ainsi que les langues choisies. Pour ce projet, il s'agit donc du modèle NLLB-200 avec les versions nllb-200-distilled-600M et nllb-200-distilled-1.3B et des langues luxembourgeoise et allemande.

Pour la suite, il a été décidé que les corpus devront être adaptés ou transformés pour être rapprochés du luxembourgeois et de l'allemand afin d'atteindre des meilleurs résultats avec le modèle ci-dessus. C'est à ce moment également que le modèle ChatGPT a été ajouté.
Avec des scripts développés pour cette recherche, les transformations ont été effectués. Une fois les corpus alsaciens doublés et transformés (transformation vers l'allemand et vers le luxembourgeois), une deuxième évaluation a eu lieu. Ces évaluations ont eu le même cadre, c'est-à-dire les mêmes métriques, que l'évaluation initiale.
Après ce tour d'évaluation et une analyse de résultats, on peut clairement dire que la stratégie de modification de corpus a abouti. Par rapport au score BLEU plutôt faible, ceci a pu être augmenté. Il est à noter qu'ici, il s'agit du score atteint par le modèle multilingue. Le modèle ChatGPT a atteint dès le début des bons scores. Or, on constate que le score de ChatGPT est presque le double que celui d'un modèle multilingue entraîné pour la traduction.
Une première conclusion ou tendance de performance de modèle peut être observé alors. On constate que le modèle ChatGPT arrive mieux à traduire que les modèles multilingues.

Afin d'augmenter encore plus les résultats obtenus, et en regard du succès de la première transformation, la stratégie de combiner des transformations auparavant employées a été adapté.
Puisque la transformation *Unaccented transformation*, qui consiste à enlever/supprimer les accents non-existants en allemand et en luxembourgeois, est facile à employer et a atteint des bons résultats, elle fait partie des transformations combinées.
Pour les transformations combinées, les deux transformations *Rule based transformation* et *Vocabulary based transformation* ont été respectivement employées en combinaison avec la transformation *Unaccented transformation*. Ensuite, les modèles ont traduit ces corpus transformés et une troisième évaluation avec les mêmes métriques a eu lieu.

Après avoir analysé les résultats de cette troisième évaluation, nous pouvons confirmer la première conclusion observée. Quel que soit le corpus, la transformation ou la langue, le modèle ChatGPT est clairement et sans aucun doute supérieur aux modèles multilingues dans toutes les quatre métriques.
Si un modèle multilingue entraîné et adapté aux dialectes alsaciens, ou du moins à une version normée, serait capable de surpasser un *large-language-model* tel que ChatGPT reste à découvrir.

6.1 Commentaire et réflexion

L'utilisation de ChatGPT comme système de traduction automatique pourrait avoir plusieurs contraintes. Tout d'abord, l'interface utilisateur n'est pas disponible pour tout le monde. Il faut s'inscrire à la communauté OpenAI[46] et créer un compte. Certaines personnes pourraient ne pas vouloir que l'on accède à leurs données personnelles.
Un autre problème est la limite de caractère que l'interface a. Ainsi, les textes longs risquent de ne pas être entièrement traduits s'ils dépassent cette limite de caractère. Par ailleurs, cette limite n'est pas connue. L'utilisation d'une interface de programmation d'applications (API) pourrait être une solution, mais elle n'est pas accessible au grand public.
Le plus gros problème de ChatGPT est que l'algorithme utilisé n'est pas transparent. Nous ne savons pas comment ce modèle a pu obtenir de tels résultats pour cette langue à faibles ressources, surtout si l'on considère la quantité de données nécessaires à l'entraînement d'un tel modèle pour une langue de grande taille. Mais sachant que l'alsacien est une langue à faibles ressources et qu'il existe un manque important de données, des questions sur ChatGPT sont soulevées.
Il n'est pas à négliger le fait que ce modèle est entraîné avec les ressources de l'internet et par conséquent des textes biaisé et/ou manquant d'un certain niveau de langue attendu. Il se peut qu'avec des textes plus complexes et grammaticalement plus exigeants, une langue plutôt courante est générée ce qui ne correspondrait pas au besoin.

[46] https://openai.com/

En même temps, il faut dire que les modèles multilingues ont aussi leurs limites et leurs défis. Les modèles multilingues nécessitent un grand serveur. Cela signifie que les traductions peuvent prendre du temps et ne peuvent pas être effectuées par un utilisateur arbitraire. Pour cette recherche spécifique, la construction d'une interface utilisateur publique ou d'une API n'est pas concevable. Mais d'autres chercheurs peuvent tout à fait suivre et développer ce projet en s'appuyant sur le nôtre.

6.2 Perspectives

Il serait intéressant de pouvoir faire du fine-tuning sur un modèle multilingue. Pour cela, il faudrait avoir les ressources financières et de données nécessaires.
Pour le cas de la traduction automatique des dialectes alsaciens, il faudrait augmenter les données numériques afin de pouvoir entraîner un tel modèle du fine-tuning.
L'idée d'affiner les transformations afin d'éviter des erreurs pour un rapprochement plus précis peut toujours être exploité avec les mêmes modèles multilingues utilisés.

Sinon, il pourrait également être intéressant de comparer dans un autre projet les performances de plusieurs LLM. La recherche autour de ce modèle de transformer est très active et avancée en ce moment. ChatGPT n'étant qu'un seul modèle, il serait tout à fait possible de comparer la justesse des traductions, le niveau de langue et la capacité du prompt par exemple.
Des possibles modèles pourront être par exemple GPT4 de OpenAI (la version suivante du modèle utilisé dans ce papier), LLAMA3 de Meta, Gemini Pro de Google, Mixtral 8x22B instruct de Mistral AI ou encore Claude-3 OPUS de Anthropic qui vient de sortir il y a peu de temps.

Ces modèles peuvent avoir des grandes différences liées aux choix du tokéniseur, la taille du vocabulaire et les algorithmes de tokénisation. Des expériences sur la longueur du prompt d'entrée seront imaginables aussi. Est-ce qu'un prompt long, en donnant plusieurs phrases du corpus en même temps génère une meilleure traduction qu'un prompt qui traite qu'une seule phrase du corpus à la fois ?

Dans tous les cas, il est clair que des possibilités et inspirations autour des dialectes alsaciens ne sont pas épuisées et que la recherche peut toujours mener à des percées dans le traitement automatique des dialectes alsaciens.

7. Références bibliographiques

Bernhard, D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources : The Example of Alsatian. *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, 23-29. https://hal.science/hal-00966820

Bernhard, D. (2021). *Lexique multilingue alsacien – français – allemand relié aux synsets de BabelNet.* [jeu de données]. https://doi.org/10.34847/nkl.3f9b2i11.

Bernhard, D., & Ligozat, A.-L. (2013). Es esch fàscht wie Ditsch, oder net? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. *TALARE 2013*, 209-220. https://hal.archives-ouvertes.fr/hal-00838355

Bernhard, D., Todirascu, A., Martin, F., Erhart, P., Steible, L., Huck, D., & Rey, C. (2017). Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard. *DiLiTAL 2017*, 14-23. https://hal.archives-ouvertes.fr/hal-01539160

Bouhrim, N., & Zenkouar, L. (2017). État de l'art de la traduction automatique des langues approches & méthodes. *Études et Documents Berbères*, *38*(2), 91-104. https://doi.org/10.3917/edb.038.0091

Bulletin de l'enseignement (1ère année, n°2; p. 37-47). (1920). Département du Bas-Rhin. https://journals.openedition.org/histoire-education/1085

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., … Wang, J. (2022). *No Language Left Behind : Scaling Human-Centered Machine Translation* (arXiv:2207.04672). arXiv. https://doi.org/10.48550/arXiv.2207.04672

Denis, M.-N. (2003). Le dialecte alsacien : État des lieux. *Ethnologie francaise*, *33*(3), 363-371.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2020). *Beyond English-Centric Multilingual Machine Translation* (arXiv:2010.11125). arXiv. http://arxiv.org/abs/2010.11125

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). *Building Large Monolingual Dictionaries at the Leipzig Corpora Collection : From 100 to 200 Languages*.

