Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

À propos des difficultés de traduire automatiquement de longs documents

Ziqian Peng[1, 2] Rachel Bawden[2] François Yvon[1]
(1) Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France
(2) Inria, Paris, France
prénom.nom@isir.upmc.fr, prénom.nom.@inria.fr

RÉSUMÉ

Les nouvelles architectures de traduction automatique sont capables de traiter des segments longs et de surpasser la traduction de phrases isolées, laissant entrevoir la possibilité de traduire des documents complets. Pour y parvenir, il est nécessaire de surmonter un certain nombre de difficultés liées à la longueur des documents à traduire. Dans cette étude, nous discutons de la traduction des documents sous l'angle de l'évaluation, en essayant de répondre à une question simple : comment mesurer s'il existe une dégradation des performances de traduction avec la longueur des documents ? Nos analyses, qui évaluent des systèmes encodeur-décodeur et un grand modèle de langue à l'aune de plusieurs métriques sur une tâche de traduction de documents scientifiques, suggèrent que traduire d'un bloc des documents longs reste un problème difficile.

ABSTRACT

Document Level Machine Translation: does length matter?

Today's machine translation architectures can process long segments and go beyond the translation of isolated sentences, opening up the possibility of translating full documents. To achieve this goal, it is necessary to overcome several difficulties related to the length of source documents. In this work, we discuss document-level machine translation from an evaluation perspective, trying to answer a simple question: how can we measure whether translation performance degrades with document length? Our analysis, which compares encoder-decoder systems and a large language model using multiple metrics on a scientific document translation task, suggests that translating long documents holistically remains a challenging problem.

MOTS-CLÉS : Traduction Automatique, Évaluation de la traduction, Traitement de documents.

KEYWORDS: Machine Translation, Evaluation of Machine translation, Document-level processing.

1 Introduction

Les évolutions récentes des modèles *Transformer* (Vaswani *et al.*, 2017) permettent de traiter (c'est-à-dire d'encoder et de décoder) de très longs contextes contenant des centaines, voire des milliers d'unités : voir (Tay *et al.*, 2023) pour quelques-unes des méthodes qui rendent ce calcul possible. Cette capacité à encoder un long contexte qui va conditionner le processus de génération ouvre de nouvelles perspectives pour de nombreuses applications du traitement automatique des langues (TAL). Nous nous intéressons ici à la traduction automatique (TA), en nous interrogeant sur les difficultés de

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 2–21.

traduire des documents [1] de manière holistique.

La *traduction holistique* d'un document consiste à l'encoder intégralement, puis à générer d'un trait toute la traduction, à l'instar, par exemple, de (Zhang *et al.*, 2018; Junczys-Dowmunt, 2019; Liu *et al.*, 2020a). Cette stratégie revient à traiter les documents comme sont traitées les phrases dans les systèmes de traduction standard. Elle se distingue des méthodes qui exploitent un contexte étendu aux phrases précédentes pour traduire la phrase courante (Post & Junczys-Dowmunt, 2023), comme de celles qui traduisent d'un bloc des segments étendus (correspondant à des fragments de taille fixe, ou bien à des paragraphes) (Tiedemann & Scherrer, 2017; Bawden *et al.*, 2018; Lopes *et al.*, 2020; Ma *et al.*, 2021) [2]. Avec le déploiement de grands modèles de langue multilingues dotés de capacité de traduction et capables d'interpréter des contextes très longs, cette stratégie devient de plus en plus commune (Hendy *et al.*, 2023; Zhang *et al.*, 2023), ce qui implique d'en analyser les principes et d'en diagnostiquer le fonctionnement.

Comme discuté §2, traduire holistiquement induit un changement notable par rapport à la traduction de phrases, aussi bien du point de vue des calculs réalisés que du point de vue des métriques, qui ne peuvent plus s'appuyer sur une comparaison phrase-à-phrase des sorties des systèmes et des références.

Dans la suite de cette contribution, nous nous intéressons plus particulièrement à une question essentielle pour le succès de ces approches, à savoir la capacité à traiter des documents de longueur variable. Plus précisément, notre contribution principale est de nature méthodologique et met à l'épreuve les méthodes expérimentales utilisées pour répondre à cette question. Nous pointons, dans un premier temps, les problèmes des comparaisons automatiques réalisées dans les articles de l'état de l'art, qui reposent sur des *métriques globales* calculées au niveau du corpus (le score BLEU (Papineni *et al.*, 2002) ou sa variante d-BLEU, voir §2.3). Nous présentons ensuite deux manières alternatives d'aborder les questions de longueur, en nous appuyant sur des *métriques locales*, qui individualisent les scores au niveau des documents. Nos expérimentations sont menées sur des documents de taille modeste (de quelques phrases à quelques dizaines de phrases), avec 7 systèmes de traduction automatique. Elles nous permettent néanmoins de conclure que la longueur des documents reste un problème, les scores de traduction ayant une tendance à se dégrader avec le nombre d'unités, un effet qui se manifeste clairement en fin de document.

2 La traduction automatique holistique de documents

2.1 Vers les systèmes traduisant des documents

Comme discuté, par exemple par Sun *et al.* (2022), l'approche holistique (aussi nommée « Doc2Doc ») se distingue de la plupart des travaux sur la traduction de documents qui le plus souvent consistent à augmenter le contexte (surtout côté source) des quelques phrases précédentes - avec potentiellement des encodeurs distincts pour la phrase courante et l'historique (Libovický *et al.*, 2018), tout en continuant de traduire phrase par phrase (« Sent2Sent »). Ces variantes sont appelées « Doc2Sent »

1. C'est-à-dire dont la longueur varie entre quelques phrases et quelques centaines de phrases, soit un résumé comme pour la campagne d'évaluation sur la TA biomédicale à WMT (Neves *et al.*, 2023), soit un exposé, comme pour la campagne d'évaluation de IWSLT 2023 (Salesky *et al.*, 2023), voire un article complet.

2. La terminologie « Traduction Automatique pour les Documents » (*Document-Level Machine Translation* ») ne distingue pas ces approches, dont certaines ne traitent pas des documents complets, mais plutôt des contextes élargis.

par Sun *et al.* (2022) et seraient plus correctement décrites comme des méthodes de traduction avec contexte étendu. Une étude de l'état de l'art, qui ne distingue pas ces deux méthodes, est dans (Maruf *et al.*, 2021). Les résultats expérimentaux qui implémentent ces techniques sont contrastés, ce qui a conduit certains à remettre en cause le bénéfice de contextes étendus (Kim *et al.*, 2019).

L'approche Doc2Doc, qui est notre principal sujet d'étude, est conceptuellement simple, mais elle introduit de multiples changements par rapport à la situation de référence où chaque phrase est encodée et décodée séparément des autres. Nous examinons ci-dessous les principaux changements pour des architectures encodeur-décodeur, sachant que les mêmes observations valent pour les approches à base de grands modèles de langue, lorsqu'on les emploie à des fins de traduction (Wang *et al.*, 2023; Karpinska & Iyyer, 2023). Traduire des documents de manière holistique signifie en particulier que :

— l'encodeur traite l'entièreté du document D source, composé de L phrases $D = (s_1 \dots s_L)$, comme une longue séquence, avec ou sans identification préalable des frontières de phrases ;
— pour générer la $l^{\text{ème}}$ phrase cible, le décodeur a accès à l'intégralité de D, ainsi qu'à toutes les phrases cibles déjà produites $t_{<l} = t_1 \dots t_{l-1}$.

Ces changements ont de nombreuses conséquences, certaines positives (4), d'autres négatives (1)-(3) :

1. les séquences à traiter sont plus longues, entraînant un surcoût computationnel car l'attention dans l'encodeur et le décodeur sont quadratiques en la longueur de l'entrée. Des implémentations approximatives efficaces de ce calcul permettent de conserver des temps de traitement raisonnables (Tay *et al.*, 2023) pour des longues séquences (jusqu'à quelques milliers d'unités).
2. lors du décodage des mots correspondant à la phrase source s_t, le décodeur ne peut plus s'appuyer sur un alignement explicite entre phrases et repose donc entièrement sur l'attention croisée, qui doit calculer une forme d'alignement de mots sur l'intégralité de D [3]. Cet effet est en particulier analysé par Bao *et al.* (2021).
3. décoder des séquences plus longues augmente les effets liés à l'accumulation des erreurs et au biais d'exposition (*exposure bias*) - dû au fait que l'apprentissage du modèle ne considère que des contextes cibles $t_{<l}$ corrects, alors qu'à l'inférence ils peuvent être erronés (Ranzato *et al.*, 2016; Mihaylova & Martins, 2019). Décoder des séquences plus longues réduit également la diversité des hypothèses représentées dans le faisceau de recherche.
4. l'allongement des contextes sources et cibles permet d'intégrer des dépendances plus longues, qui aident à désambiguïser des ambiguïtés lexicales ou des références pronominales.
5. les segments générés ne sont plus nécessairement en correspondance un-pour-un avec les segments sources, ce qui complique, voire obère, le calcul des métriques usuelles (voir §2.3).

