Reconnaissance des procédés de traduction sous-phrastiques : des ressources aux validations

Yuming Zhai

université PARIS-SACLAY

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Limsi

NNT : 2019SACLS489

Thèse de doctorat

Reconnaissance des procédés de traduction sous-phrastiques :

des ressources aux validations

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

Ecole doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Orsay, le 19 décembre 2019, par

YUMING ZHAI

Composition du Jury :

Alexandre Allauzen Professeur, École supérieure de physique et de chimie industrielles de la ville de Paris	Président
Amalia Todirascu Professeure, Université de Strasbourg (LiLPa)	Rapporteure
Mathieu Lafourcade Maître de conférences, Université de Montpellier (LIRMM)	Rapporteur
Emmanuelle Esperança-Rodier Maître de conférences, Université Grenoble Alpes (LIG)	Examinatrice
Philippe Langlais Professeur, Université de Montréal (RALI)	Examinateur
Anne Vilnat Professeure, Université Paris-Sud (LIMSI)	Directrice de thèse
Gabriel Illouz Maître de conférences, Université Paris-Sud (LIMSI)	Co-encadrant, examinateur

Résumé

Les procédés de traduction constituent un sujet important pour les traductologues et les linguistes. Face à un certain mot ou segment difficile à traduire, les traducteurs humains doivent appliquer les solutions particulières au lieu de la traduction littérale, telles que l'équivalence idiomatique, la généralisation, la particularisation, la modulation syntaxique ou sémantique, etc.

En revanche, ce sujet a reçu peu d'attention dans le domaine du Traitement Automatique des Langues (TAL). Notre problématique de recherche se décline en deux questions : est-il possible de reconnaître automatiquement les procédés de traduction ? Certaines tâches en TAL peuvent-elles bénéficier de la reconnaissance des procédés de traduction ?

Notre hypothèse de travail est qu'il est possible de reconnaître automatiquement les différents procédés de traduction (par exemple littéral versus non littéral). Pour vérifier notre hypothèse, nous avons annoté un corpus parallèle anglais-français en procédés de traduction, tout en établissant un guide d'annotation. Notre typologie de procédés est proposée en nous appuyant sur des typologies précédentes, et est adaptée à notre corpus. L'accord inter-annotateur (0,67) est significatif mais dépasse peu le seuil d'un accord fort (0,61), ce qui reflète la difficulté de la tâche d'annotation. En nous fondant sur des exemples annotés, nous avons ensuite travaillé sur la classification automatique des procédés de traduction. Même si le jeu de données est limité, les résultats expérimentaux valident notre hypothèse de travail concernant la possibilité de reconnaître les différents procédés de traduction. Nous avons aussi montré que l'ajout des traits sensibles au contexte est pertinent pour améliorer la classification automatique.

En vue de tester la généricité de notre typologie de procédés de traduction et du guide d'annotation, nos études sur l'annotation manuelle ont été étendues au couple de langues anglais-chinois. Ce couple de langues partagent beaucoup moins de points communs par rapport au couple anglais-français au niveau linguistique et culturel. Le guide d'annotation a été adapté et enrichi. La typologie de procédés de traduction reste identique à celle utilisée pour le couple anglais-français, ce qui justifie d'étudier le transfert des expériences menées pour le couple anglais-français au couple anglais-chinois.

Dans le but de valider l'intérêt de ces études, nous avons conçu un outil d'aide à la compréhension écrite pour les apprenants de français langue étrangère. Une expérience sur la compréhension écrite avec des étudiants chinois confirme notre hypothèse de travail et permet de modéliser l'outil. D'autres perspectives de recherche incluent l'aide à la construction de ressource de paraphrases, l'évaluation de l'alignement automatique de mots et l'évaluation de la qualité de la traduction automatique.

Abstract

Translation techniques constitute an important subject in translation studies and in linguistics. When confronted with a certain word or segment that is difficult to translate, human translators must apply particular solutions instead of literal translation, such as idiomatic equivalence, generalization, particularization, syntactic or semantic modulation, etc.

However, this subject has received little attention in the field of Natural Language Processing (NLP). Our research problem is twofold : is it possible to automatically recognize translation techniques ? Can some NLP tasks benefit from the recognition of translation techniques ?

Our working hypothesis is that it is possible to automatically recognize the different translation techniques (e.g. literal versus non-literal). To verify our hypothesis, we annotated a parallel English-French corpus with translation techniques, while establishing an annotation guide. Our typology of techniques is proposed based on previous typologies, and is adapted to our corpus. The inter-annotator agreement (0.67) is significant but slightly exceeds the threshold of a strong agreement (0.61), reflecting the difficulty of the annotation task. Based on annotated examples, we then worked on the automatic classification of translation techniques. Even if the dataset is limited, the experimental results validate our working hypothesis regarding the possibility of recognizing the different translation techniques. We have also shown that adding context-sensitive features is relevant to improve the automatic classification.

In order to test the genericity of our typology of translation techniques and the annotation guide, our studies of manual annotation have been extended to the English-Chinese language pair. This pair shares far fewer linguistic and cultural similarities than the English-French pair. The annotation guide has been adapted and enriched. The typology of translation techniques remains the same as that used for the English-French pair, which justifies studying the transfer of the experiments conducted for the English-French pair to the English-Chinese pair.

With the aim to validate the benefits of these studies, we have designed a tool to help learners of French as a foreign language in reading comprehension. An experiment on reading comprehension with Chinese students confirms our working hypothesis and allows us to model the tool. Other research perspectives include helping to build paraphrase resources, evaluating automatic word alignment and evaluating the quality of machine translation.

Remerciements

Je tiens tout d'abord à remercier le jury d'examiner mes travaux, et de discuter avec moi de façon approfondie pendant la soutenance.

Je remercie sincèrement ma directrice de thèse, Anne Vilnat, pour nos réunions régulières malgré son agenda serré, sa pédagogie, sa bonne humeur et son encadrement bienveillant. Du côté des encadrants, Aurélien Max m'a guidée pas à pas au début de la thèse. Je le remercie pour notre discussion scientifique, son caractère intègre et ses idées originales. Je remercie aussi chaleureusement Gabriel Illouz pour son retour toujours très rapide, sa poursuite à la perfection, sa bienveillance, ses conseils pertinents et sa relecture rigoureuse.

Je suis très chanceuse d'avoir passé mes trois années de thèse au LIMSI, où l'environnement de travail est parfait pour des jeunes doctorants. Au groupe ILES, je remercie tous les permanents pour leur attention et conseil sur mes travaux. Cyril m'a beaucoup aidée à améliorer le guide d'annotation et a relu mon manuscrit de thèse en entier avec Patrick. Sophie, Aurélie et Sahar ont donné de nombreux conseils précis lors d'une répétition de soutenance. Thomas m'a expliqué patiemment les cours de système, pour faciliter mon enseignement de TP.

La vie au laboratoire est aussi colorée grâce aux nombreux doctorants et stagiaires que j'ai connus avec un grand plaisir. Mes collègues de bureau Rachel, Arnaud et Hicham, les invités habituels de bureau Swen, Sanjay, Zheng, Christopher et Léon-Paul, et mes chères copines Elise et Tsanta. Ayant passé plus d'un an dans le bâtiment S, j'ai pu aussi tisser des liens avec des amis du groupe TLP, que ce soit ceux qui ont déjà quitté le LIMSI pour continuer leur carrière : Matthieu, Lauriane, Julia, Elena, Franck, Pierre, Ruiqing, Charlotte ; ou ceux qui sont arrivés plus récemment : Aina, Aman, Léo, Marc, François, Margot, Minh Quang, Khoa, Hugues, Syrielle, Paul, Benjamin, Jitao, Sharleyne, Soyoung, Robin ... J'ai toujours aimé notre discussion, notre entraide, et nos soirées après le travail. Je remercie également Pooyan, Lufei, Xinyi et Yaqiu pour leur travail de collaboration important.

L'environnement propice au travail du LIMSI est aussi rendu possible grâce au personnel de soutien à la recherche : Bénédicte, Sophie, Isabelle, Laurence, Blanche, Pascal, Nicolas, Olivier, Jean-Claude ... Un grand merci à leur travail quotidien !

Mes pensées vont aussi à mes camarades de master à l'INALCO, où nous avons découvert ensemble le TAL avec l'aide de nos chers professeurs. Je remercie surtout Catherine et Lucille qui sont venues à la soutenance malgré la galère des transports, et Genevieve qui a relu et corrigé toutes mes soumissions scientifiques en anglais.

Enfin j'exprime toute ma gratitude à mes parents et mon copain, Yuan, qui sont toujours mon plus grand soutien. Merci pour leur amour inconditionnel, leur écoute et leur encouragement.

Table des matières

Chapitre 1

Introduction

Sommaire

1.1 Problématique de recherche

La traduction est sans doute pratiquée depuis que les langues existent. Des traces en sont présentes depuis cinq mille ans dès l'apparition de l'écriture, dans l'Égypte ancienne ou en Mésopotamie. Il existe des routes de la traduction comme il existe des routes de la soie. À l'initiative de la conception de Barbara Cassin, nous pouvons consulter un dispositif interactif accessible en ligne nommé « Les routes de la traduction » depuis 2017. [1] Ce dispositif, conçu comme un plan de métro, propose de découvrir le voyage de différentes œuvres en suivant leurs traductions au cours du temps. L'ambition est de montrer comment notre civilisation s'est constituée via la traduction des œuvres de Luther, Aristote, Euclide, Marx, etc.

De nos jours, la traduction reste un moyen indispensable pour permettre la communication entre différentes langues et cultures. Concernant sa définition, voici celle donnée par le dictionnaire Larousse [2] :

> Énonciation dans une autre langue (ou langue cible) de ce qui a été énoncé dans une langue (la langue source), en conservant les équivalences sémantiques et stylistiques.

Nous indiquons également les définitions fournies par la page de Wikipédia dédiée à la traduction [3] :

> La traduction est le fait de faire passer un texte rédigé dans une langue (« langue source », ou « langue de départ ») dans une autre langue (« langue cible », ou « langue d'arrivée »). Elle met en relation au moins deux langues et deux cultures, et parfois deux époques.

1. https://routes-traductions.huma-num.fr/
2. https://www.larousse.fr/dictionnaires/francais/traduction/
3. https://fr.wikipedia.org/wiki/Traduction

> Une traduction représente toujours un texte original (ou « texte source », ou « texte de départ »); en cela, elle comporte un certain degré d'équivalence, bien que le concept d'équivalence stricte entre les langues soit désormais dépassé en traductologie.
>
> Le concept de traduction repose depuis longtemps sur des dichotomies telles que « fidélité » versus « liberté », « fidélité à la lettre » versus « fidélité à l'esprit », « traduction sourcière » versus « traduction cibliste », etc.

La traductologie est une discipline universitaire récente qui date de la seconde moitié du vingtième siècle. La thèse de Lemaire (2017) a présenté un cadrage théorique et méthodologique via trois articles fondateurs sur la traductologie. En tant qu'une forme de communication, la traduction est en fait une fusion culturelle et linguistique. L'essence de la traduction consiste à **comprendre** le texte original dans la langue de départ, et **réexprimer** en conservant les équivalences sémantiques et stylistiques dans la langue d'arrivée. Par exemple, la traduction de la poésie chinoise classique dans des langues européennes est un processus créatif et interculturel. Elle exige des capacités approfondies dans ces deux aspects (compréhension et réécriture). Concernant la traduction de la poésie chinoise classique, la thèse de Ruvidic (2006) présente des pièges théoriques et des obstacles dans la pratique. L'article de Froeliger (2008) s'intéresse, quant à lui, au problème de la nuance en traduction pragmatique (avec une visée de communication et non esthétique).

Dans cette thèse, nous étudions un sujet important dans le domaine de la traductologie : les procédés de traduction. Les procédés de traduction ont d'abord été étudiés par Vinay et Darbelnet (1958) du point de vue de la linguistique comparative, et ont ensuite été revisités par des chercheurs tels que Newmark (1981), Chuquet et Paillard (1989), Molina et Hurtado Albir (2002), etc. Ces travaux ont proposé des typologies différentes de procédés de traduction, qui, à gros grain, consistent à distinguer la traduction littérale de celle non littérale.

Si nous observons les traductions humaines de plus près, nous voyons que consciemment ou non, des humains ont recours aux différents moyens de traduction en dehors de la traduction littérale, par exemple l'équivalence idiomatique, la généralisation, la particularisation, la modulation sémantique, etc.

Prenons les exemples dans le tableau 1.1 : la première traduction préserve exactement le sens, où l'expression figée « *à la hauteur de* » possède un sens figuré « *avoir la compétence, les qualités nécessaires* » ; en revanche, la deuxième traduction est plus compliquée, où il existe une inférence textuelle entre le mot source « *scar* » (cicatrice) et le mot cible « traumatisme » ; dans la troisième traduction en français, le traducteur utilise le mot « *inonde* », qui est une traduction non littérale mais conserve l'image métaphorique ; et enfin dans la traduction en chinois, le traducteur donne une explication plus longue au lieu de chercher une expression idiomatique chinoise équivalente à celle anglaise « *trial and error* ». Dans le deuxième chapitre, nous présentons une définition plus précise sur les procédés de traduction.

Bien que les procédés de traduction aient été largement étudiés par des linguistes et des traductologues, ils ont reçu peu d'attention dans le domaine du Traitement Automatique des Langues (TAL*)[4]. Notre étude se concentre sur la reconnaissance automatique des procédés de traduction sous-phrastiques, et sur la validation de la contribution de cette étude dans d'autres cadres de recherche en TAL, tels que la construction de ressources de

4. Les termes soulignés et suivis d'une étoile sont indexés. Les lecteurs peuvent les retrouver dans l'index à la fin de la thèse. L'auteure a préféré garder les acronymes classiques du domaine en anglais pour garder des références claires, par exemple NMT, SMT, etc.

(1.EN) a solution ***that's big enough to solve*** *our problems* *(1.FR) une solution* ***à la hauteur de*** *nos problèmes*
(2.EN) and ***that scar has stayed with him*** *for his entire life* *(2.FR) et que, toute sa vie,* ***il a souffert de ce traumatisme***
(3.EN) The Sun begins to ***bathe*** *the slope of the landscape.* *(3.FR) Le soleil qui* ***inonde*** *les flancs de ce paysage.*
*(4.EN) well, we use that great euphemism, "****trial and error****"* *(4.ZH)* 我们普通人会做各种各样的实验不断地犯错误 *(En tant que personnes normales,* ***nous ferions continuellement diverses expériences et commettrions des fautes****.)*

Tableau 1.1 – Exemples de traduction non littérale au niveau sous-phrastique

paraphrases, l'évaluation de l'alignement automatique de mots, l'évaluation de la qualité de la traduction automatique, et la conception d'un outil pour aider l'apprentissage du français langue étrangère.

Notre motivation de recherche tire son origine de deux problématiques. La première concerne la méthode d'extraction de paraphrases dans des corpus parallèles bilingues ; la deuxième concerne l'apprentissage des langues étrangères par les apprenants adultes.

Pour la première problématique, la méthode la plus utilisée pour extraire des paraphrases dans des corpus parallèles bilingues est appelée « méthode par pivot ». L'hypothèse est que si deux segments dans la même langue partagent une ou plusieurs traductions communes (considérées comme des "pivots") dans une ou plusieurs langues étrangères, ils sont potentiellement des paraphrases. Cette méthode a été mise en œuvre pour construire la ressource de paraphrases PPDB (*ParaPhrase DataBase*) [5], aujourd'hui la plus grande ressource de paraphrases disponible pour 23 langues (Ganitkevitch *et al.*, 2013; Ganitkevitch et Callison-Burch, 2014; Pavlick *et al.*, 2015b). Le travail de Pavlick *et al.* (2015a) a pourtant montré qu'il existe d'autres relations sémantiques que l'équivalence stricte (paraphrase) dans une telle ressource (*Implication (dans les deux sens), Exclusion, Autrement lié et Indépendant*) [6].

Une estimation réalisée sur la plus grande taille de PPDB 2.0 montre qu'il existerait tout au plus seulement 10% de paraphrases strictes. Nous pouvons donc en conclure qu'une meilleure représentation sémantique est nécessaire pour améliorer cette technique, que ce soit pour obtenir des paraphrases strictes ou pour obtenir de manière contrôlée d'autres types de variantes.

Puisque la méthode par pivot implique au moins deux chemins de traduction (segment source $\rightarrow$ traduction pivot $\rightarrow$ un autre segment source (paraphrase candidate)), notre hypothèse de travail est que typer automatiquement les procédés de traduction entre deux segments bilingues (par exemple : littéral *versus* non littéral) permet de mieux contrôler sémantiquement la recherche de paraphrases. Parce que certains procédés de traduction peuvent faire dévier le sens du segment originel, ainsi l'équivalence sémantique entre le segment source et sa paraphrase candidate peut être influencée.

La deuxième problématique concerne l'apprentissage des langues étrangères. Ce sujet est important pour les étudiants, surtout ceux qui veulent poursuivre des études dans un pays étranger, où les études et l'intégration dans la société nécessitent un niveau de langue

5. `http://paraphrase.org`

6. Exclusion : X est le contraire de Y ; X et Y s'excluent mutuellement. Autrement lié : X est lié à Y d'une certaine manière (*ex. country/patriotic*). Indépendant : X n'est pas lié à Y.

intermédiaire voire avancé.

Prenons l'exemple des étudiants chinois qui étudient en France. L'anglais est la première langue étrangère pour une majorité d'entre eux. De l'école primaire jusqu'au master, l'anglais est une discipline importante lors des examens. En revanche, dans la plupart des situations, il leur manque un environnement quotidien pour communiquer en anglais. Parmi les quatre grandes compétences de la langue (compréhension orale, compréhension écrite, production orale, production écrite), beaucoup d'apprenants maîtrisent mieux la compréhension écrite, parce qu'ils passent beaucoup de temps à la lecture et à préparer l'examen. La production écrite est plus difficile parce qu'elle nécessite plus d'accumulation de connaissance et de pratique. Au cours de l'apprentissage du français comme une autre langue étrangère, avoir recours à sa langue maternelle (chinois) ou à une autre langue plus proche et déjà apprise (anglais) est une pratique courante. La traduction est ainsi la méthode la plus utilisée pour mieux assimiler et comparer les connaissances sur la langue.

Une autre méthode d'apprentissage est la reformulation paraphrastique (Rossari, 1994; Martinot, 2012). Une telle compétence aide les apprenants à élargir leur vocabulaire et leur répertoire d'expressions, à prendre l'habitude de réfléchir dans la langue étrangère au lieu de la langue maternelle. Elle est aussi importante pour la production et la compréhension écrite : pour simplifier ou rendre plus complexe leurs énoncés (Chachu, 2017; Chen *et al.*, 2013) ; pour transformer les mots du texte en ses propres termes, ce qui est crucial pour vérifier leur compréhension et lier cette compréhension avec leurs connaissances préalables en vue de faire des inférences (Kletzien, 2009).

Dans cette thèse, nous proposons un cadre de validation de nos études fondé sur un outil d'aide à la compréhension écrite pour les apprenants de français langue étrangère. Pour des mots ou segments difficiles à comprendre pour des apprenants, la tâche visée est de proposer des réécritures en contexte. La proposition des réécritures s'appuiera sur une exploitation des corpus parallèles bilingues via la méthode par pivot, et nous pouvons montrer aux apprenants les traces des traductions pivots parcourues. Pour chaque paire de traductions, nous ajouterons la classification automatique de procédé de traduction, pour alerter les apprenants sur une traduction non littérale utilisée dans le corpus parallèle. Ainsi nous combinons la traduction et la réécriture dans un même outil pour apporter de l'aide à la compréhension écrite.

À part ces deux aspects qui visent des applications pratiques, du point de vue de l'analyse linguistique comparative, il est également important de comprendre les différentes méthodes de traduction humaine ainsi que les raisons de ces choix spécifiques.

Dans notre recherche, nous nous focalisons sur les phénomènes de traduction au niveau sous-phrastique. Parce que les procédés de traduction étudiés par des linguistes et des traductologues sont utilisés par des humains au niveau sous-phrastique. De plus, les deux aspects d'application que nous venons de présenter concernent les segments au niveau sous-phrastique.

1.2 Contributions

Étant donné notre motivation de recherche, la tâche la plus importante consiste en la classification automatique des procédés de traduction au niveau sous-phrastique. Pour étudier si cette tâche est réalisable, nous avons besoin d'un jeu de données pour entraîner un classifieur supervisé. Puisque des données de ce genre n'existaient pas, nous avons annoté manuellement un corpus parallèle anglais-français en procédés de traduction.

En nous basant sur différentes typologies de procédés de traduction existantes, nous avons établi notre propre typologie au fur et à mesure que nous annotions le corpus parallèle anglais-français. Nous avons annoté un corpus de *TED Talks* [7] qui contient des discours préparés. Le corpus contient au total 2 436 lignes de phrases parallèles (51k tokens anglais, 53k tokens français), et les présentations correspondantes ont une durée totale de 4,6 heures. Nous avons choisi ce genre de corpus pour la diversité des domaines et la diversité des phénomènes de traduction présentes. Le guide d'annotation s'est enrichi tout au long de cette annotation. L'accord inter-annotateur (0,67) est significatif mais dépasse peu le seuil d'un accord fort (0,61), ce qui reflète la difficulté de la tâche d'annotation.

Après avoir construit ce jeu de données, nous avons pu travailler sur la classification automatique. Les expériences sont menées dans un scénario simplifié [8], où nous fournissons au classifieur des paires de segments bilingues avec la frontière donnée, et le but du classifieur est de prédire le procédé utilisé. Avec l'ingénierie des traits linguistiques et des classifieurs statistiques, nous avons pu valider notre hypothèse de travail concernant la possibilité de reconnaître les différents procédés de traduction, même si le jeu de données est limité. Nous avons également testé des architectures en réseaux neuronaux, avec seuls des plongements lexicaux comme entrées. Les résultats ne sont pas meilleurs par rapport aux classifieurs statistiques à cause de la taille de données. Enfin nous avons montré que l'ajout des traits sensibles à la phrase de contexte est pertinent pour améliorer la classification automatique.

En vue de tester la généricité de notre typologie de procédés de traduction et du guide d'annotation, les études sur le couple anglais-français ont été étendues au couple anglais-chinois, qui partagent beaucoup moins de points communs au niveau linguistique et culturel. Nous avons constitué un corpus de onze genres différents pour ce couple de langues. Le guide d'annotation a été adapté et enrichi, mais la typologie de procédés de traduction reste identique à celle utilisée pour le couple anglais-français. Cela justifie d'étudier le transfert des expériences menées pour le couple anglais-français au couple anglais-chinois.

Dans le but de valider l'intérêt de ces recherches, nous avons conçu un outil d'aide à la compréhension écrite pour les apprenants de français langue étrangère. Une expérience sur la compréhension écrite avec la participation des étudiants chinois confirme notre hypothèse de travail et nous permet de modéliser l'outil. D'autres perspectives de recherche incluent l'aide à la construction de ressource de paraphrases, l'évaluation de l'alignement automatique de mots et l'évaluation de la qualité de la traduction automatique.

1.3 Structure du manuscrit

Ce manuscrit est organisé de la manière suivante. Les deux premiers chapitres sont consacrés à l'état de l'art. Après avoir posé les définitions pour notre thèse, nous passons en revue les travaux sur les procédés de traduction et présentons les différentes typologies de procédés de traduction (chapitre 2).