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. https://doi.org/10.1162/coli_a_00446

Huck, D. (2022). *Les parlers dialectaux en Alsace*. https://hal.science/hal-03662138

Kübler, N., Ea, C., & Paris-Diderot, U. (2007). La traduction automatique : Traduction machine? *2007*, 14.

Lambrecht, L., Schneider, F., & Waibel, A. (2022). Machine Translation from Standard German to Alemannic Dialects. *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, 129-136. https://aclanthology.org/2022.sigul-1.17

LOD - Lëtzebuerger Online Dictionnaire—LOD. (s. d.). Lëtzebuerger Online Dictionnaire (LOD). Consulté 20 avril 2024, à l'adresse https://lod.lu/

Mohammadshahi, A., Nikoulina, V., Berard, A., Brun, C., Henderson, J., & Besacier, L. (2022). *SMaLL-100 : Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages* (arXiv:2210.11621). arXiv. https://doi.org/10.48550/arXiv.2210.11621

Morin, C. (2020, mars 1). *Entre langue et dialecte, une distinction arbitraire ?* The Conversation. http://theconversation.com/entre-langue-et-dialecte-une-distinction-arbitraire-131721

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu : A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318. https://doi.org/10.3115/1073083.1073135

Popović, M. (2017). chrF++ : Words helping character n-grams. *Proceedings of the Second Conference on Machine Translation*, 612-618. https://doi.org/10.18653/v1/W17-4770

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). Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-6319

Pré-traduction (TMS). (s. d.). Phrase. Consulté 5 juin 2024, à l'adresse https://support.phrase.com/hc/fr/articles/5709717749788-Pr%C3%A9-traduction-TMS

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 Computing Surveys*, *55*(11), 229:1-229:37. https://doi.org/10.1145/3567592

Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). *ChatGPT MT : Competitive for High- (but not Low-) Resource Languages* (arXiv:2309.07423). arXiv. http://arxiv.org/abs/2309.07423

Schmid, H. (1997). Probabilistic part-of-speech tagging using decision trees. In *New Methods In Language Processing* (p. 44-49s). Routledge.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), 1715-1725. https://doi.org/10.18653/v1/P16-1162

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223-231. https://aclanthology.org/2006.amta-papers.25

Song, Y., Ezzini, S., Klein, J., Bissyande, T., Lefebvre, C., & Goujon, A. (2023). *Letz Translate : Low-Resource Machine Translation for Luxembourgish* (arXiv:2303.01347). arXiv. https://doi.org/10.48550/arXiv.2303.01347

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – Building open translation services for the World. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. http://hdl.handle.net/10138/327852

Volkart, L. (2018). *Traduction automatique statistique vs. neuronale : Comparaison de MTH et DeepL à La Poste Suisse* [University of Geneva]. https://archive-ouverte.unige.ch/unige:113749

Xu, H. (2021). *Transformer-based NMT : Modeling, training and implementation* [doctoralThesis, Saarländische Universitäts- und Landesbibliothek]. https://doi.org/10.22028/D291-34998

Zeidler, E., & Crévenat-Werner, D. (2016). *Le système ORTHAL 2016—Ortographie alsacienne*. https://www.orthal.fr/ORTHAL_2016.pdf

III. Annexes

Tableau des évaluations des transformations

Tableau 21 - Evaluation de transformation

Model	Corpus	Transformation	Language	BLEU-default	spBLEU_flores200	chrF++	TER
nllb-200-distilled-1,3B	smallcorpus	lexique_deu	ltz_Latn	17,67	17,67	36,01	76,28
nllb-200-distilled-1,3B	small	lexique_deu	lim_Latn	3,61	5,60	27,58	91,03
nllb-200-distilled-1,3B	small	lexique_deu	deu_Latn	10,03	6,46	29,49	85,26
nllb-200-distilled-1,3B	small	lexique_deu	afr_Latn	4,27	9,74	30,36	83,97
nllb-200-distilled-1,3B	small	lexique_ltz	ltz_Latn	14,75	15,68	34,87	82,05
nllb-200-distilled-1,3B	small	lexique_ltz	lim_Latn	1,72	4,44	24,08	99,36
nllb-200-distilled-1,3B	small	lexique_ltz	deu_Latn	4,30	6,29	27,26	87,18
nllb-200-distilled-1,3B	small	lexique_ltz	afr_Latn	1,67	4,73	22,25	102,56
nllb-200-distilled-1,3B	small	regles_deu	ltz_Latn	12,76	12,47	31,26	85,90
nllb-200-distilled-1,3B	small	regles_deu	lim_Latn	14,23	12,86	29,83	88,46
nllb-200-distilled-1,3B	small	regles_deu	deu_Latn	13,47	14,73	31,89	82,05
nllb-200-distilled-1,3B	small	regles_deu	afr_Latn	7,69	9,18	27,16	91,03

nllb-200-distilled-1,3B	small	regles_ltz	ltz_Latn	17,49	18,92	36,30	80,77
nllb-200-distilled-1,3B	small	regles_ltz	lim_Latn	17,45	14,58	28,65	83,33
nllb-200-distilled-1,3B	small	regles_ltz	deu_Latn	12,35	12,63	30,16	78,85
nllb-200-distilled-1,3B	small	regles_ltz	afr_Latn	13,86	14,67	29,73	84,62
nllb-200-distilled-1,3B	small	unaccented	ltz_Latn	24,98	24,07	40,51	69,87
nllb-200-distilled-1,3B	small	unaccented	lim_Latn	11,78	9,09	25,80	89,10
nllb-200-distilled-1,3B	small	unaccented	deu_Latn	8,87	10,58	30,24	79,49
nllb-200-distilled-1,3B	small	unaccented	afr_Latn	4,82	3,69	22,07	85,90
nllb-200-distilled-1,3B	small	closed_class_transposition	ltz_Latn	30,75	27,67	46,07	58,33
nllb-200-distilled-1,3B	small	closed_class_transposition	lim_Latn	8,99	9,42	31,59	76,28
nllb-200-distilled-1,3B	small	closed_class_transposition	deu_Latn	19,72	18,28	38,69	67,95
nllb-200-distilled-1,3B	small	closed_class_transposition	afr_Latn	9,07	12,28	32,94	76,28
nllb-200-distilled-1,3B	small	open_class_transposition	ltz_Latn	9,01	8,52	30,35	81,41
nllb-200-distilled-1,3B	small	open_class_transposition	lim_Latn	12,14	9,45	28,55	89,74
nllb-200-distilled-1,3B	small	open_class_transposition	deu_Latn	7,97	8,74	27,81	80,13
nllb-200-distilled-1,3B	small	open_class_transposition	afr_Latn	12,07	9,28	27,24	83,33

nllb-200-distilled-1,3B	small	original	ltz_Latn	13,05	12,14	33,02	76,92
nllb-200-distilled-1,3B	small	original	lim_Latn	9,70	7,49	25,52	89,10
nllb-200-distilled-1,3B	small	original	deu_Latn	5,57	5,67	25,85	85,90
nllb-200-distilled-1,3B	small	original	afr_Latn	5,16	5,07	23,75	85,90
nllb-200-distilled-1,3B	TestPairs	lexique_deu	ltz_Latn	23,78	26,51	46,07	67,34
nllb-200-distilled-1,3B	TestPairs	lexique_deu	lim_Latn	17,28	19,65	39,50	71,25
nllb-200-distilled-1,3B	TestPairs	lexique_deu	deu_Latn	20,29	23,21	42,34	71,97
nllb-200-distilled-1,3B	TestPairs	lexique_deu	afr_Latn	16,41	18,10	39,02	73,84
nllb-200-distilled-1,3B	TestPairs	lexique_ltz	ltz_Latn	23,79	26,78	46,08	67,65
nllb-200-distilled-1,3B	TestPairs	lexique_ltz	lim_Latn	14,03	16,20	35,72	77,55
nllb-200-distilled-1,3B	TestPairs	lexique_ltz	deu_Latn	19,87	22,42	40,97	73,24
nllb-200-distilled-1,3B	TestPairs	lexique_ltz	afr_Latn	11,65	13,39	32,62	81,11
nllb-200-distilled-1,3B	TestPairs	regles_deu	ltz_Latn	17,44	20,90	41,09	76,79
nllb-200-distilled-1,3B	TestPairs	regles_deu	lim_Latn	11,45	13,37	33,43	82,58
nllb-200-distilled-1,3B	TestPairs	regles_deu	deu_Latn	18,17	20,75	39,41	76,18
nllb-200-distilled-1,3B	TestPairs	regles_deu	afr_Latn	11,03	12,68	32,08	85,73