Les impacts de ces changements ont été le plus souvent ignorés par les approches « mono-encodeur » qui traduisent simplement des segments longs comme s'il s'agissait de phrases isolées — en s'assurant toutefois que la longueur des segments d'apprentissage est cohérente avec celle qui sera vue au test [4].

2.2 Architectures pour la traduction de documents

Plus récemment, toutefois, divers travaux ont proposé des modifications significatives de l'architecture encodeur-décodeur de base portant notamment sur :

— la stratégie d'apprentissage, qui doit inclure des segments de longueur variable allant de la phrase isolée à des groupes de phrases (appelé « apprentissage multi-résolution » par Sun *et al.*

3. Alors que l'alignement des phrases est en général simple et monotone, contrairement à l'alignement de mots.
4. Et que la configuration du système est appropriée, c-à-d. qu'elle permet effectivement d'encoder des longs segments, que l'encodage des positions est correctement réalisé, etc.

(2022)). Le besoin de prendre en compte des documents d'apprentissage suffisamment longs est également pointé par (Zhuocheng *et al.*, 2023; Wu *et al.*, 2024) ;
— l'architecture du réseau, qui doit être plus profonde (augmentation de capacité) pour modéliser plus de phrases (Junczys-Dowmunt, 2019; Post & Junczys-Dowmunt, 2023), et également mieux régularisée (Kim *et al.*, 2019; Sun *et al.*, 2022) ;
— l'encodage des positions au sein du document source (Li *et al.*, 2022; Lupo *et al.*, 2023) ;
— la structure de l'auto-attention et de l'attention croisée avec des contraintes qui aident à localiser les phrases parallèles au sein d'un document (Bao *et al.*, 2021; Zhuocheng *et al.*, 2023; Herold & Ney, 2023) ;
— les méta-paramètres utilisés par le décodeur durant la génération (pénalité de longueur, largeur du faisceau, etc.) ;

L'enjeu principal de ces études est de parvenir à montrer que (a) les solutions techniques proposées parviennent effectivement à résoudre les problèmes de longueur (« *length bias* ») causés par la longueur des documents, et qu'une fois ces problèmes résolus (b) les méthodes holistiques surpassent les méthodes à traduisant phrase-à-phrase. Avant d'analyser de manière critique les arguments avancés pour prouver que la longueur n'est plus un problème (§3.1), nous nous intéressons dans un premier temps aux méthodes et métriques utilisées pour prouver (b).

2.3 La question de l'évaluation, nuances de BLEU

Répondre à cette question requiert des métriques permettant de comparer des traductions holistiques avec des traductions de phrases : comme le nombre de segments produits par les premières peut différer du nombre de segments sources, ces métriques doivent pouvoir comparer des documents ayant des longueurs différentes. La plupart des travaux en traduction de documents utilisent le score BLEU (Papineni *et al.*, 2002), ou plutôt une variante baptisée *d-BLEU* par Liu *et al.* (2020a) [5], et ceci en dépit des limites de cette métrique (Callison-Burch *et al.*, 2006; Reiter, 2018; Mathur *et al.*, 2020).

Le calcul de BLEU repose sur le décompte, phrase par phrase, du nombre de n-grammes (pour $1 \leq n \leq 4$) partagés par l'hypothèse de traduction et la référence humaine ; ces scores sont agrégés et normalisés, enfin moyennés (géométriquement) au niveau du corpus ; une pénalité de longueur, est enfin appliquée pour dégrader le score lorsque la longueur (agrégée) des hypothèses est plus courte que celle des références. BLEU est donc un score global, qui repose sur des alignements de phrases.

d-BLEU est également un score global, qui s'affranchit toutefois des appariements phrase-à-phase, et effectue les décomptes des n-grammes partagés au niveau des documents. Une conséquence est que d-BLEU, qui repose sur des correspondances élargies, tend à être plus élevé que BLEU, puisque les opportunités de trouver des n-grammes sont plus grandes dans une fenêtre plus large. Cet effet est connu et on peut le visualiser par exemple dans (Koehn & Knowles, 2017, Fig. 1), où l'on observe que BLEU augmente quand on considère des groupes de phrases de longueur croissante (au moins pour une certaine plage de longueur), là où on s'attendrait à une baisse (la longueur est souvent liée à la complexité syntaxique et donc à la difficulté de traduction). Il est facile de reproduire cette observation en calculant d-BLEU pour des systèmes qui traduisent phrase-à-phrase (voir le tableau 2).

Une alternative à d-BLEU consiste à réaligner traduction automatique et référence, par exemple avec

5. Hendy *et al.* (2023) considèrent également une variante de COMET (Rei *et al.*, 2022) ; Zhuocheng *et al.* (2023) introduisent, par analogie avec d-BLEU, la métrique d-chrF, qui est une variante de chrF (Popović, 2015).

l'algorithme de Wicks & Post (2022) [6], tel qu'il soit plus comparable avec les scores BLEU calculé au niveau de phrases reportés dans les publications. Ce problème se pose à l'identique en traduction de parole, l'algorithme de réalignement le plus utilisé dans ce contexte étant dû à Matusov *et al.* (2005). On se ramène ainsi au cas où référence et traduction ont même longueur et où BLEU peut être calculé. Avec cette approche, les scores obtenus dépendront des heuristiques utilisées pour l'alignement.

3 Évaluer les effets de longueur

Une question récurrente dans les études sur la traduction au niveau des documents est la question de la longueur. Un système entraîné uniquement avec des phrases parallèles aura tendance à sous-traduire des documents, stoppant typiquement le processus de génération après la première phrase. Comme expliqué §2.2, d'autres sources de problèmes pour les documents longs sont relatives aux encodages des positions (pour des positions non observées à l'apprentissage), ou encore à des méta-paramètres du système (longueurs maximales acceptées par l'encodeur et le décodeur, etc).

Ces problèmes affectent négativement la pénalité de longueur (BP) du score d-BLEU. Dans nos expériences (§5), le système FT SCIPAR, entraîné sur des phrases isolées, obtient un d-BLEU de 45,0 (BP=1) sur le corpus **THE** traduit phrase par phrase ; lorsqu'on utilise ce système pour traduire des documents complets, d-BLEU tombe à 0,9 (BP$= 0,02$). Ces problèmes sont assez faciles à corriger : dans nos expériences, en affinant le modèle avec des documents, d-BLEU remonte à 45,6 (BP$= 1$).

3.1 Les courbes BLEU / longueur

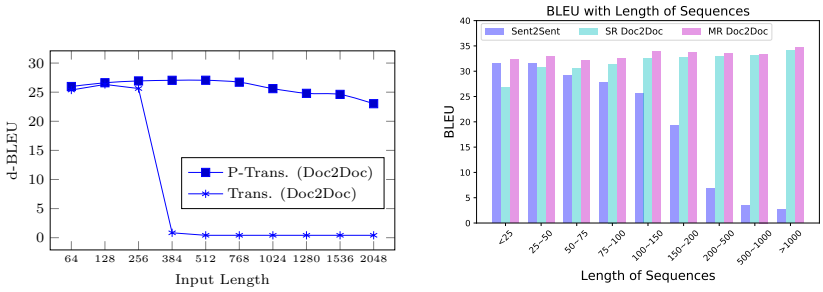

FIGURE 1 – Graphes BLEU / longueur extraits de (Li *et al.*, 2022, fig. 7) (à gauche) et de (Sun *et al.*, 2022, fig. 1) (à droite).

Pour mettre en évidence ces problèmes de longueur, ou pour prouver qu'ils ont été résolus, une première approche consiste à segmenter les documents en fragments de taille fixe, qui sont traduits séparément puis réassemblés pour calculer d-BLEU (Bao *et al.*, 2021, fig. 7), (Li *et al.*, 2022, fig. 7) et (Zhuocheng *et al.*, 2023, fig. 1 et 5). Le graphe 1 (gauche), extrait de Li *et al.* (2022) illustre la difficulté des modèles Sent2Sent à traduire des longs fragments et l'amélioration réalisée par des modèles Doc2Doc correctement entraînés. Cette comparaison occulte un biais qui joue en défaveur de la segmentation en fragments plus courts, qui (a) bénéficie de moins de contexte ; (b) traduit des fragments moins cohérents (démarrant ou s'achevant en milieu de phrases).

6. L'approche de Junczys-Dowmunt (2019) inclut un ensemble de tags qui contraignent les entrées et sortie à avoir le même nombre de phrases, voir aussi (Li *et al.*, 2022).

Une alternative consiste à regrouper les documents en fonction de leur taille et à calculer le score BLEU pour chacun des groupes de longueur – comme sur la figure 1 (droite), tirée de (Sun *et al.*, 2022). Cette manière de procéder évite l'écueil précédent, mais induit une légère confusion, puisqu'elle conduit à aligner sur un même graphique des scores incomparables, puisqu'ils sont calculés sur des corpus de test différents [7]. En se basant de telles analyses, il semble difficile de répondre avec certitude aux questions posées §2.2, ce qui motive les propositions développées ci-dessous.

