HU Jinfeng

Projet TAL : Extraction des Expressions Idiomatiques Chinoises et Leurs Traductions Françaises Correspondantes dans le Cadre du Projet DiCoP

1 Présentation de la tâche

Mon projet TAL est mis en œuvre dans le cadre du projet DiCoP (Dictionnaire et Corpus de la Phraséologie) proposé par Madame LIU Lian. Le projet DiCoP, acronyme de « Dictionnaire et Corpus de la Phraséologie », a pour objectif principal le développement d'un dictionnaire électronique multilingue axé sur les unités phraséologiques, qui est actuellement focalisé sur le français-chinois et le chinois-français. La recherche se concentre principalement sur la traduction du chinois vers le français, en mettant particulièrement l'accent sur les unités phraséologiques chinoises et leurs équivalentes françaises.

Ma tâche consiste simplement à trouver un outil ou modifier d'un outil similaire qui puisse d'abord identifier et extraire automatiquement les unités phraséologiques chinoises (principalement des expressions idiomatiques à quatre lettres ou de plus) dans la version chinoise d'un œuvre, puis le relier automatiquement aux traductions françaises correspondantes dans la version de traduction française, et enfin aboutir à une liste contenant les unités phraséologiques chinoises et leurs traductions françaises. Cette tâche réduit la charge de travail de l'annotation manuelle et fournit un corpus aligné pour le projet DiCoP. Voici un exemple du corpus et le traitement manuel de ma tâche.

Corpus chinois :

"红色联合"对"四·二八兵团"总部大楼的攻击已持续了两天，他们的旗帜在大楼周围躁动地飘扬着，仿佛渴望干柴的火种。"红色联合"的指挥官心急如焚，他并不惧怕大楼的守卫者，那二百多名"四·二八"战士，

与诞生于 l966 年初、经历过大检阅和大串联的“红色联合”相比要稚嫩许多。他怕的是大楼中那十几个大铁炉子，里面塞满了烈性炸药，用电雷管串联起来，他看不到它们，但能感觉到它们磁石般的存在，开关一合，玉石俱焚，而“四·二八”的那些小红卫兵们是有这个精神力量的。比起已经在风雨中成熟了许多的第一代红卫兵，新生的造反派们像火炭上的狼群，除了疯狂还是疯狂。

Corpus français :

L’assaut de l’Union rouge contre le quartier général de la brigade du 28 Avril durait depuis déjà deux jours.
Tout autour de l’édifice, les drapeaux de la brigade claquaient au vent, telles des torches attendant d’être ravivées.
L’officier en chef de l’Union rouge était terriblement inquiet, non qu’il ait peur des miliciens en faction dans la tour – les deux cents et quelques hommes de la brigade du 28 Avril ne pesaient pas bien lourd face aux gardes rouges de l’Union, fondée au début de l’année 1966 et dont la puissance économique et militaire était sans rivale.
Ce qu’il craignait, c’étaient les quelques dizaines de fours de métal en fusion entreposés à l’intérieur du bâtiment.
Ceux-ci, reliés par des détonateurs électriques, étaient pleins à ras bord d’explosifs.
Il ne pouvait les voir, mais il pouvait sentir leur présence magnétique.
Un doigt sur un bouton et tous seraient réduits en cendres.
Les jeunes gardes rouges de la brigade du 28 Avril étaient capables d’ une telle folie.

Pour ce corpus aligné, il faut d’abord extraire les unités phraséologiques chinoises, ici, ce sont « 心急如焚 » et « 玉石俱焚 ». Puis, trouver les traductions françaises correspondantes : « était terriblement inquiet » et « tous seraient réduits en cendres ».

S’il existe un outil utile pour faire l’extraction des unités phraséologiques chinoises et françaises, je vais évaluer cet outil est bon ou pas et donner des conseils pour l’amélioration.