nllb-200-distilled-1,3B	TestPairs	regles_ltz	ltz_Latn	17,35	20,28	38,93	75,42
nllb-200-distilled-1,3B	TestPairs	regles_ltz	lim_Latn	9,79	12,24	31,47	84,51
nllb-200-distilled-1,3B	TestPairs	regles_ltz	deu_Latn	13,38	16,30	34,47	79,74
nllb-200-distilled-1,3B	TestPairs	regles_ltz	afr_Latn	7,63	9,33	28,71	88,27
nllb-200-distilled-1,3B	TestPairs	unaccented	ltz_Latn	28,81	31,00	49,50	60,59
nllb-200-distilled-1,3B	TestPairs	unaccented	lim_Latn	20,30	22,14	42,05	69,02
nllb-200-distilled-1,3B	TestPairs	unaccented	deu_Latn	24,64	26,57	44,88	66,07
nllb-200-distilled-1,3B	TestPairs	unaccented	afr_Latn	15,14	16,83	38,47	73,44
nllb-200-distilled-1,3B	TestPairs	closed_class_transposition	ltz_Latn	30,03	31,48	49,99	60,89
nllb-200-distilled-1,3B	TestPairs	closed_class_transposition	lim_Latn	19,88	21,54	41,63	68,00
nllb-200-distilled-1,3B	TestPairs	closed_class_transposition	deu_Latn	23,47	25,45	44,48	66,12
nllb-200-distilled-1,3B	TestPairs	closed_class_transposition	afr_Latn	17,13	18,71	39,25	71,00
nllb-200-distilled-1,3B	TestPairs	open_class_transposition	ltz_Latn	25,23	28,32	46,58	65,21
nllb-200-distilled-1,3B	TestPairs	open_class_transposition	lim_Latn	18,07	20,77	39,98	73,39
nllb-200-distilled-1,3B	TestPairs	open_class_transposition	deu_Latn	20,32	23,20	41,45	70,85
nllb-200-distilled-1,3B	TestPairs	open_class_transposition	afr_Latn	14,30	17,04	36,35	77,40

nllb-200-distilled-600M	small	lexique_deu	ltz_Latn	5,83	10,78	29,75	94,87
nllb-200-distilled-600M	small	lexique_deu	lim_Latn	1,72	1,35	22,23	95,51
nllb-200-distilled-600M	small	lexique_deu	deu_Latn	2,30	4,65	23,67	98,72
nllb-200-distilled-600M	small	lexique_deu	dan_Latn	3,84	6,53	26,84	91,03
nllb-200-distilled-600M	small	lexique_deu	nno_Latn	1,89	8,43	29,59	91,67
nllb-200-distilled-600M	small	lexique_deu	nob_Latn	1,84	5,05	30,22	94,23
nllb-200-distilled-600M	small	lexique_deu	nld_Latn	1,99	4,89	24,96	85,26
nllb-200-distilled-600M	small	lexique_deu	fao_Latn	6,07	11,82	30,87	86,54
nllb-200-distilled-600M	small	lexique_deu	isl_Latn	4,19	7,84	28,44	85,90
nllb-200-distilled-600M	small	lexique_deu	afr_Latn	1,92	8,00	27,70	87,82
nllb-200-distilled-600M	small	lexique_deu	swe_Latn	3,69	5,27	26,45	94,87
nllb-200-distilled-600M	small	lexique_deu	eng_Latn	1,74	5,56	25,35	98,72
nllb-200-distilled-600M	small	lexique_ltz	ltz_Latn	4,93	8,10	27,69	96,79
nllb-200-distilled-600M	small	lexique_ltz	lim_Latn	1,45	3,12	26,12	104,49
nllb-200-distilled-600M	small	lexique_ltz	deu_Latn	3,79	6,07	26,88	98,08
nllb-200-distilled-600M	small	lexique_ltz	dan_Latn	1,13	0,94	20,00	114,10

nllb-200-distilled-600M	small	lexique_ltz	nno_Latn	1,36	1,68	25,04	100,64
nllb-200-distilled-600M	small	lexique_ltz	nob_Latn	1,39	1,20	23,06	108,97
nllb-200-distilled-600M	small	lexique_ltz	nld_Latn	2,57	5,15	25,58	89,10
nllb-200-distilled-600M	small	lexique_ltz	fao_Latn	1,42	1,18	22,63	96,79
nllb-200-distilled-600M	small	lexique_ltz	isl_Latn	1,65	1,41	23,80	98,08
nllb-200-distilled-600M	small	lexique_ltz	afr_Latn	1,97	5,26	26,57	89,10
nllb-200-distilled-600M	small	lexique_ltz	swe_Latn	1,37	1,30	20,49	105,13
nllb-200-distilled-600M	small	lexique_ltz	eng_Latn	1,27	2,43	17,55	110,26
nllb-200-distilled-600M	small	regles_deu	ltz_Latn	7,15	7,05	25,80	87,82
nllb-200-distilled-600M	small	regles_deu	lim_Latn	12,29	13,12	30,95	87,18
nllb-200-distilled-600M	small	regles_deu	deu_Latn	9,20	8,52	26,46	100,64
nllb-200-distilled-600M	small	regles_deu	dan_Latn	4,71	5,01	23,93	107,05
nllb-200-distilled-600M	small	regles_deu	nno_Latn	8,73	11,94	32,13	94,23
nllb-200-distilled-600M	small	regles_deu	nob_Latn	5,05	8,03	30,85	91,03
nllb-200-distilled-600M	small	regles_deu	nld_Latn	6,38	4,99	23,76	100,64
nllb-200-distilled-600M	small	regles_deu	fao_Latn	9,35	9,83	31,42	89,10

nllb-200-distilled-600M	small	regles_deu	isl_Latn	6,32	5,53	25,04	119,23
nllb-200-distilled-600M	small	regles_deu	afr_Latn	3,82	7,93	27,90	90,38
nllb-200-distilled-600M	small	regles_deu	swe_Latn	7,18	7,94	27,10	99,36
nllb-200-distilled-600M	small	regles_deu	eng_Latn	2,72	6,97	23,08	116,03
nllb-200-distilled-600M	small	regles_ltz	ltz_Latn	8,68	7,81	28,27	83,97
nllb-200-distilled-600M	small	regles_ltz	lim_Latn	15,05	15,55	30,18	84,62
nllb-200-distilled-600M	small	regles_ltz	deu_Latn	8,58	9,29	25,87	98,08
nllb-200-distilled-600M	small	regles_ltz	dan_Latn	2,96	2,37	20,00	102,56
nllb-200-distilled-600M	small	regles_ltz	nno_Latn	4,87	7,52	29,47	89,74
nllb-200-distilled-600M	small	regles_ltz	nob_Latn	1,58	2,98	26,35	94,23
nllb-200-distilled-600M	small	regles_ltz	nld_Latn	6,73	5,51	23,70	103,85
nllb-200-distilled-600M	small	regles_ltz	fao_Latn	7,36	8,63	28,31	91,67
nllb-200-distilled-600M	small	regles_ltz	isl_Latn	3,81	5,21	21,03	96,15
nllb-200-distilled-600M	small	regles_ltz	afr_Latn	5,10	9,76	29,89	84,62
nllb-200-distilled-600M	small	regles_ltz	swe_Latn	4,14	5,28	22,78	101,92
nllb-200-distilled-600M	small	regles_ltz	eng_Latn	1,45	1,64	18,95	137,18