3.2 Nuances de BLEU : nos analyses

Un premier changement méthodologique consiste à établir un lien direct entre longueur du document et métrique automatique, en calculant un score pour chaque document, plutôt qu'un score unique pour un groupe de documents. Nous utilisons pour cette analyse une variante du score s-BLEU. s-BLEU (*sentence-level BLEU*) est attribué à Lin & Och (2004) et consiste essentiellement à appliquer la métrique BLEU à chaque phrase, puis (éventuellement) à moyenner les scores au niveau d'un groupe de phrases ou de tout le corpus. Le calcul de s-BLEU impose toutefois de lisser les précisions n-grammes afin d'éviter les valeurs nulles. Il existe de multiples manières de réaliser ce lissage (Chen & Cherry, 2014), sans qu'aucune ne parvienne à faire de ce score une bonne évaluation de la qualité au niveau des phrases (Reiter, 2018). Notons que ce lissage est d'autant plus nécessaire que les séquences sont courtes et que la TA est de mauvaise qualité [8].

Calculer s-BLEU pour des documents (et des systèmes de bonne qualité) limite ces problèmes et permet d'obtenir un score par document, appelé dans la suite *ds-BLEU*. Dans nos analyses, nous traitons ces scores comme des réalisations d'une variable aléatoire, dont nous pouvons alors étudier la distribution et les relations avec d'autres variables, comme la longueur des documents sources.

4 Protocole expérimental

4.1 Données pour l'apprentissage et le test

Nos expérimentations considèrent des résumés et des transcriptions d'exposés dans le domaine du traitement automatique des langues, et se focalisent sur la traduction de l'anglais vers le français.

SciPar (Roussis *et al.*, 2022) est un corpus multilingue de phrases parallèles extraites de documents scientifiques collectés sur le Web, dont nous conservons uniquement la partie en–fr. Elle comprend 1,1M phrases, desquelles nous extrayons aléatoirement 3000 phrases pour la validation et le test.

TAL est constitué de résumés d'articles et de thèses dans le domaine du TAL, comprenant d'une part 1701 résumés de thèses récupérés de theses.fr et 1357 résumés d'articles extraits de ISTEX. Ces documents ont été segmentés avec Trankit (Nguyen *et al.*, 2021) et alignés phrase-à-phrase avec hunalign [9] (Varga *et al.*, 2005). Pour la traduction holistique, les phrases parallèles au sein de chaque document sont concaténées et traduites d'un bloc.

7. On constate d'ailleurs que d-BLEU semble augmenter (pour les systèmes Doc2Doc) avec la longueur des documents.
8. s-BLEU est récemment utilisé dans un contexte de TA de documents, par exemple dans Bao *et al.* (2021, tab. 2) : comme ce score est calculé phrase par phrase, il nécessite, comme pour BLEU, de réaligner références et traductions automatiques.
9. https://github.com/danielvarga/hunalign

Jeux de test : **THE** (dev et test) contient deux échantillons aléatoires de 101 et 100 résumés dans le domaine du TAL extraits également de theses.fr sans recouvrement avec **TAL**. **rTAL** contient enfin 246 résumés parallèles d'articles publiés dans la revue TAL. Ces articles sont alignés au niveau des phrases avec la même méthode que pour **TAL** ; ils ont également fait l'objet d'un filtrage avec TransQuest (Ranasinghe *et al.*, 2020) et d'une révision manuelle des alignements. Les statistiques du tableau 1 décrivent ces différents corpus. Pour analyser plus précisément les effets de longueur, nous avons enfin constitué un ensemble de 53 pseudo-documents, désigné par **IWSLT**, en segmentant 10 présentations orales transcrites, puis traduites, préparées pour la campagne IWSLT 2023 (Salesky *et al.*, 2023). La méthode utilisée pour construire ces pseudo-documents est décrite dans l'annexe B.

	SciPar			TAL				
	appr.	valid.	test	appr.	valid.	**THE**	**rTAL**	**IWSLT**
Nb. phrases	1116325	3000	3000	2858	101	100	246	53
Longueur moyenne des documents	37	38	37	265	317	327	129	402
Longueur moyenne des phrases dans un document	-	-	-	34	35	33	32	24

TABLE 1 – Statistiques des données d'apprentissage (appr.), de validation (valid.) et de test. La longueur est donnée en nombre de tokens calculés par le modèle BPE de MBART50(1-M).

4.2 Systèmes comparés

MBART50 (1-M) est un modèle encodeur-décodeur « classique » dérivé du modèle multilingue BART en poursuivant l'apprentissage avec des données parallèles associant anglais en source avec 49 langues en cible (Liu *et al.*, 2020b; Tang *et al.*, 2021). Ce modèle est dans un temps premier affiné avec les phrases parallèles du corpus **SciPar** (FT SCIPAR), puis adapté au TAL en présentant les exemples de **TAL** soit phrase-par-phrase (FT TAL-S), soit document-par-document (FT TAL-D). Un second système Doc2Doc (FT TAL-MR) est obtenu en augmentant **TAL-D** avec des pseudo-documents contenant des sous-parties des documents originaux (reproduisant l'apprentissage *multi-résolution* de Sun *et al.* (2022)). Les détails concernant l'affinage de MBART50 sont dans l'annexe C.

TOWERBASE (Alves *et al.*, 2024) est un grand modèle de langue dérivé de LLAMA2 (Touvron *et al.*, 2023) en continuant le préapprentissage avec des données multilingues en 10 langues, contenant une large proportion de données parallèles. Nous utilisons la version 7B, et l'amorce suivante : « English : SRC \n French : » en mode « zéro-exemple » [10]. Nous avons également analysé les traductions de TOWERINSTRUCT, les détails sont dans l'annexe E. À des fins de comparaison, nous utilisons la version professionnelle de DEEPL [11], sans adaptation, pour traduire les jeux de test **THE** et **rTAL**. Ce système est supposé fournir une indication de l'état-de-l'art. Dans nos expériences, nous fournissons au système des documents complets ; la sortie contient toujours autant de segments que l'entrée.

4.3 Méthodes

4.3.1 Analyse de l'impact de la longueur

Pour mesurer l'impact de la longueur sur le score de traduction, nous calculons pour chaque système la corrélation statistique entre ds-BLEU et la longueur du document : une corrélation négative indique

10. Nous avons également testé 3 exemples et 5 exemples sur le jeu de validation du **TAL**, sans amélioration de BLEU.
11. https://deepl.com

l'existence certaine d'un problème de longueur ; l'absence de corrélation indiquant qu'on ne peut pas conclure sur cette question. Un contraste intéressant pour cette analyse est de considérer des systèmes qui traduisent phrase par phrase et dont on peut considérer qu'ils n'ont aucun problème intrinsèque de longueur. Leur co-variation avec ds-BLEU donne donc une indication de l'impact des différents effets listés ci-dessus. Pour compléter cette analyse, nous reportons en annexe F une analyse des corrélations du score ds-BLEU avec la longueur moyenne des phrases dans un document.

4.3.2 Analyse des effets de position

La dégradation des scores BLEU avec la longueur peut être uniforme au sein d'un document, ou bien affecter surtout les phrases qui sont en fin de document. Analyser plus spécifiquement cet effet requiert de pouvoir comparer les scores de traduction pour des phrases au sein d'un même document. Pour ce faire, nous procédons comme suit : pour chaque document d, nous comparons trois situations impliquant une traduction holistique : (a) d est traduit isolément ; (b) d est traduit dans un pseudo-document, précédé de d' ; (c) d est traduit dans un pseudo-document, suivi de d', où d' un document sélectionné aléatoirement [12]. Les situations (b) et (c) sont éventuellement répétées pour plusieurs sélections de d'. Comparer ds-BLEU des situations (a) et (b) permet de mettre en évidence l'impact de la position. Dans la mesure où (b) introduit aussi un bruit aléatoire (lié à l'introduction du contexte d'), nous comparons également avec (c), dans lequel ce bruit est présent, mais sans changement de position par rapport à (a). Cette méthode demande simplement d'identifier dans la sortie de la traduction de dd' la frontière entre les deux documents, que nous obtenons par réalignement.

Une autre méthode consiste à recopier la première phrase de chaque document en position finale, puis de traduire de manière globale pour évaluer à quel point le changement de position (début vs. fin) modifie le texte généré. Cette méthode demande de réaligner pour identifier les frontières de phrases.