Ensuite, nous présentons un état de l'art sur l'étude des paraphrases en TAL (chapitre 3), en le centrant sur l'extraction de paraphrases dans les corpus parallèles bilingues. Nous justifions les raisons pour lesquelles nous étudions la reconnaissance des procédés

7. `https://www.ted.com`

8. Par rapport à une configuration où le système doit aussi prédire la frontière des traductions non littérales alignées.

de traduction pour valider leur contribution dans d'autres cadres de recherche en TAL.

Après avoir exposé nos problématiques de recherche, nous présentons les travaux réalisés au cours de cette thèse (partie 2).

Ayant justifié notre choix du corpus et notre méthodologie d'annotation, nous présentons notre typologie de procédés de traduction utilisée pendant l'annotation, et des exemples typiques pour chaque catégorie (chapitre 4).

La mise en œuvre de l'annotation manuelle est détaillée dans le chapitre 5. Nous présentons les caractéristiques du corpus anglais-français, l'outil d'annotation, le guide d'annotation qui a permis cette annotation, ainsi que diverses statistiques calculées. Pour vérifier la généricité de notre démarche, nous avons choisi de l'appliquer aussi au couple de langues anglais-chinois.

La classification automatique des procédés de traduction est présentée au chapitre 6. Nous présentons le jeu des données issu de l'annotation et nos différentes expériences. Les résultats expérimentaux sont présentés et analysés.

Pour valider notre étude, d'autres cadres sont examinés (chapitre 7). Cela inclut quatre pistes de recherche : l'aide à la construction de ressources de paraphrases, l'évaluation de l'alignement automatique de mots, l'évaluation de la qualité de la traduction automatique, et la conception d'un outil pour aider l'apprentissage du français langue étrangère.

Nous concluons et discutons des perspectives dans le chapitre 8.

1.4 Publications liées à la thèse

Publications dans les actes de conférence :

1. Classification automatique des procédés de traduction, **Yuming Zhai**, Gabriel Illouz, et Anne Vilnat, 26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2019). Toulouse, France
2. Conception d'un outil d'aide à la compréhension écrite pour les apprenants de français langue étrangère, **Yuming Zhai**, Gabriel Illouz, et Anne Vilnat, 9ème Conférence Environnements Informatiques pour l'Apprentissage Humain (EIAH 2019). Paris, France
3. Towards Recognizing Phrase Translation Processes : Experiments on English-French, **Yuming Zhai**, Pooyan Safari, Gabriel Illouz, Alexandre Allauzen, and Anne Vilnat, 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019). La Rochelle, France
4. Construction of a Multilingual Corpus Annotated with Translation Relations, **Yuming Zhai**, Aurélien Max and Anne Vilnat, First Workshop on Linguistic Resources for Natural Language Processing@COLING (LR4NLP 2018). Santa Fe, New Mexico, USA
5. Construction d'un corpus multilingue annoté en relations de traduction, **Yuming Zhai**, 20ème REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2018). Rennes, France

Communication orale (sans acte) :

1. Construction of a Multilingual Corpus Annotated with Translation Relations, **Yuming Zhai**, Atelier du Consortium CORLI "Analyse cross-lingue et annotation de corpus multilingues parallèles et comparables : tendances actuelles et futures", Université Paris Diderot, France

Première partie

Contexte de travail

Chapitre 2

Procédés de traduction

Sommaire

Dans ce chapitre, nous présentons la définition générale des procédés de traduction que nous avons adoptée, ainsi qu'un état de l'art sur plusieurs typologies de ces procédés.

2.1 Introduction

La traduction mot à mot peut être très éloignée du sens originel, surtout quand la traduction littérale ignore le contexte de la phrase. Durant la guerre froide, les Américains essayaient déjà de traduire automatiquement les documents russes. Une histoire humoristique et apocryphe raconte qu'ils auraient testé leur système avec la phrase biblique « *The spirit is strong but the flesh is weak.* » (« *L'esprit est fort, mais la chair est faible.* ») qui a donné après traduction automatique en russe, puis de nouveau en anglais : « *The vodka is good, but the meat is rotten.* » (« *La vodka est forte, mais la viande est pourrie.* ») (Hutchins, 1995). Depuis, de nombreux travaux ont été réalisés pour traiter le problème de la polysémie dans la traduction automatique (Mihăltz, 2005; Civera et Juan, 2007; Apidianaki, 2008; Bradford, 2010).

Dans la réalité, la traduction littérale a été la cause principale de l'accident aérien de Tenerife [1], à ce jour l'accident le plus meurtrier de l'histoire de l'aviation commerciale causé par une mauvaise communication entre les pilotes et la tour de contrôle. En effet, Par une journée brumeuse de mars en 1977, deux Boeing 747, l'un de la KLM et l'autre de la Pan Am, circulaient sur la piste de l'aéroport de Los Rodeos. Peu de temps après, l'avion de la KLM a commencé à décoller, accélérant le long de la piste à l'insu de l'autre avion de la Pan Am, et est entré en collision avec celui-ci, tuant 583 personnes. Le contrôleur a été stupéfait. Il n'avait pas donné l'autorisation de décoller. Pourquoi le capitaine de la KLM a-t-il accéléré sur la piste ?

En fin de compte, c'était une erreur simple mais lourde de conséquence : une traduction littérale. Quelques secondes avant la catastrophe, le copilote néerlandais de la KLM

1. https://fr.wikipedia.org/wiki/Accident_aérien_de_Tenerife

a annoncé en anglais : « *We are now at take-off.* » Il avait l'intention de faire savoir que l'avion était déjà en train de décoller, un sens qui serait mieux exprimé en anglais par la construction progressive « *We are now taking off.* » La phrase anglaise, traduite par un natif néerlandais, était simplement une traduction mot à mot de la construction progressive néerlandaise « *We (we) zijn (are) nu (now) aan (at) het (the) opstijgen (take-off).* » Mais comme la syntaxe anglaise fonctionne différemment, le sens progressif s'est perdu. Le contrôleur de vol a compris le message de la KLM comme « *We are now at take-off (position).* », c'est-à-dire que l'avion était en bout de piste, attendant l'autorisation de décoller. N'ayant pas donné cette autorisation, une simple confirmation « *OK* » a déclenché la catastrophe.

Hormis les différences au niveau de la syntaxe, il existe bien d'autres difficultés en traduction, pour lesquelles la traduction mot à mot ne peut pas pas être utilisée. *« Traduttore, traditore »* est une expression italienne signifiant littéralement : « *Traducteur, traître* », ou encore : « *Traduire, c'est trahir* ». Il s'agit d'une paronomase, expression qui joue sur la ressemblance entre les deux mots. Le fait de comparer un traducteur à un traître signifie que la traduction d'un texte d'une langue dans une autre ne peut jamais respecter parfaitement le texte de l'œuvre originale. Ce concept évoque l'*intraduisibilité*, qui est le caractère d'un texte ou d'un énoncé dans une langue source, auquel on ne peut faire correspondre aucun texte ou aucun énoncé dans une langue cible. Souvent, quand un texte ou un énoncé est considéré comme « *intraduisible* », il s'agit plutôt d'une « *lacune* », c'est-à-dire l'absence littérale d'un mot, d'une expression ou d'une tournure de la langue d'origine dans la langue cible. Pour combler cette lacune et ne pas "trahir", le traducteur peut avoir recours à différents *procédés de traduction*, que nous présenterons en détail dans ce chapitre.

2.2 Travaux précédents

Les procédés de traduction sont analysés dans des études qui observent leur usage dans les cas pratiques de traduction, qui est un aspect moins central pour notre étude (Gibová, 2012; Shojaei, 2012; Fernández Guerra, 2012; Ye, 2014).

Après avoir défini les principaux concepts liés aux procédés de traduction, nous présentons un état de l'art des procédés de traduction ainsi que les différentes typologies.

2.2.1 Typologies de procédés de traduction

La typologie de différents moyens de traduction a été étudiée depuis longtemps par des traducteurs humains et des linguistes, en les désignant par le terme « *procédés de traduction* » (Vinay et Darbelnet, 1958; Newmark, 1981, 1988; Chuquet et Paillard, 1989; Molina et Hurtado Albir, 2002). La traduction littérale est distinguée des autres procédés de traduction.

Il existe des désaccords parmi les chercheurs sur les procédés de traduction, au niveau non seulement terminologique mais aussi conceptuel. Il existe aussi un manque de consensus sur la terminologie : « *procédé, procédure, technique, stratégie, méthode*, etc. », et parfois ils sont confondus avec d'autres concepts. Ici nous retenons les définitions suivantes pour trois termes, qui ont été clarifiées par Molina et Hurtado Albir (2002) :

Définition ***Méthode de traduction**** *: moyen avec lequel un processus particulier de traduction est effectué, selon l'objectif du traducteur. C'est un choix global qui influence la traduction de tout le texte, par exemple : interprétatif-communicatif, littéral,*

libre, philologique, etc.
Ainsi, la traduction interprétative et communicative consiste à comprendre et à recréer le texte original, sans apporter de changements radicaux ; c'est généralement le cas de la traduction simultanée et consécutive. Les changements de style ne sont pas tolérables.
Quant à la traduction philologique, le traducteur peut ajouter des notes de nature philologique et historique à la traduction, dans le but non seulement de bien comprendre des termes et des mots correctement, mais aussi de clarifier des significations familières ; dans ce cas, le texte source est souvent soumis à un examen et la traduction est destinée à un public spécialisé ou aux étudiants.

Définition ***Stratégie de traduction**** *: c'est un élément essentiel dans la résolution de problèmes pendant la traduction. Elle désigne la procédure analytique (consciente ou inconsciente, verbale ou non verbale) qu'un traducteur peut appliquer pour faciliter la compréhension du texte source (ex. distinguer les idées principales et secondaires, établir des relations conceptuelles, chercher des informations) ou pour la reformulation (ex. traduction inverse (back translation), paraphrase, dire à haute voix, etc.). Les stratégies peuvent aider les traducteurs à trouver des procédés de traduction adéquats pour résoudre les problèmes.*

Définition ***Procédés de traduction**** *(le terme anglais correspondant est « translation techniques » ou parfois « translation procedures ») : ce sont des solutions particulières appliquées pendant le processus de traduction, quand un certain mot ou segment difficile à traduire est rencontré (ex. emprunt, calque, modulation, transposition, etc.)*

En se basant sur la linguistique comparative, Vinay et Darbelnet (1958) ont proposé la première typologie de procédés de traduction avec un but clairement méthodologique. Leur livre « *Stylistique Comparée du Français et de l'Anglais* » est un travail pionnier sur ce sujet, et nous y ferons référence avec l'acronyme SCFA* par la suite. Ils ont utilisé le terme « *procédés techniques auxquels se ramène la démarche du traducteur* », et ont défini sept procédés basiques sur trois grands domaines : lexique, agencement (morphologie et syntaxe) et message. Les procédés sont classifiés en direct (littéral) et oblique, ce qui correspond à leur distinction entre la traduction directe (littérale) et oblique (voir tableau 2.1). Cette dernière est utilisée lorsqu'une traduction littérale est inacceptable, ou lorsque les asymétries structurelles ou conceptuelles entre la langue source et la langue cible ne sont pas négligeables. Par la traduction oblique, le sens peut demeurer intact ou se trouver modifié. En complément des sept procédés basiques, des procédés supplémentaires ont été définis (voir tableau 2.2). Sauf *Compensation* et *Inversion*, ils ont tous été classifiés comme des paires opposées. Les paires *Amplification-Économie* et *Étoffement-Dépouillement* sont considérées comme des sous-types du procédé *Transposition*.

Catford (1965) a proposé les deux notions suivantes :

— Correspondance formelle : lorsqu'une structure syntaxique du texte source peut être conservée dans le texte cible avec la même valeur sémantique et pragmatique, ceci correspond à la traduction directe de SCFA.

— Équivalence textuelle : s'il faut avoir recours à une structure distincte mais de fonction homologique dans la langue cible, par exemple *English spoken* → *On parle anglais*. Ceci correspond à la traduction oblique de SCFA.

Dans ce chapitre, nous utiliserons le terme *traduction oblique*, *traduction non littérale* et *traduction libre* de façon interchangeable.

Le livre *« Approche linguistique des problèmes de traduction anglais-français »* de Chuquet et Paillard (1989) vise à apporter des éléments permettant d'analyser les contrastes entre le français et l'anglais, motivé par leur expérience d'enseignement de grammaire

Nom du procédé	Définition et exemple
	Procédés de traduction littéraux
Emprunt	Mot qu'une langue emprunte à une autre sans le traduire. (*Le **chef** cuit les pommes de terre à la vapeur.* → *The **chef** steams potatoes.*)
Calque	Emprunt d'un syntagme étranger avec traduction littérale de ses éléments. (*hard disk* → *disque dur*)
Traduction littérale	Traduction mot-à-mot. (*The ink is on the table* → *L'encre est sur la table.*)
	Procédés de traduction obliques
Transposition	Procédé par lequel un signifié change de catégorie grammaticale. (*He **soon** realized.* → *Il **ne tarda pas à** se rendre compte.*) S'il existe des changements entre deux signifiés, il s'agit d'un « *chassé-croisé* ». (*He **limped across** the street.* → *Il a **traversé** la rue **en boitant**.*)
Modulation	Variation obtenue en changeant de point de vue, d'éclairage et très souvent de catégorie de pensée. Onze types de modulation ont été listés concernant les changements entre : abstrait et concret, cause et effet, moyen et résultat, partie et tout, changement géographique, etc. (*encre de Chine* → *Indian ink*) Modulation figée : celle qu'enregistrent les dictionnaires bilingues. (*tooled leather* → *cuir repoussé*) Modulation libre : celle que les dictionnaires n'enregistrent pas encore, mais à laquelle les traducteurs ont recours lorsque la LC rejette la traduction littérale.
Équivalence	Rendre compte de la même situation que dans la LS, en ayant recours à une rédaction entièrement différente, par exemple la traduction des proverbes ou des expressions idiomatiques. *(comme un chien dans un jeu de quilles* → *like a bull in a china shop)*
Adaptation	Un élément culturel en LS est remplacé par un élément culturel en LC, soit un élément familier aux locuteurs de la LC. (*1 **tsp** white truffle paste* → *une **cuillerée à café** de beurre blanc aux truffes*)

Tableau 2.1 – Les sept procédés de traduction basiques proposés par Vinay et Darbelnet (1958), LS : langue source, LC : langue cible

anglaise, de linguistique et de traduction. Le but est d'intégrer les observations de la stylistique comparée et les concepts linguistiques les plus susceptibles d'éclairer la pratique intuitive de la traduction. Les auteurs ont d'abord fait un réexamen des sept procédés basiques proposés dans SCFA, puis fait apparaître le lien étroit entre l'utilisation de ces procédés et les différences de fonctionnement des deux langues sur le plan grammatical et syntaxique. Ils ont aussi introduit certaines notions lexicologiques indispensables à l'analyse des décalages réguliers entre les deux langues.

Chuquet et Paillard (1989) ont restreint la portée d'étude sur les procédés *Transposition* et *Modulation*, sur la base de deux observations :

— un grand nombre de procédés renvoie à une problématique grammaticale ou lexicale beaucoup plus générale.

Nom du procédé	Définition et exemple
Compensation	Un morceau d'information ou un effet stylistique dans la LS qui ne peut pas être reproduit au même endroit dans la LC est introduit ailleurs. *(I was seeking **thee**, Flathead. → C'est **toi** que je cherche, **O** Tête-Plate.)*
Concentration vs. Dilution	Concentration : la LC exprime un signifié avec moins de signifiants. *(Elle lui jeta un coup d'oeil. → She glanced at him.)* Dilution : répartition d'un signifié sur plusieurs signifiants. *(Patten waved to the crowd. → Patten salua la foule de la main.)*
Amplification vs. Économie	La LC utilise plus de signifiants pour combler les lacunes syntaxiques ou lexicales. *(He talked himself out of a job. → Il a perdu sa chance pour avoir trop parlé.)* L'économie est le procédé inverse. *(Nous ne pourrons plus vendre si nous sommes trop exigeants. → We'll price ourselves out of the market.)*
Étoffement vs. Dépouillement	Ce sont des variantes de *Amplification* et *Économie* qui sont caractéristiques pour l'anglais et le français. L'étoffement consiste à ajouter en français un syntagme nominal ou verbal pour traduire une préposition, un pronom ou un adverbe interrogatif en anglais. Le dépouillement est le procédé inverse, qui se rencontre surtout du français à l'anglais. *(a campaign by Redmond, Washington-based Microsoft → une campagne **orchestrée** par Microsoft, **société ayant son siège social à** Redmond, **dans l'État de** Washington)*
Explicitation vs. Implicitation	Explicitation : introduire dans la LC des précisions qui restent implicites dans la LS, mais qui se dégagent du contexte ou de la situation. *(his patient → son patient / sa patiente)* Implicitation : laisser au contexte ou à la situation le soin de préciser certains détails explicites dans la LS. *(Go out / Come out → Sortez)*
Généralisation vs. Particularisation	Généralisation : traduire un terme particulier (ou concret) par un terme plus général (ou abstrait). *(Guichet, fenêtre, devanture → window)* La particularisation est le procédé inverse.
Inversion	Déplacer un mot ou un segment à un autre endroit dans une phrase ou un paragraphe, pour que la traduction se lise de façon naturelle. *(Pack separately [...] for convenient inspection → Pour faciliter la visite de la douane mettre à part [...])*
Articulation vs. Juxtaposition	Utilisation de charnières qui ponctuent le raisonnement dans le déroulement de l'énoncé. La juxtaposition est le procédé inverse. *In all **this** immense variety of conditions [...] → **Et cependant**, **malgré** la diversité des conditions [...]*
Grammaticalisation vs. Lexicalisation	Grammaticalisation : remplacer les signes lexicaux par des signes grammaticaux. La lexicalisation est le procédé inverse. *Peut-être [...] → Il se peut que [...]*

Tableau 2.2 – Les procédés supplémentaires proposés par Vinay et Darbelnet (1958), LS : langue source, LC : langue cible

— *Emprunt* et *Calque* sont rarement des procédés de traduction à proprement par-

ler, mais se trouvent généralement intégrés au lexique ; *Équivalence* n'est pas autre chose qu'une modulation figée, bien illustrée notamment dans la correspondance entre les proverbes d'une langue à l'autre ; quant à *Adaptation*, il paraît difficile de l'isoler en tant que procédé de traduction, dans la mesure où elle fait entrer en jeu des facteurs socio-culturels et subjectifs autant que linguistiques.

Reprenons les contributions de Molina et Hurtado Albir (2002), pour étudier la façon dont la traduction fonctionne et analyser les traductions, il faut prendre en compte trois catégories : la catégorie textuelle (qui décrit les mécanismes de cohérence, cohésion et progression thématique), la catégorie contextuelle (tous les éléments extra-textuels liés au contexte du texte source et de la traduction produite) et la catégorie du processus (la méthode et les stratégies choisies pendant la traduction). Par contre, il est aussi important de considérer des micro-unités textuelles*, c'est-à-dire comment le résultat de la traduction fonctionne par rapport aux unités correspondantes dans le texte source. Pour cela, il faut étudier les procédés de traduction. Molina et Hurtado Albir ont pris conscience de ce besoin lors de leur étude sur le traitement des éléments culturels dans les traductions de l'espagnol en arabe du roman « *Cent ans de solitude* ». Ils ont besoin des procédés de traduction qui permettent de décrire des étapes réelles effectuées par des traducteurs pour chaque micro-unité textuelle, afin d'obtenir des données claires sur l'option méthodologique générale choisie.

Il existe différentes typologies de procédés et les intersections entre typologies sont nombreuses. Molina et Hurtado Albir ont présenté un réexamen sur les typologies précédentes de Vinay et Darbelnet (1958), Nida (1964), Margot (1979), Nida et Taber (1969), Vázquez-Ayora (1977), Newmark (1988) et Delisle (1993), et ont proposé leur propre définition et typologie utilisés dans leur travail.

Dans les études sur la traduction biblique, Nida, Taber et Margot se sont concentrés sur les questions de transfert culturel. Ils ont proposé plusieurs procédés à utiliser quand aucune équivalence n'existe dans la langue cible. Voici les propositions de Nida (1964) :

— Addition : différentes situations nécessitent de fournir une traduction plus complète : clarifier une expression elliptique, éviter l'ambiguïté dans la langue cible, changer une catégorie grammaticale (correspond à *Transposition* dans SCFA), développer des éléments implicites (correspond à *Explicitation* dans SCFA), ajouter des connecteurs logiques (correspond à *Articulation* dans SCFA), etc.

  Par exemple, une traduction littérale de « *they tell him of her* » en mazatèque a besoin d'être amplifiée en « *the people there told Jesus about the woman* ». Puisque cette langue ne distingue pas le nombre ni le genre des affixes pronominaux, il existerait 36 interprétations différentes.

— Soustraction : quatre situations où ce procédé doit être utilisé : répétition non nécessaire, références précisées, conjonctions et adverbes.

— Modification (*Alteration*) : ces modifications doivent être faites en raison des incompatibilités entre deux langues.
  1) Des changements dus aux problèmes causés par la translittération quand un nouveau mot est introduit depuis la langue source.
  2) Des changements dus aux différences structurales entre deux langues, par exemple les changements de l'ordre des mots, des catégories grammaticales, etc. (comparable à *Transposition* dans SCFA)
  3) Des changements dus aux inadaptations sémantiques, surtout avec les expressions idiomatiques. Une suggestion est d'utiliser un « *équivalent descriptif* », c'est-

à-dire un équivalent satisfaisant pour des objets, évènements ou attributs qui n'ont pas un terme standard dans la langue cible.

— Naturalisation : ce concept est introduit après que Nida utilise le terme « *naturel* » pour définir l'équivalence dynamique (*l'équivalent naturel le plus proche au message de la langue source*). (comparable à *Adaptation* dans SCFA)

— La note de bas de page a été incluse comme un procédé, qui a deux fonctions principales : pour expliquer des différences linguistiques et culturelles ; pour ajouter des informations supplémentaires sur le contexte historique et culturel du texte en question.

Margot (1979) a présenté des critères pour justifier l'adaptation culturelle, et il les considère comme des différences essentielles, voici ses propositions :

— Pour des entités inconnues dans la culture cible, comme Nida le fait, il suggère d'ajouter un mot à côté pour le qualifier, par exemple : ***the city of*** *Jerusalem* ; ou utiliser un équivalent culturel. (comparable à *Adaptation* dans SCFA)

— Le concept de la redondance : supprimer l'information quand des éléments de la langue source sont redondants pour les lecteurs cibles. (comparable à *Implicitation* dans SCFA)

— Il a aussi inclus la note de bas de page, comme une aide pour l'adaptation culturelle.

Nida, Taber et Margot sont d'accord sur la distinction entre la paraphrase légitime* et illégitime. La paraphrase légitime est un changement lexical qui rend la traduction plus longue que la source mais sans changer le sens (comparable à *Amplification/Dilution* dans SCFA). La paraphrase illégitime* rend des entités explicites dans la traduction, et ils pensent que ce n'est pas le travail des traducteurs, puisque cela peut introduire de la subjectivité.

Vázquez-Ayora (1977) a combiné l'approche normative de Vinay et Darbelnet (1958) et l'approche descriptive des chercheurs de la traduction de la Bible (Nida, Taber, Margot, entre autres). Il a introduit ces nouveaux procédés :

— Omission : omettre la redondance et répétition qui est caractéristique en langue source.