nllb-200-distilled-600M	small	unaccented	ltz_Latn	14,10	13,46	32,04	78,21
nllb-200-distilled-600M	small	unaccented	lim_Latn	13,08	10,45	28,01	82,69
nllb-200-distilled-600M	small	unaccented	deu_Latn	12,29	12,90	30,11	76,92
nllb-200-distilled-600M	small	unaccented	dan_Latn	7,39	5,92	24,51	96,15
nllb-200-distilled-600M	small	unaccented	nno_Latn	6,46	10,62	30,10	87,82
nllb-200-distilled-600M	small	unaccented	nob_Latn	5,91	9,30	29,41	94,23
nllb-200-distilled-600M	small	unaccented	nld_Latn	8,37	7,00	25,06	89,10
nllb-200-distilled-600M	small	unaccented	fao_Latn	12,68	14,27	31,36	82,69
nllb-200-distilled-600M	small	unaccented	isl_Latn	6,05	6,09	23,48	96,15
nllb-200-distilled-600M	small	unaccented	afr_Latn	7,01	11,20	30,02	80,77
nllb-200-distilled-600M	small	unaccented	swe_Latn	5,29	4,50	20,34	101,92
nllb-200-distilled-600M	small	unaccented	eng_Latn	0,75	2,30	16,29	221,15
nllb-200-distilled-600M	small	closed_class_transposition	ltz_Latn	22,01	20,09	39,72	62,18
nllb-200-distilled-600M	small	closed_class_transposition	lim_Latn	14,33	12,23	32,14	78,21
nllb-200-distilled-600M	small	closed_class_transposition	deu_Latn	15,49	15,26	32,79	71,79
nllb-200-distilled-600M	small	closed_class_transposition	dan_Latn	4,25	6,26	28,61	85,90

nllb-200-distilled-600M	small	closed_class_transposition	nno_Latn	4,51	8,36	34,08	80,77
nllb-200-distilled-600M	small	closed_class_transposition	nob_Latn	3,72	3,57	32,82	81,41
nllb-200-distilled-600M	small	closed_class_transposition	nld_Latn	19,24	18,07	35,43	69,87
nllb-200-distilled-600M	small	closed_class_transposition	fao_Latn	12,69	12,68	34,51	75,00
nllb-200-distilled-600M	small	closed_class_transposition	isl_Latn	8,05	7,30	31,96	83,97
nllb-200-distilled-600M	small	closed_class_transposition	afr_Latn	12,29	13,11	35,58	72,44
nllb-200-distilled-600M	small	closed_class_transposition	swe_Latn	3,88	5,94	29,06	84,62
nllb-200-distilled-600M	small	closed_class_transposition	eng_Latn	12,43	13,76	32,18	76,28
nllb-200-distilled-600M	small	open_class_transposition	ltz_Latn	8,25	7,97	26,12	87,82
nllb-200-distilled-600M	small	open_class_transposition	lim_Latn	9,91	7,84	28,13	85,26
nllb-200-distilled-600M	small	open_class_transposition	deu_Latn	7,33	5,62	26,21	81,41
nllb-200-distilled-600M	small	open_class_transposition	dan_Latn	5,69	4,38	23,83	92,95
nllb-200-distilled-600M	small	open_class_transposition	nno_Latn	1,70	2,44	26,78	91,67
nllb-200-distilled-600M	small	open_class_transposition	nob_Latn	1,68	2,54	28,20	92,95
nllb-200-distilled-600M	small	open_class_transposition	nld_Latn	9,91	9,33	27,30	87,18
nllb-200-distilled-600M	small	open_class_transposition	fao_Latn	14,85	15,44	34,65	76,28

nllb-200-distilled-600M	small	open_class_transposition	isl_Latn	3,93	3,12	24,36	92,95
nllb-200-distilled-600M	small	open_class_transposition	afr_Latn	3,01	3,43	27,16	84,62
nllb-200-distilled-600M	small	open_class_transposition	swe_Latn	6,23	4,91	24,10	91,67
nllb-200-distilled-600M	small	open_class_transposition	eng_Latn	1,74	4,99	23,31	99,36
nllb-200-distilled-600M	small	original	ltz_Latn	6,10	4,82	27,76	80,77
nllb-200-distilled-600M	small	original	lim_Latn	12,50	9,74	27,03	84,62
nllb-200-distilled-600M	small	original	deu_Latn	7,81	5,95	25,29	82,05
nllb-200-distilled-600M	small	original	dan_Latn	5,45	4,70	22,33	101,28
nllb-200-distilled-600M	small	original	nno_Latn	1,68	2,84	26,37	90,38
nllb-200-distilled-600M	small	original	nob_Latn	1,67	2,53	27,41	91,67
nllb-200-distilled-600M	small	original	nld_Latn	10,82	10,32	27,14	90,38
nllb-200-distilled-600M	small	original	fao_Latn	11,46	11,41	28,77	81,41
nllb-200-distilled-600M	small	original	isl_Latn	5,08	4,11	21,73	98,08
nllb-200-distilled-600M	small	original	afr_Latn	3,56	3,98	25,37	85,26
nllb-200-distilled-600M	small	original	swe_Latn	5,33	4,55	20,67	103,21
nllb-200-distilled-600M	small	original	eng_Latn	1,49	5,43	20,62	105,13

nllb-200-distilled-600M	TestPairs	lexique_deu	ltz_Latn	15,07	17,46	37,35	77,76
nllb-200-distilled-600M	TestPairs	lexique_deu	lim_Latn	11,89	13,20	34,93	79,89
nllb-200-distilled-600M	TestPairs	lexique_deu	deu_Latn	11,95	13,91	35,36	88,22
nllb-200-distilled-600M	TestPairs	lexique_deu	dan_Latn	6,16	8,04	29,78	94,16
nllb-200-distilled-600M	TestPairs	lexique_deu	nno_Latn	6,70	7,06	32,61	84,56
nllb-200-distilled-600M	TestPairs	lexique_deu	nob_Latn	7,50	7,82	33,03	94,06
nllb-200-distilled-600M	TestPairs	lexique_deu	nld_Latn	9,40	11,16	32,46	85,58
nllb-200-distilled-600M	TestPairs	lexique_deu	fao_Latn	7,53	8,74	31,46	90,20
nllb-200-distilled-600M	TestPairs	lexique_deu	isl_Latn	6,94	8,82	30,95	88,22
nllb-200-distilled-600M	TestPairs	lexique_deu	afr_Latn	9,88	11,00	33,94	81,87
nllb-200-distilled-600M	TestPairs	lexique_deu	swe_Latn	7,23	8,29	30,50	91,87
nllb-200-distilled-600M	TestPairs	lexique_deu	eng_Latn	5,77	7,47	27,49	94,82
nllb-200-distilled-600M	TestPairs	lexique_ltz	ltz_Latn	15,00	17,03	37,00	77,55
nllb-200-distilled-600M	TestPairs	lexique_ltz	lim_Latn	9,59	12,14	32,00	84,46
nllb-200-distilled-600M	TestPairs	lexique_ltz	deu_Latn	11,20	12,23	33,10	93,30
nllb-200-distilled-600M	TestPairs	lexique_ltz	dan_Latn	2,28	3,18	22,18	130,47

