Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Vers une fouille de phrases parallèles pour les langues régionales de France métropolitaine

Shu Okabe[1, 2] Alexander Fraser[1,2,3]
(1) Technische Universität München (TUM), D-80333 Munich, Allemagne
(2) Munich Center for Machine Learning, D-80333 Munich, Allemagne
(3) Munich Data Science Institute, D-85748 Garching, Allemagne
shu.okabe@tum.de

RÉSUMÉ

La fouille de phrases parallèles vise à extraire des paires de traduction à partir de corpus monolingues. Si des paires de langues relativement bien dotées ont déjà été étudiées par le passé, cet article a pour objectif d'étendre cette tâche, dans un premier temps, à six langues régionales de France métropolitaine, appariées au français : le breton, le corse, le basque, l'alsacien, l'occitan et le picard. Afin de pouvoir évaluer la qualité des outils de fouille, nous générons des corpus synthétiques en introduisant des phrases parallèles dans des corpus monolingues. Nos expériences suggèrent que les quatre modèles de langue considérés représentent ces langues de manière variable, reflétant le niveau de ressources numériques disponibles et la proximité linguistique avec les langues de pré-entraînement. Nous avons également étudié deux types d'approches pour améliorer l'alignement multilingue qui ne requièrent aucune phrase parallèle incluant les six langues étudiées.

ABSTRACT

Towards the Parallel Sentence Mining for Regional Languages of Metropolitan France.

Parallel sentence mining aims to extract translation pairs from monolingual corpora. While relatively well-resourced language pairs have already been studied previously, the objective of this article is to extend this task to six regional languages of metropolitan France, paired with French, as a first step : Breton, Corsican, Basque, Alsatian, Occitan, and Picard. We generate synthetic corpora to be able to evaluate the quality of our mining tools by introducing parallel sentences in monolingual corpora. Our experiments suggest that the four language models considered represent these languages in varying ways, reflecting the level of available digital resources and the linguistic proximity to the pre-training languages. We also studied two types of approaches to improve multilingual alignment that do not require any parallel sentences featuring the six studied languages.

MOTS-CLÉS : Fouille de phrases parallèles, langues régionales de France, représentation de phrases.

KEYWORDS: Parallel sentence mining, regional languages of France, sentence representation.

1 Introduction

La fouille de phrases parallèles consiste en l'identification de paires de traduction au sein de deux corpus monolingues. La principale motivation réside dans la constitution d'un corpus parallèle pour entraîner des modèles de traitement automatique des langues (TAL), le plus évident étant la traduction

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 688–701.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

automatique. Il s'agit d'une tâche plus complexe que celle connexe d'appariement (« *matching* ») de phrases parallèles, qui intervient notamment pour évaluer la qualité de modèles de langue multilingues, car les phrases ont alors nécessairement un équivalent dans l'autre corpus.

Différentes éditions de défis partagés se sont intéressées à la fouille de phrases parallèles, le plus notable étant BUCC (Zweigenbaum *et al.*, 2017; Pierre Zweigenbaum & Rapp, 2018), qui a étudié quatre paires de langues bien dotées : le français, l'allemand, le chinois et le russe, tous associés à l'anglais. Un défi partagé similaire porte sur le filtrage de corpus parallèles (Koehn *et al.*, 2019, 2020), où des langues peu dotées d'Asie (du sous-continent indien et d'Asie du Sud-Est) ont été considérées. L'objectif est de ne conserver que les paires de traduction, en supprimant les paires bruitées, de moindre qualité, notamment obtenues à travers une fouille de phrases. Cependant, leurs corpus de départ à filtrer comptent déjà des millions de paires de phrases, ce qui reste une barre élevée à atteindre pour d'autres langues peu dotées.

Dans ce contexte, si le français est bien représenté dans le TAL (notamment pour la fouille), nous nous intéressons aux autres langues parlées en France. Nous considérons, dans cette première étude, six langues régionales de France métropolitaine, appariées avec le français : le breton, le corse, le basque, l'alsacien, l'occitan et le picard. L'objectif est, d'une part, d'avoir un aperçu de leur représentation dans les modèles de langue actuels (encodeurs) et, d'autre part, d'évaluer leur influence sur la qualité de la fouille. Il s'agit ici de langues parmi les plus présentes en ligne après le français, comparativement.

Notre approche repose principalement sur l'utilisation de modèles de langue multilingues entraînés avec des données *monolingues*, afin de pouvoir inclure plus simplement davantage de langues par la suite, en particulier les moins dotées. De fait, les encodeurs de phrases compétitifs tels que LaBSE (Feng *et al.*, 2022) nécessitent des phrases parallèles pour l'entraînement, ce qui reste coûteux, voire difficilement accessible en quantité suffisante en fonction des paires de langues. Par exemple, l'approche contrastive de Tan *et al.* (2023) a besoin de dizaines de milliers de phrases *parallèles* pour entraîner leur modèle dans le cas de langues « extrêmement peu dotées ».

Nous présentons les six langues et les corpus utilisés en section 2, puis nous détaillons nos expériences de fouille de phrases en section 3. Nous rendons accessibles les corpus d'évaluation générés, BELOPSEM_FRANCE [1], ainsi que les modifications effectuées sur le système de fouille [2].