	Scores	MBART50(1-M)	FT SciPar	FT TAL-S	FT TAL-D	FT TAL-MR	TowerBase	DeepLPro
THE	BLEU*	29,0 (0,82)	1,1 (0,03)	1,7 (0,05)	43,3 (1,00)	43,0 (1,00)	39,9 (1,00)	44,5 (1,00)
	d-BLEU	32,0 (0,81)	0,9 (0,02)	1,7 (0,03)	45,6 (1,00)	45,3 (1,00)	42,2 (1,00)	46,8 (1,00)
	dS-BLEU	29,2 (0,77)	3,4 (0,07)	3,9 (0,08)	43,3 (0,98)	43,4 (0,98)	39,9 (0,97)	45,0 (0,98)
rTAL	BLEU*	21,1 (0,75)	3,9 (0,12)	4,7 (0,14)	34,9 (0,99)	35,0 (1,00)	31,9 (0,99)	36,0 (1,00)
	d-BLEU	23,2 (0,74)	4,1 (0,11)	5,0 (0,13)	36,5 (0,99)	36,7 (0,99)	33,5 (0,99)	37,7 (1,00)
	dS-BLEU	21,5 (0,73)	6,7 (0,19)	7,4 (0,21)	33,9 (0,95)	34,0 (0,96)	30,9 (0,95)	34,9 (0,96)
IWSLT	BLEU*	31,8 (0,79)	nan	nan	48,1 (0,97)	49,4 (0,98)	48,3 (0,98)	52,9 (1,00)
	d-BLEU	33,7 (0,78)	0,8 (0,02)	1,1 (0,02)	50,2 (0,96)	51,3 (0,98)	50,1 (0,98)	54,2 (1,00)
	dS-BLEU	32,8 (0,71)	3,4 (0,07)	4,0 (0,08)	49,9 (0,96)	50,8 (0,97)	50,6 (0,94)	52,6 (0,99)

TABLE 2 – Variantes du score BLEU (et pénalité de longueur), calculés sur les résumés de **THE**, de **rTAL** et les transcriptions de **IWSLT**. * indique qu'un réalignement est nécessaire pour l'évaluation.

5 Résultats et analyses

Le tableau 2 donne les scores BLEU [13] pour les trois corpus de test, traduits de manière holistique [14]. On constate en premier lieu, comme attendu, que d-BLEU est toujours supérieur au score BLEU

12. Modulo les contraintes sur la longueur cumulée de d et d' qui doit rester compatible avec les limites de l'encodeur.
13. Tous les scores BLEU et variantes sont calculés avec SacreBLEU (Post, 2018) version 2.4.0.
14. Pour **IWSLT**, certains scores sont absents, à cause de l'impossibilité de réaligner une sortie trop courte avec la référence.

(obtenu par réalignement), avec un écart de deux à trois points. Les systèmes entraînés à traduire des phrases ont des scores médiocres, en particulier causés par la très faible pénalité de longueur, démontrant leur inadéquation pour cette tâche [15]. Les systèmes entrainés pour les documents, y compris TOWERBASE, obtiennent des performances bien meilleures, proches de DEEPLPRO.

5.1 Impact de la longueur des documents

	DEEPLPRO	MBART50(1-M)	FT SCIPAR	FT TAL-S	FT TAL-D	FT TAL-MR	TOWERBASE
THE	0,100 (0,323)	0,139 (0,169)	**-0,416** (0,000)	**-0,309** (0,002)	0,100 (0,322)	0,078 (0,442)	0,080 (0,431)
rTAL	**0,214** (0,001)	**0,136** (0,032)	**-0,469** (0,000)	**-0,417** (0,000)	**0,237** (0,000)	**0,220** (0,001)	**0,249** (0,000)
IWSLT	0,099 (0,480)	0,010 (0,943)	**-0,532** (0,000)	**-0,500** (0,000)	0,002 (0,987)	-0,055 (0,694)	-0,151 (0,279)
THE diff	-	0,080 (0,429)	**-0,234** (0,019)	**-0,233** (0,020)	-0,094 (0,353)	-0,123 (0,221)	-0,162 (0,107)
rTAL diff	-	**-0,136** (0,033)	**-0,471** (0,000)	**-0,467** (0,000)	-0,009 (0,885)	-0,031 (0,625)	-0,021 (0,740)
IWSLT diff	-	-0,059 (0,673)	-0,164 (0,241)	-0,199 (0,154)	-0,093 (0,509)	-0,105 (0,453)	-0,252 (0,069)

TABLE 3 – Corrélation de Spearman (et *p-values*) entre ds-BLEU et la longueur des sources (L_s) (haut) ; corrélation de Spearman entre L_s et la différence des scores ds-BLEU entre chaque système et DEEPLPRO (bas).

Le tableau 3 présente les corrélations de Spearman relevées entre ds-BLEU et la longueur des documents. Pour DEEPLPRO, notre référence, on observe une légère corrélation positive (significative pour **rTAL**), induite par le biais de ds-BLEU en faveur des documents plus longs (§2.3). Des résultats similaires sont obtenus pour les systèmes holistiques (FT TAL-D, FT TAL-MR et TOWERBASE), laissant penser qu'ils n'ont pas plus de problèmes de longueur que DEEPLPRO. Cette tendance s'inverse pour **IWSLT**, dont la distribution de longueur est plus équilibrée. Les déficiences des systèmes entraînés avec des phrases isolées se traduisent par de fortes corrélations négatives. La partie inférieure du tableau vise à neutraliser les effets liés à la difficulté intrinsèque de chaque document, en soustrayant de chaque valeur de ds-BLEU le score obtenu par DEEPLPRO. Nous observons alors que toutes les corrélations pour les systèmes holistiques deviennent négatives (de manière non significative) : plus les documents sont longs, plus les scores ds-BLEU s'écartent de ceux de DEEPLPRO, ce qui suggère que la longueur reste un facteur de difficulté pour la traduction des documents holistiques.

5.2 Analyse des changements de position

Les graphes de la figure 2 mettent en évidence un effet de position très clair pour les deux systèmes dérivés de MBART : décaler un document de quelques centaines de tokens dans l'encodeur cause une perte moyenne d'environ 1,5 points pour la métrique ds-BLEU, alors qu'ajouter un second document à la suite introduit un léger bruit moyen et une dégradation faible de ds-BLEU. Pour TOWERBASE, les deux transformations sont également problématiques, ce qui s'explique par une architecture différente (décodeur pur) qui est plus impactée par l'adjonction du document d'.

Ces résultats sont confirmés par le tableau 5 où nous comparons les traductions des phrases en position initiale avec celle de leur copie en position finale : les premières sont légèrement meilleures que les secondes, surtout pour le corpus **THE**, dont les documents sont plus longs que **rTAL**.

15. L'annexe D montre que lorsque l'on traduit phrase à phrase, ces systèmes obtiennent les meilleurs scores BLEU.

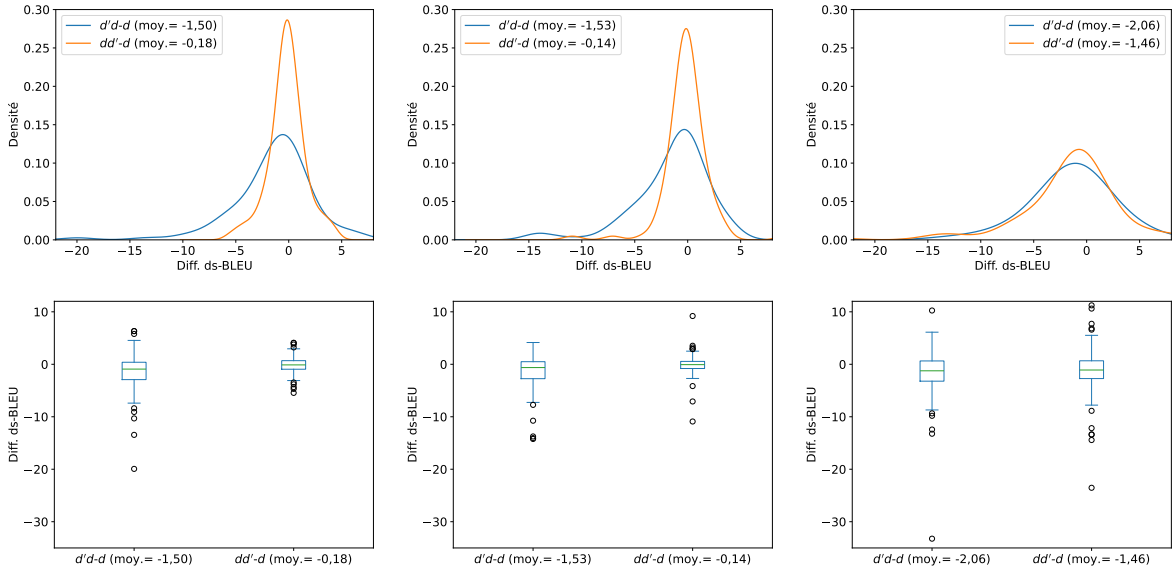

FIGURE 2 – Densité (haut) et boîtes à moustaches (bas) de la différence entre ds-BLEU des documents traduits par les méthodes (b) ou (c) et par la méthode (a) (voir §4.3) : FT TAL-D (gauche), FT TAL-MR (milieu) et TOWEBASE (droite).