Ainsi, les problématiques de ce projet peuvent être :

Comment concevoir un algorithme ou un modèle capable d'identifier avec précision les unités phraséologiques dans le texte chinois et la traduction française correspondante ?

2 État de l'art

2.1 Unité phraséologique

« Les unités phraséologiques sont des éléments de la phrase qui sont construits en transgressant les règles de sélection de leurs constituants lexicaux ou morphologiques. » (Polguère, 2008 : 164) Les unités phraséologiques, ou expressions idiomatiques, jouent un rôle essentiel dans la communication quotidienne et littéraire. Elles sont souvent chargées de significations culturelles et linguistiques spécifiques, ce qui les rend particulièrement difficiles à traduire de manière précise.

La langue chinoise est riche en unités phraséologiques, notamment les chengyu (成语 chéng yǔ) qui sont les expressions toutes faites, et les xiehouyu (歇后语 xiē hòu yǔ) qui sont les expressions en suspens, sont profondément enracinés dans la culture chinoise. (Zhou, 2018) Les chengyu constituent un type particulier d'expression fixe dont le sens n'est généralement pas compris à travers le sens littéral de ses composants. Ils sont souvent le produit de la culture, des coutumes ou de l'histoire, leur signification peut donc être difficile à comprendre pour les locuteurs non natifs. (Zhou, 2004)

Les unités phraséologiques chinoises jouent un rôle important dans la langue chinoise et sont souvent utilisées pour exprimer des concepts abstraits, des histoires ou des significations symboliques. Chaque unité phraséologique chinoise a sa propre histoire ou contexte historique, qui est généralement lié à la culture, à l'histoire, à la tradition ou aux contes populaires chinois (Zhang, 1999). Par exemple, « 千军万马[1] » （qiān jūn wàn mǎ）est un chengyu dont la signification littérale est « des milliers de soldats » et « dix mille chevaux », mais sa signification réelle est de décrire une force

[1] L'exlication de ce chengyu vient de *Dictionary of Chinese Idioms* par NATIONAL ACADEMY for EDUCATIONAL RECHEARCH : https://dict.idioms.moe.edu.tw/idiomView.jsp?ID=962&webMd=2&la=0

forte et la force du nombre. Il y a une histoire historique derrière ce chengyu, qui raconte la magnifique scène de Guan Yu, un célèbre général militaire de la Chine ancienne, il a donc une signification symbolique dans la culture chinoise.

2.2 Dictionnnaire numérique

Un dictionnaire est un « recueil des mots d'une langue ou d'un domaine de l'activité humaine, réunis selon une nomenclature d'importance variable et présentés généralement par ordre alphabétique, fournissant sur chaque mot un certain nombre d'informations relatives à son sens et à son emploi et destiné à un public défini ». (CNRTL) C’est-à-dire, il est un recueil systématique de mots ou d'expressions d'une langue donnée, accompagnés de leurs définitions, de leur prononciation, de leur étymologie et d'autres informations linguistiques pertinentes. Ainsi qu’il sert de référence essentielle pour comprendre, apprendre et utiliser une langue.

Il existe principalement deux types de dictionnaire : le dictionnaire monolingue et le dictionnaire multilingue. Les dictionnaires monolingues fournissent des définitions, des exemples d'utilisation et des informations linguistiques dans une seule langue. Ils aident les locuteurs natifs à approfondir leur compréhension de leur langue maternelle et aident les apprenants à améliorer leurs compétences linguistiques. En ce qui concerne les dictionnaires multilingues, ils offrent des traductions entre plusieurs langues. Ils sont essentiels pour la traduction et la communication interculturelle, fournissant des équivalents lexicaux dans différentes langues.