2 Constitution des corpus comparables

2.1 Langues étudiées

Nous nous intéressons à six langues régionales de France métropolitaine dans cette étude : le breton (code ISO 639-3 : `bre`, Glottocode : `bret1244` ; langue celtique), le corse (`cos`, `cors1241` ; langue romane), le basque (`eus`, `basq1248` ; isolat), l'alsacien (dialecte de l'alémanique [3], `gsw`, `alsa1241` ; langue germanique), l'occitan (`oci` ; langue romane) et le picard (`pcd` ; langue romane), toutes appariées avec le français.

Deux colonnes du tableau 1 indiquent la vitalité de la langue selon la classification d'Ethnologue

1. https://github.com/shuokabe/Belopsem
2. https://github.com/shuokabe/PaSeMiLL
3. L'alsacien est groupé avec le suisse allemand pour le code ISO `gsw` (alémanique).

langue	Ethnologue	ressource	OPUS	XLM-R	mmBERT	Glot500-m	LaBSE
breton	en danger	1	1,5M	✓	✓	✓ (749k)	✗
corse	en danger	1	855	✗	✓	✓ (3,0M)	✓
basque	institutionnel	4	2,6M	✓	✓	✓ (13M)	✓
alsacien*	stable	0	26	✗	✓	✓ (449k)	✗
occitan	en danger	1	11M	✗	✓	✓ (1,4M)	✗
picard	en danger	1	360	✗	✓	✓ (32,9k)	✗

TABLE 1 – Informations complémentaires sur les six langues étudiées : la vitalité de la langue selon Ethnologue, la quantité de ressources disponibles selon Joshi *et al.* (2020) (allant de 0 à 5), le nombre de paires de phrases parallèles dans la collection OPUS (Tiedemann, 2012) et la présence de la langue dans les données de pré-entraînement des quatre modèles respectifs. * les informations pour l'alsacien correspondent au code ISO `gsw` (qui inclut le suisse allemand), hormis pour le niveau de ressources.

(Eberhard *et al.*, 2026) ainsi que le groupe affecté par l'étude de Joshi *et al.* (2020) sur la quantité de ressources disponibles en ligne, avec une échelle allant de 0 (langue très peu dotée) à 5 (langue bien dotée). Notre étude porte notamment sur quatre langues catégorisées comme étant en danger ainsi que sur l'alsacien, un dialecte de l'alémanique classé « vulnérable » par l'UNESCO (2010). De plus, toutes, sauf le basque, sont considérées comme (très) peu dotées pour le TAL (catégories 0 et 1), avec peu de données en ligne comparativement.

2.2 Méthodologie pour la création de corpus synthétiques

Nous évaluons la représentation des langues par les modèles encodeurs ainsi que la qualité de la fouille de phrases en suivant la méthodologie de Okabe *et al.* (2025), qui se sont intéressés à la même tâche pour trois langues peu dotées (dont l'occitan), en créant un jeu de données d'évaluation (BELOPSEM). Nous constituons donc un corpus synthétique d'évaluation pour la fouille, sur le modèle du défi partagé BUCC (Zweigenbaum *et al.*, 2017). Étant donné un corpus parallèle existant pour une paire de langues, nous injectons aléatoirement ses phrases dans deux corpus monolingues distincts. L'objectif est alors de retrouver ces mêmes paires de phrases. Cette approche permet d'évaluer automatiquement la performance des outils de fouille, tout en se rapprochant d'un cadre réaliste.

Source des corpus Pour créer nos jeux de données synthétiques, nous avons besoin de paires de phrases traduites ainsi que de corpus monolingues comparables [4] (mais non parallèles), afin que la tâche ne soit pas trop aisée. Tout d'abord, pour les textes *monolingues*, nous avons recours aux corpus monolingues de Wikipédia (2021), accessibles à travers la Leipzig Corpora Collection (Goldhahn *et al.*, 2012). Nous utilisons 30 000 phrases pour toutes les langues (français inclus), sauf le corse, pour lequel nous concaténons les 10 000 phrases disponibles de 2014, 2016 et 2021 (toujours de Wikipédia), pour aboutir à 30 000 phrases. Ce choix signifie également que nous utilisons le même texte monolingue en français pour les six paires.

Pour les phrases *parallèles* à injecter, nous utilisons principalement les corpus disponibles dans la collection OPUS (Tiedemann, 2012). La colonne OPUS du tableau 1 présente, pour chaque langue

4. Nous utilisons le terme « comparable » pour désigner un domaine et contexte similaire pour deux langues.

régionale étudiée, le nombre *total* de phrases parallèles avec le français, accessibles sur le site en mars 2026 [5]. Nous pouvons observer que les caractéristiques mentionnées précédemment (niveaux de vitalité et de ressources pour les langues) ne correspondent pas nécessairement à la réalité des données disponibles pour notre tâche. Si le breton ou l'occitan sont considérés comme étant en danger et peu dotés, ils présentent tous deux plus d'un million de phrases parallèles, tandis que le corse et le picard présentent moins de mille paires de traduction, malgré une classification similaire.

Nous sélectionnons, toutefois, une seule source de corpus par paire de langues. Pour le breton, un corpus parallèle spécifique et conséquent était disponible : le corpus OfisPublik (Tyers, 2009). Pour le basque, nous utilisons le corpus parallèle TED2020. Comme les données parallèles disponibles sur OPUS sont de taille plus réduite pour le corse et le picard (voir tableau 1), nous concaténons les corpus Wikimedia et Tatoeba (ce dernier ne contenant que huit phrases pour le corse). Pour l'occitan, nous utilisons le corpus Wikimedia correspondant. Enfin, pour l'alsacien, comme OPUS ne présente pas même 50 paires de phrases pour le suisse allemand, nous choisissons une source de corpus différente : le corpus théâtral parallèle pour l'alsacien, issu du projet DIVITAL (Bernhard, 2026).