— Déplacement et inversion : la définition correspond à *Inversion* dans SCFA.

Newmark (1988) a ajouté les procédés ci-dessous :

— Traduction reconnue : une traduction d'un terme qui est déjà officielle ou largement acceptée, même si elle ne serait pas la plus adéquate.

— Équivalent fonctionnel : utiliser un mot neutre au niveau culturel, et ajouter un terme qui le spécifie, par exemple : *baccalauréat* = *French secondary school leaving exam* (le mot « *French* » spécifie la traduction). (correspond à *Adaptation* dans SCFA)

— Naturalisation : adapter un mot de la langue source aux normes phonétiques et morphologiques de la langue cible.

— Étiquette de traduction (*translation label*) : une traduction provisoire, souvent pour un nouveau terme, pour lequel une traduction littérale pourrait être acceptable.

Newmark a aussi considéré la possibilité de résoudre un problème de traduction par la combinaison de deux ou plus de ces procédés.

Delisle (1993) a simplifié les paires opposées *Étoffement-Dépouillement* et *Amplification-Économie*, et les a réduites à une seule paire : *Étoffement-Économie*. Il a ensuite

distingué trois types sous l'étoffement : dilution, explicitation et périphrase (correspond à *Amplification* dans SCFA) ; et trois types sous l'économie : concentration, implicitation et concision (correspond à *Économie* dans SCFA).

Il a aussi introduit d'autres catégories :

— Addition vs. Omission : il les définit comme des périphrases et concisions injustifiées, et les considère comme des erreurs de traduction. L'addition consiste à introduire des éléments stylistiques injustifiés et des informations non présentes dans la source. L'omission est la suppression injustifiée des éléments de la source.

— Paraphrase : usage excessif de paraphrases qui complique la traduction sans justification stylistique ou rhétorique. C'est aussi considéré comme une erreur de traduction.

— Création discursive : pendant la traduction, une équivalence non lexicale est établie, qui fonctionne seulement en contexte, par exemple la traduction d'un nom de film.

Molina et Hurtado Albir (2002) ont souligné que l'approche de SCFA est limitée à la typologie des différences entre les systèmes de langue. Par exemple, les procédés *Emprunt, Transposition, Inversion* dans SCFA et *Omission* de Vázquez-Ayora (1977) sont contraints par les caractéristiques de la paire de langues, et ne sont pas des options ouvertes aux traducteurs. En outre, selon eux, la plupart des études sur les procédés de traduction ne sont pas adaptées à la nature dynamique de l'équivalence de traduction. Un procédé de traduction étant le résultat de choix effectué par un traducteur, sa validité doit dépendre des questions variées liées au contexte, du but de la traduction, des attentes de l'audience, etc. Un procédé ne peut être jugé avec du sens qu'au sein d'un contexte particulier. Ainsi la paire opposée *Addition-Omission* jugée comme une erreur de traduction par Delisle (1993) ne fait pas de sens pour eux. Molina et Hurtado Albir ont défini leur propre liste de procédés de traduction afin d'analyser comment fonctionne l'équivalence de traduction. Le tableau 2.3 montre ce résultat.

Plus récemment, Ballard (2006) a critiqué les procédés de traduction tels qu'ils ont été envisagés par Vinay et Darbelnet (1958). Selon lui, les sept procédés basiques ne sont pas suffisants parce qu'ils ne couvrent pas l'ensemble des opérations de traduction, et leur terminologie n'est pas clairement définie, car souvent inadéquate pour l'objet décrit. Par exemple, *Équivalence* et *Adaptation* ne sont pas nettement distinguées, et la définition de *Modulation* est trop vague.

Au-delà des différences de typologie des procédés, nous observons également une différence conceptuelle autour des unités de traduction. Vinay et Darbelnet (1958) ont défini l'« *unité de traduction* » comme suit : *le plus petit segment de l'énoncé dont la cohésion des signes est telle qu'ils ne doivent pas être traduits séparément, ex : "prendre son élan", "battre à coups précipités". Les unités de traduction permettent d'effectuer le découpage d'un texte.* Selon Ballard (2006), cette définition est trop restrictive et seulement axée sur le texte source. Il a utilisé cet exemple pour illustrer son argument :

Texte original : *Sally asked her to join a party going to the Slade dance. Paul detested dances.* ***After some pleading she went alone, and arrived back at six in the morning. Dora was unable to be exact about time or anything else.***

Traduction littérale : *Sally lui demanda de se joindre à un groupe qui se rendait au bal de la Slade. Paul détestait danser. Après l'avoir supplié de venir, elle y alla seule et rentra à six heures du matin. Dora était incapable d'exactitude en ce qui concerne le temps ou toute autre chose.*

Traduction par un traducteur humain : *Un jour, son amie la pria d'assister à une soirée organisée à son ancienne école. Paul exécrait la danse.* ***Après l'avoir imploré, elle obtint***

Nom du procédé	Définition
Calque	identique à celle de la SCFA
Traduction littérale	identique à celle de la SCFA
Adaptation	identique à celle de la SCFA
Compensation	identique à celle de la SCFA
Généralisation	identique à celle de la SCFA
Particularisation	identique à celle de la SCFA
Modulation	identique à celle de la SCFA
Transposition	identique à celle de la SCFA
Équivalence établie	correspond à *Équivalence* et *Traduction littérale* de SCFA
Emprunt	*Emprunt* pur dans SCFA et emprunt naturalisé de Newmark.
Amplification	Introduire des détails qui ne sont pas formulés dans la langue source : information, paraphrase explicative. Ce procédé inclut *Explicitation* dans SCFA, *Addition* et *Périphrase* de Delisle, *Paraphrase* (légitime et illégitime) de Margot, et la note de bas de page.
Réduction	Ce procédé inclut *Implicitation* de SCFA, *Concision* de Delisle, et *Omission* de Vázquez-Ayora.
Description	Remplacer un terme ou une expression par une description de sa forme et/ou sa fonction.
Création discursive	identique à celle de Delisle
Amplification linguistique	Ajouter des éléments linguistiques, souvent utilisé dans l'interprétation consécutive et le doublage.
Compression linguistique	Synthétiser des éléments linguistiques, souvent utilisé dans l'interprétation simultanée et la traduction de sous-titres.
Substitution	Remplacer des éléments linguistiques par des éléments paralinguistiques (intonation, geste), ou vice versa. C'est surtout utilisé dans l'interprétation.
Variation	Changer des éléments linguistiques et paralinguistiques qui influencent des aspects de variation linguistique : changement de style, ton, dialecte, etc.

Tableau 2.3 – La typologie de procédés selon Molina et Hurtado Albir (2002)

l'autorisation de s'y rendre seule, mais incapable de la moindre discipline, n'en revient qu'à six heures du matin.

Pour obtenir une construction plus idiomatique en français, le traducteur humain réorganise et modifie les deux dernières phrases (en gras). Cet exemple montre que l'on ne peut pas prédécouper des unités de traduction dans le texte source. La traduction humaine relève d'une traduction non littérale et souple, faisant intervenir la créativité, mais qui ne peut pas être analysée en termes des sept procédés basiques de Vinay et Darbelnet (1958). Pour Ballard, c'est par rapport à la reformulation, autour de l'équivalence, tout autant que de la réécriture, que se construit l'unité de traduction. Puisque cette réalisation passe par le cerveau du traducteur, il faut aussi intégrer la subjectivité.

Par rapport aux procédés de traduction, Ballard (2006) préfère parler d'opérations qui reflètent la triple démarche de la traduction, en l'occurrence, des opérations d'interprétation, de paraphrase et d'ajustement. Cela s'explique par le processus suivant : étant donné un texte, après une interprétation des formes, le traducteur met en rapport une unité constituante du texte source avec le système de la langue cible en vue de produire une

équivalence acceptable. Cette dernière est susceptible de contribuer à la réécriture d'un texte, dont l'équivalence globale par rapport au texte source doit s'accommoder d'ajustements internes dictés par sa cohérence et sa lisibilité.

À travers un réexamen dans la littérature des procédés de traduction disponibles, Đorđević (2017) a étudié l'applicabilité des procédés existants pour la traduction des textes non littéraires. L'auteure considère ce travail comme une contribution initiale dans ce domaine. Les exemples cités dans l'article proviennent d'un corpus de traduction non littérale authentique, compilés pendant 22 ans de travail de traduction officielle.

Dans la seconde moitié du vingtième siècle (1972), les études de traduction ont été reconnues comme un domaine de recherche indépendant, séparé de la linguistique ou de la littérature. Deux genres de traduction* ont été identifiés : traduction des textes littéraires et des textes non littéraires. Avec le développement des approches plus contemporaines de la traduction des textes autres que des romans, la traduction non littéraire s'est propagée progressivement à un grand nombre de genres : gouvernemental, légal, diplomatique, administratif, scientifique, manuel d'utilisation, encyclopédie, documentaires, journaux, etc. Les textes non littéraires ont une fonction essentiellement instructive et pédagogique, et les textes littéraires ont des valeurs esthétiques et artistiques, qui visent à évoquer des émotions et apporter un certain niveau de divertissement. Pour la traduction non littéraire, le registre de langue et des termes dans la langue source doivent être respectés dans la traduction ; le contexte culturel et une certaine façon de comprendre le monde doivent être adaptés pour les lecteurs de la langue cible.

Les procédés de traduction sont importants puisqu'ils fournissent des solutions spécifiques aux problèmes qu'un traducteur rencontre pendant la traduction, par exemple des termes spécifiques à une culture, ou la terminologie liée à un certain domaine d'expertise. La contribution principale peut être attribuée à Vinay et Darbelnet (1958), qui ont proposé la première typologie de procédés de traduction avec un but clairement méthodologique. D'autres typologies ou mêmes de nouveaux procédés, proposés en particulier par Nida (1964), Newmark (1988) et Munday (2016) peuvent être considérés comme complémentaires puisqu'ils étendent les procédés de SCFA aux éléments de culture et à une terminologie spécifique, qui se présentent au sein d'une certaine paire de langues ou dans un domaine d'expertise.

Le tableau 2.4 liste les procédés de traduction que Đorđević (2017) a considéré comme applicables pour traduire des textes non littéraires de l'anglais en serbe. Pour certains procédés, la définition a été adaptée. Elle a souligné qu'un traducteur doit avoir recours à des procédés de traduction variés, appliquer différentes stratégies et apprendre à combiner des approches différentes pour produire des traductions correctes.

2.2.2 Études spécifiques sur la paire anglais-chinois

Lu et Fang (2012) ont fait un réexamen de la théorie sur la traduction littérale de Newmark (1988), qui est censée être bénéfique pour la pratique et l'enseignement de la traduction, pour voir à quel point il est applicable pour la traduction entre l'anglais et le chinois.

Depuis la fin des années 1970, des théoriciens de traduction ont été influencés par le déconstructionnisme. Le premier paradigme orienté langue cible a été établi par des théoriciens de traduction fonctionnaliste, représentés par Vermeer (1987) et Nord (1997). Ce paradigme considère la langue source comme une simple "offre d'information" ou la "matière première" des traducteurs, et met plus accent sur des facteurs sociaux qui

Nom du procédé	Définition
Emprunt	Un traducteur choisit consciemment ce procédé quand aucun autre équivalent de traduction n'est disponible.
Transposition	Identique à celle de la SCFA, la transposition peut aussi être utilisée pour résoudre le manque de correspondance au niveau de grammaire, syntaxe et morphème.
Modulation	Identique à celle de la SCFA.
Adaptation	Identique à celle de la SCFA.
Compensation	Ce procédé permet au traducteur d'exprimer un certain effet stylistique qui est présent dans le texte source mais qu'il est difficile de reproduire dans le texte cible. Par exemple, ce procédé est nécessaire pour traduire des articles scientifiques ou académiques de l'anglais en serbe. L'anglais utilise la première personne du singulier ou la voix passive pour présenter les résultats de recherche, or le serbe utilise plutôt la première personne du pluriel.
Amplification	Un traducteur ajoute des détails qui ne sont pas présents ou exprimés en langue source mais sont nécessaires en langue cible à la compréhension du texte.
Économie	Un procédé inverse de l'amplification, où moins de détails sont fournis en langue cible que ceux présents en langue source. Ce procédé est plus souvent utilisé dans la traduction du serbe en anglais.

Tableau 2.4 – Les procédés de traduction applicables pour la traduction non littéraire selon Đorđević (2017)

influencent la traduction. Un autre paradigme orienté langue cible « *the Manipulation School* » (Snell-Hornby, 1995) a été créé pour l'étude de la traduction littéraire, qui a un but essentiellement descriptif et explicatif. Les deux paradigmes ont remis en question la position centrale du texte source, en attachant une importance excessive aux facteurs extralinguistiques dans la traduction.

Le paradigme « *the Manipulation School* » (1995) a apporté un virage culturel dans les études de traduction, avec des résultats fructueux dans la description, l'explication, et la prédiction des phénomènes de traduction. En revanche, pour des chercheurs qui sont plus dévoués à la pratique et l'enseignement, on s'éloigne trop du pragmatisme de traduction avec ce paradigme. Le but de Lu et Fang (2012) est de rectifier cette extrémité après ce virage culturel, en prêtant plus attention sur les études de traduction appliquées.

Peter Newmark, le grand théoricien des études de traduction, a mis un accent spécifique sur la traduction littérale dans son ouvrage « *A Textbook of Translation* » en 1988. L'opinion de Newmark est ancrée dans des débats traditionnels avec une longue histoire en Occident et en Chine sur deux méthodes de traduction basiques : la traduction littérale et la traduction libre. Puisque Newmark n'a jamais pratiqué la traduction entre l'anglais et le chinois, les difficultés spécifiques à cette paire de langues sont inimaginables pour lui. Ainsi Lu et Fang (2012) ont dû modifier une partie de sa théorie comme suit :

1) Une traduction littérale est une traduction qui suit de près la forme et le sens du texte source, alors qu'une traduction « *mécanique* » ou « *morte* » suit de près seulement la forme au détriment du sens. Elle peut être atteinte par une traduction mot-à-mot utilisée de façon absolue :

Texte source : 你们俩从小在一起长大，可算是**青梅竹马**了。

Traduction mot-à-mot : *You two have been growing up together since you were little things. You certainly* ***have green plums and bamboo horses***.

Traduction libre : *You two have been growing up together ever since you were a little boy and a little girl. You certainly* ***have had intimate childhood friendship***.

La traduction mot-à-mot est clairement incompréhensible, et seulement une traduction libre peut rendre la traduction correcte. Un traduction libre abandonne la forme du texte source et garde seulement le sens visé. Parfois la traduction libre est obligatoire parce que la traduction littérale est impossible, parfois les deux sont possibles et le traducteur choisit l'une des deux selon des facteurs différents. Une traduction clairement inexacte est causée par une traduction mot-à-mot « *mécanique* » ou une traduction excessivement libre.

2) Lu et Fang ont défini la « vérité » de traduction par le fait d'être fidèle au texte source au niveau factuel, stylistique et culturel. Dans la plupart des cas, la traduction littérale a une bonne performance à cet égard, mais quant à la vérité culturelle, le traducteur doit prendre en compte la compréhension des lecteurs cibles et utiliser la traduction littérale méticuleusement :

Texte source : 他是帝国主义的**走狗**。

Traduction littérale : *He was a* ***running dog*** *of imperialism.*

Traduction libre : *He was a* ***lackey*** *of imperialism.* [2]

Utilisé avec un sens métaphorique, le terme《走(marcher)狗(chien)》signifie «laquais» au sens figuré. Cet exemple montre que quand il existe une grande divergence culturelle, la traduction littérale peut apporter de la confusion.

3) L'exactitude est l'aspect le plus important de la traduction, qui en aucun cas ne peut être sacrifiée par un traducteur. Lu et Fang ont démontré que la traduction littérale ne garantit pas forcément l'exactitude. Le traducteur doit prendre en compte des facteurs tels que les habitudes d'expression dans la langue cible, les lecteurs cibles, la fonction esthétique, etc.

教练车 (littéralement : *coach vehicle*) : *training vehicle / student vehicle*

九折 (littéralement : *ninety percent discount*) : *10 percent discount*

En résumé, pour la traduction entre l'anglais et le chinois, les traducteurs doivent s'abstenir d'utiliser la traduction littérale dans les cas suivants :

— Pour des mots sources généraux, il n'existe pas d'équivalents en langue cible de traduction mot-à-mot qui soient satisfaisants, même si la traduction est déjà généralisée.

— Quand la divergence entre deux cultures est trop grande et que la traduction littérale génère de la confusion ou des malentendus.

— Quand la langue cible est trop différente de la langue source sur certaines habitudes d'expression, pour que la traduction littérale soit possible.

— Quand le traducteur pense que les lecteurs cibles ne vont pas apprécier une version littérale.

— Quand une traduction libre est plus belle et compréhensible et qu'elle ne déforme pas le sens.

Wang (2017) a combiné les théories de traduction existantes pour analyser les raisons qui causent la non équivalence entre le chinois et l'anglais au niveau culturel. Pour cette paire de langues, l'équivalence est un concept relatif basé sur la non équivalence. La

2. Lackey : sens figuré de « *laquais* ».

non équivalence signifie que la langue cible n'a pas d'équivalent direct pour un mot ou un segment dans le texte source (Baker, 1993). À cause des différences dans le contexte culturel, les habitudes linguistiques et l'idéologie qui existent entre la Chine et les pays où l'anglais est la première langue, il n'existe pas d'équivalence absolue, et nous ne pouvons pas atteindre l'équivalence complète (Lian, 2006).

En tant qu'une forme de communication, la traduction* est en fait une fusion culturelle et linguistique. Dans la traduction, les facteurs de culture sont plus importants que les différences linguistiques. L'auteur a listé des exemples des différences dans la culture historique, la culture régionale, les coutumes et les façons de penser, qui sont des facteurs qui peuvent causer la non équivalence. Face à ces questions, l'auteur a proposé les cinq « *stratégies* » principales (considérés comme des procédés par nous) ci-dessous, afin d'essayer d'atteindre une équivalence maximale.

— Conversion : la définition correspond à celle de la *Transposition* dans SCFA. Par exemple :

*The camel is **characterized** by the ability to go for long periods without water.*
→ 骆驼的**特点**是能够长期行走而不用喝水。("La caractéristique du chameau est [...]") (verbe traduit par un nom)

— Négation : la traduction adopte un point de vue inverse du texte source, cela ressemble à *Modulation* dans SCFA.

I hear everything. → 什么都瞒不过我。("Rien ne peut m'être dissimulé.")

— Amplification : aussi appelé *Addition*, qui consiste à ajouter certains mots pour expliquer la phrase qui est facilement compréhensible pour des lecteurs de la langue source, mais pas pour des lecteurs cibles (Newmark, 2001).

We have not lost control of our time, but every little things wasting time will cause our fear. The shortsighted people has sold his birthright for present, but the smart people will concern longer affection.
→ 我们尚未丧失对时间的控制，不过任何浪费时间的小事都引发人们对未来的担忧，**为了眼前的利益**，目光短浅的人在浪费现在的时间，但是聪明的人会为了更长远的未来考虑。("pour des intérêts immédiats")

Ici le segment chinois a été rajouté pour que des lecteurs chinois aient une meilleure compréhension du texte.

— Domestication : un style transparent et coulant est adopté pour minimiser l'étrangeté du texte source pour les lecteurs cibles (Venuti, 2004; Lian, 2006).

*She is a bit out of my class, don't you think ? If I did try to do anything, I'd only **get sent off with a flea in my ear**.* ("envoyer promener quelqu'un")
→ 你没看到她和我不怎么相称吗？我要干点什么，也只会**碰一鼻子灰**。
(littéralement "se cogner le nez contre les cendres", le sens figuré est "essuyer une rebuffade").

L'expression anglaise utilise une métaphore avec une ironie, qui est similaire en sens à l'expression chinoise. Ainsi la traduction paraît plus native.

— Dépaysement (*Foreignization*) : s'efforcer de préserver le style étranger le plus possible, viser à transmettre les caractéristiques de la culture source, en transplantant les nouvelles images et les nouveaux concepts dans la langue et la culture cible (Venuti, 2004).

美人卷珠帘，深坐蹙**蛾眉**。(poésie chinoise classique)

→ *A lovely woman rolls up the delicate bamboo blind.*
She sits deep within, knitting her ***moth eyebrows****.*

《蛾眉》 est une image caractéristique dans la culture chinoise. Il désigne littéralement les sourcils fins, incurvés et délicats, considérés comme beaux en Chine. Les poètes ont fréquemment utilisé cette image pour décrire l'apparence d'une femme. 《蹙蛾眉》 ("froncer les beaux sourcils") décrit de façon vivante l'apparence d'une belle femme avec une humeur triste. Ici par une traduction littérale de 《蛾眉》 , le traducteur garde l'image artistique du texte original, et introduit en même temps une nouvelle expression aux lecteurs cibles, laissant de la place à l'imagination.

En conclusion, l'auteur a souligné que maîtriser les styles de langues et le contexte culturel entre les deux langues peut réduire certaines différences culturelles pendant la traduction.

Afin de savoir si les divergences de traduction posent des défis à la tentative de concevoir des représentations sémantiques partagées cross-lingues, Deng et Xue (2017) ont étudié les divergences présentes dans la traduction chinois-anglais à l'aide d'un schéma d'alignement hiérarchique sur un corpus arboré parallèle. Ils ont utilisé le corpus décrit dans l'article de Li *et al.* (2012), et ont aligné manuellement environ 10 000 paires de phrases. Ce corpus contient des blogs en ligne, des messages de forum et des journaux. Il a été analysé manuellement en syntaxe sur les deux langues. L'alignement manuel dans leur travail est hiérarchique, c'est-à-dire au niveau du mot et au niveau du constituant. Cela est réalisé de façon à éliminer les conflits et les redondances entre les alignements de mots et d'arbres syntaxiques, afin d'éviter d'extraire des fausses divergences de traduction.

Dans ce cadre, Deng et Xue (2017) ont défini la divergence de traduction* (DT) comme suit :

Si n_c (chinois) et n_e (anglais) sont deux nœuds non-terminaux alignés, alors il existe une divergence de traduction entre n_c et n_e, si et seulement si au moins une des deux conditions suivantes est remplie :

— au moins un de tous les enfants immédiats de n_c ou n_e est non-aligné ou aligné à plus d'un nœud.

— tous les enfants immédiats de n_c et n_e sont alignés un à un, mais les nœuds enfants apparaissent avec différents ordres de mots sous n_e et n_e.

À partir de cette définition, ils ont pu trouver 62 809 instances de DT dans leur corpus HACEPT* (*Hierarchically Aligned Chinese-English Parallel Treebank*), toutes causées par des différences translingues, ou par des traductions non littérales. Dorr (1994) ignore les DT causées par les traductions non littérales, sans doute parce que c'est évitable si la traduction non littérale est changée en une traduction littérale. Deng et Xue (2017), au contraire, prennent ces instances de DT en considération parce que les traductions non littérales abondent dans les corpus réels, et les DT ainsi causées ne peuvent pas être ignorées et ont besoin d'être traitées. En utilisant l'alignement entre des nœuds non terminaux, ils extraient semi-automatiquement et catégorisent les divergences de traduction en sept types (voir tableau 2.5), et ils quantifient chaque type pour mieux guider la recherche en traduction automatique. Ils ont montré que leur corpus HACEPT peut être utilisé pour extraire des paires de segments hiérarchiques qui capturent les divergences de traduction identifiées dans le corpus. Par contre, les corpus arborés existants avec une structure plate ne sont pas bien adaptés pour ce genre d'extraction.