		BLEU	s-BLEU	d-BLEU	ds-BLEU
FT TAL-D	$d'd$	40,1 (0,98)	37,6 (0,92)	42,9 (0,98)	41,8 (0,97)
	dd'	43,0 (1,00)	40,7 (0,94)	45,2 (1,00)	43,1 (0,98)
FT TAL-MR	$d'd$	40,6 (1,00)	38,1 (0,93)	43,3 (1,00)	41,9 (0,97)
	dd'	42,8 (1,00)	40,6 (0,95)	45,1 (1,00)	43,3 (0,98)
TOWERBASE	$d'd$	38,4 (0,98)	35,9 (0,91)	40,6 (0,98)	37,9 (0,94)
	dd'	37,8 (0,98)	35,0 (0,89)	40,2 (0,98)	38,5 (0,94)

TABLE 4 – Variantes de BLEU, calculées sur les documents d en position initiale (i.e. dd'), ou finale (i.e. $d'd$). Les scores sont moyennés sur 6 répétitions en variant le choix de d'.

		MBART50(1-M)	FT TAL-D	FT TAL-MR
THE	début	36,7 (0,94)	44,5 (0,99)	45,0 (0,99)
	fin	14,7 (0,45)	43,1 (0,99)	43,7 (0,98)
	début :fin	33.1 (0,50)	90,9 (0,99)	91,5 (0,99)
rTAL	début	25,7 (0,92)	34,7 (0,98)	35,6 (0,98)
	fin	4,8 (0,30)	34,8 (0,98)	35,6 (0,98)
	début :fin	17,2 (0,36)	95,0 (1,00)	95,6 (1,00)

TABLE 5 – Score BLEU calculé sur l'ensemble des premières phrases, traduites respectivement au début et à la fin. "début :fin" évalue (avec BLEU) la distance entre les deux traductions.

6 Conclusion

La traduction au niveau du document semble à notre portée, mais il reste des défis à relever, notamment en ce qui concerne les longs documents. Dans cette étude, nous avons examiné certains des avantages et inconvénients théoriques de la traduction automatique holistique et exploré les différentes manières dont le score BLEU peut être utilisé pour évaluer la traduction au niveau du document. Notre étude expérimentale indique qu'il existe toujours un effet négatif visible de la longueur du document sur la qualité de la traduction, comme le montre le score BLEU, et cet effet négatif semble croître lorsque la longueur du document augmente. Parmi les pistes de travail, mentionnons l'étude des effets du choix de l'encodage positionnel pour les documents longs et l'impact de la longueur des documents vus à l'apprentissage sur la capacité du modèle à traduire des documents de test de longueur variable.

Remerciements

Nous adressons nos remerciements à Mathilde Huguin pour l'extraction des données brutes de ISTEX et à Maxime Bouthors pour les données brutes de theses.fr. Nous remercions également Jean-François Nominé pour les traductions avec DeepLPro. Nous remercions enfin Paul Lerner pour ses retours sur une version préliminaire de cet article. Ces travaux sont financés par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet MaTOS. La contribution de R. Bawden a été partiellement financée par sa chaire à l'institut PRAIRIE financé par l'agence nationale française ANR dans le cadre du programme "Investissements d'avenir" sous la référence ANR-19- P3IA-0001.

Références

ALVES D. M., POMBAL J., GUERREIRO N. M., MARTINS P. H., ALVES J., FARAJIAN A., PETERS B., REI R., FERNANDES P., AGRAWAL S., COLOMBO P., DE SOUZA J. G. C. & MARTINS A. F. T. (2024). Tower : An open multilingual large language model for translation-related tasks.

BAO G., ZHANG Y., TENG Z., CHEN B. & LUO W. (2021). G-transformer for document-level machine translation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 3442–3455, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.267.

BAWDEN R., SENNRICH R., BIRCH A. & HADDOW B. (2018). Evaluating discourse phenomena in neural machine translation. In M. WALKER, H. JI & A. STENT, Éds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1304–1313, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1118.

CALLISON-BURCH C., OSBORNE M. & KOEHN P. (2006). Re-evaluating the role of Bleu in machine translation research. In D. MCCARTHY & S. WINTNER, Éds., *11th Conference of the European Chapter of the Association for Computational Linguistics*, p. 249–256, Trento, Italy : Association for Computational Linguistics.

CHEN B. & CHERRY C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In O. BOJAR, C. BUCK, C. FEDERMANN, B. HADDOW, P. KOEHN, C. MONZ, M. POST & L. SPECIA, Éds., *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 362–367, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/v1/W14-3346.

HENDY A., ABDELREHIM M., SHARAF A., RAUNAK V., GABR M., MATSUSHITA H., KIM Y. J., AFIFY M. & AWADALLA H. H. (2023). How good are GPT models at machine translation ? a comprehensive evaluation.

HEROLD C. & NEY H. (2023). Improving long context document-level machine translation. In M. STRUBE, C. BRAUD, C. HARDMEIER, J. J. LI, S. LOAICIGA & A. ZELDES, Éds., *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, p. 112–125, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.codi-1.15.

JUNCZYS-DOWMUNT M. (2019). Microsoft translator at WMT 2019 : Towards large-scale document-level neural machine translation. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, A. MARTINS, C. MONZ,

M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Fourth Conference on Machine Translation (Volume 2 : Shared Task Papers, Day 1)*, p. 225–233, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5321.

KARPINSKA M. & IYYER M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. In P. KOEHN, B. HADDOW, T. KOCMI & C. MONZ, Éds., *Proceedings of the Eighth Conference on Machine Translation*, p. 419–451, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wmt-1.41.

KIM Y., TRAN D. T. & NEY H. (2019). When and why is document-level context useful in neural machine translation ? In A. POPESCU-BELIS, S. LOÁICIGA, C. HARDMEIER & D. XIONG, Éds., *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, p. 24–34, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-6503.

KOEHN P. & KNOWLES R. (2017). Six challenges for neural machine translation. In T. LUONG, A. BIRCH, G. NEUBIG & A. FINCH, Éds., *Proceedings of the First Workshop on Neural Machine Translation*, p. 28–39, Vancouver : Association for Computational Linguistics. DOI : 10.18653/v1/W17-3204.

LI Y., LI J., JIANG J., TAO S., YANG H. & ZHANG M. (2022). P-Transformer : Towards Better Document-to-Document Neural Machine Translation. arXiv :2212.05830 [cs].

LIBOVICKÝ J., HELCL J. & MAREČEK D. (2018). Input combination strategies for multi-source transformer decoder. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 253–260, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6326.

LIN C.-Y. & OCH F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 605–612, Barcelona, Spain. DOI : 10.3115/1218955.1219032.

LIU C., ZHANG Q., ZHANG X., SINGH K., SARAF Y. & ZWEIG G. (2020a). Multilingual graphemic hybrid ASR with massive data augmentation. In D. BEERMANN, L. BESACIER, S. SAKTI & C. SORIA, Éds., *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, p. 46–52, Marseille, France : European Language Resources association.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020b). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

LOPES A., FARAJIAN M. A., BAWDEN R., ZHANG M. & MARTINS A. F. T. (2020). Document-level neural MT : A systematic comparison. In A. MARTINS, H. MONIZ, S. FUMEGA, B. MARTINS, F. BATISTA, L. COHEUR, C. PARRA, I. TRANCOSO, M. TURCHI, A. BISAZZA, J. MOORKENS, A. GUERBEROF, M. NURMINEN, L. MARG & M. L. FORCADA, Éds., *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, p. 225–234, Lisboa, Portugal : European Association for Machine Translation.

LUPO L., DINARELLI M. & BESACIER L. (2023). Encoding sentence position in context-aware neural machine translation with concatenation. In S. TAFRESHI, A. AKULA, J. SEDOC, A. DROZD, A. ROGERS & A. RUMSHISKY, Éds., *The Fourth Workshop on Insights from Negative Results in NLP*, p. 33–44, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.insights-1.4.

MA Z., EDUNOV S. & AULI M. (2021). A comparison of approaches to document-level machine translation.

MARUF S., SALEH F. & HAFFARI G. (2021). A Survey on Document-Level Neural Machine Translation : Methods and Evaluation. *ACM Comput. Surv.*, **54**(2). Place : New York, NY, USA Publisher : Association for Computing Machinery, DOI : 10.1145/3441691.

MATHUR N., BALDWIN T. & COHN T. (2020). Tangled up in BLEU : Reevaluating the evaluation of automatic machine translation evaluation metrics. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4984–4997, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.448.

MATUSOV E., LEUSCH G., BENDER O. & NEY H. (2005). Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

MIHAYLOVA T. & MARTINS A. F. T. (2019). Scheduled sampling for transformers. In F. ALVA-MANCHEGO, E. CHOI & D. KHASHABI, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 351–356, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-2049.

NEVES M., JIMENO YEPES A., NÉVÉOL A., BAWDEN R., DI NUNZIO G. M., ROLLER R., THOMAS P., VEZZANI F., VICENTE NAVARRO M., YEGANOVA L., WIEMANN D. & GROZEA C. (2023). Findings of the WMT 2023 biomedical translation shared task : Evaluation of ChatGPT 3.5 as a comparison system. In P. KOEHN, B. HADDOW, T. KOCMI & C. MONZ, Éds., *Proceedings of the Eighth Conference on Machine Translation*, p. 43–54, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.wmt-1.2.