Traitement des données Nous avons tout d'abord vérifié manuellement certains des corpus parallèles *bruts* (pour le corse et l'occitan), du fait de la proximité linguistique avec le français, l'espagnol ou l'italien, afin de contrôler partiellement les phrases qui seront injectées. En effet, ces deux derniers corpus présentaient à l'origine des erreurs manifestes, comme, par exemple, une différence claire de longueur des phrases, avec des éléments non traduits dans l'autre langue.

Nous avons, de ce fait, pré-traité les corpus bilingues pour les six paires de langues afin de supprimer le bruit et d'améliorer leur qualité. Nous avons filtré les phrases dupliquées, trop courtes ou longues, en dessous de 10 mots (ou de 5 pour le corse et le picard, afin de conserver au moins 200 paires) et au-dessus de 80. Pour les phrases parallèles, nous écartons, de plus, les paires présentant un écart significatif de longueur (mauvais alignement ou traduction partielle) ou un chevauchement trop élevé de caractères entre les parties source et cible (par exemple, dû à des fragments cités ou une abondance de noms propres transparents). Enfin, nous procédons également à une identification de langue avec l'outil GlotLID (Kargaran *et al.*, 2023), spécifiquement entraîné pour les langues peu dotées. Cependant, comme les scores rapportés par Kargaran *et al.* (2023) pour l'alsacien et le picard sont légèrement inférieurs à ceux des quatre autres langues, nous ne leur appliquons pas ce filtre.

Une fois ce pré-traitement effectué, nous créons les corpus en maintenant une proportion aux alentours de 5-7 % de phrases parallèles injectées aléatoirement dans les textes monolingues. Puis, nous affectons 25 % des phrases pour l'entraînement et les 75 % restants pour le test, afin de reproduire une situation réelle de fouille de phrases. Un jeu de données existant mais de petite taille (ici, d'entraînement) permet de calibrer l'outil de fouille à déployer sur un corpus plus grand (ici, de test). Le tableau 2 présente la taille de chacun des corpus constitués.

5. Lorsque plusieurs versions d'un corpus existent, nous comptabilisons uniquement la plus grande.

		bre-fr	cos-fr	eus-fr	gsw-fr	oci-fr	pcd-fr
entraînement	source	5 881	2 749	5 309	5 234	5 996	1 091
	cible (FR)	5 981	3 276	5 968	6 216	5 956	1 298
	dont parallèle	374	136	361	324	349	59
test	source	17 647	8 260	15 934	15 704	17 976	3 279
	cible (FR)	17 948	9 835	17 911	18 650	17 873	3 900
	dont parallèle	1 124	413	1 087	974	1 049	182

TABLE 2 – Statistiques sur les corpus d'évaluation synthétiques pour les six paires de langues.

3 Fouille de phrases parallèles

3.1 Système de fouille de phrases

Nous utilisons le système de fouille de phrases présenté par Okabe & Fraser (2025), qui étend l'approche de Hangya & Fraser (2019) par des plongements contextuels (et non statiques) de phrases. Le système de fouille en lui-même permet d'obtenir des paires de phrases de qualité, lorsque les langues sont bien alignées entre elles. La performance dépend, de fait, de la *représentation* des langues par les modèles utilisés lors de l'encodage des phrases.

Une fois les phrases source (langue régionale) et cible (français) représentées par des plongements grâce au modèle choisi, le système calcule la similarité entre deux représentations de phrases grâce au score CSLS (« *Cross-Domain Similarity Local Scaling* » en anglais ; Conneau *et al.*, 2018) qui permet notamment de mieux différencier les paires de traduction des phrases simplement similaires. L'implantation de cette partie repose sur la bibliothèque `faiss` (Johnson *et al.*, 2019). Enfin, une paire de phrases est considérée comme étant parallèle lorsque ce score est supérieur à un seuil τ, à définir sur la base d'entraînement. Suivant Hangya & Fraser (2019), nous définissons ce seuil de manière dynamique dans l'équation (1), en fonction de la moyenne M et de l'écart-type σ des similarités sur le corpus :

$$\tau = M + \lambda \times \sigma, \tag{1}$$

où λ est le méta-paramètre à ajuster (sa valeur par défaut est de 2,0).

Représentation des phrases Afin d'évaluer l'impact d'un pré-entraînement uniquement *monolingue* sur la langue étudiée, nous comparons tout d'abord deux modèles de langue : XLM-RoBERTa ou XLM-R (base) (Conneau *et al.*, 2020) ainsi que Glot500-m (Imani *et al.*, 2023). Ce dernier peut être considéré comme une extension du premier pour davantage de langues (plus de 500), notamment peu dotées. De plus, nous avons étudié le modèle mmBERT (base) (Marone *et al.*, 2025), multilingue et plus récent, qui a été entraîné sur plus de 1 800 langues. Pour ces trois modèles de langue, nous considérons la moyenne des représentations de mots afin de représenter une phrase. Par ailleurs, nous présentons également les résultats obtenus avec LaBSE (Feng *et al.*, 2022), un encodeur de phrases multilingue de l'état de l'art [6], afin d'avoir une référence. Grâce à son entraînement sur des corpus parallèles avec un objectif contrastif, ce modèle obtient les meilleurs scores de fouille pour des paires

6. Nos expériences préliminaires avec LaBSE ont montré une meilleure performance que LASER2 (Heffernan *et al.*, 2022) et multilingual-e5-base (Wang *et al.*, 2024).

de langues peu dotées sur le jeu de données d'évaluation BELOPSEM (Okabe *et al.*, 2025). Pour une comparaison juste entre les modèles, nous nous concentrons uniquement sur les encodeurs produisant des représentations de phrase de taille 768.