Les dictionnaires numériques multilingues ont des fonctions pour l'apprentissage d'une langue étrangère et la traduction. Ils sont des outils essentiels pour les apprenants de langues. Ils fournissent des définitions, des exemples d'utilisation, des synonymes, des antonymes et des informations grammaticales pour aider à développer des compétences linguistiques. De nombreux dictionnaires numériques intègrent des fonctionnalités de prononciation audio pour aider les utilisateurs à améliorer leur

prononciation. Les dictionnaires numériques facilitent la traduction entre différentes langues parce qu'ils offrent des traductions rapides et des informations contextuelles.

2.3 Absence d'un dictionnaire numérique pour les unités phraséologies chinois-français

L'évolution des dictionnaires numériques dans le domaine des unités phraséologiques chinois-français est un développement significatif, mais il est important de noter qu'il existe encore une lacune majeure : l'absence d'un dictionnaire dédié spécifiquement à la traduction chinois-français pour les unités phraséologiques. Bien que des dictionnaires unité phraséologique chinois et français ainsi que des dictionnaires bilingues chinois-français soient disponibles, cette absence a des implications importantes pour la qualité des traductions et la compréhension des expressions idiomatiques. (LIU, 2023) Les dictionnaires existants peuvent être utiles pour la compréhension individuelle des langues, mais l'absence d'une ressource dédiée à la traduction dans le contexte phraséologique constitue une limitation majeure pour les traducteurs, ainsi que la qualité de la traduction automatique entre le chinois et le français.

2.4 Méthodes pour identifier les unités phraséologiques

Pour compléter un dictionnaire numérique pour les unités phraséologiques chinois-français, il faut d'abord avoir une grande quantité de corpus de textes parallèles chinois et français pour extraire les unités phraséologies chinoises et la traduction française dans la version française correspondante. (CHEN, 2023)

Les expressions polylexicales peuvent être identifiées dans le corpus parallèle en utilisant différentes techniques d'alignement pour les apparier. (Bouamor, 2013) Les techniques d'extraction monolingue d'expressions polylexicales tournent autour de trois approches : des méthodes symboliques reposant sur des patrons morphosyntaxiques, des méthodes statistiques utilisant des mesures d'association pour classer les

expressions candidates et des méthodes hybrides combinant les deux premières approches. (Bouamor, 2013)

Les méthodes symboliques reposant sur des patrons morphosyntaxiques consistent à extraire des expressions polylexicales à partir de patrons morphosyntaxiques prédéfinis. Ces patrons peuvent être basés sur des règles linguistiques ou sur des statistiques. Les méthodes statistiques utilisent des mesures d'association pour classer les expressions candidates. Ces mesures peuvent être basées sur la fréquence, la co-occurrence ou d'autres critères. Les méthodes hybrides combinent les deux premières approches en utilisant des patrons morphosyntaxiques pour extraire des expressions candidates, puis en utilisant des mesures d'association pour classer ces expressions et sélectionner les meilleures. (Bouamor, 2013)

En conclusion, des unités phraséologiques dans la communication quotidienne et littéraire, ainsi que des dictionnaires numériques dans ce domaine dans la compréhension, l'apprentissage et la traduction des langues sont importants. L'évolution des dictionnaires numériques, de leur origine dans les dictionnaires papier à leur adaptation aux technologies numériques modernes, a facilité l'accès à des informations linguistiques riches et variées. Cependant, l'absence d'un dictionnaire numérique spécifiquement dédié à la traduction des unités phraséologiques chinoises en français constitue toujours une lacune importante. Pour combler cette lacune, il est nécessaire de collecter et d'extraire des unités phraséologiques chinoises à partir de corpus de textes parallèles chinois-français. Les techniques d'extraction peuvent varier, de l'utilisation de patrons morphosyntaxiques à des approches statistiques ou hybrides. La création d'un dictionnaire numérique spécialisé permettrait d'améliorer la qualité des traductions et de faciliter la compréhension interculturelle entre le chinois et le français.