Divergence de traduction	Définition et exemple
Encodage lexical	Un élément lexical dans une langue est traduit par une chaîne de mots continue ou discontinue dans une autre langue. – Différence en lexicalisation : *prioritize NP* → 安排 NP 的 优先 顺序 ("to arrange the priority order of NP") 崛起 ("rise") → *rise to prominence* *under* *the table* → 在 桌子 下 – Traduction non littérale : *a few days ago* (traduction littérale : 几天前) → 日("day")前("ago")
Différence en transitivité	Un verbe dans une langue et son équivalent lexical ou traduction dans une autre langue diffèrent en transitivité. *complain about NP* → 抱怨("complain") NP 责怪("blame") NP VP → *blame NP for VP*
Absence de mots grammaticaux	Des mots grammaticaux spécifiques à une langue n'existent pas dans une autre langue. *the* *capital* → 首府("capital") *has* *become* → 成为("become") 一 个(measure word) 月 → *one month*
Différence en catégorie de segment	Un segment et sa traduction diffèrent en catégorie de segment, qui implique l'usage des mots grammaticaux et le changement de structure. *the further expansion of NP* → 进一步 扩大 NP ("further expand NP") 主管("oversee") NP → *in charge of NP*
Différence dans l'ordre de mots	再("again") 伤害("hurt") 我们("us") → *hurt us again* 发展("development") 速度("speed") → *the rate of development*
Éléments omis	Un constituant dans un segment n'a pas de correspondance manifeste, dû à une traduction non littérale ou certaines raisons grammaticales indépendantes, autres que l'absence de mots grammaticaux. 在("at") NP 方面("aspect") → *in NP* 停止("stop") VP → *They* *had stopped VP* 我("I") 喜欢("like") → *I like* *it* (le chinois permet l'omission de pronoms en sujet et en objet)
Paraphrases structurelles	Un segment et sa traduction sont des paraphrases structurelles l'un de l'autre. La différence en structure et le manque d'alignement de mots entre les deux causent la divergence de traduction. 是("be") 农村("coutryside") 的(particle) 孩子("kid") ("a child from a rural area") → *grew up in the countryside* Cas extrême : traduction des expressions idiomatiques : 面("front") 朝("face") 黄土("yellow soil") 背("back") 朝("face") 天("sky") → *toiling on the land*

Tableau 2.5 – Les sept types de divergence de traduction entre le chinois et l'anglais, catégorisés par Deng et Xue (2017). Le chinois désigne le mandarin dans ce cas

La plupart des travaux que nous avons présentés ci-dessus concernent les procédés de traduction au niveau sous-phrastique. Nous admettons que ces procédés ne couvrent pas toutes les opérations réalisées pendant la traduction. Les exemples de Ballard (2006) cités plus tôt montrent bien que des réorganisations au delà de la frontière d'une phrase sont aussi importantes, surtout pendant la traduction littéraire.

Dans cette thèse, nous nous limitons à l'étude des procédés de traduction au niveau sous-phrastique au sein d'une phrase. En nous basant sur les typologies précédentes, nous proposons la nôtre adaptée à notre besoin de recherche, qui sera présentée dans le prochain chapitre.

2.2.3 Études sur la traduction non littérale

Récemment, Ahrenberg (2017) et Chen *et al.* (2018) ont souligné l'importance d'étudier les procédés de traduction de façon plus fine et automatique. Nous inscrivons notre travail de thèse dans cette démarche. Nous résumons leurs contributions ci-après.

La qualité de la traduction automatique est assez bonne pour comprendre l'essentiel d'un texte, ou pour faciliter la conversation entre des locuteurs qui ne partagent pas une même langue. Par contre, la traduction humaine a souvent un but plus ambitieux, en l'occurrence, générer des textes qui satisfont aux normes linguistiques d'une langue cible et qui sont adaptés aux connaissances des lecteurs cibles. Quand la qualité de traduction est visée, par exemple pour publier des sous-titres, des journaux ou des articles de qualité, la participation des traducteurs humains est toujours nécessaire.

Bien que la Traduction Automatique (TA) et les études de traduction manquent de concepts partagés et d'une terminologie commune, un point d'intérêt partagé porte cependant sur l'évaluation de la qualité de traduction. Afin d'étudier quelles sont les différences entre la TA et la Traduction Humaine (TH), Ahrenberg (2017) a effectué une étude pour répondre à ces questions :

— Par rapport à la traduction humaine, quelles sont les caractéristiques d'un texte traduit automatiquement ?
— Quelles sont les opérations de traduction qui sont seulement effectuées par les traducteurs humains ?
— Quelles sont les actions nécessaires pour combler cette lacune entre TA et TH ?

Le texte source est un article d'opinion en anglais publié en 2017, et l'étude compare deux versions de traductions en suédois, une traduite par un traducteur humain, l'autre par *Google Translate* (NMT* : *Neural Machine Translation*). Pour cette comparaison, à part des données statistiques, il a utilisé les procédés de traduction de façon descriptive. Le problème lié à la taxonomie de procédés est de définir comment les appliquer en pratique. Étant donné cette complexité, l'auteur a étudié des phénomènes de traduction à gros grains : traduction littérale (*unshifted translation*), c'est-à-dire seulement les procédés obligatoires ou standards pour la langue cible sont utilisés, par opposition à la traduction avec des glissements (*shifts*). Des glissements sémantiques et structurels sont notés séparément autant que possible, et ils sont identifiés au niveau de la phrase et de la proposition.

Concernant les statistiques basiques, la traduction humaine est plus longue en nombre de mots et de caractères. Elle contient aussi plus de phrases, car le traducteur humain a divisé huit phrases en deux ou trois phrases plus courtes. Au niveau de la monotonicité, la TH montre quasiment deux fois plus de changements dans l'ordre de mots que la TA. Cela concerne des phrases plus longues et couvre des distances plus longues.

L'auteur a utilisé la TH comme référence, et calculé le score BLEU (Papineni *et al.*, 2002) et TER (Snover *et al.*, 2006) de la TA par rapport à cette référence. Les résultats sont montrés dans la figure 2.1 selon les sections différentes de traduction (littérale (*unshifted*) ou avec des glissements (*shifted*)). Nous pouvons voir que sur les traductions humaines non littérales, la TA est beaucoup moins performante.

Section	BLEU	Bleu(1)	Bleu(2)	TER
Unshifted	42.79	69.0	48.7	0.374
Shifted	16.84	48.2	23.6	0.662
All	23.27	59.6	30.7	0.621

FIGURE 2.1 – Tableau extrait de l'article de Ahrenberg (2017) : scores BLEU et TER pour différentes sections de traduction (traduction littérale versus traductions avec des glissements), en prenant la traduction humaine comme référence

À ce jour, il semble évident que la TA n'a pas atteint la qualité attendue permettant une publication, l'auteur a listé le nombre de chaque type d'édition qu'il juge nécessaire pour améliorer la traduction jusqu'au niveau de la publication. L'édition la plus fréquente est « *édition sur les mots* », qui consiste à remplacer des mots pour des raisons de style et d'exactitude. Il n'y a pas de garantie qu'après les post-éditions, la qualité ou l'expérience de lecture sera la même que celle offerte par la traduction humaine. Par contre, si le but est seulement de comprendre l'essentiel de l'article, la qualité de TA est adéquate.

Au niveau de la longueur, de l'ordre de l'information et de la structure, la TA est plus similaire au texte source que la TH, mais elle présente un répertoire plus restreint de procédés utilisés. Voici des procédés que le traducteur humain a utilisé et qui semblent hors de portée du système de TA :

— division de phrase : ce procédé apporte aussi l'insertion de nouveaux matériels, tels qu'un adverbe ou un sujet

— changement de catégorie ou fonction : par exemple un attribut adjectif traduit par une proposition relative, un adverbe traduit par un adjectif, etc.

— explicitation : commenter des noms dont les référents ne sont pas forcément connus par des lecteurs cibles ; ajouter des mots outils grammaticaux et des articles indéfinis (la TH le fait plus souvent que la TA)

— modulation : traduction avec un changement de point de vue

— paraphrase : la sémantique n'est pas exactement la même, mais le contenu est assez similaire pour préserver le message

En conclusion, l'auteur suggère que la prédiction automatique des procédés de traduction est un sujet de recherche futur, comme par exemple les tâches partagées sur la prédiction des efforts de post-édition ou celles concernant la qualité de traduction.

Dans la traduction, consciemment ou non, la paraphrase est utilisée comme une stratégie de traduction. Quand la fluidité, en plus de la fidélité, concerne la traduction automatique ou humaine, accéder aux équivalents sensibles au contexte est essentiel pour différentes tâches du traitement automatique des langues (Madnani et Dorr, 2010). L'étude de Chen *et al.* (2018) est motivée par le besoin d'extraire des exemples de traduction libre, en vue de servir comme référence pour des traducteurs humains, et d'améliorer la fluidité de la traduction automatique.

Bien que la ressource de paraphrases PPDB contienne trois sortes de paraphrases : lexical, sous-phrastique et syntaxique, il existe une restriction sur la catégorie syntaxique,

c'est-à-dire que cette dernière est partagée par une paire de paraphrases (voir la description des travaux sur la PPDB dans la section 3.2.3). En revanche, les paraphrases de différentes catégories et constructions syntaxiques sont particulièrement importantes dans la traduction. Hormis la fidélité, la fluidité est plus importante surtout pour une paire de langues avec des propriétés linguistiques nettement différentes, pour lesquelles la traduction littérale n'est pas toujours la meilleure option ou même parfois pas une option naturelle.

Pour identifier des traductions libres dans les phrases parallèles anglais-chinois, Chen *et al.* (2018) ont utilisé des scores de mécanisme d'attention d'un système neuronal d'une façon innovante. Leur hypothèse de travail est la suivante : dans un système NMT, l'encodeur génère une séquence de vecteurs de contexte pour une phrase source, ensuite le décodeur calcule une distribution de probabilité sur la traduction en utilisant un mécanisme d'attention sur les vecteurs de contexte (Bahdanau *et al.*, 2014; Luong *et al.*, 2015). Pour la plupart des travaux en NMT, les priorités sont mises sur des paires corrélées de façon plus forte dans l'alignement généré par le mécanisme d'attention, qui indiquent souvent des traductions fidèles ou standards. Pour d'autres traductions possibles et peut-être plus fluides que fidèles, la corrélation entre la source et la cible sera plus faible. À condition que des corpus parallèles bilingues soient de bonne qualité, ces parties corrélées de façon faible correspondent potentiellement à des traductions libres mais fluides. En excluant des parties connues en traduction littérale, les parties restantes qui sont moins bien alignées peuvent contribuer à des expressions paraphrastiques, bien qu'elles puissent être moins fréquentes et probablement restreintes en termes de style littéraire ou de but communicatif.

Puisque de grands corpus parallèles ont tendance à contenir du bruit (faux alignements, traduction erronée, etc.), les auteurs ont d'abord utilisé le score de la fonction de perte d'entropie croisée pour un classement des paires de phrases selon la qualité de traduction. Leur approche a été testée sur des corpus anglais-chinois et chinois-anglais avec des données d'apprentissage de NIST12 OpenMT, et des données de validation de NIST 2006. Le système est implémenté en se basant sur la boîte à outil `Nematus` (Sennrich *et al.*, 2017).

Ils ont présenté les résultats avec une perspective plus qualitative. Leur méthode montre bien qu'il existe au moins deux facteurs qui décident si une paire de phrases est utile pour fournir un exemple de traduction libre :

— le degré global de traduction littérale de la phrase entière : si la phrase entière est traduite plus ou moins de façon littérale ou très fidèle, le modèle NMT trouvera une probabilité relativement grande pour la phrase cible, donc la perte d'entropie croisée sera basse. Les phrases avec une perte extrêmement basse sont trop fidèles pour contenir un exemple de traduction libre, inversement celles avec une perte très élevée sont trop bruitées pour être des traductions correctes.

— la corrélation des segments révélée par le score d'attention au niveau sous-phrastique : plus le score est bas, plus c'est une traduction libre, ou plus elle est dépendante du contexte.

La différence sur le degré de traduction littérale pourrait être moins nette quand une traduction totalement littérale n'est pas trouvée dans le corpus, ou lorsqu'elle n'est simplement pas possible. Cela est plus souvent observé dans la traduction du chinois vers l'anglais, surtout pour les expressions idiomatiques de quatre caractères. Quand la fluidité est visée, il est souvent nécessaire d'exprimer le sens dans une autre forme syntaxique, qui pourrait paraître plus naturelle et idiomatique dans la langue cible. Par exemple :

因此，我可证明接受政府的提案，并非**一时冲动的决定**，而是已经我们商讨已久。

→ *Therefore, I can prove that I have not **rashly** accepted the government's proposal because we have discussed the issue for a long time.*

La phrase nominale 《一时冲动的决定》("une décision basée sur l'impulsion") est traduite par une phrase verbale « *(have not) rashly (accepted)* ». Les auteurs ont pour but d'extraire ces exemples, qui ont de la valeur pour aider les traducteurs humains à se rendre compte des relations subtiles entre la fidélité et la fluidité.

Les résultats préliminaires et leurs observations ont servi comme preuve de concept, qui montre que leur approche de détection de traduction libre en utilisant les scores d'attention est réalisable en pratique et potentiellement effective. L'utilisation de ces scores doit être affinée pour indiquer la frontière précise des segments qui correspondent aux traductions non littérales. Une application possible de l'extraction et du classement des traduction libres par les scores d'attention, serait de compléter, d'une façon systématique et organisée, les phrases exemples pour des dictionnaires. Les perspectives de travail incluent l'extension de la méthode pour regrouper ensemble les exemples qui utilisent des stratégies similaires de traduction ; étudier l'efficacité de la méthode pour viser des observations plus profondes à l'égard de la nature des phénomènes de traduction entre le chinois et l'anglais.

2.3 Conclusion

D'un point de vue linguistique, hormis des traductions littérales mot-à-mot, différentes versions de traductions humaines reflètent la richesse des expressions langagières, où des procédés de traductions variés sont utilisés. En raison des différences existantes entre les langues et les cultures, des traductions non littérales sont indispensables pour produire des textes corrects et naturels.

Dans ce chapitre, nous avons présenté un état de l'art ainsi que différentes typologies de procédés de traduction existantes et nous avons comparé ces typologies. Notre hypothèse de travail est **qu'il est possible de reconnaître automatiquement les différents procédés de traduction**. Cela correspond également à une partie des perspectives des articles de Ahrenberg (2017) et de Chen *et al.* (2018), que nous avons présentés. Dans le chapitre 4, nous reviendrons à ces typologies, sur la base desquelles nous proposons notre propre typologie pour l'annotation de corpus.

Dans l'introduction générale, nous avons signalé que l'extraction de paraphrases est un cadre possible de validation de nos études. Pour cette raison, dans le prochain chapitre, nous présentons un état de l'art autour de la paraphrase en TAL, ce qui permettra de poser notre problématique de recherche.

Chapitre 3

Étude des paraphrases en traitement automatique des langues naturelles

Sommaire

Dans le chapitre précédent, nous avons présenté les différentes typologies de procédés de traduction. Afin de valider la contribution principale de cette thèse sur la reconnaissance des procédés de traduction, un cadre de validation possible en TAL est l'extraction de paraphrases dans des corpus parallèles bilingues.

Pour fonder notre approche, nous présentons dans ce chapitre différentes études sur la paraphrase en TAL. Nous discutons d'abord les définitions et typologies variées de la paraphrase en linguistique et en TAL. Ensuite, nous présentons les techniques d'extraction et de génération automatique de paraphrases. Les tâches en aval, qui bénéficient des connaissances sur la paraphrase, sont également passées en revue.

Ayant alors montré le rôle central de la paraphrase en TAL, nous expliquons pourquoi les procédés de traduction semblent nécessaires pour améliorer la qualité des ressources de paraphrases.

3.1 Définitions et typologies de la paraphrase

L'exégèse biblique s'est développée comme un véritable genre littéraire pendant le Moyen Âge. À l'époque de la Renaissance, le mot « *paraphrase* » apparaît dans la langue française dans le sens de « *développement explicatif d'un texte* » chez Lefèvre d'Étaples en 1525. La paraphrase était une traduction amplifiée (en latin comme en langue vulgaire) de l'original biblique et « *une sorte de commentaire* » (Daunay, 2004). Le mot « *paraphrase* » dérive du mot latin « *paraphrasis* », et ce dernier est emprunté du mot

grec « παράφρασις » (*paraphrazein*). Il est composé de *para-* (à côté de) et de *phrasis* (discours).

Différentes définitions associées à la paraphrase dans le domaine de la linguistique ont été proposées (Harris, 1957; Honeck, 1971; Martin, 1976; Mel'cuk, 1988, 1992; Fuchs, 1982, 1994; Milićević, 2007), en élaborant des modèles de différents niveaux de complexité, ou en proposant une définition à un niveau très abstrait. En revanche, les modèles définis en linguistique théorique sont difficiles à appréhender et ne peuvent être mis en œuvre que si l'on utilise des représentations complexes. Les définitions plus abstraites sont difficiles à employer dans une tâche informatique (Bouamor, 2012).

Les théories linguistiques s'accordent sur le fait que paraphraser est un genre de reformulation*. Fuchs (1994) a distingué la reformulation explicative de la reformulation à visée imitative. La première consiste à interpréter un texte T via un autre texte T', avec un but d'expliciter le sens (par exemple l'exégèse biblique). La deuxième est tournée vers la production d'un texte, où on cherche à construire les formes d'expression à partir du sens d'un texte original (par exemple reformuler des textes en rhétorique classique).

Ces théories linguistiques attribuent aussi à la paraphrase des propriétés qui en font une relation d'équivalence sémantique. Paraphraser* est une opération qui change la forme d'un texte (les mots, la syntaxe), tout en préservant son sens (la sémantique). Elle consiste à produire un texte cible à partir d'un texte source afin de clarifier, expliciter ou développer certains aspects (Bouamor, 2012). De ce point vue, la paraphrase et la traduction partagent des points communs sur la préservation sémantique et les changements de forme. La paraphrase et la traduction peuvent être considérées ensemble comme la réécriture*, la première étant monolingue, la deuxième étant bilingue.

En traitement automatique des langues, des définitions variées ont été associées à la paraphrase, qui ont en commun de se fonder sur le principe de l'équivalence sémantique*. La paraphrase est définie comme un moyen alternatif exprimant, dans une même langue, le même contenu sémantique, la même information ou la même idée que la forme originale (Barzilay et McKeown, 2001; Fujita, 2005; Callison-Burch, 2007; Bhagat, 2009; Zhao *et al.*, 2009; Madnani et Dorr, 2010).

La notion d'équivalence sémantique a été modélisée selon plusieurs points de vue : en suivant les règles de la logique (Dras, 1999) ; en considérant la paraphrase comme une forme de traduction avec une même langue source et cible (Quirk *et al.*, 2004; Zhao *et al.*, 2008b) ; ou comme un phénomène d'implication textuelle bidirectionnelle (Malakasiotis, 2011). Bhagat (2009) affirme que, contrairement aux définitions basées sur des notions de logique très restrictives, même si certaines paraphrases potentielles ne sont pas équivalentes au sens logique, elles doivent être considérées comme paraphrases ou "quasi-paraphrases" pour des raisons purement pratiques. Par conséquent, les auteurs sur la paraphrase en TAL font souvent l'hypothèse que les paraphrases sont des constructions approximativement équivalentes sur le plan sémantique. Généralement, la définition de la paraphrase dépend fortement de l'application visée (par exemple : recherche de réponses à des questions, évaluation de la traduction automatique, aide à la rédaction, etc.). D'un point de vue applicatif, l'utilité des paraphrases se mesure uniquement par la capacité des systèmes à bien les exploiter dans un contexte spécifique.

Il existe plusieurs typologies de paraphrases en linguistique et en TAL. En linguistique, ces typologies ont été introduites généralement au sein d'une théorie abordant le phénomène paraphrastique (Martin, 1976; Mel'cuk, 1988; Milićević, 2007). Par conséquent, elles sont strictement liées à un cadre particulier et sont difficilement exploitables dans d'autres contextes. En TAL, Bouamor (2012) a passé en revue les anciennes typo-

logies proposées (Dras, 1999; Barzilay, 2003; Kozlowski *et al.*, 2003; Fujita, 2005; Max, 2010; Vila *et al.*, 2011), et a conclu que ces catégories peuvent être résumées dans une typologie fondée sur deux axes principaux : le niveau de granularité du texte et les différents niveaux d'analyse de la langue.

La paraphrase peut se produire à trois niveaux de granularité textuelle :

Paraphrase lexicale* (ou synonyme) : substitution synonymique qui laisse intact le cadre syntaxique de la phrase source (ex. *bouquin* ↔ *livre*).

Paraphrase sous-phrastique* : recouvre aussi bien une paire de mots qu'une paire de groupe de mots (syntagmes ou fragments textuels quelconques), dont la taille peut être aussi grande que nécessaire dans la limite d'une phrase et qui sont en relation d'équivalence sémantique dans un contexte donné (ex. *envisage-t-elle* ↔ *a-t-elle l'intention*). Ces unités textuelles peuvent, dans certains cas, se présenter sous forme de patrons liant des éléments textuels variables tels que *« X ne doute pas de Y »* ↔ *« X est sûr de Y »*.

Paraphrase phrastique* : deux phrases véhiculant le même contenu sémantique. Cela comprend aussi la situation où le locuteur doit maîtriser les aspects stylistiques : *« Vous n'êtes même pas en mesure de me donner ce renseignement. »* ↔ *« Tu n'es même pas fichu de me passer ce tuyau. »*

Concernant les différents niveaux d'analyse de la langue sur la paraphrase, Martin (1976) a introduit la distinction entre la paraphrase linguistique* (ou sémantique) (Cristea, 2001) et la paraphrase non linguistique* (pragmatique ou référentielle) (Fuchs, 1982). Cette distinction passe par la compréhension de la différence entre deux concepts linguistiques importants : phrase* et énoncé*. Nous reprenons les définitions de ces deux concepts proposées par Bouamor (2012) :

Définition ***Phrase*** *: Une phrase est un groupe stable et constant de constituants structurés pour exprimer une idée ou fournir une signification. Une phrase est construite selon des règles structurales de la syntaxe et selon des critères de grammaticalités bien définis.*

Définition ***Énoncé*** *: Lorsqu'une phrase est prononcée dans un certain contexte (circonstances, lieu, moment) et dans un certain cotexte (son entourage linguistique), elle devient un énoncé unique. L'énoncé est un phénomène variable lié à l'activité langagière. Il est relié à un contexte et il fournit le sens en fonction de la compréhension et de l'interprétation de ce dernier.*

La paraphrase linguistique est ainsi indépendante de la situation. Généralement un tel système de paraphrase comporte des règles de deux types :

— paraphrase syntaxique (Cristea, 2001) :
L'eau est limpide*, cela permet de voir les algues.* ↔ ***La limpidité de l'eau*** *permet de voir les algues.*

— paraphrase lexico-syntaxique (Bouamor, 2012) :
Pierre a prêté des disques à Jean. ↔ *Jean a emprunté des disques à Pierre.*

La paraphrase non linguistique comprend ces deux sous-catégories :

— paraphrase pragmatique (Martin, 1976) : P_j est une paraphrase pragmatique de P_i si, dans une situation donnée, P_j se réfère à la même intention que P_i.
Il y a un courant d'air. ↔ *Je veux que l'on ferme la fenêtre.*

— paraphrase référentielle (Fuchs, 1982) : il est nécessaire de connaître la référence de certains termes pour déterminer s'il existe une relation de paraphrase.
Il est allé te voir là-bas le mois dernier. ↔ *Jack est allé te voir à la maison le mois dernier.*

La nature multi-facettes et sans borne des phénomènes de paraphrase pose des difficultés pour son traitement automatique. La variété des visions sur la paraphrase est encore plus large si on considère l'analyse du discours et la psycholinguistique, qui ont aussi abordé ce sujet ; ou d'un point de vue diachronique, la rhétorique ou l'exégèse biblique. Les linguistes informaticiens ne parviennent pas à une définition définitive sur la paraphrase, d'où viennent ces définitions vagues : "exprimer une chose en d'autres termes" (Shinyama *et al.*, 2002), "façon alternative pour transmettre la même information" (Barzilay, 2003), "phrases ou segments qui transmettent approximativement le même sens en utilisant différent mots de surface" (Bhagat, 2009).