Les quatre dernières colonnes du tableau 1 indiquent les langues présentes dans les données de pré-entraînement de chacun des quatre modèles. Nous avons également choisi ces six langues car elles sont présentes de manière variable dans les corpus d'entraînement de ces modèles, notamment le breton pour XLM-R ou le corse pour LaBSE. Notons que les six langues étudiées font partie des langues de pré-entraînement de Glot500-m et mmBERT. Pour ce dernier, la plupart des langues (toutes, hormis le basque et l'alémanique) font uniquement partie de l'ultime étape du pré-entraînement : la phase de décroissance (« *decay phase* »). Il s'agit d'une partie hautement multilingue, portant le nombre total de langues observées de 110 jusqu'alors (deuxième phase, où le basque et l'alémanique sont présents) à 1 833, incluant un large éventail de langues peu dotées.

3.2 Résultats expérimentaux de fouille

En suivant le protocole du défi partagé BUCC (Zweigenbaum *et al.*, 2017), nous évaluons les modèles grâce aux métriques usuelles de précision (P), rappel (R) et F-score (F). Nous interpréterons cependant principalement le F-score dans cet article. Le tableau 3 présente les scores obtenus pour les six paires de langues à l'aide des quatre modèles présentés en section 3.1 sur la partie test de nos corpus. Le méta-paramètre optimal λ lié au seuil minimum de similarité τ a été défini séparément pour chaque modèle et paire de langues, en maximisant le F-score sur la base d'entraînement.

	bre-fr	cos-fr	eus-fr	gsw-fr	oci-fr	pcd-fr	moy.
mmBERT	1,98	25,21	6,98	12,30	65,88	39,71	25,34
XLM-R	5,20	16,06	41,14	14,44	51,98	29,75	26,43
Glot500-m	13,47	49,94	23,56	15,60	86,85	44,59	39,00
LaBSE	**42,07**	**90,32**	**98,48**	**69,24**	**97,43**	**65,81**	**77,23**

TABLE 3 – Résultats (F-scores) et moyennes obtenus sur les six corpus synthétiques de test pour la fouille de phrases parallèles, selon le modèle de base choisi pour la représentation des phrases.

Tout d'abord, pour les représentations de phrases moyennées, nous observons que Glot500-m parvient à atteindre de meilleurs F-scores que XLM-R, hormis en basque. Le pré-entraînement spécifique pour les langues peu dotées semble donc bénéfique, tandis que le basque semble en être impacté négativement. Le modèle mmBERT, malgré son pré-entraînement plus ambitieux pour les langues peu dotées, 1 833 contre 511 pour Glot500-m, ne parvient pas à être compétitif avec ce dernier pour la fouille de phrases parallèles.

LaBSE obtient ici systématiquement le meilleur score pour les six paires de langues, avec parfois plus de 50 points de différence avec Glot500-m. Le modèle semble avoir été aidé par son entraînement explicitement au niveau des phrases et par son objectif d'alignement de phrases parallèles. Ce résultat est également en accord avec Okabe *et al.* (2025), où LaBSE apparaissait comme premier choix pour la fouille de phrases pour des paires de langues proches (notamment de la même famille linguistique).

Analyse par langue Bien que les six corpus ne soient pas comparables entre eux, l'occitan apparaît comme la langue la mieux représentée à travers les modèles de langue étudiés, avec en moyenne 76 points de F-score. Cette langue romane est en effet proche de l'espagnol et du catalan, tous deux présents lors du pré-entraînement des quatre modèles.

Trois langues présentent un F-score aux alentours de 40 en moyenne : le corse, le basque et le picard, malgré des tendances diverses. Le corse est représenté de manière médiocre par les modèles de langue standard, alors que LaBSE obtient un F-score au-delà de 90, grâce à son entraînement sur des données aussi bien monolingues que bilingues dans la langue (voir tableau 1).

Le basque, bien que présent dans les langues de pré-entraînement des quatre modèles, semble être une paire plus difficile pour les modèles standards. On observe, de fait, un lien négatif entre le nombre de langues lors du pré-entraînement et le score de fouille. Nous pouvons supposer qu'en tant qu'isolat, il présente un mauvais alignement cross-lingue et souffre plus particulièrement de la « malédiction du multilinguisme » (Conneau *et al.*, 2020). Cela ne semble pas être le cas pour LaBSE, en revanche, grâce à son pré-entraînement sur des phrases principalement bilingues anglais-basque.

Le picard présente enfin la plus faible variance des F-scores parmi ces trois langues. On observe, de fait, un score nettement plus bas obtenu par LaBSE, mais la langue semble bénéficier de sa proximité avec le français, même pour XLM-R. Il s'agit là du troisième score le plus élevé obtenu par XLM-R après le basque (vu lors du pré-entraînement) et l'occitan (proche de plusieurs langues romanes).