3 Présentation des données

Dans mon projet TAL, je vais utiliser le corpus du roman « Le Problème à Trois Corps » en version originale chinoise de Liu Cixin et sa version française traduite par Gwennaël Gaffric pour extraire les unités phraséologiques chinoises et leurs traductions françaises dans le cadre du projet DiCoP par Madame LIU Lian.

Le roman « Le Problème à Trois Corps » (三体, Sān tǐ) de l'auteur chinois Liu Cixin est une œuvre majeure de la science-fiction contemporaine. « Le Problème à Trois Corps » est une œuvre qui explore des thèmes complexes tels que la communication interstellaire, la politique, et les avancées scientifiques. L'intrigue tourne autour de la découverte d'une civilisation extraterrestre et des conséquences inattendues de cette rencontre. La traduction de ce roman pose des défis uniques en raison de la richesse de la langue chinoise en culture et en idéogrammes. Gwennaël Gaffric, le traducteur en français, a joué un rôle essentiel dans la transmission des subtilités de l'œuvre originale.

Le nombre de mots dans ce roman est 185595 mots en version chinoise et 132115 en version français, qui est une quantité grande pour le projet. La raison pour laquelle j’ai choisi ce texte comme unique corpus dans le cadre de mon projet en linguistique repose sur plusieurs considérations méthodologiques et linguistiques. En se concentrant sur un seul corpus, je vise à créer une analyse approfondie et ciblée des expressions idiomatiques, des tournures phraséologiques et des constructions linguistiques propres au genre de la science-fiction chinoise. La nature complexe de "Le Problème à Trois Corps" offre un terrain fertile pour l'identification et l'extraction d'unités phraséologiques spécifiques, en mettant en évidence la richesse de la langue chinoise dans un contexte littéraire. La comparaison entre la version chinoise originale et sa traduction française contribuera à mettre en lumière les défis spécifiques liés à la transposition interlinguistique des phraséologies. L'analyse des choix de traduction effectués par Gwennaël Gaffric permettra d'identifier les variations et les adaptations

nécessaires pour préserver l'intégrité des expressions phraséologiques tout en les rendant accessibles au lectorat francophone.

4 Descriptif de la méthode

Afin de générer un fichier contenant des unités phraséologiques chinoises accompagnées de leurs traductions françaises correspondantes, plusieurs étapes clés seont mises en œuvre dans le cadre de mon projet TAL. Initialement, je dispose de deux fichiers TXT contenant les versions chinoise et française de l'œuvre « Le Problème à Trois Corps », fournis par Madame Lian Chen, constituant ainsi le corpus brut de départ de mon projet. La figure I est le schéma du processus de traitement du projet TAL.