À l'égard de ces difficultés, Vila *et al.* (2014) ont fait un bilan des travaux en linguistique et en linguistique computationnelle portant sur les paraphrases. L'approche d'analyse linguistique vise à expliquer et formaliser les phénomènes de la paraphrase ; or en linguistique computationnelle, des méthodes et techniques sont développées pour traiter ces phénomènes dans leur système ou application.

Chaque technique en TAL traite une facette concrète de la paraphrase, qui est généralement partielle et *ad-hoc*. Compte tenu des lacunes de ces travaux, Vila *et al.* ont travaillé sur la caractérisation linguistique de la paraphrase, pour fournir au TAL des bases plus solides pour le développement des méthodes et des systèmes. Ils ont analysé les cas limites de la paraphrase avec des phénomènes linguistiques connexes (la coréférence et l'implication textuelle). En s'appuyant sur d'autres typologies existantes dans la littérature, ils ont proposé une nouvelle typologie de paraphrase, qui est incluse dans une structure hiérarchique. Cette typologie a été validée empiriquement par une annotation de plus de 5 700 paires de phrases dans trois corpus de paraphrases, et a été testée dans la détection automatique du plagiat (Barrón-Cedeño *et al.*, 2013).

Selon Vila *et al.*, les phénomènes de paraphrase existent dans un continuum qui va de l'identité absolue à l'absence de la similarité sémantique. En conséquence, où dessiner la limite entre les paraphrases et les non paraphrases est problématique. Ils considèrent qu'une frontière fixe et précise n'existe pas, elle dépend au contraire de la tâche et de l'objectif. Par exemple, la notion du « seuil de distortion » proposée par Fuchs (1994) est variable selon les sujets et les situations. Trois cas limites de paraphrase ont été discutés : perte de contenu (lié fortement à l'implication textuelle), connaissance pragmatique (connaissance encyclopédique et connaissance situationnelle) et changements de traits grammaticaux (personne, nombre, temps).

Nous voyons que la paraphrase est un phénomène linguistique complexe. Elle contient au moins ces différentes dimensions :

— niveau de langue : linguistique et non linguistique

— taille des unités couvertes : lexical, sous-phrastique, phrastique

— types de transformations : morphologique, lexico-syntaxique, syntaxique, sémantique

— types de connaissances requises

— registre de langue : spécialisé, non spécialisé ; familier, courant, soutenu

— degré d'équivalence sémantique

Dans cette thèse, nous nous focalisons sur la paraphrase linguistique au niveau sous-phrastique et nous utiliserons le terme « réécriture » pour désigner les reformulations textuelles qui ne gardent pas strictement la même sémantique que la phrase originale.

Après avoir vu la complexité des phénomènes de paraphrase, nous allons présenter plus en détail des travaux sur l'extraction et la génération automatique de paraphrases en TAL.

3.2 Extraction de paraphrase

3.2.1 Exploitation de corpus monolingues

Avant le recours massif aux approches du traitement statistique des langues naturelles, l'extraction de paraphrases s'appuyait souvent sur des transformations syntaxiques, des formes logiques, et des représentations sémantiques (McKeown, 1979; Muraki, 1982; Meteer et Shaked, 1988; Shemtov, 1996). Quelques travaux en génération de paraphrases ont utilisé la grammaire TAG* synchrone (*Tree-Adjoining Grammar*) non probabiliste (Shieber et Schabes, 1990; Dras, 1997, 1999; Kozlowski *et al.*, 2003).

Les premières approches pilotées par les données ont exploité des corpus monolingues. Les méthodes proposées ont notamment reposé sur l'analyse des contextes environnants (Barzilay et McKeown, 2001), des calculs de similarité fondée sur des arbres de dépendances (Lin et Pantel, 2001; Ibrahim *et al.*, 2003), la fusion d'arbres de constituants (Pang *et al.*, 2003), ou le regroupement de documents selon des critères de dates et de sujets (Dolan *et al.*, 2004). Les patrons lexico-syntaxiques ont été utilisés pour apprendre les hyperonymes ou les hyponymes (Hearst, 1992; Snow *et al.*, 2004).

Corpus monolingue brut Lin et Pantel (2001) ont proposé un algorithme non supervisé, nommé DIRT (*Discovery of Inference Rules from Text*), pour découvrir des règles d'inférences dans des textes monolingues non parallèles, par exemple « X est l'auteur de Y ≈ X a écrit Y ». Ils utilisent le concept de "règle d'inférence" car ils veulent inclure des paires qui ne sont pas exactement des paraphrases, mais sont pourtant liées et potentiellement utiles pour des systèmes de recherche d'information.

Les auteurs ont étendu l'hypothèse distributionnelle de Harris (1954), selon laquelle les mots qui apparaissent dans les mêmes contextes linguistiques partagent des significations similaires. Ils supposent que si des chemins extraits dans des arbres de dépendance ont tendance à connecter des ensembles de mots similaires, alors ces chemins partagent des significations similaires. Un chemin représente une relation sémantique binaire et indirecte entre deux mots pleins, à savoir deux entités nommées dans leur étude. Leur algorithme cherche des chemins similaires pour générer des règles d'inférence.

Bien que ces règles générées automatiquement soient facilement reconnaissables par des humains, il est difficile de constituer manuellement une telle liste riche de règles. Pour un système de recherche d'information, cet algorithme peut donc faciliter la construction manuelle des règles. Une question à résoudre à l'issue de leur travail implique de reconnaître la polarité des relations d'inférence, par exemple « X aggrave Y » est parmi les règles les plus similaires de « X résout Y ».

Corpus monolingue parallèle Le corpus parallèle monolingue est le type de corpus le plus naturel pour collecter les paraphrases (Bouamor, 2012). Par exemple, de multiples traductions des mêmes romans (Barzilay et McKeown, 2001), de multiples traductions de référence des mêmes phrases sources pour évaluer la traduction automatique, etc. puisqu'en général les traductions préservent le sens des phrases sources, mais en utilisant parfois différents mots ou expressions.

En se basant sur des traits contextuels et lexico-syntaxiques, Barzilay et McKeown (2001) ont présenté un algorithme d'apprentissage non supervisé qui extrait automatique-

ment des paraphrases au niveau lexical, sous-phrastique et syntaxique. Des traductions multiples en anglais des romans étrangers ont été utilisées comme données.

L'hypothèse est que des segments dans des paires de phrases alignées et se trouvant dans des contextes locaux similaires sont des paraphrases. Afin d'inférer automatiquement des contextes qui sont des bons indicateurs de paraphrases, des contextes qui entourent des mots identiques dans des phrases alignées ont été extraits et filtrés selon leur puissance de prédiction.

Un algorithme de « *co-training* » a été utilisé, qui entraîne le système sur deux ensembles de traits indépendants : un qui décrit la paire de paraphrases elle-même, l'autre qui décrit le contexte où les paraphrases apparaissent. Des paraphrases apprises sont ensuite appliquées au corpus pour apprendre de nouvelles règles de contexte. Cet algorithme itératif continue jusqu'à ce qu'aucune nouvelle paraphrase ne soit découverte.

Dans le guide d'annotation utilisé par des juges humains lors de l'évaluation, la paraphrase est définie comme une « équivalence conceptuelle approximative ». Les paires de paraphrases ont été évaluées sans contexte ou avec une seule phrase de contexte.

Pang *et al.* (2003) ont proposé des nouvelles représentations de paraphrases basées sur des automates à états finis (*Finite State Automata* (FSA)), ce qui permet d'encoder de façon compacte un grand nombre de paraphrases. Ils ont proposé des algorithmes d'alignement basés sur la syntaxe pour construire automatiquement ces FSAs (sous forme de treillis de mots), à partir de multiples traductions en anglais des textes étrangers. Des paraphrases lexicales et structurales peuvent être capturées, par exemple {*last week's fighting, the battle of last week*}. Ces représentations permettent aussi l'extraction des contextes dans lesquels ces paraphrases sont correctes.

Combiner les approches En vue de surmonter les limites des travaux précédents, Ibrahim *et al.* (2003) ont proposé une approche qui combine les techniques développées par Barzilay et McKeown (2001) et Lin et Pantel (2001). Étant donné les désaccords résumés par Dras (1999) sur la définition exacte de la paraphrase, Ibrahim *et al.* ont utilisé une définition opérationnelle : les paraphrases structurales sont grossièrement interchangeables au sein d'une configuration spécifique de structures syntaxiques. Leur méthode non supervisée est utilisée pour acquérir des paraphrases structurales à partir des corpus monolingues alignés au niveau de la phrase. Les paraphrases générées ont tendance à être plus longues en moyenne, et permettent de capturer des dépendances à plus longue distance.

Corpus monolingue comparable Dolan *et al.* (2004) ont étudié des techniques non supervisées pour acquérir des paraphrases au niveau phrastique, en utilisant des articles de journaux collectés en ligne et rassemblés selon des critères de date et de sujet.

Deux techniques ont été utilisées : (1) la simple distance d'édition entre chaînes de caractères et (2) une stratégie heuristique qui associe des paires de phrases initiales (qui sont sans doute des résumés) dans différents articles au sein d'un même groupe. Les deux jeux de données obtenus sont évalués via la mesure d'erreur d'alignement de mots (*Word Alignment Error Rate*). Une analyse manuelle est menée sur un échantillon de 100 paires de phrases. Les résultats montrent que la première technique basée sur la distance d'édition identifie des paires de phrases qui contiennent des changements lexicaux ou sous-phrastiques, ce qui facilite l'alignement de mots de façon considérable. Par contre, des variations plus complexes manquent dans ces données. Au contraire, la stratégie heuristique extrait des données avec plus de variétés, mais qui sont plus difficiles à aligner au niveau du mot. Ces données contiennent plus de paraphrases non triviales qui sont de plus grand intérêt pour la tâche d'acquisition de paraphrases phrastiques.

Le manque de corpus de paraphrases phrastiques annoté, de taille importante, et disponible au public a été un obstacle pour la recherche en identification et génération automatique de paraphrases. Face à ce besoin, Dolan et Brockett (2005) ont créé un corpus de paraphrases nommé `Microsoft Research Paraphrase Corpus`*, qui contient 5 801 paires de phrases. Chaque paire est annotée manuellement avec un jugement binaire pour déterminer si la paire correspond à des paraphrases.

Ce corpus est créé en utilisant des techniques d'extraction heuristiques, et un classifieur basé sur SVM (*Support-Vector Machine*) pour sélectionner des paraphrases phrastiques possibles à partir de larges corpus de journaux rassemblés selon les sujets. Ces paires de phrases sont ensuite soumises aux juges humains, qui confirment que 67% des paires sont en effet sémantiquement équivalentes (selon le critère "si les deux phrases signifient la même chose").

3.2.2 Exploitation de corpus parallèles bilingues

Nous avons passé en revue des travaux représentatifs qui exploitent des corpus monolingues. Une autre approche importante dans ce domaine exploite des corpus bilingues parallèles, disponibles en abondance pour certaines paires de langues et certains domaines. L'approche la plus étudiée repose sur l'équivalence de traduction entre segments [1]. Nous ferons référence à cette approche par le terme de méthode par pivot*, et la figure 3.1 illustre un exemple.

Définition ***Méthode par pivot*** *: l'hypothèse est que si deux segments dans la même langue partagent une ou plusieurs traductions communes (considérées comme des "pivots") dans une ou plusieurs langues étrangères, ils sont potentiellement des paraphrases.*

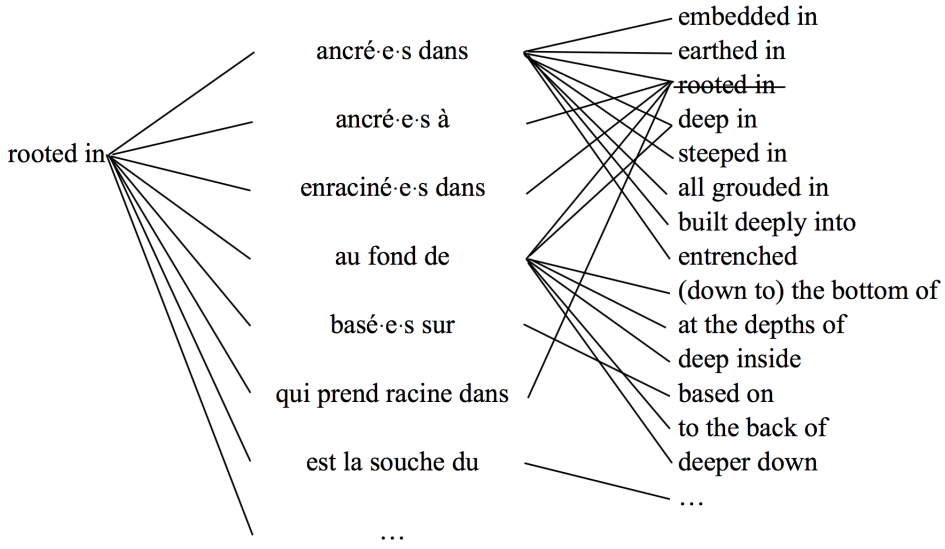

FIGURE 3.1 – Exemple d'extraction de paraphrases par pivot en français pour le segment anglais « *rooted in* »

Première proposition de l'approche Bannard et Callison-Burch (2005) sont les premiers à avoir proposé cette approche. Par rapport à l'approche qui exploite des corpus

1. Dans cette section, pour des raisons de simplicité, le terme « segment » va désigner des segments sous-phrastiques composés d'un ou plusieurs mots.

parallèles monolingues, l'utilisation des données parallèles bilingues permet de créer un inventaire plus large de paraphrases candidates, qui peuvent être utilisées dans plus de contextes variés. L'essence de leur travail consiste à aligner des segments dans un corpus parallèle bilingue, et à considérer différents segments anglais alignés avec des mêmes segments étrangers comme des paraphrases. C'est une extension des travaux sur la traduction automatique statistique basée sur des segments (PBSMT*, *Phrase-Based Statistical Machine Translation*) (Koehn *et al.*, 2003), et plus spécifiquement une exploitation des techniques d'alignement bilingue. Ils utilisent l'heuristique d'alignement de segments décrite dans les travaux de Och et Ney (2003), par une construction incrémentale de segments plus longs à partir de mots et segments qui ont des points d'alignement contigus.

Une différence importante avec l'approche qui s'appuie sur des corpus parallèles monolingues est que la méthode par pivot extrait fréquemment plus d'une seule paraphrase pour chaque segment (souvent à travers plusieurs traductions pivot), et la liste des paraphrases potentielles peut être longue. De plus, les paraphrases varient en qualité parce que les alignements de mots automatiques sont bruités. Ainsi les auteurs assignent une probabilité à chaque paraphrase candidate, ce qui permet de classer les paraphrases.

La probabilité de paraphrase* $p(e_2|e_1)$ est définie en termes de probabilités de modèle de traduction $p(f|e_1)$ (le segment original e_1 est traduit par un segment particulier f dans une autre langue), et $p(e_2|f)$ (la paraphrase candidate e_2 est une traduction du segment étranger f). Puisque e_1 peut être traduit par de multiples segments étrangers, une somme sur toutes les occurrences de différents f est faite pour calculer l'estimation de la probabilité de paraphrase :

$$\hat{e_2} = argmax_{e_2 \neq e_1} p(e_2|e_1) \tag{3.1}$$

où

$$p(e_2|e_1) = \Sigma_f p(f|e_1) p(e_2|f, e_1) \tag{3.2}$$

$$\approx \Sigma_f p(f|e_1) p(e_2|f) \tag{3.3}$$

$p(f|e_1)$ et $p(e_2|f)$ peuvent être calculés directement en utilisant l'estimation du maximum de vraisemblance, via une énumération de toutes les paires bilingues de segments cohérentes avec l'alignement de mots dans des paires de phrases (Koehn *et al.*, 2003; Och et Ney, 2004). Par exemple, la probabilité qu'un segment étranger f soit traduit par un segment anglais e est calculée comme suit : compter le nombre d'occurrences où e et f sont alignés, et le diviser par la somme des nombres d'occurrences où f est aligné avec d'autres segments e.

$$p(e|f) = \frac{count(e, f)}{\Sigma_e count(e, f)} \tag{3.4}$$

L'équation 3.1 définit la meilleure paraphrase $\hat{e_2}$, insensible au contexte où e_1 est utilisé. Puisque la meilleure paraphrase pourrait varier selon l'information contextuelle, la probabilité de paraphrase a été étendue pour inclure la phrase S :

$$\hat{e_2} = argmax_{e_2 \neq e_1} p(e_2|e_1, S) \tag{3.5}$$

où S permet de re-classer des paraphrases candidates en se basant sur des informations contextuelles supplémentaires. Bannard et Callison-Burch (2005) ont inclus une probabilité de modèle de langue sur la phrase S où e_1 est remplacé par e_2.

Ces configurations différentes ont été comparées :

1. alignements manuels de mots : correction de l'alignement automatique pour des segments à paraphraser (segment original → traduction pivot → paraphrase candidate)
2. alignements automatiques
3. pour améliorer l'exactitude de l'alignement : alignements automatiques réalisés sur d'autres corpus parallèles avec différentes langues cibles. L'équation 3.1 a été modifiée comme suit :

$$\hat{e_2} = argmax_{e_2 \neq e_1} \Sigma_C \Sigma_{f\_in\_C} p(f|e_1) p(e_2|f) \tag{3.6}$$

   où C est un corpus parallèle qui appartient à un ensemble de corpus parallèles.
4. la somme sur des traductions pivot différentes incluent peut-être des pivots de sens différents, ce qui va influencer le résultat d'évaluation. En complément de tous les composants ci-dessus, les paraphrases candidates sont limitées à celles de mêmes sens que le segment original, avec pour contrainte que les paraphrases candidates soient alignées avec les mêmes segments pivot dans une langue étrangère.
5. tous les composants ci-dessus, avec en plus un re-classement de paraphrases par un modèle de langue tri-gramme

Pour évaluer cette approche, Bannard et Callison-Burch (2005) ont extrait aléatoirement 46 expressions multi-mots anglaises à paraphraser dans WordNet (Miller, 1995). Le corpus parallèle utilisé à la base est la portion allemand-anglais du corpus `Europarl` version 2 (Koehn, 2005). Pour la troisième configuration, des corpus parallèles français-anglais, espagnol-anglais et italien-anglais ont été ajoutés. L'alignement automatique de mots est réalisé par `Giza++` (Och et Ney, 2003). En vue de mettre en évidence l'importance des contextes variés, pour chaque segment original, les auteurs ont remplacé son ensemble de paraphrases dans 2 à 10 phrases qui contiennent ce segment original. Deux locuteurs natifs anglais ont évalué la qualité des paraphrases classées en premier. Si le sens est préservé et que la phrase après la substitution reste grammaticale, la paraphrase est correcte ; et elle est incorrecte si un critère parmi les deux n'est pas satisfait. Ils ont aussi mesuré la préservation du sens sans tenir compte de la grammaticalité.

La meilleure exactitude est obtenue avec des alignements manuels (74,9%), mais le re-classement par le modèle de langue diminue le résultat au lieu de l'améliorer (71,7%). La mesure sur la préservation du sens seul est plus haute (84,7%). Cela suggère que les contextes variés influencent la grammaticalité plus que le sens de la phrase. L'utilisation des alignements automatiques fait baisser le résultat de façon significative (55,3% avec le re-classement selon le modèle de langue), et l'ajout des informations d'alignement dans d'autres corpus parallèles l'améliore un peu (57,4% avec le re-classement). Le contrôle des sens améliore les résultats de façon plus importante (61,9% avec le re-classement), ce qui montre l'importance de contrôler les pivots de sens différents.

Ajout de contraintes syntaxiques Bien que toutes les applications qui utilisent des paraphrases exigent de préserver le sens, certaines exigent en plus que la phrase après la substitution reste grammaticale. Callison-Burch (2008) a affiné l'approche par pivot avec l'introduction des contraintes syntaxiques. La qualité des paraphrases générées s'améliore de façon significative quand les paraphrases doivent obligatoirement avoir la même catégorie syntaxique que le segment original.

L'heuristique d'alignement de segments utilisée dans la traduction automatique (Och et Ney, 2004) permet aux mots non alignés d'être inclus dans les frontières des segments source ou cible (voir l'explication dans la figure 3.2). Cette inclusion peut apporter des effets non désirables pour l'extraction de paraphrases. Selon la définition de Bannard et Callison-Burch (2005), dans ce cas, « *equal* », « *create equal* » et « *to create equal* » peuvent être considérés comme des paraphrases, parce qu'ils sont alignés au même segment étranger. Ces résultats d'extraction de paraphrases candidates qui sont une partie du segment original (*sub-phrase*) ou qui contiennent le segment original (*super-phrase*) rendraient en général la phrase après la substitution non grammaticale. De plus, Callison-Burch (2008) a démontré que la quantité de ces phénomènes est non négligeable dans les résultats d'extraction.

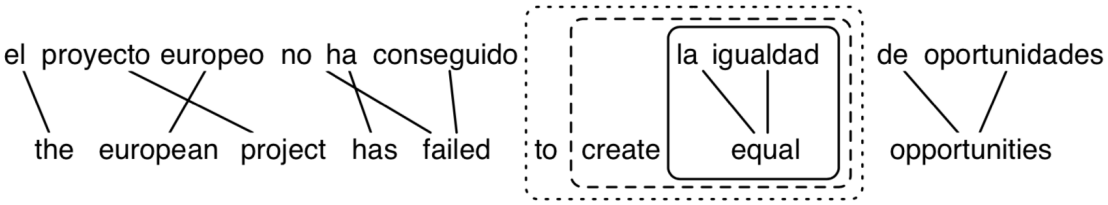

FIGURE 3.2 – Figure extraite de l'article de Callison-Burch (2008) : l'interaction entre l'heuristique d'extraction de segments et les mots non alignés signifie que le segment espagnol « *la igualdad* » ("l'égalité") peut être aligné avec « *equal* », « *create equal* » et « *to create equal* »

L'essence du travail de Callison-Burch (2008) consiste à exiger que les paraphrases aient la même catégorie syntaxique que le segment original. Deux méthodes d'application des contraintes syntaxiques sont possibles (lors de l'extraction de segments ou lors de la substitution de paraphrases dans la phrase originale).