Toutefois, nous observons que la proximité linguistique ne suffit pas à obtenir une représentation robuste, car les performances sont plus faibles pour l'alsacien, proche de l'allemand. Il s'agit également de la paire pour laquelle l'avantage de Glot500-m est le moins prononcé par rapport à XLM-R. Bien qu'il s'agisse d'une langue indo-européenne, cette baisse pourrait être expliquée par les différences linguistiques, entre langues romanes et germaniques.

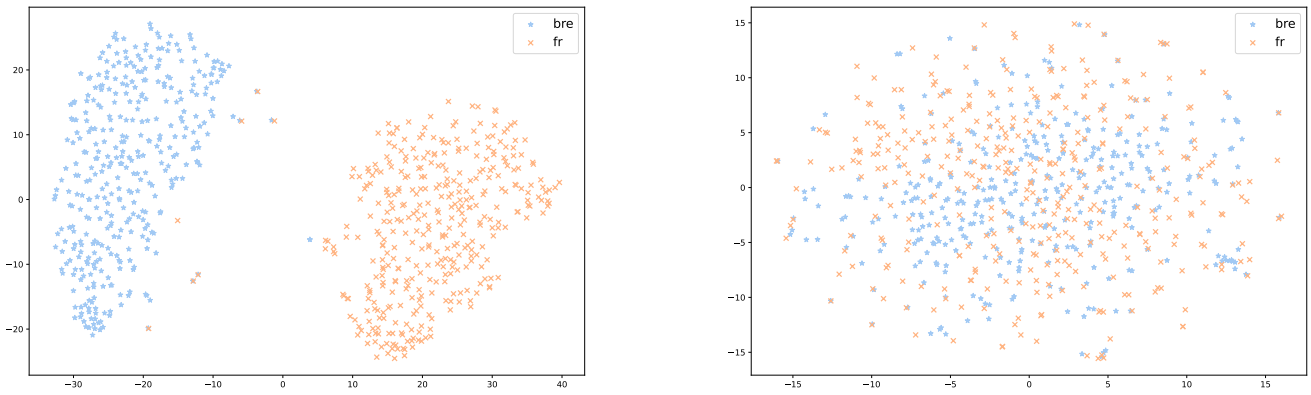
Enfin, le breton est la langue la moins bien représentée par les différents modèles, où même LaBSE atteint moins de 50 points. Le score est également bas pour XLM-R qui a pourtant été pré-entraîné sur la langue. Comme il s'agit d'une langue celtique, qui ne présente pas de langues très bien dotées (contrairement aux langues romanes ou germaniques), la représentation semble en être affectée.

En somme, nous constatons que le pré-entraînement (monolingue) permet d'améliorer la qualité de représentation des paires de phrases (Glot500-m), lorsque des transferts multilingues s'opèrent. La proximité linguistique au sein même de la paire (picard-français) ou avec des langues mieux dotées du pré-entraînement semble également bénéfique pour la tâche de fouille. Enfin, LaBSE, par son entraînement au niveau des phrases et son objectif d'alignement cross-lingue, permet de bien apparier les phrases dans des langues même absentes de son pré-entraînement. Nous observons également que, bien que les domaines des corpus parallèles soient différents des corpus monolingues extraits de Wikipédia, pour le breton et l'alsacien, notre jeu de données d'évaluation reste d'une difficulté appropriée.

3.3 Vers un meilleur alignement cross-lingue

3.3.1 Amélioration de l'isotropie

Afin d'améliorer la qualité de représentation des phrases obtenues en calculant la moyenne des mots, nous appliquons une transformation CBIE (« *cluster-based isotropy enhancement* » ; Rajaee &



(a) Glot500-m

(b) Glot500-m après la transformation CBIE

FIGURE 1 – Représentation t-SNE des phrases parallèles breton-français, avant et après la transformation CBIE des plongements obtenus avec Glot500-m.

Pilehvar, 2021) ou un blanchiment (« *whitening* » ; Huang *et al.*, 2021). En effet, Hämmerl *et al.* (2023) notent une amélioration de l'isotropie à travers ces procédés, aboutissant à un meilleur alignement multilingue et appariement de phrases. Nous étendons donc cette approche à la tâche plus complexe de fouille de phrases, étant donné que l'approche CBIE a déjà été bénéfique sur le jeu de données d'évaluation de trois langues peu dotées (Okabe *et al.*, 2025).

Nous appliquons la transformation CBIE aux représentations de phrases obtenues avec Glot500-m. Celle-ci détermine tout d'abord des groupements (« *clusters* ») sur l'ensemble du corpus, puis supprime les 12 premières directions dominantes *par groupement* (et non sur l'ensemble du corpus). Nous utilisons alors ces nouveaux vecteurs dans l'outil de fouille pour calculer la similarité entre les phrases. De même, nous utilisons le blanchiment sur les plongements de phrases de Glot500-m. Nous considérons ces deux transformations de manière séparée.

Visualisation La figure 1 présente la visualisation t-SNE obtenue pour les phrases parallèles d'entraînement pour la paire breton-français [7]. À gauche, les plongements de Glot500-m apparaissent comme deux groupements distincts, avec très peu de paires alignées visibles, tandis qu'à droite, après la transformation CBIE, nous observons deux groupements qui se chevauchent, suggérant un meilleur alignement cross-lingue.