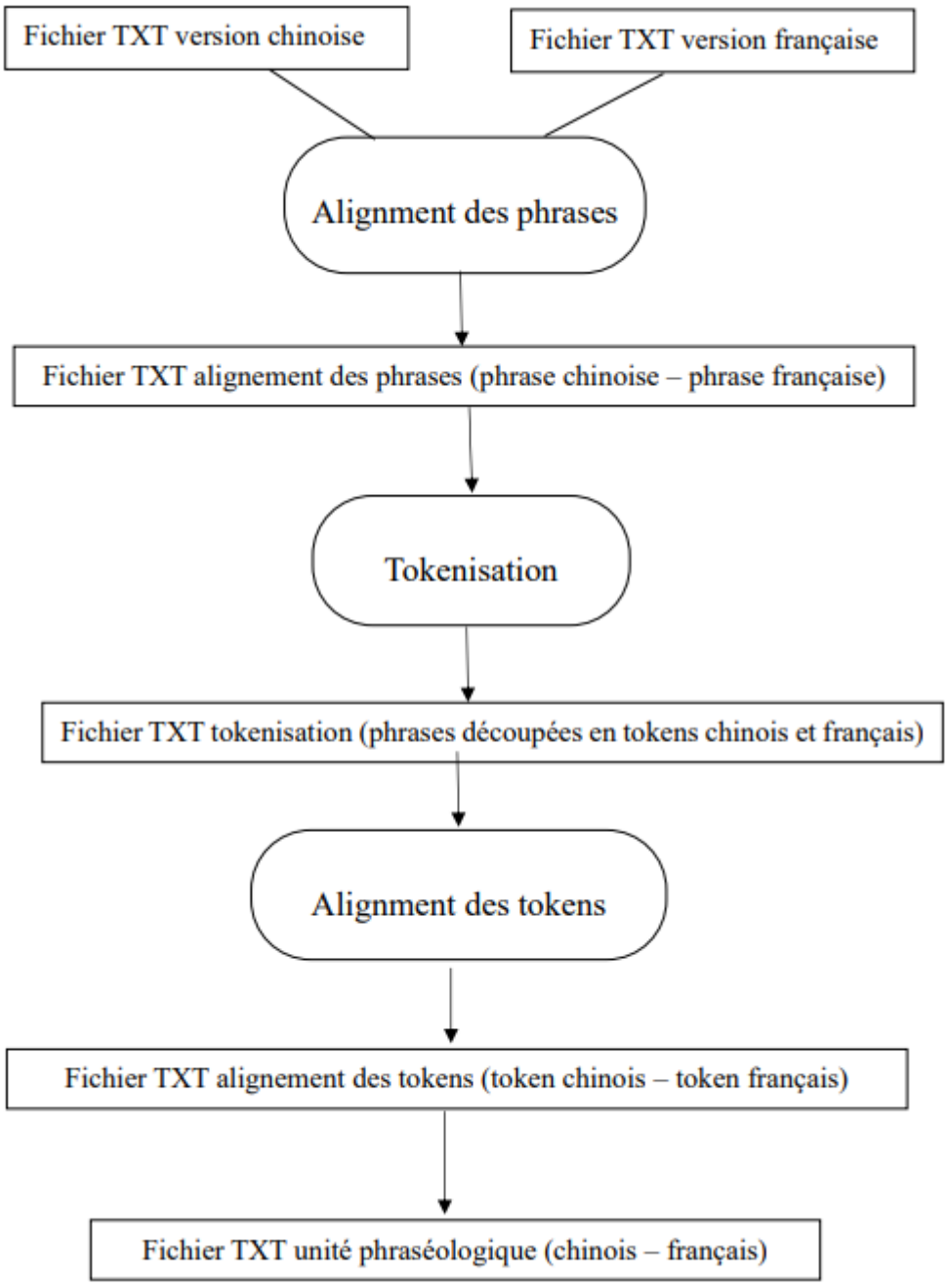

Figure I : Schéma du processus de traitement du projet TAL

La première étape cruciale consiste à aligner les phrases des deux langues afin de réduire le nombre d'unités à traiter lors des étapes ultérieures. À cette fin, l'utilisation de Bertalign[2] se révèle pertinente, étant un outil d'alignement de phrases multilingues.

[2] Lien vers Bertalign : https://github.com/bfsujason/bertalign

Suite à l'alignement des phrases, j'exporterai les résultats obtenus pour procéder à une évaluation rigoureuse. La méthodologie d'évaluation reposera sur les critères d'accuracy (Powers, 2007). Plusieurs phrases, judicieusement sélectionnées au début, au milieu et à la fin du texte aligné, seront examinées afin d'analyser la pertinence du résultat. Cette évaluation permettra de mesurer la qualité de l'alignement des phrases et de confirmer la fiabilité des données traitées.

Par la suite, le processus se poursuivra avec la découpe des tokens pour chaque phrase française, utilisant l'outil spaCy[3]. L'utilisation de spaCy pour la tokenization des phrases françaises permet d'obtenir une segmentation précise du texte en unités linguistiques significatives, facilitant ainsi les étapes suivantes du projet.