Pour la première méthode, l'auteur redéfinit la probabilité de paraphrase :

$$\hat{e_2} = argmax_{e_2 \neq e_1 \wedge s(e_2)=s(e_1)} p(e_2|e_1, s(e_1)) \tag{3.7}$$

où $s(e_1)$ signifie la catégorie syntaxique de e_1, et la probabilité $p(e_2|e_1, s(e_1))$ peut être calculée approximativement comme suit :

$$\Sigma_{c \in C} \frac{\Sigma_f p(f|e_1, s(e_1)) p(e_2|f, s(e_1))}{|C|} \tag{3.8}$$

où c est un corpus parallèle qui appartient à un ensemble de corpus parallèles C.

Un nouvel algorithme d'extraction de segments qui s'appuie sur l'arbre d'analyse syntaxique du côté anglais est défini, ce qui limite l'adaptation de cette approche aux langues qui disposent d'analyseurs syntaxiques. Par la redéfinition de la probabilité de paraphrase, l'espace des paraphrases possibles est divisé selon leur catégorie syntaxique. Après ce changement, le pourcentage des meilleures paraphrases qui sont des *sub-phrase* diminue significativement (de 73% à 24%). Le pourcentage total des paraphrases qui sont des *sub-phrase* ou *super-phrase* du segment original diminue de 34% à 12%.

En vue de garder la couverture large de la méthode pivot basique, cette approche n'est pas limitée aux segments qui sont des constituants syntaxiques valides. Pour cela, l'auteur a introduit les étiquettes syntaxiques complexes en changeant encore l'heuristique d'extraction de segments, en se basant sur la grammaire CCG* *(Combinatory Categorial Grammar)* (Steedman et Baldridge, 2011).

Prenons cette phrase comme exemple : « *How do we create equal rights ?* ». Pour le segment « *create equal* » qui n'est pas un constituant syntaxique valide, une parmi ses trois étiquettes syntaxiques complexes en notation CCG est VP/(NP/NNS). Cela signifie que « *create equal* » est un segment verbal (VP) auquel il manque un segment nominal (NP) à droite. Et pour ce segment nominal, il manque à son tour un nom au pluriel (NNS) à sa droite.

La deuxième méthode applique les contraintes syntaxiques lors de la substitution de paraphrases dans la phrase originale.

Dans cette configuration, les paraphrases sont limitées à celles avec la même catégorie syntaxique que le segment à remplacer. Puisque chaque segment possède un ensemble d'étiquettes CCG différentes au sein d'une phrase (au lieu d'un seul symbole non terminal), Callison-Burch (2008) ont proposé trois façons de choisir l'étiquette :

1) Choisir simultanément la meilleure paraphrase et la meilleure étiquette pour le segment dans l'arbre syntaxique P de la phrase de test :

$$\hat{e_2} = argmax_{e_2 \neq e_1} argmax_{s \in CCG-labels(e_1,P)} p(e_2|e_1, s) \tag{3.9}$$

2) Effectuer une moyenne sur toutes les étiquettes générées pour le segment dans l'arbre syntaxique :

$$\hat{e_2} = argmax_{e_2 \neq e_1} \Sigma_{s \in CCG-labels(e_1,P)} p(e_2|e_1, s) \tag{3.10}$$

3) Puisque la suite des étiquettes CCG dans une phrase donnée est très spécifique, le nombre des paraphrases compatibles est restreint. Ainsi les auteurs proposent d'utiliser la paraphrase ayant la plus grande probabilité, quelle que soit la catégorie syntaxique.

$$\hat{e_2} = argmax_{e_2 \neq e_1} argmax_{s \in \cap \forall T\_in\_C CCG-labels(e_1,T)} p(e_2|e_1, s) \tag{3.11}$$

Le corpus d'entraînement contient dix corpus parallèles de l'anglais vers dix langues étrangères. L'ensemble de test contient 380 segments venant d'une tâche partagée sur la traduction automatique. Huit configurations différentes ont été testées (quatre configurations principales, avec un re-classement selon un modèle de langue en plus pour chaque configuration). Les annotateurs jugent la qualité des paraphrases candidates selon une échelle à cinq points (de 1 à 5) sur le degré de la préservation du sens et de la grammaticalité. Comme dans le travail de Bannard et Callison-Burch (2005), plusieurs phrases de contexte ont été choisies pour chaque ensemble de paraphrases, vu que la qualité de paraphrase varie selon le contexte de phrase (Szpektor *et al.*, 2007). Les résultats montrent que l'ajout de contraintes syntaxiques améliore la préservation du sens et la grammaticalité. Selon le critère strict de l'exactitude de paraphrases (le sens et la grammaire sont tous corrects), la meilleure configuration (une moyenne sur toutes les étiquettes CCG plus un re-classement par le modèle de langue) améliore le résultat de 19% par rapport à la méthode basique par pivot.

Extraction des paraphrases syntaxiques Les paraphrases syntaxiques ont un potentiel plus grand pour la généralisation et pour capturer des transformations paraphrastiques intéressantes, par exemple :

NP_1's $NP_2 \rightarrow$ the NP_2 of the NP_1

(*the committee's second proposal* $\rightarrow$ *the second proposal of the committee*)

Bien que des paraphrases syntaxiques riches aient été extraites dans des corpus parallèles monolingues, la couverture est contrainte par la disponibilité limitée de ce genre de données.

Dans le but d'obtenir une meilleure généralisation, Zhao *et al.* (2008b) ont utilisé des arbres de dépendance du côté anglais dans un corpus parallèle anglais-chinois, pour apprendre des patrons de paraphrase qui incluent des variables de partie du discours. Ils ont proposé un modèle log-linéaire pour calculer la probabilité de deux patrons et exploité des fonctions de traits basés sur MLE (*Maximum Likelihood Estimation*) et la pondération lexicale (*lexical weighting*).

En vue de savoir à quel point des paraphrases au niveau phrastique peuvent être extraites des corpus parallèles, Ganitkevitch *et al.* (2011) ont étendu l'approche bilingue par pivot qui intègre des contraintes syntaxiques (Callison-Burch, 2008) à une méthode qui inclut le modèle SCFG* (*Synchronous Context-Free Grammar*) (Aho et Ullman, 1972). Le modèle SCFG est de nouveau utilisé pour la traduction automatique statistique (SMT* : *Statistical Machine Translation*) dans les travaux de Chiang (2005).

Une grammaire SCFG probabiliste $\mathcal{G}$ est définie par :

$$\mathcal{G} = \langle \mathcal{N}, \mathcal{T}_{\mathcal{S}}, \mathcal{T}_{\mathcal{T}}, \mathcal{R}, \mathcal{S} \rangle \tag{3.12}$$

où $\mathcal{N}$ est un ensemble de symboles non terminaux ; $\mathcal{T}_{\mathcal{S}}$ et $\mathcal{T}_{\mathcal{T}}$ sont des vocabulaires de langue source et cible ; $\mathcal{R}$ est un ensemble de règles, et $\mathcal{S} \in \mathcal{N}$ est le symbole de la racine.

Les règles dans $\mathcal{R}$ sont de la forme suivante :

$$C \rightarrow \langle \gamma, \alpha, \sim, w \rangle \tag{3.13}$$

où le côté gauche de la règle $C \in \mathcal{N}$ est un non terminal ;

$\gamma \in (\mathcal{N} \cup \mathcal{T}_{\mathcal{S}})^*$ et $\alpha \in (\mathcal{N} \cup \mathcal{T}_{\mathcal{T}})^*$ sont des chaînes de symboles terminaux et non terminaux, ayant un nombre équivalent de non terminaux ;

$\sim$ constitue une fonction de correspondance un-à-un entre les non terminaux dans γ et α ;

un poids positif w est assigné à chaque règle, qui reflète la probabilité de la règle.

Dans l'approche de la traduction automatique basée sur des segments, des paires de segments bilingues sont extraits à partir de phrases parallèles alignées automatiquement au niveau du mot (Och et Ney, 2004). Ces heuristiques d'extraction de segments ont été étendues pour extraire des règles de grammaire synchrone, dans un corpus parallèle analysé en syntaxe et aligné au niveau du mot (Chiang, 2005).

Ganitkevitch *et al.* (2011) ont adopté la méthode d'extraction de règles SCFG pour tous les segments, y compris ceux qui ne sont pas des constituants syntaxiques valides. Suivant le travail de Zhao *et al.* (2008a), au lieu d'assigner un seul poids w, ils ont défini un ensemble de fonctions de traits $\vec{\varphi} = \{\varphi_1 ... \varphi_N\}$ qui sont combinés dans un modèle log-linéaire, pour calculer le coût d'application d'une règle :

$$w = -\sum_{i=1}^{N} \lambda_i log_{\varphi_i} \tag{3.14}$$

où les pondérations $\vec{\lambda}$ de ces fonctions de traits sont ajustées pour maximiser une fonction objective comme BLEU (Papineni *et al.*, 2002), en utilisant la procédure MERT (*Minimum Error Rate Training*) (Och, 2003). Les traits typiques utilisés en SMT incluent les probabilités de traduction de phrase, la pondération lexicale, la pénalité d'application de règles et la probabilité du modèle de langue (Koehn *et al.*, 2003).

Pour créer une grammaire de paraphrases à partir d'une grammaire de traduction, Ganitkevitch *et al.* (2011) ont étendu l'approche par pivot sensible à la grammaire proposée par Callison-Burch (2008) au modèle SCFG. Ils supposent d'abord une grammaire qui

traduit d'une langue donnée vers l'anglais, puis, pour chaque paire de règles de traduction où le côté gauche C et la chaîne étrangère γ correspond à :

$$C \rightarrow \langle \gamma, \alpha_1, \sim_1, \vec{\varphi_1} \rangle \tag{3.15}$$

$$C \rightarrow \langle \gamma, \alpha_2, \sim_2, \vec{\varphi_2} \rangle \tag{3.16}$$

en suivant l'intuition que deux chaînes anglaises α_1 et α_2 traduites par la même chaîne étrangère γ peuvent être considérées comme ayant le même sens, ils créent une règle de paraphrase :

$$C \rightarrow \langle \alpha_1, \alpha_2, \sim, \vec{\varphi} \rangle \tag{3.17}$$

où la relation de correspondance $\sim$ est ajustée pour refléter l'alignement combiné des non terminaux :

$$\sim \; = \; \sim_1^{-1} \circ \sim_2$$

Le côté source commun γ implique que α_1 et α_2 partagent le même ensemble de symboles non terminaux. Dans le calcul des traits $\vec{\varphi}$ à partir de $\vec{\varphi_1}$ et $\vec{\varphi_2}$, ils ont suivi l'équation 3.1, qui génère des traits sur la probabilité de paraphrase lexicale et sous-phrastique. De plus, ils ont ajouté un trait booléen qui indique si la paraphrase est identique à la phrase source ; et un indicateur de changement d'ordre des non terminaux, afin de promouvoir des paraphrases plus complexes qui exigent le changement d'ordre structural. Enfin ces règles de paraphrase sont appliquées dans le décodeur `Joshua` (Li *et al.*, 2010) de SMT pour générer des paraphrases syntaxiques.

Une analyse approfondie des types de paraphrases structurales obtenues a été réalisée. À travers ces expériences, ils ont pu ainsi démontrer que le corpus parallèle peut être utilisé pour générer des paraphrases phrastiques.

La génération de paraphrases phrastiques est au cœur de plusieurs tâches de génération texte-à-texte (simplification textuelle, résumé automatique, compression de phrases, etc.). Ganitkevitch *et al.* (2011) ont proposé d'utiliser leur grammaire de paraphrase pour générer de nouvelles phrases en utilisant le décodeur de SMT et les techniques d'optimisation des paramètres. Leur cadre de travail a été adapté pour la tâche de compression de phrases, en ajoutant de nouveaux traits pour le décodeur et proposant une nouvelle fonction objective en imitant la mesure BLEU. Un corpus de référence a été créé pour le développement et le test, et l'évaluation humaine sur une échelle à cinq points a été utilisée. Des résultats comparables à ceux de l'état-de-l'art ont été obtenus avec cette adaptation.

Intégration des traits sur la similarité distributionnelle monolingue Nous avons vu qu'il existe des approches qui s'appuient sur des données parallèles bilingues, et d'autres sur des corpus monolingues via des méthodes distributionnelles.

Chan *et al.* (2011) ont amélioré la méthode par pivot en utilisant la similarité distributionnelle monolingue pour reclasser des paraphrases candidates. Des données brutes monolingues fournissent une source d'information complémentaire, ce qui peut réduire les erreurs générées par des méthodes bilingues par pivot. Des résultats expérimentaux montrent que le score monolingue des paraphrases extraites par pivot possède une corrélation plus forte (de façon significative) par rapport à des jugements humains sur la grammaticalité.

Après les études préliminaires de Chan *et al.* (2011) et en se basant sur leur travail précédent (Ganitkevitch *et al.*, 2011), Ganitkevitch *et al.* (2012) ont combiné ces deux sources d'information complémentaires. Ils ont intégré directement des traits basés sur la similarité distributionnelle monolingue dans le modèle log-linéaire de leur système d'extraction de paraphrases.

Les méthodes qui s'appuient sur des corpus monolingues mesurent la similarité des segments en se basant sur des traits contextuels. Le vecteur $\vec{s_{e,i}}$ définit un ensemble de traits qui capturent le contexte de la i-ème occurrence d'un segment e dans le corpus. Une agrégation sur toutes les occurrences de e constitue une signature distributionnelle du segment e : $\vec{s_e} = \Sigma_i \vec{s_{e,i}}$. Suivant l'intuition que des segments avec des sens similaires apparaissent dans des contextes similaires, nous pouvons quantifier la qualité d'un segment e' comme la paraphrase d'un segment e par le calcul de la similarité cosinus entre leur signature distributionnelle :

$$sim(e, e') = \frac{\vec{s_e} \cdot \vec{s_{e'}}}{|\vec{s_e}||\vec{s_{e'}}|} \tag{3.18}$$

Ganitkevitch *et al.* (2012) ont aussi défini la similarité distributionnelle pour des patrons de paraphrases discontinues qui contiennent des trous au niveau des constituants. Voici un exemple de ce genre de patrons :

NN 's NP in the long term $\rightarrow$ the long-term NP of NN

(*IBM's goals in the long term* $\rightarrow$ *the long-term goals of IBM*)

Les auteurs décomposent les segments en sous-phrases continues $\mathcal{P}(\mathbf{a}) = \{\langle e, e' \rangle\}$ si elles sont cohérentes avec l'alignement monolingue des mots $\{a\}$. Le score final pour une paire de patrons est la moyenne des scores de similarité sur toutes les paires extraites.

$$sim = \frac{1}{|\mathcal{P}(\mathbf{a})|} \Sigma_{(e,e') \in \mathcal{P}(\mathbf{a})} sim(e, e') \tag{3.19}$$

Ainsi leur travail consiste à ajouter l'information monolingue sur des paires de paraphrases déjà existantes. Deux traits supplémentaires sim_{ngram} et sim_{syn} sont ajoutés dans le vecteur de fonctions de traits $\vec{\varphi}$:

— Modèle n-grammes : à partir du corpus `Web 1T 5-gram` (Brants et Franz, 2006; Lin *et al.*, 2010), pour chaque segment p, ils cherchent des n-grammes de forme wp et pv, où w et v sont des mots simples. Le trait est le nombre d'occurrences des n-grammes wp et de pv.

— Modèle syntaxique : traits lexicaux et morpho-syntaxiques sensibles à la position, en unigramme et bigramme, extraits à partir d'une fenêtre de trois mots à gauche et à droite du segment ; traits basés sur les dépendances syntaxiques à l'intérieur et à l'extérieur du segment, avec des mots et des étiquettes morpho-syntaxiques correspondants ; tous les constituants qui gouvernent le segment ; des étiquettes CCG gouvernantes et manquantes avec la direction (gauche ou droite) indiquée.

Ils ont évalué cette approche combinée dans la tâche de compression de phrase, en comparant avec ILP (*Integer Linear Programming*) (Clarke et Lapata, 2008), dont la méthode est basée sur la suppression. Leur modèle basique qui s'appuie seulement sur la méthode par pivot montre des faiblesses sur la préservation de la grammaticalité dans la compression ; en ajoutant des informations distributionnelles monolingues venant du modèle n-gramme ou syntaxique, la grammaticalité et la préservation du sens sont améliorées. Comparée avec ILP, leur approche est meilleure sur la préservation du sens mais moins bonne sur la grammaticalité.

D'autres travaux basés sur la méthode par pivot Kok et Brockett (2010) ont introduit un modèle à base de graphes présenté sous le nom de *HTP (Hitting Time Paraphraser)*. Cette approche repose sur des parcours aléatoires et sur le temps d'atteinte (*hitting time*) afin d'extraire des paraphrases à partir de corpus parallèles multilingues. À la différence de la méthode par pivot classique, cette approche parcourt des chemins de longueur supérieure à 2, en utilisant l'information entre les nœuds représentant des segments dans la langue d'origine et dans des langues étrangères. Leur méthode permet aussi d'intégrer facilement des connaissances monolingues sous forme de nœuds spéciaux.

Mallinson *et al.* (2017) ont présenté le système PARANET, qui est un modèle neuronal de paraphrases basé sur la méthode par pivot bilingue. Ils ont revisité cette approche dans le contexte de traduction neuronale (NMT). Leur modèle basé purement sur des réseaux neuronaux représente des paraphrases dans un espace continu. Le système estime le degré de ressemblance sémantique entre des segments textuels de longueur arbitraire, ou génère des paraphrases candidates pour n'importe quelle entrée source. Cette approche utilise les scores en mécanisme d'attention pour identifier des parties sémantiquement équivalentes, et elle n'est pas dépendante des connaissances sur la syntaxe. Des paraphrases de niveaux variés sont capturées : mots, segments ou phrases, sans le besoin de créer explicitement une table de traduction de phrases. Les résultats expérimentaux obtenus dans le cadre de plusieurs tâches (prédiction de similarité, identification et génération de paraphrase) montrent que leur approche est plus performante que les approches conventionnelles par pivot basées sur les segments.

Depuis la première proposition de la méthode par pivot par Bannard et Callison-Burch (2005), plusieurs travaux ont continué à étendre et à améliorer cette approche. Nous avons récapitulé les travaux principaux dans cette direction. Les trois travaux présentés en détail sont les plus importants (Callison-Burch, 2008; Ganitkevitch *et al.*, 2011, 2012), qui ont servi comme techniques de base à la construction d'une ressource de paraphrases largement utilisée, que nous présentons dans la section suivante.

3.2.3 Travaux sur la ressource de paraphrases PPDB

La méthode par pivot a été mise en œuvre pour la construction de la ressource PPDB* (*ParaPhrase DataBase*) [2], aujourd'hui la plus grande ressource de paraphrases disponible, couvrant le niveau lexical, sous-phrastique et syntaxique.

PPDB 1.0 La première version de PPDB a été publiée par Ganitkevitch *et al.* (2013). Elle contient des paraphrases pour l'anglais et l'espagnol.

La partie anglaise contient plus de 220 millions de paires de paraphrases (73 millions au niveau sous-phrastique (une chaîne continue de mots), 8 millions au niveau lexical, et 140 millions de patrons paraphrastiques (transformations syntaxiques préservant le sens, qui contiennent des mots et des non terminaux)), dont 170 millions sont des paraphrases non identiques aux segments originaux. La construction de cette ressource a été rendue possible par l'utilisation d'un corpus parallèle de plus de 106 millions de paires de phrases, contenant plus de 2 milliards de mots anglais et couvrant 22 langues pivot. Ces diverses langues ont été traitées comme une langue pivot unique pour le calcul des scores des paraphrases anglaises. En procédant ainsi, il n'est pas possible de déterminer quelle langue pivot est meilleure pour générer des paraphrases. Par contre, le volume de paraphrases produites est supérieur à celui obtenu avec une seule langue pivot, et cette mé-

2. `http://paraphrase.org`

thode permet une augmentation de la qualité des paraphrases (Bannard et Callison-Burch, 2005).

La partie espagnole contient 196 millions de paraphrases, dont la construction s'est appuyée sur un corpus contenant 15 millions de paires de phrases.

Ganitkevitch *et al.* (2013) décrivent la collection de paraphrases comme une grammaire pondérée hors contexte synchrone (*weighted SCFG*) (voir la description plus haute de l'article de Ganitkevitch *et al.* (2011)). Des règles de paraphrases obtenues en utilisant cette méthode permettent de générer des transformations syntaxiques généralisées qui préservent le sens, par exemple :

$NP \rightarrow$ the NP_1 of NNS_2 | the NNS_2 's NP_1

Ainsi chaque paire de paraphrases dans PPDB partage une même catégorie syntaxique. Un mot peut être la paraphrase d'un segment sous-phrastique, et vice versa, tant qu'ils partagent la même catégorie syntaxique.

En poursuivant leurs travaux de 2012 (décrits dans la section précédente), Ganitkevitch *et al.* (2013) ont intégré un ensemble de scores de similarité distributionnelle comme traits dans le modèle log-linéaire (voir équation 3.14). Des traits contextuels en se basant sur les n-grammes ont été calculés pour 200 millions de n-grammes les plus fréquents dans le corpus de `Google n-gram` (Brants et Franz, 2006; Lin *et al.*, 2010) ; et des traits linguistiquement riches pour 175 millions de segments dans le corpus `Gigaword annoté` (Napoles *et al.*, 2012). Par rapport à leurs travaux en 2012, Ganitkevitch *et al.* (2013) ont ajouté des traits monolingues basés sur des lemmes et des entités nommées.

Pour estimer l'utilité de PPDB comme ressource pour les tâches d'étiquetage de rôles sémantiques ou d'analyse syntaxique, les auteurs ont analysé la couverture de prédicats et de paires prédicat-argument de Propbank (Kingsbury et Palmer, 2002) dans PPDB. Pour cela, l'analyse syntaxique du Penn Treebank (Marcus *et al.*, 1993) a été utilisée pour faire la projection des annotations de Propbank vers les patrons syntaxiques de PPDB, et les arguments ont été remplacés par des non terminaux syntaxiques, afin de pouvoir chercher des paraphrases qui correspondent au prédicat annoté. Afin de quantifier le compromis entre la précision et le rappel de PPDB dans ce contexte, ils ont choisi une estimation simple comme score pour chaque paire de paraphrases, une combinaison uniforme de la probabilité de paraphrase et des scores de similarité monolingue. Les résultats montrent un rappel de 52% pour les types de prédicats distincts ; et 27% pour les relations complètes prédicat-argument avec au plus deux arguments (*ex.* $S \rightarrow NNS$ expect S). Ces deux taux ne changent pas même si les auteurs ne gardent que des paraphrases de haute précision.

En vue d'évaluer la qualité des paraphrases, les auteurs ont jugé 1 900 paires de paraphrases des verbes présents dans Propbank (aléatoirement extraites), sur une échelle de 1 à 5 (5 étant le meilleur). Même avec une approche simple de pondération (combinaison uniforme), les scores de PPDB montre une corrélation claire avec les jugements humains.

Extension multilingue de PPDB Il y a eu d'autres efforts pour extraire des paraphrases non anglaises pour des applications diverses en TAL. Par exemple, des tables de paraphrases de cinq langues différentes ont été extraites comme une partie de METEOR-NEXT, une extension multilingue de la mesure METEOR pour évaluer la traduction automatique (Denkowski et Lavie, 2010). De façon similaire, des paraphrases extraites de façon automatique en arabe et en chinois ont été utilisées pour améliorer les systèmes de traduction anglais-arabe (Denkowski *et al.*, 2010) et chinois-japonais (Zhang et Yamamoto, 2002, 2005). Mizukami *et al.* (2014) ont travaillé sur la création d'une ressource de paraphrases japonaises. Bien que de bons résultats aient été obtenus par ces travaux,

à notre connaissance, seulement les collections de paraphrases générées par Denkowski *et al.* (2010) et Mizukami *et al.* (2014) sont disponibles au public.