Résultats de la fouille Nous observons dans le tableau 4, l'impact de ces deux post-traitements sur les représentations de Glot500-m. Tout d'abord, la transformation CBIE, peu coûteuse, permet d'améliorer la qualité des phrases parallèles obtenues avec Glot500-m de manière notable pour les six paires. Ce résultat confirme la tendance constatée par Okabe *et al.* (2025), où l'amélioration de l'isotropie pour ces modèles de langue se traduit par de meilleurs appariements de phrases.

De plus, le blanchiment se révèle être légèrement plus efficace que le traitement CBIE pour renforcer l'alignement cross-lingue, avec une marge de plus de 5 points de F-score en moyenne. Le bénéfice dépend toutefois des langues source : le picard et l'occitan présentent une hausse d'au plus 7 points en moyenne, tandis que le basque et le corse observent une amélioration de plus de 20 points. Nous

7. Nous observons le même phénomène pour le picard, l'alsacien et le basque, mais pas pour le corse et l'occitan.

	bre-fr	cos-fr	eus-fr	gsw-fr	oci-fr	pcd-fr	moy.
Glot500-m	13,47	49,94	23,56	15,60	86,85	44,59	39,00
+ CBIE	26,28	69,80	41,70	22,06	93,65	46,65	50,02
+ blanchiment	35,44	72,14	48,30	31,83	94,11	51,76	55,60
LaBSE	42,07	**90,32**	**98,48**	**69,24**	97,43	**65,81**	**77,23**
Glot500+LaBSE	**53,12**	86,90	89,97	66,60	**97,73**	63,46	76,30

TABLE 4 – Résultats (F-scores) obtenus sur les six corpus synthétiques de test pour la fouille de phrases parallèles à travers une amélioration de l'isotropie (soit par le traitement CBIE, soit par blanchiment) et la distillation de connaissances.

notons ici que ces deux transformations sont plus efficaces pour les paires de langues plus distantes : le français, langue romane occidentale (tout comme l'occitan et le picard), est de fait plus éloigné du breton, de l'alsacien ou du basque. Par ailleurs, LaBSE reste plus performant pour les six paires de langues étudiées.

3.3.2 Distillation de connaissances

Nous employons également une approche de distillation de connaissances (Reimers & Gurevych, 2020), en considérant LaBSE comme modèle enseignant et Glot500-m comme modèle élève. Nous nommerons cette configuration Glot500+LaBSE. Notre objectif ici est d'évaluer si les atouts des deux modèles peuvent être combinés : une représentation robuste et cross-lingue de phrases par LaBSE ainsi qu'une couverture réelle et meilleure de langues peu dotées par le pré-entraînement *monolingue* de Glot500-m. Pour la distillation, un corpus parallèle est nécessaire, or, pour trois de nos langues, leur taille est minime (voir tableau 1). Nous n'utiliserons donc pas de phrases parallèles *directes* (comme le corse-français), mais indirectes (comme l'italien-français), en nous appuyant sur les langues voisines et mieux dotées. Ceci aboutirait à un modèle LaBSE étendu aux langues peu dotées, en nécessitant uniquement des phrases monolingues (Glot500-m). Nous utilisons pour cela 100 000 paires de phrases du corpus `sentence-transformers/parallel-sentences-talks`, où l'anglais est utilisé comme pivot avec le français, l'allemand (pour sa proximité avec l'alsacien), l'espagnol (pour l'occitan) et l'italien (pour le corse).

Résultats de la fouille Dans le tableau 4, nous observons que la distillation améliore significativement le résultat de la fouille pour les six paires de langues par rapport à Glot500-m. En moyenne, le F-score est augmenté de 37 points, allant de 10 pour l'occitan, déjà élevé, à 66 pour le basque. Nous observons également les mêmes tendances que pour l'amélioration de l'isotropie : les langues romanes occidentales en bénéficient moins que les langues plus éloignées du français. La distillation permet donc de mieux aligner ces représentations et agit positivement pour la fouille de phrases parallèles. Néanmoins, les résultats restent sensiblement en deçà de LaBSE (moins d'un point de différence en moyenne) sur notre jeu d'évaluation. Notons que pour le breton, la différence est notable et en faveur de Glot500+LaBSE. Il s'agit, en effet, d'une paire éloignée du français, n'ayant pas de langue voisine bien dotée (contrairement à l'alsacien) et non représentée par LaBSE (contrairement au basque). Dans ce cas de figure précis, l'approche par distillation permet de surpasser les performances

de base de LaBSE, sans même avoir recours à des phrases parallèles avec le breton.

3.4 Analyse qualitative

cos	A catedrale hè iscritta in u 1929 cum'è munumentu storicu è classata in u 2000 .
XLM-R	Repris, il est réintégré 12 ventôse an II sur ordre de l'administration de police.
Glot500-m	Cette fontaine, construite au et restaurée en 2015, alimentait l'ancien lavoir communal.
+ blanchiment	L'édifice est classé au titre des monuments historiques en 2000.
Glot500+LaBSE & référence	Après une inscription en 1929, fait l'objet d'un classement au titre des monuments historiques depuis le 3 février 2000.

FIGURE 2 – Exemple de paire de phrases retrouvée avec XLM-R ou les variantes de Glot500-m. La traduction de la phrase corse en français correspond à la phrase identifiée par Glot500+LaBSE.