nllb-200-distilled-600M	TestPairs	lexique_ltz	nno_Latn	3,55	4,35	26,57	94,21
nllb-200-distilled-600M	TestPairs	lexique_ltz	nob_Latn	3,34	4,94	26,66	95,73
nllb-200-distilled-600M	TestPairs	lexique_ltz	nld_Latn	4,64	7,40	26,09	97,05
nllb-200-distilled-600M	TestPairs	lexique_ltz	fao_Latn	4,60	5,55	26,64	96,75
nllb-200-distilled-600M	TestPairs	lexique_ltz	isl_Latn	3,19	4,54	23,54	107,11
nllb-200-distilled-600M	TestPairs	lexique_ltz	afr_Latn	5,04	6,99	27,47	97,92
nllb-200-distilled-600M	TestPairs	lexique_ltz	swe_Latn	3,76	5,27	24,23	102,29
nllb-200-distilled-600M	TestPairs	lexique_ltz	eng_Latn	3,26	4,28	21,61	119,91
nllb-200-distilled-600M	TestPairs	regles_deu	ltz_Latn	11,04	13,19	33,89	87,61
nllb-200-distilled-600M	TestPairs	regles_deu	lim_Latn	5,74	8,13	29,17	102,34
nllb-200-distilled-600M	TestPairs	regles_deu	deu_Latn	10,51	12,35	33,42	91,16
nllb-200-distilled-600M	TestPairs	regles_deu	dan_Latn	3,49	4,38	24,35	129,36
nllb-200-distilled-600M	TestPairs	regles_deu	nno_Latn	4,78	4,35	29,11	92,69
nllb-200-distilled-600M	TestPairs	regles_deu	nob_Latn	5,49	4,90	30,10	90,25
nllb-200-distilled-600M	TestPairs	regles_deu	nld_Latn	5,85	7,68	27,54	93,35
nllb-200-distilled-600M	TestPairs	regles_deu	fao_Latn	4,16	5,05	27,19	97,71

nllb-200-distilled-600M	TestPairs	regles_deu	isl_Latn	3,12	4,56	25,51	101,73
nllb-200-distilled-600M	TestPairs	regles_deu	afr_Latn	5,80	6,36	28,44	94,11
nllb-200-distilled-600M	TestPairs	regles_deu	swe_Latn	5,99	6,62	27,21	101,22
nllb-200-distilled-600M	TestPairs	regles_deu	eng_Latn	2,17	3,68	23,37	121,33
nllb-200-distilled-600M	TestPairs	regles_ltz	ltz_Latn	7,68	10,70	30,37	85,88
nllb-200-distilled-600M	TestPairs	regles_ltz	lim_Latn	5,63	7,38	27,42	103,30
nllb-200-distilled-600M	TestPairs	regles_ltz	deu_Latn	8,89	9,68	30,01	88,01
nllb-200-distilled-600M	TestPairs	regles_ltz	dan_Latn	2,91	3,72	22,79	123,77
nllb-200-distilled-600M	TestPairs	regles_ltz	nno_Latn	3,11	3,33	26,10	103,50
nllb-200-distilled-600M	TestPairs	regles_ltz	nob_Latn	3,59	3,90	26,78	102,23
nllb-200-distilled-600M	TestPairs	regles_ltz	nld_Latn	4,15	5,34	24,65	97,77
nllb-200-distilled-600M	TestPairs	regles_ltz	fao_Latn	2,89	3,56	24,56	100,66
nllb-200-distilled-600M	TestPairs	regles_ltz	isl_Latn	1,81	3,30	23,04	100,51
nllb-200-distilled-600M	TestPairs	regles_ltz	afr_Latn	5,22	6,15	26,78	93,60
nllb-200-distilled-600M	TestPairs	regles_ltz	swe_Latn	3,34	3,63	23,21	109,55
nllb-200-distilled-600M	TestPairs	regles_ltz	eng_Latn	2,70	3,40	20,47	127,78

nllb-200-distilled-600M	TestPairs	unaccented	ltz_Latn	24,36	25,55	45,29	65,26
nllb-200-distilled-600M	TestPairs	unaccented	lim_Latn	15,60	16,86	38,40	73,64
nllb-200-distilled-600M	TestPairs	unaccented	deu_Latn	18,47	19,85	40,86	72,07
nllb-200-distilled-600M	TestPairs	unaccented	dan_Latn	7,53	9,27	30,90	88,07
nllb-200-distilled-600M	TestPairs	unaccented	nno_Latn	8,11	9,19	33,75	82,53
nllb-200-distilled-600M	TestPairs	unaccented	nob_Latn	8,31	8,93	34,35	81,82
nllb-200-distilled-600M	TestPairs	unaccented	nld_Latn	11,21	12,19	33,62	82,33
nllb-200-distilled-600M	TestPairs	unaccented	fao_Latn	8,62	9,79	33,84	84,36
nllb-200-distilled-600M	TestPairs	unaccented	isl_Latn	9,71	10,35	32,72	84,05
nllb-200-distilled-600M	TestPairs	unaccented	afr_Latn	11,76	12,34	36,34	77,30
nllb-200-distilled-600M	TestPairs	unaccented	swe_Latn	8,04	8,94	31,70	96,14
nllb-200-distilled-600M	TestPairs	unaccented	eng_Latn	6,99	8,83	29,71	89,08
nllb-200-distilled-600M	TestPairs	closed_class_transposition	ltz_Latn	22,33	23,47	43,44	66,33
nllb-200-distilled-600M	TestPairs	closed_class_transposition	lim_Latn	12,71	13,95	36,83	87,71
nllb-200-distilled-600M	TestPairs	closed_class_transposition	deu_Latn	18,26	19,70	39,66	71,20
nllb-200-distilled-600M	TestPairs	closed_class_transposition	dan_Latn	6,95	8,59	30,27	88,57

nllb-200-distilled-600M	TestPairs	closed_class_transposition	nno_Latn	6,65	6,62	32,47	81,82
nllb-200-distilled-600M	TestPairs	closed_class_transposition	nob_Latn	7,88	8,34	33,68	80,14
nllb-200-distilled-600M	TestPairs	closed_class_transposition	nld_Latn	9,89	11,65	32,81	81,06
nllb-200-distilled-600M	TestPairs	closed_class_transposition	fao_Latn	8,32	8,79	31,51	85,07
nllb-200-distilled-600M	TestPairs	closed_class_transposition	isl_Latn	6,81	8,16	29,31	96,24
nllb-200-distilled-600M	TestPairs	closed_class_transposition	afr_Latn	9,90	10,51	34,24	83,80
nllb-200-distilled-600M	TestPairs	closed_class_transposition	swe_Latn	6,40	7,61	29,25	94,77
nllb-200-distilled-600M	TestPairs	closed_class_transposition	eng_Latn	6,38	8,08	28,56	93,45
nllb-200-distilled-600M	TestPairs	open_class_transposition	ltz_Latn	16,72	19,38	38,71	75,88
nllb-200-distilled-600M	TestPairs	open_class_transposition	lim_Latn	11,08	13,68	34,04	89,49
nllb-200-distilled-600M	TestPairs	open_class_transposition	deu_Latn	14,72	17,03	36,18	77,50
nllb-200-distilled-600M	TestPairs	open_class_transposition	dan_Latn	5,41	7,49	28,27	93,04
nllb-200-distilled-600M	TestPairs	open_class_transposition	nno_Latn	5,18	6,56	30,81	86,39
nllb-200-distilled-600M	TestPairs	open_class_transposition	nob_Latn	6,91	8,19	31,74	85,83
nllb-200-distilled-600M	TestPairs	open_class_transposition	nld_Latn	7,35	10,60	30,25	86,59
nllb-200-distilled-600M	TestPairs	open_class_transposition	fao_Latn	7,02	8,08	30,27	88,78

nllb-200-distilled-600M	TestPairs	open_class_transposition	isl_Latn	4,50	6,88	27,05	101,17
nllb-200-distilled-600M	TestPairs	open_class_transposition	afr_Latn	8,16	10,09	32,16	88,37
nllb-200-distilled-600M	TestPairs	open_class_transposition	swe_Latn	6,51	8,30	28,89	91,98
nllb-200-distilled-600M	TestPairs	open_class_transposition	eng_Latn	5,71	8,59	26,89	87,66
chatGPT	TestPairs	lexique_deu	deu	50,75	53,97	71,75	39,82
chatGPT	TestPairs	lexique_ltz	deu	50,50	51,80	71,45	40,43
chatGPT	TestPairs	regles_deu	deu	53,09	53,80	72,13	37,33
chatGPT	TestPairs	regles_ltz	deu	54,03	55,48	71,86	36,47
chatGPT	small	lexique_deu	deu	46,53	49,29	63,75	44,51
chatGPT	small	lexique_ltz	deu	36,54	33,92	52,69	63,41
chatGPT	small	regles_deu	deu	52,88	57,83	70,55	35,37
chatGPT	small	regles_ltz	deu	40,30	42,79	57,74	50,61