Ainsi que pour la tokenisation des phrases chinoises, l'outil THULAC[4] sera utilisé. THULAC (Li & Sun, 2009), développé par le groupe de recherche en linguistique informatique de l'Université de Tsinghua, est un analyseur lexical chinois qui se distingue par sa rapidité et sa précision. Le chinois, contrairement à d'autres langues latines que l'on peut utiliser des outils tels que spaCy et NLTK pour la tokenisation, l'utilisation de cet outil THULAC sur du texte chinois permet d'identifier automatiquement des unités phraséologiques à partir de tokens dans les phrases chinoises, ce qui est important pour ce projet de recherche.

L'alignement des mots sert à établir des correspondances entre les tokens des deux langues, créant ainsi des paires de mots associées. Cette tâche complexe est simplifiée grâce à l'utilisation de l'outil awesome-align[5] (Dou & Neubig, 2021), reconnu pour son efficacité dans l'alignement multilingue. En tirant parti des caractéristiques distinctives de chaque langue, awesome-align identifie les correspondances entre les mots chinois et français, constituant ainsi la base essentielle pour l'extraction ultérieure des unités

[3] Lien vers spaCy : https://spacy.io/
[4] Lien vers THULAC : https://github.com/thunlp/THULAC-Python
[5] Lien vers awesome-align : https://github.com/neulab/awesome-align

phraséologiques chinoises et de leurs traductions françaises.

Cependant, il convient de noter qu'avec l'outil awesome-align, les unités phraséologiques chinoises peuvent indiquer plusieurs mots français simultanément. En d'autres termes, ces unités chinoises ne sont pas directement liées à des expressions françaises spécifiques. Pour surmonter cette limitation, il est nécessaire de développer un algorithme dédié à l'extraction de tous les mots français correspondants à une unité phraséologique chinoise.

De plus, une même unité phraséologique chinoise peut être traduite de différentes manières en expressions françaises. Par conséquent, à la sortie des résultats, le format final sera un fichier TXT contenant un dictionnaire où les unités phraséologiques chinoises agissent en tant que clés et leurs diverses traductions françaises sont répertoriées comme éléments associés à ces clés. Cette approche offre une représentation complète des nuances de traduction pour chaque unité phraséologique chinoise, reflétant la richesse sémantique de la langue.

5 Résultat de performance

J’ai séparé chaque étape en script différent pour évaluer la performance plus facile.

Dans l’évaluation de l’alignement des phrases, Madame LIU m’a aidé à annoter manuellement 60 phrases en tant que référence. Dans ces 60 phrases, il y a 56 phrases ont bien alignés alors que 4 phrases n’ont pas aligné correctement. Ainsi, j’ai obtenu un score d’accuracy élevé, soit 0.93, qui manifeste qu’en utilisant l’outil Bertalign on a l’alignement des phrases entre le chinois et le français précis.

En ce qui concerne l’extraction des unités phraséologiques, ici, j’ai extrait plutôt celles en chinois. J’ai donc annoté 30 unités phraséologiques, mais dans les mêmes phrases alignées, avec l’outil awesome-align, il a identifié 44 unités phraséologiques,

c’est-à-dire la machine a mal compris ce qui sont des unités phraséologiques par des tokens chinois. D’ailleurs, la machine a bien identifié 18 unités phraséologiques en comparaison avec l’annotation manuelle, soit 60% de correction. De plus, on ignore un peu la traduction cent pour cent correcte, sinon il n’y a pas de traduction française correspondante, on a 9 traductions correcte, soit 30%. (Voir Tableau I) En analysant cela, j’ai observé que les erreurs de relier les unités phraséologiques sont d’abord, des tokens français ont parfois enlevé automatique par la machine, comme stopwords, ce qui rend la traduction pas très complète. Ensuite, des unités phraséologiques dans le corpus peuvent apparaître plusieurs fois, mais la machine a tout collecté des tokens correspondants et affiché dans le fichier, ce qui n’est pas une manière intelligente. En plus, c’est le problème de l’alignement des tokens qui cause le problème pour l’étape à sortie de la liste des unités phraséologiques reliées.