La figure 2 présente une paire de phrases corse-français, où la phrase associée par le modèle ne correspond pas à la référence lorsque XLM-R ou des variantes de Glot500-m sont utilisés. La traduction en français est correctement retrouvée par Glot500+LaBSE (et LaBSE). En comparant les différentes phrases en français, nous remarquons que les phrases extraites se rapprochent progressivement de la traduction. XLM-R, qui n'a pas été pré-entraîné sur le corse, a seulement en commun avec la traduction, la présence d'une année. Les variantes de Glot500-m sans ajustement traitent toutes deux du thème d'un site patrimonial ; le blanchiment permet, par ailleurs, de retrouver une phrase portant exactement sur un classement aux monuments historiques en l'an 2000. Nous notons ici que la version française n'est pas une traduction exacte mais très proche ; nous avons conservé de telles paires du fait de la taille du corpus parallèle de référence en corse.

4 Conclusion

Nous avons effectué une première étude de la tâche de fouille de phrases parallèles pour six langues régionales de France métropolitaine parmi les mieux représentées en TAL, appariées avec le français. Pour mieux évaluer la qualité des outils utilisés, nous avons généré des corpus synthétiques en introduisant dans des textes monolingues, des phrases issues de corpus parallèles répertoriés. Puis, nous utilisons une version mise à jour d'un système de fouille de phrases parallèles, en comparant quatre modèles de langue multilingues pour représenter les phrases et calculer leur similarité.

Nous notons que le pré-entraînement monolingue standard sur les langues source est bénéfique lorsqu'un transfert cross-lingue est possible, comme pour Glot500-m. Les représentations moyennées sont, toutefois, moins performantes que LaBSE pour notre jeu d'évaluation. Nous considérons également deux méthodes pour améliorer l'alignement entre les langues dans les représentations de Glot500-m, sans nécessiter de phrases parallèles *directes* : deux transformations pour rendre les plongements plus isotropes (CBIE et blanchiment) ainsi que la distillation de connaissances entre LaBSE et Glot500-m. Les trois stratégies permettent d'atteindre de meilleurs scores pour les six paires de langues par rapport à la performance de base de Glot500-m. Cependant, même la distillation,

qui aboutit pourtant au F-score le plus élevé en moyenne pour un modèle fondé sur Glot500-m, ne permet pas de surpasser LaBSE. Notre approche est, néanmoins, bénéfique pour le breton.

Limites de notre approche Le tableau 3 suggère que la tâche de fouille de phrases parallèles semble relativement aisée sur nos corpus générés actuels (hormis pour la paire breton-français) pour les meilleurs modèles. La différence de domaines entre les textes monolingues et parallèles peut également être un facteur partiellement explicatif. Cependant, nous observons que dans les cas les plus divergents, comme pour l'alsacien (Wikipédia et texte théâtral), la fouille reste difficile pour ces modèles.

De plus, les modèles que nous considérons représentent les phrases en calculant la moyenne des plongements de mots, ce qui reste sous-optimal par rapport aux encodeurs de phrases. De fait, ces derniers sont entraînés avec un objectif explicite sur des corpus parallèles conséquents.

Par ailleurs, l'une des particularités essentielles que nous ne prenons pas en compte dans notre approche réside dans les variations orthographiques ou géographiques. Une étude dédiée nécessite, cependant, une annotation des phrases selon les variétés (comme les dialectes de l'occitan), peu disponible.

Travaux futurs Ce travail constitue une première étape vers une extension à d'autres langues régionales de France, telles que le poitevin-saintongeais, qui fait l'objet de travaux récents de la communauté TAL, comme ParCoLab (Stosic *et al.*, 2024). Ce dernier comporte par ailleurs un corpus parallèle pour quatre langues régionales de France : l'alsacien (que nous utilisons), le corse, l'occitan et le poitevin-saintongeais. D'autre part, nous complexifierons les corpus générés, en reproduisant une approche similaire à Chen *et al.* (2023), qui introduisent des phrases volontairement proches, mais fondamentalement différentes des vraies traductions, en remplaçant, par exemple, les nombres ou les entités nommées. En effet, dans notre cas, les F-scores pour trois des six paires sont déjà au-delà de 90 avec LaBSE. Un jeu de données d'évaluation plus complexe nous permettrait de mieux évaluer la robustesse de nos approches.

Remerciements

Nous remercions les relecteurs anonymes pour leurs commentaires ainsi que Delphine BERNHARD pour son aide concernant l'alsacien. Ce travail a été en partie financé par l'Union européenne (ERC, EPICAL, 101141712). Les points de vue et opinions exprimés n'engagent toutefois que leurs auteurs et ne reflètent pas nécessairement ceux de l'Union européenne ou du Conseil européen de la recherche. Ni l'Union européenne ni l'autorité de financement ne sauraient en être tenues pour responsables.

Références

BERNHARD D. (2026). Theatrical parallel corpus for alsatian - divital project. DOI : 10.34847/NKL.8DBD134Z.

CHEN M., HEFFERNAN K., ÇELEBI O., MOURACHKO A. & SCHWENK H. (2023). xSIM++ : An improved proxy to bitext mining performance for low-resource languages. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 101–109, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-short.10.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

CONNEAU A., LAMPLE G., RANZATO M., DENOYER L. & JÉGOU H. (2018). Word translation without parallel data. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

EBERHARD D. M., SIMONS G. F. & ROBINSON A. J. (2026). *Ethnologue : Languages of the World*. Dallas, Texas : SIL International, 29e édition. Version en ligne.