Tableau des évaluations des transformations combinées

Tableau 22 - Evaluation de transformation combinée

model	corpus	transformation	language	BLEU_default	spBLEU_flores200	chrF++	TER
chatGPT	small	rule_deu	Translation_ChatGPT	62,28	62,87	73,48	32,32
chatGPT	small	rule_ltz	Translation_ChatGPT	43,82	45,12	60,26	47,56
chatGPT	small	vocab_ltz	Translation_ChatGPT	54,98	53,30	64,58	34,15
chatGPT	small	vocab_deu	Translation_ChatGPT	45,84	50,66	64,93	43,90
chatGPT	TestPairs	vocab_deu	Translation_ChatGPT	56,84	58,74	74,73	35,35
chatGPT	TestPairs	vocab_ltz	Translation_ChatGPT	58,94	59,79	76,21	32,50

chatGPT	TestPairs	rule_ltz	Translation_ChatGPT	54,85	56,56	72,70	36,47
chatGPT	TestPairs	rule_deu	Translation_ChatGPT	53,64	55,45	72,16	38,55
nllb-200-distilled-1,3B	small	rule_deu	ltz_Latn	16,58	17,11	36,09	79,88
nllb-200-distilled-1,3B	small	rule_deu	lim_Latn	15,49	13,77	30,74	89,02
nllb-200-distilled-1,3B	small	rule_deu	deu_Latn	16,09	18,54	34,02	80,49
nllb-200-distilled-1,3B	small	rule_deu	dan_Latn	2,10	3,50	21,76	96,95
nllb-200-distilled-1,3B	small	rule_deu	nno_Latn	6,91	8,59	30,64	92,68
nllb-200-distilled-1,3B	small	rule_deu	nob_Latn	9,83	9,28	31,71	91,46
nllb-200-distilled-1,3B	small	rule_deu	nld_Latn	14,30	11,85	29,65	83,54
nllb-200-distilled-1,3B	small	rule_deu	fao_Latn	13,92	14,76	32,77	83,54
nllb-200-distilled-1,3B	small	rule_deu	isl_Latn	9,91	9,41	25,73	89,02
nllb-200-distilled-1,3B	small	rule_deu	afr_Latn	6,34	8,14	28,30	87,80
nllb-200-distilled-1,3B	small	rule_deu	swe_Latn	8,60	9,07	25,67	97,56
nllb-200-distilled-1,3B	small	rule_deu	eng_Latn	8,48	9,98	27,12	95,73
nllb-200-distilled-1,3B	small	rule_ltz	ltz_Latn	18,01	20,62	36,37	78,66
nllb-200-distilled-1,3B	small	rule_ltz	lim_Latn	17,94	15,32	29,81	84,15

nllb-200-distilled-1,3B	small	rule_ltz	deu_Latn	13,89	16,36	32,98	82,93
nllb-200-distilled-1,3B	small	rule_ltz	dan_Latn	2,86	5,22	20,11	106,71
nllb-200-distilled-1,3B	small	rule_ltz	nno_Latn	6,51	6,77	27,33	96,95
nllb-200-distilled-1,3B	small	rule_ltz	nob_Latn	3,09	3,39	26,23	101,83
nllb-200-distilled-1,3B	small	rule_ltz	nld_Latn	16,32	14,02	28,78	89,02
nllb-200-distilled-1,3B	small	rule_ltz	fao_Latn	4,22	7,60	28,80	87,80
nllb-200-distilled-1,3B	small	rule_ltz	isl_Latn	8,86	9,98	24,86	90,24
nllb-200-distilled-1,3B	small	rule_ltz	afr_Latn	9,68	12,15	28,29	84,76
nllb-200-distilled-1,3B	small	rule_ltz	swe_Latn	6,96	5,48	22,62	100,61
nllb-200-distilled-1,3B	small	rule_ltz	eng_Latn	14,04	13,56	28,63	82,93
nllb-200-distilled-1,3B	small	vocab_ltz	ltz_Latn	22,44	24,00	38,65	73,78
nllb-200-distilled-1,3B	small	vocab_ltz	lim_Latn	1,82	4,76	26,89	96,95
nllb-200-distilled-1,3B	small	vocab_ltz	deu_Latn	20,87	20,44	36,59	74,39
nllb-200-distilled-1,3B	small	vocab_ltz	dan_Latn	3,16	5,34	20,52	102,44
nllb-200-distilled-1,3B	small	vocab_ltz	nno_Latn	5,01	6,59	27,12	97,56
nllb-200-distilled-1,3B	small	vocab_ltz	nob_Latn	6,30	5,09	27,57	95,73

nllb-200-distilled-1,3B	small	vocab_ltz	nld_Latn	4,21	3,33	21,79	98,17
nllb-200-distilled-1,3B	small	vocab_ltz	fao_Latn	3,49	5,74	23,48	92,07
nllb-200-distilled-1,3B	small	vocab_ltz	isl_Latn	8,08	8,84	25,40	86,59
nllb-200-distilled-1,3B	small	vocab_ltz	afr_Latn	3,81	5,69	24,03	101,83
nllb-200-distilled-1,3B	small	vocab_ltz	swe_Latn	7,92	7,63	25,30	98,78
nllb-200-distilled-1,3B	small	vocab_ltz	eng_Latn	6,15	5,15	21,31	96,34
nllb-200-distilled-1,3B	small	vocab_deu	ltz_Latn	19,18	21,30	39,41	75,00
nllb-200-distilled-1,3B	small	vocab_deu	lim_Latn	4,71	7,37	29,71	87,20
nllb-200-distilled-1,3B	small	vocab_deu	deu_Latn	15,03	17,25	35,95	79,88
nllb-200-distilled-1,3B	small	vocab_deu	dan_Latn	4,13	9,43	28,61	88,41
nllb-200-distilled-1,3B	small	vocab_deu	nno_Latn	5,85	12,53	34,53	79,88
nllb-200-distilled-1,3B	small	vocab_deu	nob_Latn	5,61	8,65	30,18	87,80
nllb-200-distilled-1,3B	small	vocab_deu	nld_Latn	3,96	7,43	29,32	86,59
nllb-200-distilled-1,3B	small	vocab_deu	fao_Latn	6,53	11,54	31,83	85,37
nllb-200-distilled-1,3B	small	vocab_deu	isl_Latn	3,37	5,56	24,67	92,07
nllb-200-distilled-1,3B	small	vocab_deu	afr_Latn	4,07	9,58	29,70	81,71