NGUYEN M. V., LAI V. D., POURAN BEN VEYSEH A. & NGUYEN T. H. (2021). Trankit : A light-weight transformer-based toolkit for multilingual natural language processing. In D. GKATZIA & D. SEDDAH, Éds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : System Demonstrations*, p. 80–90, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-demos.10.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. In W. AMMAR, A. LOUIS & N. MOSTAFAZADEH, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, p. 48–53, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-4009.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

POPOVIĆ M. (2015). chrF : character n-gram F-score for automatic MT evaluation. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, B. HADDOW, C. HOKAMP, M. HUCK, V. LOGACHEVA & P. PECINA, Éds., *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 392–395, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/W15-3049.

POST M. (2018). A call for clarity in reporting BLEU scores. In O. BOJAR, R. CHATTERJEE, C. FEDERMANN, M. FISHEL, Y. GRAHAM, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, C. MONZ, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POST, L. SPECIA, M. TURCHI & K. VERSPOOR, Éds., *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-6319.

POST M. & JUNCZYS-DOWMUNT M. (2023). Escaping the sentence-level paradigm in machine translation. arXiv :2304.12959 [cs].

RANASINGHE T., ORASAN C. & MITKOV R. (2020). TransQuest : Translation quality estimation with cross-lingual transformers. In D. SCOTT, N. BEL & C. ZONG, Éds., *Proceedings of the 28th International Conference on Computational Linguistics*, p. 5070–5081, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.445.

RANZATO M., CHOPRA S., AULI M. & ZAREMBA W. (2016). Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico.

REI R., C. DE SOUZA J. G., ALVES D., ZERVA C., FARINHA A. C., GLUSHKOVA T., LAVIE A., COHEUR L. & MARTINS A. F. T. (2022). COMET-22 : Unbabel-IST 2022 submission for the metrics shared task. In P. KOEHN, L. BARRAULT, O. BOJAR, F. BOUGARES, R. CHATTERJEE, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, A. FRASER, M. FREITAG, Y. GRAHAM, R. GRUNDKIEWICZ, P. GUZMAN, B. HADDOW, M. HUCK, A. JIMENO YEPES, T. KOCMI, A. MARTINS, M. MORISHITA, C. MONZ, M. NAGATA, T. NAKAZAWA, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POPEL, M. TURCHI & M. ZAMPIERI, Éds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 578–585, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

REITER E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, **44**(3), 393–401. DOI : 10.1162/coli_a_00322.

ROUSSIS D., PAPAVASSILIOU V., PROKOPIDIS P., PIPERIDIS S. & KATSOUROS V. (2022). SciPar : A collection of parallel corpora from scientific abstracts. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 2652–2657, Marseille, France : European Language Resources Association.

SALESKY E., DARWISH K., AL-BADRASHINY M., DIAB M. & NIEHUES J. (2023). Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In E. SALESKY, M. FEDERICO & M. CARPUAT, Éds., *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, p. 62–78, Toronto, Canada (in-person and online) : Association for Computational Linguistics. DOI : 10.18653/v1/2023.iwslt-1.2.

SUN Z., WANG M., ZHOU H., ZHAO C., HUANG S., CHEN J. & LI L. (2022). Rethinking document-level neural machine translation. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Findings of the Association for Computational Linguistics : ACL 2022*, p. 3537–3548, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.279.

TANG Y., TRAN C., LI X., CHEN P.-J., GOYAL N., CHAUDHARY V., GU J. & FAN A. (2021). Multilingual translation from denoising pre-training. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3450–3466, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.304.

TAY Y., DEHGHANI M., BAHRI D. & METZLER D. (2023). Efficient Transformers : A Survey. *ACM Computing Surveys*, **55**(6), 1–28. DOI : 10.1145/3530811.

TIEDEMANN J. & SCHERRER Y. (2017). Neural machine translation with extended context. In B. WEBBER, A. POPESCU-BELIS & J. TIEDEMANN, Éds., *Proceedings of the Third Workshop on Discourse in Machine Translation*, p. 82–92, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/W17-4811.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S., BIKEL D., BLECHER L., FERRER C. C., CHEN M., CUCURULL G., ESIOBU D., FERNANDES J., FU J., FU W., FULLER B., GAO C., GOSWAMI V., GOYAL N., HARTSHORN A., HOSSEINI S., HOU R., INAN H., KARDAS M., KERKEZ V., KHABSA M., KLOUMANN I., KORENEV A., KOURA P. S., LACHAUX M.-A., LAVRIL T., LEE J., LISKOVICH D., LU Y., MAO Y., MARTINET X., MIHAYLOV T., MISHRA P., MOLYBOG I., NIE Y., POULTON A., REIZENSTEIN J., RUNGTA R., SALADI K., SCHELTEN A., SILVA R., SMITH E. M., SUBRAMANIAN R., TAN X. E., TANG B., TAYLOR R., WILLIAMS A., KUAN J. X., XU P., YAN Z., ZAROV I., ZHANG Y., FAN A., KAMBADUR M., NARANG S., RODRIGUEZ A., STOJNIC R., EDUNOV S. & SCIALOM T. (2023). Llama 2 : Open foundation and fine-tuned chat models.

VARGA D., HALAÁCSY P., KORNAI A., NAGY V., NÉMETH L. & TRÓN V. (2005). Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, p. 590–596.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is All you Need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems 30*, p. 5998–6008 : Curran Associates, Inc.

WANG L., LYU C., JI T., ZHANG Z., YU D., SHI S. & TU Z. (2023). Document-level machine translation with large language models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 16646–16661, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.1036.

WICKS R. & POST M. (2022). Does sentence segmentation matter for machine translation ? In P. KOEHN, L. BARRAULT, O. BOJAR, F. BOUGARES, R. CHATTERJEE, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, A. FRASER, M. FREITAG, Y. GRAHAM, R. GRUNDKIEWICZ, P. GUZMAN, B. HADDOW, M. HUCK, A. JIMENO YEPES, T. KOCMI, A. MARTINS, M. MORISHITA, C. MONZ, M. NAGATA, T. NAKAZAWA, M. NEGRI, A. NÉVÉOL, M. NEVES, M. POPEL, M. TURCHI & M. ZAMPIERI, Éds., *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 843–854, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

WU M., WANG Y., FOSTER G. F., QU L. & HAFFARI G. (2024). Importance-aware data augmentation for document-level neural machine translation. *CoRR*, **abs/2401.15360**. DOI : 10.48550/ARXIV.2401.15360.

ZHANG B., HADDOW B. & BIRCH A. (2023). Prompting large language model for machine translation : A case study. In A. KRAUSE, E. BRUNSKILL, K. CHO, B. ENGELHARDT, S. SABATO & J. SCARLETT, Éds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 de *Proceedings of Machine Learning Research*, p. 41092–41110 : PMLR.

ZHANG J., LUAN H., SUN M., ZHAI F., XU J., ZHANG M. & LIU Y. (2018). Improving the transformer translation model with document-level context. In E. RILOFF, D. CHIANG, J. HOCKENMAIER & J. TSUJII, Éds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 533–542, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1049.

ZHUOCHENG Z., GU S., ZHANG M. & FENG Y. (2023). Addressing the length bias challenge in document-level neural machine translation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 11545–11556, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.773.

A Algorithme de réalignement

Le calcul des scores BLEU et s-BLEU pour les systèmes de traduction holistiques demande un réalignement des phrases traduites automatiquement avec les phrases de la référence. Notre stratégie de réalignement est présentée dans l'algorithme 1. Elle consiste essentiellement à aligner les traductions automatiques avec les références, puis à segmenter les sorties automatiques au niveau des frontières de phrase de la référence. L'alignement est réalisé au niveau des caractères avec la bibliothèque Python `edlib`.

Pour évaluer l'exactitude de l'algorithme 1, nous avons traduit les phrases de **THE-S** séparément et calculé un premier score BLEU. Nous avons ensuite concaténé les sorties pour former des documents complets, que nous avons réalignés sur les phrases sources pour obtenir **THE-S-D-S**. Le tableau 6 donne la différence des scores BLEU évalués sur ces deux traductions, qui est toujours très faible (inférieure à 0,2), montrant que la segmentation obtenue par réalignement est très proche de la segmentation initiale. Les mêmes observations valent pour **rTAL**.