La version multilingue de PPDB contient des paires de paraphrases pour 23 langues, dont le français (Ganitkevitch et Callison-Burch, 2014). Puisque des analyseurs syntaxiques statistiques sont disponibles pour l'anglais, et que ces annotations en anglais permettent de créer des paraphrases syntaxiques pour des langues qui ne disposent pas d'analyseur syntaxique, cette extension de ressource a été réalisée en utilisant l'anglais comme langue pivot. Ganitkevitch et Callison-Burch (2014) ont projeté la syntaxe de l'anglais sur les phrases étrangères via des alignements de mots automatiques, en suivant les travaux de Zollmann et Venugopal (2006) et de Weese *et al.* (2011). Seul le côté anglais de chaque corpus parallèle a besoin d'être analysé, tâche réalisée avec l'analyseur `Berkeley` (Petrov *et al.*, 2006).

Chaque règle de paraphrase, contenant une grammaire SCFG (*Synchronous Context-Free Grammar*), est composée de quatre composants séparés par le symbole |||. Pour cette règle de paraphrase :

[VP] ||| *contenues dans le* [JJ, 1] [NN, 2] ||| *figurant dans le* [JJ, 1] [NN, 2] ||| [une collection de traits]

Le premier champ contient le symbole non terminal à gauche de la grammaire qui domine la règle SCFG ; le deuxième champ contient le segment original ; le troisième champ contient la paraphrase, et si c'est une règle syntaxique, elle a l'ensemble de symboles non terminaux identiques à ceux du segment original, mais ils peuvent apparaître dans un ordre différent. La correspondance entre des symboles est marquée par des indices, par exemple $[NP, 1]$ et $[NP, 2]$; le quatrième champ contient une collection de traits associés à la règle.

Dans cette version de PPDB, chaque règle est associée à 31 traits, notamment la probabilité de paraphrase, des scores de similarité distributionnelle monolingue, des traits utiles pour la compression de phrase, le score de la pondération lexicale (Koehn *et al.*, 2003), la différence en nombre de mots et en caractères, etc. Pour classer les paraphrases dans cette version, Ganitkevitch et Callison-Burch (2014) ont combiné un sous-ensemble de sept traits avec des pondérations *ad hoc* fondées sur les intuitions des auteurs :

$$\begin{aligned} SCORE = 1.0 * -logp(e_1|e_2) \\ +1.0 * -logp(e_2|e_1) \\ +1.0 * -logp(e_1|e_2, LHS) \\ +1.0 * -logp(e_2|e_1, LHS) \\ +0.3 * -logp(LHS|e_1) \\ +0.3 * -logp(LHS|e_2) \\ +100 * RarityPenalty \end{aligned} \tag{3.20}$$

où e_1 est le segment original ; e_2 est la paraphrase ; LHS signifie le symbole non terminal à gauche de la règle qui domine la grammaire SCFG ; $RarityPenalty$ marque la règle qui a été vue peu de fois, calculé par $exp(1 - c(e, f))$, où $c(e, f)$ est l'estimation de la fréquence de cette paire. Les auteurs admettent qu'une approche plus scientifique est de fixer le poids de chaque trait selon les jugements humains sur un échantillon aléatoire de paraphrases.

Le nombre de paraphrases obtenues pour chaque langue est à peu près proportionnel à la taille des corpus parallèles utilisés pour extraire des paraphrases pour cette langue-là. Pour les langues bien couvertes par des corpus parallèles, par exemple l'arabe, le français, le chinois, l'espagnol et le russe, la collection de paraphrases est trop grande. Par conséquent, Ganitkevitch et Callison-Burch (2014) ont créé des collections de tailles différentes : S (petit), M (moyen), L (large), XL (extra large), XXL et XXXL. Chaque ensemble plus grand est conçu pour doubler environ le nombre de paraphrases de chaque type : lexical, sous-phrastique et syntaxique. Un ensemble plus grand contient des paraphrases qui sont déjà dans un ensemble plus petit. Avant la division, les paraphrases ont été classées selon le score calculé avec l'équation 3.20, pour garantir que des paraphrases de meilleure qualité sont incluses dans un ensemble de taille plus petite.

Beaucoup de langues couvertes dans cette ressource sont morphologiquement plus complexes que l'anglais, et l'approche pivot via l'anglais a tendance à regrouper des variantes morphologiques d'un mot étranger dans le même ensemble de paraphrases. Par exemple « grand », « grande », « grands » et « grandes » partagent tous la traduction en anglais « *tall* ». En revanche, cette méthode distingue les paraphrases quand elles ont des catégories syntaxiques différentes. Cela permet de bien séparer des mots qui ont des sens distincts selon les parties de discours différentes (par exemple, en tant que nom, « *squash* » possède « *racquetball* » comme synonyme ; en tant que verbe, son synonyme est « *crush* »). Pour l'instant les étiquettes syntaxiques utilisées sont celles du Penn Treebank, appliquées à l'anglais. Des étiquettes adaptées et spécifiques aux langues étrangères sont plus pertinentes pour diviser et distinguer les ensembles de paraphrases s'il est besoin.

PPDB 2.0 Pavlick *et al.* (2015b) ont amélioré le classement des paraphrases dans PPDB, qui atteint une corrélation plus grande avec les jugements humains par rapport aux classements heuristiques dans PPDB 1.0. De nouveaux traits ont été ajoutés pour chaque paire de paraphrases : relation d'implication à grain fin (Pavlick *et al.*, 2015a), similarité entre plongements lexicaux des paraphrases (Rastogi *et al.*, 2015) (chaque entrée est associée à un plongement lexical appris avec MVLSA (*MultiView Latent Semantic Analysis*)), information sur le changement de style (complexité et formalité) (Pavlick et Nenkova, 2015). Des traits pour chaque règle de paraphrase ont aussi été étendus par rapport à la version 1.0. La liste complète de ces traits est disponible dans le matériel supplémentaire de leur article.

Dans cette nouvelle version, un modèle de régression supervisé a été utilisé afin d'adapter les scores de paraphrase à des jugements humains. Afin d'entraîner le modèle, Pavlick *et al.* (2015b) ont collecté des jugements humains sur un échantillon de 26 455 paires de paraphrases dans PPDB. Les annotateurs ont assigné le score de préservation sémantique selon une échelle à 5 points, qui a été proposée par Callison-Burch (2008). Chaque paire a été annotée par 5 personnes, et la moyenne est calculée comme le score final. Ces 26k paires de paraphrases annotées sont à la disposition de la communauté. [3]

Avec tous les traits disponibles dans PPDB 1.0 et les 176 nouveaux traits développés, qui incluent notamment la similarité cosinus entre les plongements lexicaux, les traits sur le chevauchement lexical, les traits dérivés de WordNet, et les traits de la similarité distributionnelle (voir leur matériel supplémentaire pour plus de détails), un modèle de régression *Ridge* a été entraîné, pour lequel les paramètres de régularisation ont été optimisés avec une validation croisée sur les données d'apprentissage.

3. `http://www.seas.upenn.edu/~nlp/resources/ppdb-2.0-human-labels.tgz`

Ce nouveau classement de paraphrases a été évalué de deux façons :

— calculer la corrélation entre les scores de paraphrase (calculés automatiquement de différentes façons) et les 26k jugements humains collectés. Le résultat montre que leur nouvelle méthode de classement obtient la plus grande corrélation.

— mesurer la qualité des paraphrases classées parmi les premières positions (en termes de MRR (*mean reciprocal rank*) et AP (*averaged precision*)), pour 100 segments extraits aléatoirement de Wikipédia. Ils ont collecté des jugements humains pour la liste complète de paraphrases pour chaque segment, et comparé la capacité de chaque méthode de classer des bonnes paraphrases en premières positions. Leur méthode reste la plus performante même si les critères utilisés pour définir ce qui est une bonne paraphrase varient (recevoir un score humain moyen $\geq$ 3 ou 4 ou 4,5).

Ajout d'une sémantique interprétable à PPDB Jusqu'à ce stade, la relation entre des paires de segments dans PPDB a été définie approximativement comme équivalente. De manière importante, le travail de Pavlick *et al.* (2015a) a mis en évidence le fait qu'il existe d'autres relations sémantiques que l'équivalence stricte (paraphrase) dans une telle ressource obtenue via les équivalences de traduction. Cet article décrit une attribution automatique de diverses relations d'implication sémantique aux entrées de PPDB, pour que cette ressource devienne plus riche d'informations pour des tâches en aval, telle que la détection d'implication textuelle (RTE* : *Recognizing Textual Entailment*).

Étant donné deux phrases, souvent appelées texte (T) et hypothèse (H), un système de RTE doit déterminer si T implique H, T contredit H, ou T et H ne sont pas liés. Au contraire, l'extraction de paraphrases pilotée par les données évite de développer une définition claire sur l'équivalence du sens et se concentre plutôt sur des techniques qui peuvent produire des paraphrases (Barzilay, 2003). En revanche, l'utilité des ressources de paraphrases est limitée par la définition vague de la paraphrase. Une définition concrète est que les paraphrases sont en relation d'implication bidirectionnelle (Androutsopoulos et Malakasiotis, 2010). En réalité, il existe plus de nuances dans les paraphrases (Bhagat et Hovy, 2013), et les entrées dans la plupart des ressources de paraphrases ne correspondent pas à la définition sur l'implication bidirectionnelle. Par exemple, l'algorithme DIRT (Lin et Pantel, 2001) apprend des paires de vraies paraphrases, mais aussi des paires en contradiction, telle que « *X rises, X falls* ». L'approche par pivot apprend souvent des hyperonymes et des hyponymes comme des paraphrases, en raison des variations de structure de discours dans les deux langues (Callison-Burch (2007), section 3.3.4) ; elle apprend aussi des paires qui ne sont liées par aucune relation sémantique, à cause de faux alignements de mots ou de la polysémie dans la langue pivot.

Prenons cette paire de phrases comme exemple :

(T) Riots in Denmark were sparked by 12 editorial cartoons that were offensive to Muhammad.

(H) Twelve illustrations insulting the prophet caused unrest in Jordan.

La tâche RTE a non seulement besoin d'information sur les mots équivalents (*Muhammad* $\rightarrow$ *the prophet*), les implications asymmétriques (*editorial cartoons* est un hyperonyme de *illustrations*), mais aussi l'exclusion sémantique (*in Denmark* | *in Jordan*) pour pouvoir conclure que le texte T n'implique pas l'hypothèse H. Ces relations d'implication lexicale sont capturées par la « logique naturelle* » (*natural logic*) (Van Benthem, 1991), un formalisme qui considère le langage naturel comme une représentation

de sens, au lieu d'avoir recours à des représentations externes telle que la logique du premier ordre. Ce formalisme s'avère bien adapté pour des paraphrases automatiquement extraites, et Pavlick *et al.* (2015a) visent à annoter les paires de segments dans PPDB automatiquement avec ces relations sémantiques. Ils se focalisent sur des paraphrases lexicales et sous-phrastiques (au total 77M de règles). La classification des paraphrases syntaxiques est naïve. Elle traite les symboles non terminaux comme des mots simples et les auteurs ont reporté à de futurs travaux une représentation plus fine. Ensuite, Pavlick *et al.* (2015a) se concentrent sur des paires de paraphrases dans PPDB qui apparaissent aussi dans les données de RTE : le jeu de données SICK pour la tâche partagée de RTE de SemEval en 2014 (Marelli *et al.*, 2014). Ce jeu de données contient 10k paires de phrases, annotées en trois catégories d'implication textuelle : Implication (29%), Contradiction (15%) et Neutre (56%).

Pour toute paire de segments $\langle p_1, p_2 \rangle$ dans PPDB, où chaque segment contient au plus trois mots en longueur, Pavlick *et al.* choisissent les paires à condition qu'il existe des paires de phrases $\langle T, H \rangle$ où p_1 apparaît dans T et p_2 apparaît dans H. Cela donne une liste de 9 600 paires de segments pour le développement et le test. En se basant sur la théorie de la « logique naturelle », les relations logiques proposées par MacCartney (2009) sur l'inférence des langues naturelles ont été utilisées par les auteurs avec deux petites modifications : 1) les relations *Négation* et *Disjonction* sont regroupées et représentées par une notion d'exclusion ; 2) la relation *Indépendance* est divisée en deux cas : des paires de segments vraiment indépendants et ceux liés par une relation autre que l'implication textuelle.

Le *crowdsourcing* a été utilisé pour l'annotation de ces paires. Les différentes étiquettes avec leur définition sont présentées dans la figure 3.3. Chaque paire est montrée à cinq annotateurs, et l'étiquette choisie majoritairement est retenue.

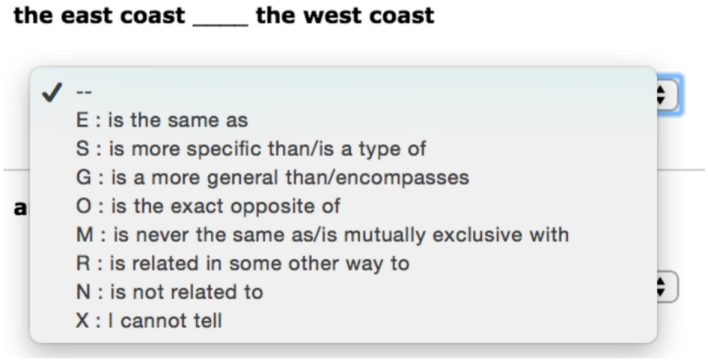

FIGURE 3.3 – Figure extraite de l'article de Pavlick *et al.* (2015a), pour chaque paire de segments hors contexte, les annotateurs doivent choisir une étiquette dans la liste donnée, pour décrire la relation du premier segment par rapport au second segment

Le but est de catégoriser automatiquement les paires parmi ces cinq classes, pour chacune nous indiquons l'étiquette correspondante (ou les étiquettes correspondantes) dans la figure 3.3 :

— Équivalence (E) : *distant* → *remote*

— Implication (les deux sens d'implication sont combinés en un seul : p_1 est plus général que p_2, soit S + G) : *building* → *tower*

— Exclusion (O + M) : *close* → *open*

— Autrement lié (R) : *country* → *patriotic*

— Indépendant (N) : *found* → *party*

Un classifieur de régression logistique est utilisé pour distinguer ces cinq classes, en exploitant de nombreux traits [4] qui s'appuient sur :

— des informations au niveau lexical

— des informations issues de WordNet

— des chemins de dépendance : les auteurs apprennent de nouveaux patrons qui différencient les relations plus subtiles que celles de Snow *et al.* (2004), qui ont seulement utilisé des patrons lexico-syntaxiques pour apprendre des hyperonymes et des hyponymes

— des valeurs de similarité distributionnelle, en se basant sur le corpus `Gigaword annoté` (Napoles *et al.*, 2012)

— des mesures de similarité symétrique (Lin, 1998) et asymétriques (Weeds *et al.*, 2004; Szpektor et Dagan, 2008; Clarke, 2009) sur des vecteurs de contexte de dépendance

— des probabilités de paraphrase

— le nombre total de traductions partagées pour chaque paire de phrases

Une validation croisée sur les données d'apprentissage montre qu'aucun des traits monolingues n'était efficace en faisant la distinction subtile entre des paires équivalentes et les autres types de relations sémantiques. En revanche, la mesure de similarité bilingue est assez précise pour l'identification des paires équivalentes, mais elle fournit moins d'information pour distinguer différents types de relations non équivalentes.

Ensuite, Pavlick *et al.* (2015a) ont effectué une évaluation intrinsèque et extrinsèque. L'évaluation intrinsèque teste la capacité de reproduire les étiquettes de référence humaine sur les données de test. La combinaison qui utilise à la fois des traits monolingues et des traits bilingues permet d'atteindre une exactitude globale de 79%. La majorité des erreurs repose sur la confusion entre la catégorie *Autrement lié* et *Indépendant*, ce qui ne présente pas un impact trop important pour la tâche RTE, parce que *Autrement lié* peut être considéré comme un cas spécial de *Indépendant*.

L'évaluation extrinsèque démontre l'utilité des relations sémantiques prédites automatiquement dans un système de RTE *Nutcracker* (Bjerva *et al.*, 2014), qui est basé sur des preuves. Étant donné une paire de phrases $\langle T, H \rangle$, *Nutcracker* produit une représentation sémantique formelle pour chaque phrase, qui est ensuite traduite en une logique standard du premier ordre. Le système utilise un outil de démonstration automatique de théorèmes et un constructeur de modèle pour chercher une implication ou contradiction logique. Si aucune des deux ne peut être trouvée, la classe majoritaire est prédite par défaut (c'est la classe *Neutre* en l'occurrence). Par conséquent, *Nutcracker* dépend beaucoup des ressources d'implication lexicale, en vue d'améliorer le rappel du système.

Différentes configurations ont été testées :

— baseline 1 : toujours prédire la classe majoritaire *Neutre*

— baseline 2 : exécuter *Nutcracker* tout seul, sans aucun axiome externe

4. Le matériel supplémentaire de leur article présente des détails sur l'annotation manuelle et les traits utilisés.

— baseline 3 : pour chaque paire de segments dans PPDB-XL, générer un axiome de synonyme

— utiliser une base de connaissances basée sur WordNet, qui contient des axiomes pour tous les synonymes, antonymes et hypernonymes dans WordNet

— PPDB+ : convertir les prédictions du classifieur en un ensemble d'axiomes : synonyme, hyperonyme et antonyme

— PPDB-Human : une base de connaissances basée sur les étiquettes de référence humaine

Les résultats montrent que PPDB+ apporte une amélioration de 4% en exactitude par rapport à la baseline qui utilise *Nutcracker* seul. Pour toutes les configurations, PPDB+ constitue une meilleure source d'axiome que WordNet. L'ajout de PPDB+ à WordNet donne une augmentation relative de 17% dans le nombre de preuves trouvées, par rapport à l'utilisation du WordNet seul. Ces preuves additionnelles permettent à *Nutcracker* d'effectuer plus de prédictions correctes pour des bonnes raisons. De plus, en utilisant PPDB+, *Nutcracker* atteint une performance très proche en comparaison de l'utilisation de PPDB-Human.

Une évaluation des relations d'implication prédites dans toute la ressource PPDB a aussi été effectuée, sur un échantillon de 1 000 instances pour chaque classe de relation sémantique. Les annotations de référence ont été obtenues par *crowdsourcing*. Les résultats de classification sont très bons pour la catégorie *Équivalence* et *Exclusion*, et plus bas pour *Implication*. La plupart des erreurs consistent à classifier faussement les instances de *Équivalence* comme *Implication*.

La figure 3.4 montre une estimation de distributions des relations d'implication dans chaque taille de PPDB. Après le classement amélioré par le travail de Pavlick *et al.* (2015b), la plus petite taille S contient la proportion la plus élevée de paraphrases strictes. Or dans la taille la plus grande XXXL, il existe tout au plus 10% de paraphrases. Nous pensons qu'une meilleure représentation sémantique est nécessaire pour améliorer cette technique, que ce soit pour obtenir des paraphrases ou pour obtenir de manière contrôlée d'autres types de variantes.

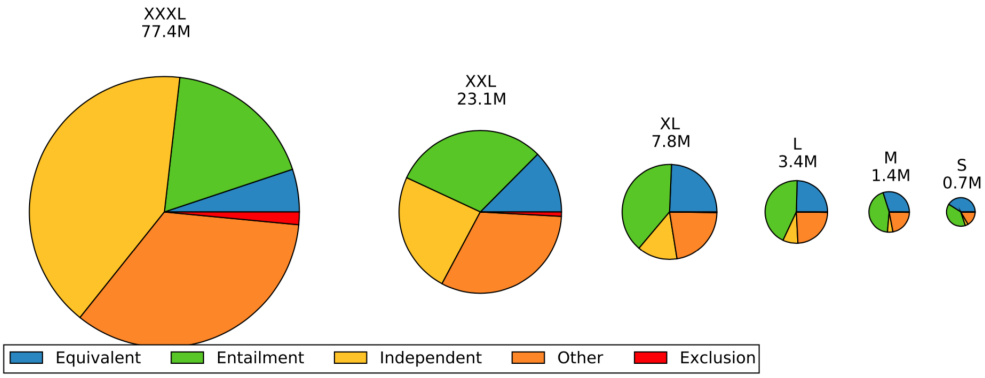

FIGURE 3.4 – Figure extraite de l'article de Pavlick *et al.* (2015a) : distributions estimées des relations d'implication dans chaque taille de PPDB 2.0 (version anglaise). L'estimation est basée sur des annotations manuelles d'un échantillon aléatoire de paires de segments. La taille la plus grande XXXL contient 77,4M de paraphrases lexicales et sous-phrastiques, où la classe majoritaire est *Indépendant* ; la plus petite taille S contient 700k paraphrases, où *Équivalence* est la classe majoritaire

Jusqu'ici, nous avons passé en revue les travaux principaux sur l'extraction de paraphrases qui exploitent les corpus monolingues et bilingues. Nous avons présenté plus en détail des recherches autour de la méthode par pivot, sur lesquelles nous reviendrons par la suite en présentant nos problématiques de recherche.

3.3 Génération de paraphrase

Une présentation sur la génération de paraphrases est importante pour compléter notre revue des différents travaux en TAL autour de la paraphrase. [5] Cette tâche consiste à générer une paraphrase pour une phrase ou un segment source dans une application donnée, et elle est importante pour plusieurs applications en TAL.

Des méthodes de génération de paraphrases traditionnelles utilisent des règles créées manuellement (McKeown, 1983), des patrons de paraphrase complexes appris automatiquement (Zhao *et al.*, 2008b), des alignements basés sur des thésaurus (Hassan *et al.*, 2007), des techniques en SMT (Quirk *et al.*, 2004), etc.

Barzilay et Lee (2003) ont proposé une approche pour la génération de paraphrases au niveau phrastique, qui est plus difficile que la génération des paraphrases lexicales ou sous-phrastiques. Des alignements de multiple séquences sont appliqués aux phrases collectées à partir de corpus comparables non annotés. Dans leur travail, des collections d'articles produits par deux agences de journaux différentes sur les mêmes évènements ont été utilisés. Cette méthode apprend un ensemble de patrons paraphrastiques qui sont similaires en structure et qui sont représentés par des treillis de mots. La méthode détermine automatiquement comment appliquer ces patrons pour réécrire de nouvelles phrases. Des représentations en treillis de mots ont déjà été utilisées par Pang *et al.* (2003) sur des corpus monolingues parallèles. Barzilay et Lee (2003) montrent qu'il est aussi possible de les induire à partir de corpus comparables qui sont plus faciles à obtenir pour certains domaines.

En vue de produire des reformulations sous-phrastiques pour l'aide à la révision, Max (2008) a proposé une approche basée sur la méthode par pivot, qui utilise des tables de traduction de segments et combine les scores de différents modèles. Les évaluations basées sur trois critères dont un concernant directement l'intérêt des reformulations proposées pour le rédacteur, permettent de dégager la contribution du modèle basé sur la conservation des relations de dépendance syntaxique entre un énoncé et ses reformulations.