FENG F., YANG Y., CER D., ARIVAZHAGAN N. & WANG W. (2022). Language-agnostic BERT sentence embedding. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 878–891, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.62.

GOLDHAHN D., ECKART T. & QUASTHOFF U. (2012). Building large monolingual dictionaries at the Leipzig corpora collection : From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 759–765, Istanbul, Turkey : European Language Resources Association (ELRA).

HÄMMERL K., FASTOWSKI A., LIBOVICKÝ J. & FRASER A. (2023). Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Findings of the Association for Computational Linguistics : ACL 2023*, p. 7023–7037, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.439.

HANGYA V. & FRASER A. (2019). Unsupervised parallel sentence extraction with parallel segment detection helps machine translation. In A. KORHONEN, D. TRAUM & L. MÀRQUEZ, Éds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1224–1234, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1118.

HEFFERNAN K., ÇELEBI O. & SCHWENK H. (2022). Bitext mining using distilled sentence representations for low-resource languages. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2022*, p. 2101–2112, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-emnlp.154.

HUANG J., TANG D., ZHONG W., LU S., SHOU L., GONG M., JIANG D. & DUAN N. (2021). WhiteningBERT : An easy unsupervised sentence embedding approach. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 238–244, Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-emnlp.23.

IMANI A., LIN P., KARGARAN A. H., SEVERINI S., JALILI SABET M., KASSNER N., MA C., SCHMID H., MARTINS A., YVON F. & SCHÜTZE H. (2023). Glot500 : Scaling multilingual corpora and language models to 500 languages. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds.,

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), p. 1082–1117, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.61.

JOHNSON J., DOUZE M. & JÉGOU H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, **7**(3), 535–547.

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

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

KOEHN P., CHAUDHARY V., EL-KISHKY A., GOYAL N., CHEN P.-J. & GUZMÁN F. (2020). Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In L. BARRAULT, O. BOJAR, F. BOUGARES, R. CHATTERJEE, M. R. COSTA-JUSSÀ, C. FEDERMANN, M. FISHEL, A. FRASER, Y. GRAHAM, P. GUZMAN, B. HADDOW, M. HUCK, A. J. YEPES, P. KOEHN, A. MARTINS, M. MORISHITA, C. MONZ, M. NAGATA, T. NAKAZAWA & M. NEGRI, Éds., *Proceedings of the Fifth Conference on Machine Translation*, p. 726–742, Online : Association for Computational Linguistics.

KOEHN P., GUZMÁN F., CHAUDHARY V. & PINO J. (2019). Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. 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 3 : Shared Task Papers, Day 2)*, p. 54–72, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5404.

MARONE M., WELLER O., FLESHMAN W., YANG E., LAWRIE D. & DURME B. V. (2025). mmbert : A modern multilingual encoder with annealed language learning.

OKABE S. & FRASER A. (2025). Bilingual sentence mining for low-resource languages : a case study on upper and Lower Sorbian. In J. LACHLER, G. AGYAPONG, A. ARPPE, S. MOELLER, A. CHAUDHARY, S. RIJHWANI & D. ROSENBLUM, Éds., *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 11–19, Honolulu, Hawaii, USA : Association for Computational Linguistics.

OKABE S., HÄMMERL K. & FRASER A. (2025). Improving parallel sentence mining for low-resource and endangered languages. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 196–205, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-short.17.

PIERRE ZWEIGENBAUM S. S. & RAPP R. (2018). Overview of the third bucc shared task : Spotting parallel sentences in comparable corpora. In R. RAPP, P. ZWEIGENBAUM & S. SHAROFF, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France : European Language Resources Association (ELRA).

RAJAEE S. & PILEHVAR M. T. (2021). A cluster-based approach for improving isotropy in contextual embedding space. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers), p. 575–584, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-short.73.

REIMERS N. & GUREVYCH I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 4512–4525, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.365.

STOSIC D., MARJANOVIĆ S., BERNHARD D., BACH X., BRAS M., KEVERS L., RETALI-MEDORI S., VERGEZ-COURET M. & WERNER C. (2024). The ParCoLab parallel corpus and its extension to four regional languages of France. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 16014–16023, Torino, Italia : ELRA and ICCL.

TAN W., HEFFERNAN K., SCHWENK H. & KOEHN P. (2023). Multilingual representation distillation with contrastive learning. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 1477–1490, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.108.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey : European Language Resources Association (ELRA).

TYERS F. M. (2009). Rule-based augmentation of training data in Breton-French statistical machine translation. In L. MÀRQUEZ & H. SOMERS, Éds., *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain : European Association for Machine Translation.

UNESCO (2010). *Atlas des langues en danger dans le monde*. Paris, France, 3e édition.

WANG L., YANG N., HUANG X., YANG L., MAJUMDER R. & WEI F. (2024). Multilingual e5 text embeddings : A technical report.

ZWEIGENBAUM P., SHAROFF S. & RAPP R. (2017). Overview of the second BUCC shared task : Spotting parallel sentences in comparable corpora. In S. SHAROFF, P. ZWEIGENBAUM & R. RAPP, Éds., *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, p. 60–67, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/W17-2512.