	THE-S	**THE-S-D-S**	**rTAL-S**	**rTAL-S-D-S**
MBART50(1-M)	36,2 (1,00)	36,0 (1,00)	29,4 (1,00)	29,2 (1,00)
FT SciPar	42,8 (1,00)	42,7 (1,00)	34,1 (0,99)	33,9 (0,99)
FT TAL-S	44,1 (1,00)	43,9 (1,00)	34,7 (0,98)	34,5 (0,98)
FT TAL-D	44,2 (1,00)	44,1 (1,00)	34,9 (0,99)	34,7 (0,99)
FT TAL-MR	43,9 (1,00)	43,8 (1,00)	35,2 (0,99)	35,0 (0,99)

TABLE 6 – Scores BLEU calculés sur **THE-S** et **THE-S-D-S**, où nous concaténons les phrases de **THE-S** pour former des documents, puis réalignons avec l'algorithme 1, De même pour **rTAL-S**.

Algorithm 1: Calcul d'un réalignement au niveau des phrases

Data: SYS : l'ensemble des traductions produites automatiquement.
Data: [sep] : un délimiteur qui sépare les phrases au sein des documents.
Data: REF : les traductions de référence, dans lesquelles SEP est inséré aux frontières de phrases
Result: SYS_{sent}, REF_{sent} : alignement de phrases entre SYS et REF
begin
  $N \longleftarrow$ nombre de documents dans REF
  $\text{SYS}_{sent} \longleftarrow$ liste vide
  $\text{REF}_{sent} \longleftarrow$ liste vide
  for $I \in \{1, \ldots, N\}$ **do**
    $D_R \longleftarrow$ le I$^{\text{ème}}$ document de REF
    $D_S \longleftarrow$ le I$^{\text{ème}}$ document de SYS
    $I_{SEP} \longleftarrow$ liste des positions de [sep] dans D_R
    $D_R \longleftarrow$ supprimer [sep] de D_R
    dériver l'alignement optimal de D_R et D_S du calcul de la distance de Levenshtein
    $I_{split} \longleftarrow$ indices des caractères de D_R alignés avec les positions de I_{SEP}
    $\text{SYS}_{sent} \longleftarrow$ segmenter D_S en phrases aux positions de I_{split}
    $\text{REF}_{sent} \longleftarrow$ segmenter D_R en phrases aux positions de I_{SEP}
  end
end

B Construction du corpus IWSLT

À partir des 10 documents du corpus **IWSLT**, nous construisons un ensemble de 53 pseudo-documents équilibrés en longueur de la manière suivante. Soit L l'ensemble des longueurs $L = \{32, 24, 16, 8, 4\}$, nous construisons pour chaque document D une permutation aléatoire de L notée $L(D)$, puis segmentons d selon $L(D)$ en affectant les $L(D)[1]$ premières phrases au pseudo-document D_1, puis les $L(D)[2]$ phrases suivantes au pseudo-document D_2, etc. Les phrases en excédent (au-delà de 84) sont affectées au pseudo-document D_5.

Nb. phrases	4	6	7	8	12	16	21	23	24	32	33
Effectif	10	2	1	8	1	9	1	1	10	9	1
Longueur moyenne des documents	99	172	118	198	490	349	573	601	584	777	657
Longueur moyenne des phrases dans un document	25	29	17	25	41	22	27	26	24	24	20

TABLE 7 – Statistiques de **IWSLT** : pour chaque longueur de documents (en nombre de phrases) nous donnons la longueur moyenne (en nombre d'unités) des documents et la longueur moyenne (en nombre d'unités) des phrases dans chaque document.

C Affinage de MBART

L'affinage de MBART50(1-M) est implémenté avec *fairseq* (Ott *et al.*, 2019). Tous les modèles sont constitués de 12 couches pour l'encodeur et le décodeur, de dimension 1024 avec 16 têtes d'attention. Les systèmes sont entraînés avec un GPU de type `NVIDIA RTX A6000 48G` et 12 CPU avec chacun 8G de mémoire. La taille des lots est fixée à 4096 avec une mise à jour tous les 4 lots pour FT SCIPAR, et à 2048 avec une mise à jour tous les deux lots pour les autres systèmes. Afin d'éviter le sur-apprentissage, nous utilisons une procédure d'arrêt précoce avec une patience de 5 époques, en fonction des scores BLEU évalués sur le jeu de validation.

La première étape consiste à entraîner FT SCIPAR en affinant MBART50(1-M) avec le corpus **SciPar** pour réaliser une adaptation au domaine scientifique. À partir de FT SCIPAR, nous avons continué à affiner le système avec les documents du domaine du TAL, présentés soit document par document (FT TAL-D) soit phrase par phrase (FT TAL-S).

Pour implanter l'approche multi-résolution (MR) de (Sun *et al.*, 2022), un jeu d'apprentissage **TAL-MR** est construit avec les documents du corpus **TAL**, avec lequel nous dérivons FT TAL-MR par affinage de FT SCIPAR. Cette approche MR consiste à constituer un jeu d'apprentissage des sous-documents de longueur équilibrée en coupant chaque document (de **TAL-D** dans notre cas) en K sous-documents plusieurs fois, avec $K \in \{1, 2, 4, 8, \dots\}$. C'est-à-dire, un document de longueur 8 est réparti en 15 sous-parties, avec un document de 8 phrases, 2 sous-documents de 4 phrases, 4 sous-documents de 2 phrases et 8 sous-documents d'une phrase.

Pour le décodage avec `fairseq-interactive`, nous utilisons les valeurs de paramètres `max-len-a`=1,5 au lieu de la valeur par défaut (1,2), et `max-len-b`=10. Ces deux paramètres servent à contrôler la longueur des phrases générées par traduction.

D Analyse des systèmes MBART

Le tableau 8 présente les résultats de traduction pour l'ensemble des systèmes dérivés de MBART50. Pour tous les systèmes, l'adaptation au registre scientifique, puis au domaine du TAL a de larges effets positifs. Traduire phrase par phrase donne des scores d-BLEU assez comparables pour les trois systèmes adaptés. Les traductions automatiques sont d'une longueur adéquate. Il en va tout autrement pour les deux systèmes entraînés sur des phrases isolées, qui ont des pénalités de longueur proches de zéro et des scores d-BLEU insignifiants. L'affinage par document suffit à corriger ce problème et conduit à des performances légèrement inférieures (pour **THE**) ou comparables (pour **rTAL**) à la traduction phrase à phrase.

	Sent2Sent		Doc2Doc	
	THE	**rTAL**	**THE**	**rTAL**
MBART50 (1-M)	38,3 (1,0)	31,0 (1,0)	32,0 (0,8)	23,3 (0,7)
FT SciPar	45,0 (1,0)	35,7 (1,0)	0,9 (0,0)	4,1 (0,1)
FT TAL-S	46,2 (1,0)	36,2 (1,0)	1,7 (0,0)	5,1 (0,1)
FT TAL-D	46,4 (1,0)	36,5 (1,0)	45,6 (1,0)	36,6 (1,0)
FT TAL-MR	46,0 (1,0)	36,8 (1,0)	45,3 (1,0)	36,8 (1,0)

TABLE 8 – d-BLEU et pénalité de longueur pour MBART50 (1-M) et l'ensemble des modèles affinés pour les données de test traduites par phrases et par documents.

E Résultats avec TowerInstruct

TowerInstruct est développé pour les tâches concernant la TA, en affinant TowerBase avec instruction sur les jeux de données **TOWERBLOCKS**. **TOWERBLOCKS** contient une grande partie des données pour la traduction au niveau des phrases, la détection des erreurs, conversation et code. Il inclut également une petite portion des corpus parallèles pour la traduction en contexte (Alves *et al.*, 2024, fig.3). Nous utilisons la version 7B de TowerInstruct, et l'amorce suivante [16] : « Translate the following text from French into English.\n English : SRC \n French : » en mode « zéro-exemple ».

Nous reportons pour tous les systèmes les scores BLEU dans le tableau 9, les corrélations de Spearman entre ds-BLEU et la longueur de source dans le tableau 10. Pour **THE** et **rTAL**, les traductions de TowerInstruct sont meilleures que TowerBase en terme des scores BLEU. La corrélation entre ds-BLEU et la longueur de la source montre la même tendance précédemment, même si elle est moins forte que pour les autres systèmes holistiques. Pour **IWSLT**, le score ds-BLEU de TowerInstruct est inférieur à celui de TowerBase. La corrélation positive (non significative) devient moins faible lorsque nous soustrayons le ds-BLEU de TowerInstruct avec celui de DeepLPro.

Le tableau 11 rassemble les évaluations des méthodes de traduction (a), (b) et (c) présentées dans §4.3 pour tous les systèmes, y compris TowerInstruct. Nous constatons que la différence entre les ds-BLEU des traductions par les méthodes (b) et (c) a diminué par rapport aux résultats de TowerBase. Ce phénomène est aussi illustré dans la figure 3. Il montre que la performance de TowerInstruct est plus stable pour traduire des textes situant aux différentes positions dans notre scénario, avec des séquences d'entrée $d'd$ ou dd' moins longues que 1024.