nllb-200-distilled-1,3B	small	vocab_deu	swe_Latn	5,11	7,27	27,20	87,80
nllb-200-distilled-1,3B	small	vocab_deu	eng_Latn	6,42	7,82	30,26	90,85
nllb-200-distilled-1,3B	TestPairs	rule_deu	ltz_Latn	24,60	26,46	46,67	65,72
nllb-200-distilled-1,3B	TestPairs	rule_deu	lim_Latn	16,87	18,87	39,41	75,01
nllb-200-distilled-1,3B	TestPairs	rule_deu	deu_Latn	20,11	22,33	42,66	73,13
nllb-200-distilled-1,3B	TestPairs	rule_deu	dan_Latn	9,40	10,50	31,71	87,91
nllb-200-distilled-1,3B	TestPairs	rule_deu	nno_Latn	8,56	9,28	35,44	83,44
nllb-200-distilled-1,3B	TestPairs	rule_deu	nob_Latn	6,22	6,99	33,17	86,90
nllb-200-distilled-1,3B	TestPairs	rule_deu	nld_Latn	11,18	13,17	34,43	82,38
nllb-200-distilled-1,3B	TestPairs	rule_deu	fao_Latn	11,58	12,07	36,03	81,06
nllb-200-distilled-1,3B	TestPairs	rule_deu	isl_Latn	8,98	9,23	31,12	86,08
nllb-200-distilled-1,3B	TestPairs	rule_deu	afr_Latn	12,52	13,67	36,25	80,09
nllb-200-distilled-1,3B	TestPairs	rule_deu	swe_Latn	8,30	8,94	31,44	91,67
nllb-200-distilled-1,3B	TestPairs	rule_deu	eng_Latn	7,61	8,64	31,00	88,01
nllb-200-distilled-1,3B	TestPairs	rule_ltz	ltz_Latn	20,58	22,94	43,23	69,88
nllb-200-distilled-1,3B	TestPairs	rule_ltz	lim_Latn	13,77	16,55	37,01	76,13

nllb-200-distilled-1,3B	TestPairs	rule_ltz	deu_Latn	15,16	18,13	37,54	76,38
nllb-200-distilled-1,3B	TestPairs	rule_ltz	dan_Latn	5,98	7,85	29,03	91,16
nllb-200-distilled-1,3B	TestPairs	rule_ltz	nno_Latn	6,74	6,90	32,26	86,64
nllb-200-distilled-1,3B	TestPairs	rule_ltz	nob_Latn	5,07	5,70	30,51	92,69
nllb-200-distilled-1,3B	TestPairs	rule_ltz	nld_Latn	8,34	9,81	30,77	86,49
nllb-200-distilled-1,3B	TestPairs	rule_ltz	fao_Latn	10,94	11,72	35,12	79,08
nllb-200-distilled-1,3B	TestPairs	rule_ltz	isl_Latn	8,56	9,40	30,67	84,66
nllb-200-distilled-1,3B	TestPairs	rule_ltz	afr_Latn	9,61	11,07	33,35	81,72
nllb-200-distilled-1,3B	TestPairs	rule_ltz	swe_Latn	6,08	7,06	28,48	97,66
nllb-200-distilled-1,3B	TestPairs	rule_ltz	eng_Latn	6,77	7,63	27,86	91,26
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	ltz_Latn	26,82	29,71	49,10	62,06
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	lim_Latn	13,87	17,03	37,90	74,45
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	deu_Latn	19,40	22,90	41,66	70,95
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	dan_Latn	6,92	8,95	28,70	90,20
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	nno_Latn	7,14	9,03	33,26	84,00
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	nob_Latn	6,25	7,81	31,80	90,71

nllb-200-distilled-1,3B	TestPairs	vocab_ltz	nld_Latn	10,36	13,22	33,21	81,62
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	fao_Latn	11,74	13,73	35,83	79,02
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	isl_Latn	7,82	10,27	29,12	87,61
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	afr_Latn	11,56	14,44	34,99	78,31
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	swe_Latn	6,32	8,41	28,72	90,60
nllb-200-distilled-1,3B	TestPairs	vocab_ltz	eng_Latn	7,65	10,35	29,64	88,07
nllb-200-distilled-1,3B	TestPairs	vocab_deu	ltz_Latn	25,28	28,35	48,10	64,80
nllb-200-distilled-1,3B	TestPairs	vocab_deu	lim_Latn	20,59	23,15	43,66	66,28
nllb-200-distilled-1,3B	TestPairs	vocab_deu	deu_Latn	20,09	23,33	42,87	71,61
nllb-200-distilled-1,3B	TestPairs	vocab_deu	dan_Latn	12,01	13,95	35,57	79,48
nllb-200-distilled-1,3B	TestPairs	vocab_deu	nno_Latn	12,84	14,63	40,43	72,63
nllb-200-distilled-1,3B	TestPairs	vocab_deu	nob_Latn	11,32	13,33	38,80	74,61
nllb-200-distilled-1,3B	TestPairs	vocab_deu	nld_Latn	16,76	18,83	39,95	73,39
nllb-200-distilled-1,3B	TestPairs	vocab_deu	fao_Latn	17,50	19,00	41,21	72,12
nllb-200-distilled-1,3B	TestPairs	vocab_deu	isl_Latn	15,69	17,10	38,44	74,25
nllb-200-distilled-1,3B	TestPairs	vocab_deu	afr_Latn	18,87	20,67	41,70	70,44

nllb-200-distilled-1,3B	TestPairs	vocab_deu	swe_Latn	10,19	12,56	35,32	79,74
nllb-200-distilled-1,3B	TestPairs	vocab_deu	eng_Latn	14,25	16,26	37,75	76,28
nllb-200-distilled-600M	small	rule_deu	ltz_Latn	15,17	13,03	30,58	81,10
nllb-200-distilled-600M	small	rule_deu	lim_Latn	13,62	14,54	31,79	86,59
nllb-200-distilled-600M	small	rule_deu	deu_Latn	9,43	8,94	25,43	101,83
nllb-200-distilled-600M	small	rule_deu	dan_Latn	7,88	7,29	26,53	100,00
nllb-200-distilled-600M	small	rule_deu	nno_Latn	11,33	12,87	33,72	90,85
nllb-200-distilled-600M	small	rule_deu	nob_Latn	9,23	10,35	33,77	88,41
nllb-200-distilled-600M	small	rule_deu	nld_Latn	8,77	7,09	26,35	102,44
nllb-200-distilled-600M	small	rule_deu	fao_Latn	15,66	13,61	35,63	83,54
nllb-200-distilled-600M	small	rule_deu	isl_Latn	8,04	6,53	26,96	114,02
nllb-200-distilled-600M	small	rule_deu	afr_Latn	7,98	9,78	29,26	90,85
nllb-200-distilled-600M	small	rule_deu	swe_Latn	9,00	8,96	27,98	98,17
nllb-200-distilled-600M	small	rule_deu	eng_Latn	3,66	7,81	24,72	113,41
nllb-200-distilled-600M	small	rule_ltz	ltz_Latn	10,55	10,01	29,07	84,76
nllb-200-distilled-600M	small	rule_ltz	lim_Latn	17,19	17,85	32,31	85,98

nllb-200-distilled-600M	small	rule_ltz	deu_Latn	5,19	6,73	25,60	96,34
nllb-200-distilled-600M	small	rule_ltz	dan_Latn	4,18	4,49	22,94	105,49
nllb-200-distilled-600M	small	rule_ltz	nno_Latn	3,33	5,38	27,41	92,68
nllb-200-distilled-600M	small	rule_ltz	nob_Latn	4,14	4,22	26,97	96,95
nllb-200-distilled-600M	small	rule_ltz	nld_Latn	11,89	11,26	27,37	93,90
nllb-200-distilled-600M	small	rule_ltz	fao_Latn	12,57	12,91	30,03	93,90
nllb-200-distilled-600M	small	rule_ltz	isl_Latn	3,64	3,09	20,10	96,34
nllb-200-distilled-600M	small	rule_ltz	afr_Latn	12,27	12,86	29,98	81,71
nllb-200-distilled-600M	small	rule_ltz	swe_Latn	7,00	5,82	25,14	106,71
nllb-200-distilled-600M	small	rule_ltz	eng_Latn	1,98	2,03	18,48	151,83
nllb-200-distilled-600M	small	vocab_ltz	ltz_Latn	23,30	23,22	40,23	76,83
nllb-200-distilled-600M	small	vocab_ltz	lim_Latn	3,26	4,60	27,67	103,05
nllb-200-distilled-600M	small	vocab_ltz	deu_Latn	14,19	15,86	35,86	85,37
nllb-200-distilled-600M	small	vocab_ltz	dan_Latn	2,78	2,58	22,26	116,46
nllb-200-distilled-600M	small	vocab_ltz	nno_Latn	2,90	2,70	25,98	98,78
nllb-200-distilled-600M	small	vocab_ltz	nob_Latn	5,18	4,51	25,15	109,76