	idiom_reference	idiom_predict	idiom_correct	idiom_traduction_correct
num	30	44	18	9
percentage	100%	146.67%	60.00%	30.00%

Tableau I : Statistique sur des données de l’évaluation de qualité d’extraction des unités phraséologiques

6 Limites et Améliorations possibles

Pour fournir la tâche et évaluer pour certaines étapes, j’ai écrit des scripts avec Python différemment dans les documents des outils différents, ce qui est une limite pour les autres à comprendre la logique à faire afin de sortir la liste des unités phraséologiques alignées. Cependant, quand j’essaie de combiner des scripts avec import relatif, j’ai rencontré des problèmes d’importer des modules des outils, donc à la fin, j’ai renoncé à cette façon. Une chose à améliorer est de créer un nouveau script avec import relatif des modules.

La séparation pour chaque étape influence le format de sortie des données. En idéal,

il consiste aux unités phraséologiques alignées et son contexte dans le corpus. Pourtant, je ne peux pas le réaliser dans ce cas-là.

Même si la performance de sortie la liste des unités phraséologiques alignées n’est pas très mauvais, je pense que ces données ne peuvent pas être utilisées directement dans le projet de DiCop car il existe quand même des fautes avec ce moyen, ce qui vont détruire la rigueur d’un dictionnaire. Dans l’avenir, il faut trouver un outil plus efficace dans l’alignement des tokens ou des mots entre le chinois et le français, ou bien concevoir un nouvel outil pour le traiter.

7 Conclusion

En conclusion, le projet TAL visant à l'extraction des unités phraséologiques chinoises et de leurs équivalents français dans le cadre du projet DiCoP a progressé significativement tout en rencontrant quelques défis. Cette initiative s'inscrit dans une démarche cruciale pour le développement d'un dictionnaire électronique multilingue axé sur les unités phraséologiques français-chinois et chinois-français.

Le processus de traitement du corpus du roman « Le Problème à Trois Corps » de LIU Cixin a impliqué plusieurs étapes clés, allant de l'alignement des phrases avec Bertalign à l'extraction des unités phraséologiques chinoises et françaises avec l'aide d'outils tels que THULAC, spaCy, et awesome-align. Les résultats obtenus jusqu'à présent révèlent des succès encourageants, mais des défis subsistent.

L'évaluation de l'alignement des phrases a démontré une précision élevée de 93%, témoignant de l'efficacité de l'outil Bertalign dans la synchronisation des phrases entre les versions chinoise et française. Cependant, des ajustements pourraient être apportés pour améliorer la couverture des phrases mal alignées.

L'étape d'extraction des unités phraséologiques a présenté des résultats encourageants, avec une précision de 60% par rapport à l'annotation manuelle. Les principales sources d'erreur résident dans la gestion des stopwords et des occurrences multiples des unités phraséologiques dans le corpus. Ces problèmes soulignent la nécessité d'une optimisation plus poussée des outils d'alignement et d'extraction pour garantir une meilleure précision.

En outre, la création d'un dictionnaire numérique spécialisé dans la traduction chinois-français pour les unités phraséologiques reste un enjeu majeur. L'absence actuelle de cette ressource dédiée peut impacter la qualité des traductions et la compréhension des expressions idiomatiques. Il est impératif de poursuivre la collecte et l'extraction des unités phraséologiques chinoises à partir de corpus de textes parallèles afin de combler cette lacune.

Des pistes d'amélioration incluent la consolidation des scripts utilisés pour chaque étape du projet, l'exploration de nouveaux outils d'alignement plus performants, et la mise en œuvre d'une méthodologie plus sophistiquée pour gérer les variations de traduction des unités phraséologiques. Le projet TAL a jeté les bases d'une approche innovante pour l'extraction et la traduction des unités phraséologiques chinoises en français. Tout en mettant en lumière les avancées réalisées, la route vers un dictionnaire électronique complet et précis nécessitera des efforts continus pour surmonter les défis spécifiques liés à la complexité linguistique et culturelle des expressions idiomatiques.