16. https://huggingface.co/Unbabel/TowerInstruct-7B-v0.1#prompt-format

	Scores	MBART50(1-M)	FT SciPar	FT TAL-S	FT TAL-D	FT TAL-MR	TowerBase	TowerInstruct	DeepLPro
THE	BLEU*	29,0 (0,82)	1,1 (0,03)	1,7 (0,05)	43,3 (1,00)	43,0 (1,00)	39,9 (1,00)	41,0 (1,00)	44,5 (1,00)
	d-BLEU	32,0 (0,81)	0,9 (0,02)	1,7 (0,03)	45,6 (1,00)	45,3 (1,00)	42,2 (1,00)	43,3 (1,00)	46,8 (1,00)
	ds-BLEU	29,2 (0,77)	3,4 (0,07)	3,9 (0,08)	43,3 (0,98)	43,4 (0,98)	39,9 (0,97)	41,6 (0,98)	45,0 (0,98)
rTAL	BLEU*	21,1 (0,75)	3,9 (0,12)	4,7 (0,14)	34,9 (0,99)	35,0 (1,00)	31,9 (0,99)	33,3 (1,00)	36,0 (1,00)
	d-BLEU	23,2 (0,74)	4,1 (0,11)	5,0 (0,13)	36,5 (0,99)	36,7 (0,99)	33,5 (0,99)	35,0 (1,00)	37,7 (1,00)
	ds-BLEU	21,5 (0,73)	6,7 (0,19)	7,4 (0,21)	33,9 (0,95)	34,0 (0,96)	30,9 (0,95)	32,3 (0,96)	34,9 (0,96)
IWSLT	BLEU*	31,8 (0,79)	nan (nan)	nan (nan)	48,1 (0,97)	49,4 (0,98)	48,3 (0,98)	48,5 (0,98)	52,9 (1,00)
	d-BLEU	33,7 (0,78)	0,8 (0,02)	1,1 (0,02)	50,2 (0,96)	51,3 (0,98)	50,1 (0,98)	50,1 (0,98)	54,2 (1,00)
	ds-BLEU	32,8 (0,71)	3,4 (0,07)	4,0 (0,08)	49,9 (0,96)	50,8 (0,97)	50,6 (0,94)	48,5 (0,97)	52,6 (0,99)

TABLE 9 – Variantes du score BLEU (et pénalité de longueur), calculés sur les résumés de **THE**, de **rTAL** et les transcriptions de **IWSLT**. * indique qu'un réalignement est nécessaire pour l'évaluation.

	DeepLPro	MBART50(1-M)	FT SciPar	FT TAL-S	FT TAL-D	FT TAL-MR	TowerBase	TowerInstruct
THE	0,100 (0,323)	0,139 (0,169)	**-0,416** (0,000)	**-0,309** (0,002)	0,100 (0,322)	0,078 (0,442)	0,080 (0,431)	0,066 (0,512)
rTAL	**0,214** (0,001)	**0,136** (0,032)	**-0,469** (0,000)	**-0,417** (0,000)	**0,237** (0,000)	**0,220** (0,001)	**0,249** (0,000)	**0,231** (0,000)
IWSLT	0,099 (0,480)	0,010 (0,943)	**-0,532** (0,000)	**-0,500** (0,000)	0,002 (0,987)	-0,055 (0,694)	-0,151 (0,279)	0,234 (0,092)
THE diff	-	0,080 (0,429)	**-0,234** (0,019)	**-0,233** (0,020)	-0,094 (0,353)	-0,123 (0,221)	-0,162 (0,107)	-0,069 (0,494)
rTAL diff	-	**-0,136** (0,033)	**-0,471** (0,000)	**-0,467** (0,000)	-0,009 (0,885)	-0,031 (0,625)	-0,021 (0,740)	-0,025 (0,697)
IWSLT diff	-	-0,059 (0,673)	-0,164 (0,241)	-0,199 (0,154)	-0,093 (0,509)	-0,105 (0,453)	-0,252 (0,069)	0,023 (0,869)

TABLE 10 – Corrélation de Spearman (et *p-values*) entre ds-BLEU et la longueur des sources (L_s) (haut) ; corrélation de Spearman entre L_s et la différence des scores ds-BLEU entre chaque système et DeepLPro (bas).

		BLEU	s-BLEU	d-BLEU	ds-BLEU
FT TAL-D	d	43,3 (1,00)	40,9 (0,94)	45,6 (1,00)	43,3 (0,98)
	$d'd$	40,1 (0,98)	37,6 (0,92)	42,9 (0,98)	41,8 (0,97)
	dd'	43,0 (1,00)	40,7 (0,94)	45,2 (1,00)	43,1 (0,98)
FT TAL-MR	d	43,0 (1,00)	41,0 (0,95)	45,3 (1,00)	43,4 (0,98)
	$d'd$	40,6 (1,00)	38,1 (0,93)	43,3 (1,00)	41,9 (0,97)
	dd'	42,8 (1,00)	40,6 (0,95)	45,1 (1,00)	43,3 (0,98)
TowerBase	d	39,9 (1,00)	37,2 (0,93)	42,2 (1,00)	39,9 (0,97)
	$d'd$	38,4 (0,98)	35,9 (0,91)	40,6 (0,98)	37,9 (0,94)
	dd'	37,8 (0,98)	35,0 (0,89)	40,2 (0,98)	38,5 (0,94)
TowerInstruct	d	41,0 (1,00)	39,0 (0,95)	43,3 (1,00)	41,6 (0,98)
	$d'd$	39,8 (1,00)	37,5 (0,94)	42,2 (1,00)	40,1 (0,98)
	dd'	39,7 (0,99)	37,3 (0,92)	42,1 (0,99)	40,3 (0,97)

TABLE 11 – Variantes de BLEU calculées sur les documents d en position initiale (i.e. dd'), ou finale (i.e. $d'd$), ou tout seul (i.e. d) et moyennés sur 6 répétitions avec des documents d' différents.

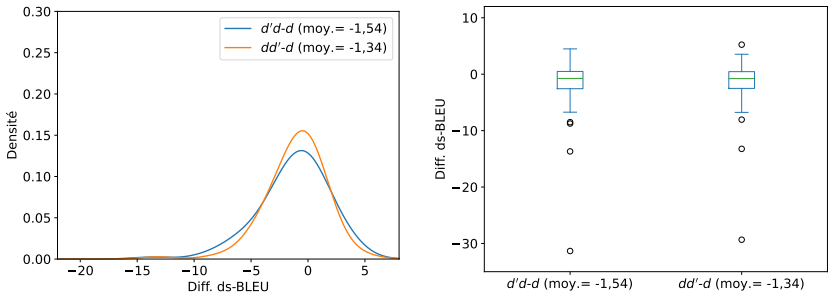

FIGURE 3 – Densité (gauche) et boîtes à moustaches (droite) de la différence entre ds-BLEU des documents traduits par les méthodes (b) ou (c) et par la méthode (a) (voir §4.3) avec TOWEINSTRUCT.

F Analyse des performances selon la longueur moyenne des documents

Le tableau 12 présente des analyses de corrélation de *la longueur moyenne* des phrases au sein du document avec le score ds-BLEU et visent à confirmer l'intérêt de l'analyse de ces corrélations. On observe des corrélations négatives (pour **THE** et **IWSLT**) ou faiblement positives (pour **TAL**), les documents ayant des phrases moyennes plus longues recevant des scores ds-BLEU en moyenne plus faibles, que l'on peut interpréter comme étant causés par la complexité plus grande des phrases à traduire.

	DEEPLPRO	MBART50(1-M)	FT SCIPAR	FT TAL-S	FT TAL-D	FT TAL-MR	TOWERBASE	TOWERINSTRUCT
THE	-0,149 (0,139)	**-0,211** (0,035)	0,021 (0,834)	0,010 (0,920)	-0,165 (0,101)	-0,177 (0,078)	-0,159 (0,115)	-0,179 (0,074)
rTAL	0,044 (0,491)	0,043 (0,501)	**0,245** (0,000)	**0,228** (0,000)	0,016 (0,800)	0,019 (0,763)	0,001 (0,989)	0,014 (0,826)
IWSLT	0,051 (0,718)	-0,047 (0,737)	0,161 (0,250)	0,095 (0,497)	-0,006 (0,969)	-0,074 (0,601)	-0,169 (0,226)	-0,042 (0,767)
THE diff	-	0,006 (0,953)	**0,217** (0,03)	**0,201** (0,045)	-0,080 (0,432)	-0,079 (0,435)	-0,002 (0,981)	-0,036 (0,724)
rTAL diff	-	0,001 (0,986)	0,091 (0,157)	0,098 (0,124)	-0,062 (0,335)	-0,087 (0,174)	-0,102 (0,109)	-0,079 (0,216)
IWSLT diff	-	-0,051 (0,719)	0,162 (0,246)	0,129 (0,356)	-0,073 (0,605)	-0,120 (0,393)	-0,254 (0,066)	-0,134 (0,337)

TABLE 12 – Corrélation de Spearman (et *p-values*) entre ds-BLEU et la longueur moyenne des phrases sources (L_m) (haut) ; corrélation de Spearman entre L_m et la différence des scores ds-BLEU entre chaque système et DEEPLPRO (bas).