nllb-200-distilled-600M	small	vocab_ltz	nld_Latn	4,10	6,14	28,16	92,07
nllb-200-distilled-600M	small	vocab_ltz	fao_Latn	3,48	2,77	25,71	95,73
nllb-200-distilled-600M	small	vocab_ltz	isl_Latn	4,28	3,63	25,11	97,56
nllb-200-distilled-600M	small	vocab_ltz	afr_Latn	4,69	6,42	28,79	95,12
nllb-200-distilled-600M	small	vocab_ltz	swe_Latn	3,06	2,97	22,56	103,66
nllb-200-distilled-600M	small	vocab_ltz	eng_Latn	2,14	2,69	21,20	142,68
nllb-200-distilled-600M	small	vocab_deu	ltz_Latn	17,37	22,00	39,98	80,49
nllb-200-distilled-600M	small	vocab_deu	lim_Latn	3,88	2,85	23,92	93,29
nllb-200-distilled-600M	small	vocab_deu	deu_Latn	8,89	11,88	32,42	89,63
nllb-200-distilled-600M	small	vocab_deu	dan_Latn	5,46	8,29	30,88	87,20
nllb-200-distilled-600M	small	vocab_deu	nno_Latn	3,67	9,37	33,44	89,02
nllb-200-distilled-600M	small	vocab_deu	nob_Latn	3,49	5,78	32,21	90,24
nllb-200-distilled-600M	small	vocab_deu	nld_Latn	3,62	5,63	24,99	87,20
nllb-200-distilled-600M	small	vocab_deu	fao_Latn	7,06	12,02	33,15	84,15
nllb-200-distilled-600M	small	vocab_deu	isl_Latn	5,81	10,10	34,43	85,37
nllb-200-distilled-600M	small	vocab_deu	afr_Latn	3,88	8,20	29,63	88,41

nllb-200-distilled-600M	small	vocab_deu	swe_Latn	5,39	6,60	30,60	85,98
nllb-200-distilled-600M	small	vocab_deu	eng_Latn	5,38	10,07	30,46	87,80
nllb-200-distilled-600M	TestPairs	rule_deu	ltz_Latn	18,95	20,05	41,61	72,52
nllb-200-distilled-600M	TestPairs	rule_deu	lim_Latn	11,05	13,02	35,12	80,65
nllb-200-distilled-600M	TestPairs	rule_deu	deu_Latn	15,00	16,66	37,83	77,96
nllb-200-distilled-600M	TestPairs	rule_deu	dan_Latn	3,85	5,02	27,43	122,35
nllb-200-distilled-600M	TestPairs	rule_deu	nno_Latn	4,13	5,01	32,18	90,25
nllb-200-distilled-600M	TestPairs	rule_deu	nob_Latn	3,61	4,74	32,14	89,59
nllb-200-distilled-600M	TestPairs	rule_deu	nld_Latn	7,52	9,72	31,47	88,88
nllb-200-distilled-600M	TestPairs	rule_deu	fao_Latn	4,99	5,46	31,64	95,12
nllb-200-distilled-600M	TestPairs	rule_deu	isl_Latn	5,18	5,77	30,82	91,01
nllb-200-distilled-600M	TestPairs	rule_deu	afr_Latn	6,87	7,46	32,61	94,06
nllb-200-distilled-600M	TestPairs	rule_deu	swe_Latn	4,89	5,75	29,50	94,36
nllb-200-distilled-600M	TestPairs	rule_deu	eng_Latn	3,77	5,01	28,10	108,33
nllb-200-distilled-600M	TestPairs	rule_ltz	ltz_Latn	14,06	16,30	37,95	75,57
nllb-200-distilled-600M	TestPairs	rule_ltz	lim_Latn	9,55	10,81	32,89	82,07

nllb-200-distilled-600M	TestPairs	rule_ltz	deu_Latn	11,82	13,87	34,65	79,84
nllb-200-distilled-600M	TestPairs	rule_ltz	dan_Latn	4,15	5,07	26,55	109,90
nllb-200-distilled-600M	TestPairs	rule_ltz	nno_Latn	4,02	4,77	29,38	92,03
nllb-200-distilled-600M	TestPairs	rule_ltz	nob_Latn	4,30	5,13	29,52	91,57
nllb-200-distilled-600M	TestPairs	rule_ltz	nld_Latn	6,50	7,57	28,63	88,52
nllb-200-distilled-600M	TestPairs	rule_ltz	fao_Latn	6,10	6,24	29,74	91,16
nllb-200-distilled-600M	TestPairs	rule_ltz	isl_Latn	4,06	4,92	27,39	93,40
nllb-200-distilled-600M	TestPairs	rule_ltz	afr_Latn	7,47	7,95	31,97	84,71
nllb-200-distilled-600M	TestPairs	rule_ltz	swe_Latn	3,69	4,62	26,28	106,15
nllb-200-distilled-600M	TestPairs	rule_ltz	eng_Latn	3,37	4,07	25,13	104,16
nllb-200-distilled-600M	TestPairs	vocab_ltz	ltz_Latn	18,23	21,10	42,43	70,19
nllb-200-distilled-600M	TestPairs	vocab_ltz	lim_Latn	11,13	13,83	35,32	78,67
nllb-200-distilled-600M	TestPairs	vocab_ltz	deu_Latn	15,29	16,97	37,65	77,65
nllb-200-distilled-600M	TestPairs	vocab_ltz	dan_Latn	4,25	5,30	25,39	112,95
nllb-200-distilled-600M	TestPairs	vocab_ltz	nno_Latn	6,16	7,70	30,10	90,55
nllb-200-distilled-600M	TestPairs	vocab_ltz	nob_Latn	2,93	5,89	28,81	93,91

nllb-200-distilled-600M	TestPairs	vocab_ltz	nld_Latn	6,74	9,57	29,52	98,22
nllb-200-distilled-600M	TestPairs	vocab_ltz	fao_Latn	6,45	8,53	30,69	86,80
nllb-200-distilled-600M	TestPairs	vocab_ltz	isl_Latn	5,15	6,44	27,30	100,36
nllb-200-distilled-600M	TestPairs	vocab_ltz	afr_Latn	5,87	8,76	30,82	91,77
nllb-200-distilled-600M	TestPairs	vocab_ltz	swe_Latn	3,41	5,22	25,12	112,65
nllb-200-distilled-600M	TestPairs	vocab_ltz	eng_Latn	3,98	5,34	24,27	106,86
nllb-200-distilled-600M	TestPairs	vocab_deu	ltz_Latn	20,73	23,07	43,92	68,11
nllb-200-distilled-600M	TestPairs	vocab_deu	lim_Latn	17,44	19,92	40,78	70,90
nllb-200-distilled-600M	TestPairs	vocab_deu	deu_Latn	16,54	18,79	40,27	73,74
nllb-200-distilled-600M	TestPairs	vocab_deu	dan_Latn	9,35	10,89	34,05	83,04
nllb-200-distilled-600M	TestPairs	vocab_deu	nno_Latn	10,86	12,46	37,34	78,42
nllb-200-distilled-600M	TestPairs	vocab_deu	nob_Latn	8,74	10,77	36,63	87,66
nllb-200-distilled-600M	TestPairs	vocab_deu	nld_Latn	12,31	14,66	36,87	83,29
nllb-200-distilled-600M	TestPairs	vocab_deu	fao_Latn	11,89	13,64	37,86	77,15
nllb-200-distilled-600M	TestPairs	vocab_deu	isl_Latn	11,26	13,20	37,45	78,77
nllb-200-distilled-600M	TestPairs	vocab_deu	afr_Latn	12,02	14,16	38,82	77,25

nllb-200-distilled-600M	TestPairs	vocab_deu	swe_Latn	7,77	8,73	33,02	94,01
nllb-200-distilled-600M	TestPairs	vocab_deu	eng_Latn	8,93	10,73	32,77	85,27