8 Références

Bolly, C. (2008). Les Unités Phraséologiques : Un Phénomène Linguistique Complexe ? Séquences (Semi-) Figées Construites Avec Les Verbes Prendre et Donner En Français Écrit L1 et L2 Approche Descriptive et Acquisitionnelle.C.

https://dial.uclouvain.be/pr/boreal/object/boreal:19625/datastream/PDF_08/view

Miguel, M. P. (2017). Les Locuteurs Natifs Parlent En Phrasèmes : Les Différents Types d'unités Phraséologiques. https://gredos.usal.es/bitstream/handle/10366/146557/TFG%20-%20MARIA%20PEREZ%20MIGUEL.pdf?sequence=1&isAllowed=y

Polguère, A. (2008). Lexicologie et Sémantique Lexicale: Notions Fondamentales. Presses de l'Université de Montréal.

Cui, X.-L. (2005) *Unités phraséologiques chinoises et représentation de l'humanité en chinois* [汉语熟语与中国人文世界 Hàn yǔ shú yǔ yǔ zhōng guó rén wén shì jiè]. Beijing: University language and culture press.

Zhou, J. (2018). La Voie et l'importance de La Recherche Sur Les Unités Phraséologiques Chinoises[汉语熟语研究的正轨与要务 hàn yǔ shú yǔ yán jiū de zhèng guǐ yǔ yào wù]. Presses de Linguistique Chinoise.

Zaklani, M. (2020). *Le Dictionnaire Numérique Ou Électronique : Des Outils Pédagogiques Favorisant l'apprentissage Lexical*. https://www.epi.asso.fr/revue/articles/a2004d.htm

Bouamor, D. (2013). *Acquisition de Lexique Bilingue d'expressions Polylexicales: Une Application à La Traduction Automatique Statistique*. https://aclanthology.org/F13-5001.pdf

Zhang, T. (1999). *Totalité et Chronologie Des Expressions Idiomatiques Chinoises[成语的数量及产生年代 chéng yǔ de shù liàng jí chǎn shēng nián dài]*. Linguistic Construction.

Zeng, Z. (2017). *Une Étude Des Paraphrases Franco-Françaises Dans Frhelper [*《法语助手》中法法释义的研究 《 fǎ yǔ zhù shǒu 》 zhōng fǎ fǎ shì yì de yán jiū]. Littérature Nordique.

Liu, C. (2006). *Le Problème à Trois Corps*. ISBN : 978-7-5366-9293-0

Liu, L. (2023). *Totalité et Chronologie Des Expressions Idiomatiques Chinoises*. ASIALEX 2023 The Asian Association for Lexicography.

Définition de dictionnaire dans CNRTL : https://www.cnrtl.fr/definition/dictionnaire

Petite histoire des dictionnaires : https://www.etudes-litteraires.com/histoire-dictionnaires.php

Explication de « 千 军 万 马 » (qiān jūn wàn mǎ) : https://dict.idioms.moe.edu.tw/idiomView.jsp?ID=962&webMd=2&la=0

Powers, David M W (2007). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*. 2 (1): 37–63.https://web.archive.org/web/20191114213255/https://www.flinders.edu.au/science_engineering

/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf

Li, Z., & Sun, M. (2009). Punctuation as Implicit Annotations for Chinese Word Segmentation. *Computational Linguistics*, *35*(4), 505-512. https://doi.org/10.1162/coli.2009.35.4.35403

Dou, Z.-Y., & Neubig, G. (2021). *Word Alignment by Fine-tuning Embeddings on Parallel Corpora* (arXiv:2101.08231). arXiv. http://arxiv.org/abs/2101.08231