Zhao *et al.* (2008a) ont présenté une nouvelle méthode sur la génération de paraphrases au niveau sous-phrastique et phrastique. Cette étude est basée sur des systèmes SMT et sur l'exploitation des informations de différentes ressources, telles qu'un thésaurus construit automatiquement, un corpus monolingue parallèle, un corpus monolingue comparable, une table de segments bilingues, les définitions des mots dans un dictionnaire et un corpus de requêtes similaires des utilisateurs. Des résultats expérimentaux indiquent que même si les contributions des ressources exploitées diffèrent beaucoup, elles sont toutes utiles pour la génération de paraphrases au niveau phrastique.

Après avoir analysé des différences entre la génération de paraphrases et d'autres recherches (surtout la traduction automatique), Zhao *et al.* (2009) sont les premiers à avoir proposé un modèle statistique spécifiquement dédié à la génération de paraphrases phras-

5. À part l'extraction et la génération, le troisième aspect concerne l'identification de paraphrases : étant donné une paire de phrases, effectuer une classification binaire sur leur relation de paraphrase. Nous renvoyons les lecteurs à l'article de Androutsopoulos et Malakasiotis (2010) pour un état de l'art détaillé.

tiques. La caractéristique de ce travail est qu'un même modèle est utilisé pour la génération de paraphrases dans plusieurs applications, telles que la compression de phrase, la simplification de phrase et le calcul de similarité entre phrases. De multiples ressources ont été utilisées (segments paraphrastiques, patrons, locutions, etc.) pour résoudre le problème du manque de données et générer des paraphrases plus variées.

Pour un état de l'art exhaustif (jusqu'en 2010) sur l'extraction et la génération de paraphrases pilotée par les données, nous renvoyons les lecteurs aux articles de Androutsopoulos et Malakasiotis (2010) et de (Madnani et Dorr, 2010).

Prakash *et al.* (2016) sont les premiers à avoir proposé une architecture de réseaux neuronaux pour la génération de paraphrase. Les auteurs ont utilisé un réseau *stacked residual LSTM (Long-Short Term Memory)*. Cette extension de l'apprentissage par une architecture séquence à séquence (Sutskever *et al.*, 2014) permet un entraînement efficace des réseaux profonds de type LSTM. Les connections résiduelles entre les couches de LSTM aident à conserver les mots importants dans la paraphrase générée. L'évaluation de leur travail sur trois jeux de données différents avec des mesures BLEU, METEOR et TER montrent que leur modèle est le plus performant en comparaison avec d'autres architectures en réseaux neuronaux.

Les modèles séquence-à-séquence pour la génération de paraphrases ont tendance à mémoriser des mots et des patrons dans le jeu d'apprentissage, au lieu d'apprendre le sens des mots. La raison principale est que la couche de sortie du décodeur ne modélise pas l'information sémantique. Pour résoudre ce problème, Ma *et al.* (2018) ont introduit un nouveau modèle basé sur une architecture encodeur-décodeur, nommée *Embedding Attention Network (WEAN)*. Ce modèle génère de mots via des requêtes sur des plongements lexicaux neuronaux, avec le but de capturer le sens des mots correspondants. Le modèle est évalué sur deux tâches liées à la paraphrase, à savoir la simplification textuelle et le résumé abstractif (résumé par compréhension) des textes courts.

Li *et al.* (2018) ont proposé un nouveau cadre pour la génération de paraphrases basé sur l'apprentissage par renforcement. Ce modèle contient un générateur et un évaluateur entraînés sur les données. Le générateur est construit avec un modèle séquence-à-séquence, qui produit une paraphrase pour une phrase donnée ; l'évaluateur peut juger si les deux phrases sont des paraphrases l'une pour l'autre. Pendant l'entraînement, le générateur est mis au point (*fine-tuning*) par le renforcement, où la récompense est donnée par l'évaluateur. Des résultats expérimentaux sur deux jeux de données montrent que leur générateur peut produire des paraphrases plus précises que des méthodes d'état-de-l'art.

Dans le but de générer des exemples antagonistes pour tromper des modèles pré-entraînés en TAL, et d'augmenter la robustesse de ces systèmes vis-à-vis de la variation syntaxique, Iyyer *et al.* (2018) ont proposé leur modèle SCPNs (*Syntactically Controlled Paraphrase Networks*), et l'ont utilisé pour générer des exemples antagonistes. Étant donné une phrase en entrée et une forme syntaxique cible (par exemple une analyse en constituant), leur modèle encodeur-décodeur est entraîné pour produire une paraphrase avec la syntaxe désirée. Les travaux de Chen *et al.* (2019) ont étendu cette tâche en remplaçant la forme syntaxique cible par une phrase seule, sans avoir besoin d'un analyseur syntaxique supervisé.

Cet aspect sur la génération de paraphrases est moins central à cette thèse. Nous nous sommes limités ici à présenter un bref état de l'art et la tendance actuelle des travaux qui s'appuient fortement sur des architectures à base de réseaux neuronaux.

3.4 Utilisation de paraphrases dans d'autres tâches

Nous avons présenté un état de l'art sur l'extraction et la génération automatique de paraphrases. Ce domaine de recherche reste actif parce que la paraphrase est un élément crucial pour l'interprétation et la génération des langues naturelles. D'un point de vue pratique, la diversité d'expression pour communiquer les mêmes informations présente un défi majeur pour beaucoup d'applications en TAL. En conséquence, des ressources de paraphrases sont utilisées dans de nombreux domaines.

Dans une interface en langue naturelle pour accéder aux bases de données, la génération des paraphrases des questions permet à l'utilisateur de comprendre si sa requête a été bien comprise par le système (McKeown, 1979). L'extraction de paraphrases a été utilisée pour améliorer les résultats des systèmes de questions-réponses (Moldovan et Rus, 2001; Ravichandran et Hovy, 2002; Duclaye *et al.*, 2003; Dong *et al.*, 2017).

Concernant la traduction automatique, la paraphrase est utilisée pour : améliorer l'exactitude de l'évaluation, qui permet une mise en correspondance plus flexible de la sortie du système par rapport aux références humaines (Kauchak et Barzilay, 2006; Callison-Burch *et al.*, 2006b; Snover *et al.*, 2009), faciliter l'entraînement des paramètres dans SMT (Madnani *et al.*, 2007), traiter des segments sources jamais vus pendant l'apprentissage ou augmenter des données d'apprentissage (Callison-Burch *et al.*, 2006a; Marton *et al.*, 2009),

Dans le domaine de l'extraction d'information, les paraphrases de surface ont été utilisées pour générer des patrons de surface, qui sont utiles pour extraire des relations entre deux entités (ex. la relation "lieu de naissance" entre une personne et son lieu de naissance) (Bhagat et Ravichandran, 2008).

La ressource de paraphrases PPDB a été utilisée pour aider l'alignement de mots monolingue (Sultan *et al.*, 2014), mesurer la similarité sémantique textuelle (Agirre *et al.*, 2016), améliorer des plongements lexicaux (Yu et Dredze, 2014; Rastogi *et al.*, 2015; Faruqui *et al.*, 2015).

La détection de paraphrases est utile pour réduire les redondances pour la tâche de résumé multi-documents (Barzilay *et al.*, 1999; Barzilay, 2003; Barzilay et McKeown, 2005; Zhou *et al.*, 2006), ou pour la détection automatique de plagiat (Barrón-Cedeño *et al.*, 2013).

La paraphrase a également été utilisée pour l'aide à la rédaction (Barreiro, 2009), la normalisation du style de la rédaction (Xu *et al.*, 2012), la simplification textuelle où le remplacement des termes spécialisés (par exemple médiaux) par des expressions non-spécialisées facilite la compréhension des non-experts (Elhadad et Sutaria, 2007; Deléger et Zweigenbaum, 2009).

Les travaux précédemment cités (non exhaustifs) nous ont permis de montrer l'importance de la paraphrase en TAL, tant pour ses ressources que pour ses techniques (extraction, génération, identification).

3.5 Problématique de recherche

En tant qu'une tâche importante en compréhension et génération de langue naturelle, la traduction automatique (MT* : *Machine Translation*) a d'abord été améliorée de façon importante par les techniques de la traduction statistique basée sur les segments (PBSMT) (Koehn *et al.*, 2003), et ensuite depuis plusieurs années par les techniques de la traduction neuronale (NMT) (Wu *et al.*, 2016). La présentation de l'état-de-l'art dans la

section 3.2.2 montre que les techniques de MT ont aussi été exploitées pour extraire des paraphrases dans des corpus parallèles bilingues.

En revanche, les travaux dans ces domaines ont très peu pris en compte les procédés de traduction entre paires de segments, alors que cela correspond à un sujet important pour les traducteurs humains. Les cas de traduction littérale sont bien exploités par l'utilisation de grands corpus et par les systèmes neuronaux ; en revanche, il existe un très grand nombre de traductions non littérales, en particulier dans des genres textuels non techniques. Ces traductions posent souvent des difficultés pour l'alignement automatique de mots (essentiel pour les méthodes statistiques), et elles peuvent faire dévier le sens originel du texte source.

Pour différentes paires de langues, il est probable que la distribution de différents procédés de traduction varie. Pour l'extraction de réécritures via la méthode par pivot, des langues pivots différentes pourraient produire des résultats très différents, ce qui n'a pas été considéré jusque-là. Des traductions pivots non littérales peuvent influencer l'équivalence stricte entre le segment source et la paraphrase candidate extraite, néanmoins cet aspect n'a pas reçu assez d'attention pendant cette exploration de corpus parallèle bilingues.

Ne pas tenir compte de ces phénomènes conduit probablement à une perte de contrôle sur la relation sémantique entre des paraphrases extraites, ce qui est attesté par les diverses relations sémantiques dans la ressource PPDB (Pavlick *et al.*, 2015b). Le travail de Pavlick *et al.* (2015a) révèle que dans une telle ressource, il existe des paires de phrases qui sont liées sémantiquement autrement que par une équivalence stricte (voir figure 3.4).

Compte tenu des problèmes observés et du rôle important de la paraphrase en TAL, nous pouvons poser ces questions : en utilisant la méthode par pivot, quelle est l'influence des traductions pivot non littérales sur la relation sémantique entre le segment source et la paraphrase candidate extraite ? Est-il réalisable d'étudier cette question selon différentes catégories de traduction non littérale ? À notre connaissance, en raison des difficultés linguistiques, la reconnaissance automatique des procédés de traduction reste un domaine de recherche quasiment inexploré. Ainsi, un cadre de validation de notre recherche consiste à tenter d'améliorer la qualité des ressources de paraphrases, en nous fondant sur les procédés de traduction.

Notre problématique de recherche peut se résumer en deux questions : est-il possible de reconnaître automatiquement les procédés de traduction ? Certaines tâches en TAL peuvent-elles bénéficier de la reconnaissance des procédés de traduction ?

Nous citons ici une phrase de Ballard (2006), qui, selon nous, renforce notre intérêt pour ces recherches :

> *La traductologie*[6] *ne peut sécréter une machine qui dispenserait de penser et d'agir, pas plus que la linguistique ou la textologie ; ces sciences permettent une meilleure conceptualisation des problèmes, de meilleures analyses, des prises de décision plus conscientes.*

3.6 Conclusion

Dans ce chapitre, nous avons passé en revue diverses définitions et typologies de la paraphrase en linguistique et en TAL. La complexité des phénomènes de paraphrase et

6. Concernant un état de l'art sur la traductologie, nous renvoyons les lecteurs au premier chapitre de la thèse de Lemaire (2017).

son importance pour diverses applications en TAL ont impliqué de nombreux efforts sur l'extraction et la génération automatique de paraphrases.

Ayant présenté le contexte de travail complet sur les procédés de traduction et sur la paraphrase, nous avons présenté notre problématique de recherche et un cadre de validation possible sur l'aide à la construction de ressource de paraphrases. De plus, nous avons deux autres cadres de validation. L'un consiste à étudier s'il est pertinent d'évaluer l'alignement automatique de mots et la qualité de la traduction automatique en utilisant un corpus parallèle annoté en procédés de traduction.

L'autre cadre concerne l'aide à l'apprentissage des langues étrangères. Les traductions non littérales présentent un défi pour les outils actuels qui aident l'apprentissage des langues étrangères. Par exemple, le concordancier bilingue en ligne *Linguee*[7] propose des dictionnaires bilingues et des paires de phrases parallèles. Pourtant, jusqu'à présent, l'information sur les procédés de traduction est absente et l'alignement de mots souligné est souvent absent ou erroné pour des traductions non littérales (voir figure 3.5). Si des traductions difficiles traduites de façon non littérale peuvent être mieux traitées et présentées, cette ressource multilingue peut fournir plus d'aide aux utilisateurs.

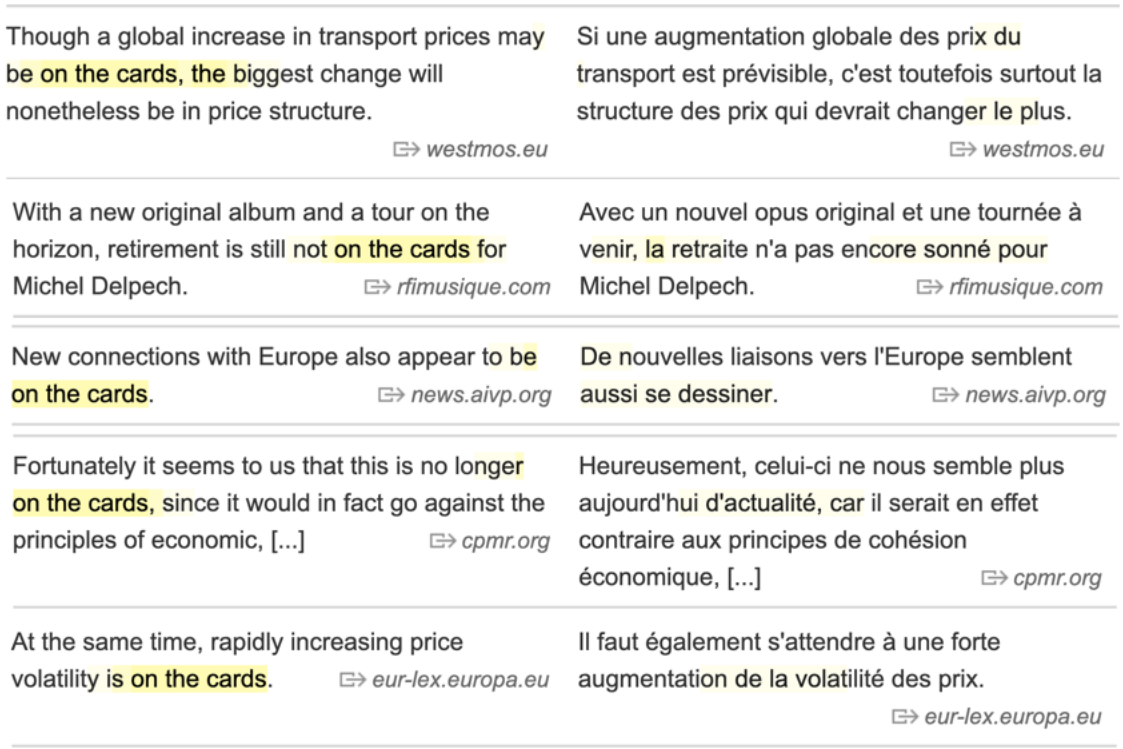

FIGURE 3.5 – Exemples de traductions non littérales du segment anglais « *on the cards* » trouvés dans *Linguee*. Les alignements de mot sont presque absents

En conséquence, ce cadre de validation concerne comment intégrer la reconnaissance des procédés de traduction dans un outil d'aide aux apprenants de langues étrangères.

Dans le cadre de notre problématique (Est-il possible de reconnaître automatiquement les procédés de traduction ? Certaines tâches en TAL peuvent-elles bénéficier de la reconnaissance des procédés de traduction ?), nous devons d'abord vérifier notre hypothèse de travail : il est possible de reconnaître automatiquement les différents procédés de traduction. Dans la deuxième partie de cette thèse, nous présentons nos solutions à

7. https://www.linguee.com/

diverses questions scientifiques rencontrées pendant ce processus : choix de corpus, annotation manuelle, développement du guide d'annotation, ingénierie des traits pour les classifieurs, etc. Notre étude constitue un cycle de travail complet en apprentissage automatique : annoter un corpus pour constituer le jeu de données, entraîner, tester et évaluer le modèle, et enfin réviser le modèle et les données.

Deuxième partie

Apports des procédés de traduction

Chapitre 4

Choix du corpus et méthodologie d'annotation

Sommaire

Dans les deux chapitres précédents, nous avons présenté un état de l'art sur les typologies existantes des procédés de traduction et sur les études de la paraphrase en TAL. Nous avons présenté nos problématiques de recherche et les trois cadres de validation de notre étude.

Dans ce chapitre, nous expliquons notre choix d'étudier les procédés de traduction dans un corpus parallèle bilingue de discours préparé, les *TED Talks*. En nous basant sur des typologies existantes de procédés de traduction, et en analysant le corpus, nous avons d'abord catégorisé les différents procédés selon notre propre typologie, qui est adaptée aux phénomènes de traductions présents dans le corpus. Ensuite, nous avons annoté le corpus tout en établissant un guide d'annotation.

Nous présentons notre typologie de procédés de traduction, les définitions ainsi que les exemples pour chaque procédé. Suite à l'annotation, notre objectif porte sur la reconnaissance automatique de ces procédés, afin d'intégrer cette information dans l'extraction de paraphrases ou de paires en relation d'implication textuelle à partir des corpus parallèles bilingues. Nous faisons l'hypothèse que cela permettra davantage de contrôle sémantique.

Une partie des travaux réalisés dans ce chapitre et le chapitre suivant ont été publiés dans ces deux articles : Zhai *et al.* (2018) et Zhai (2018).

Ils ont aussi été présentés dans une communication orale (sans acte) :

Construction of a Multilingual Corpus Annotated with Translation Relations, Yuming Zhai, *Atelier du Consortium CORLI "Analyse cross-lingue et annotation de corpus multilingues parallèles et comparables : tendances actuelles et futures"*, Université Paris Diderot, France, 2018

4.1 Examen des corpus parallèles anglais-français

Notre but est d'étudier les procédés de traduction sur plusieurs couples de langues. Pour cela, nous avons commencé par le couple de langues anglais-français. Ces deux langues sont bien dotées en termes de ressources textuelles, et il existe plusieurs corpus parallèles qui ont favorisé différentes recherches. Nous pouvons citer plusieurs plateformes qui fournissent des corpus parallèles déjà alignés au niveau de la phrase, par exemple le site de WMT (*Conference on Machine Translation*) [1], le site OPUS [2] (Tiedemann, 2012), CLARIN [3], LDC [4], etc.

Nous listons ci-dessous quelques corpus parallèles anglais-français largement utilisés dans le domaine de la traduction automatique :

— `Europarl` (Koehn, 2005) : actes de conférence du parlement européen

— `MultiUN` (Eisele et Chen, 2010), `UNPC` (*United Nations Parallel Corpus*) (Ziemski *et al.*, 2016) : comptes rendus officiels et documents parlementaires des Nations Unies

— `Paracrawl` [5] : corpus parallèle crawlé depuis le Web

— `OpenSubtitles` [6] (Lison et Tiedemann, 2016) : sous-titres traduits des films

— `News Commentary` [7] : commentaires aux journaux politiques et économiques, crawlés depuis le Web

— `Tatoeba` [8] : une collection de phrases et de traductions, construites de façon collaborative et ouverte

— `TED Talks` [9] : transcriptions et traductions humaines des sous-titres des vidéos de conférences *TED* (discours préparé)

Puisque les procédés de traduction concernent les transformations de la langue source à la langue cible, la direction de traduction est importante. Nous devons choisir les corpus pour lesquels nous sommes sûrs de cette information. Nous avons voulu choisir un corpus qui inclut plusieurs domaines pour ne pas être limité à un domaine spécifique (par exemple les corpus `Europarl` et `MultiUN` contiennent seulement des contenus du domaine politique). En même temps, pour ne pas travailler dans un cas extrême, nous avons préféré que la diversité des phénomènes de traduction soit entre ceux présents dans les corpus littéraires et ceux présents dans les corpus techniques. Selon ces critères, nous avons choisi de travailler sur le corpus de `TED Talks`.

Les conférences *TED* (*Technology, Entertainment and Design*) sont une série de conférences organisées au niveau international depuis 1984. Le but est de diffuser des idées sous forme de conférences courtes mais impressionnantes. Aujourd'hui un grand éventail de sujets y est abordé dans plus de 110 langues. La traduction des sous-titres des vidéos de ces conférences est contrôlée par des bénévoles et des coordinateurs par langue [10],

1. `http://www.statmt.org/wmt19/index.html`
2. `http://opus.nlpl.eu/`
3. `https://www.clarin.eu/portal`
4. `https://catalog.ldc.upenn.edu/`
5. `https://paracrawl.eu/index.html`
6. `http://www.opensubtitles.org/`
7. `http://www.casmacat.eu/corpus/news-commentary.html`
8. `https://tatoeba.org/fra/`
9. `http://www.ted.com/`
10. `https://www.ted.com/participate/translate/get-started`

assurant en général une traduction d'un bon niveau de qualité. En vue de faciliter les recherches en traduction automatique, l'inventaire Web *WIT*^{3} (*Web Inventory of Transcribed and Translated Talks*) [11] (Cettolo *et al.*, 2012) donne accès à des corpus de `TED Talks`, qui ont été surtout utilisés pour la campagne d'évaluation *IWSLT (International Workshop on Spoken Language Translation*) depuis 2011 (Federico *et al.*, 2011). Nous avons aussi recueilli des corpus parallèles sur cet inventaire en ligne. La segmentation en phrases est fournie par le site *WIT*^{3}, où plusieurs lignes de sous-titres peuvent être regroupés dans une seule phrase, mais des sous-titres ne sont jamais coupés.

Il existe déjà des travaux qui ont effectué des annotations sur les corpus de `TED Talks`, nous résumons ici les contributions de deux travaux représentatifs.

Face à un manque de ressources annotées manuellement en analyse syntaxique sur les corpus de discours préparé, Neubig *et al.* (2014) ont construit un corpus arboré (`NAIST-NTT TED Talk Treebank`) à partir du corpus de `TED Talks`. La première version contient 1,2k phrases et 23k mots anglais. Le corpus est multilingue. Des sous-titres ont été alignés au niveau de la phrase dans 18 langues pour chacune des 10 transcriptions de conférence annotées. Leur annotation suit le format standard du *Penn Treebank* (Marcus *et al.*, 1993). Après une première analyse automatique par `Berkeley Parser`, les annotateurs corrigent les erreurs de l'analyseur. Les incohérences sont filtrées automatiquement. Pour examiner l'interaction entre la syntaxe et le discours, toutes les phrases sont alignées automatiquement au niveau du temps avec le fichier de discours correspondant. En comparaison avec d'autres corpus, des analyses en complexité syntaxique et en différence stylistique ont été effectuées. Les expériences d'analyse syntaxique automatique ont démontré que l'addition des données de `TED Talks` aux données d'entraînement du `Wall Street Journal` permet une légère augmentation de la F-mesure (de 88,65 à 88,99) sur les données de test de `TED Talks`.

La traduction des expressions multi-mots (MWE) [12] reste un défi pour les méthodes automatiques. Par exemple, selon Isabelle *et al.* (2017), cette phrase source anglaise ne peut pas être traduite correctement en français, par des systèmes de traduction statistique ou neuronaux :

With this argument, ***the nail has been hit on the head****.*

Avec cet argument, ***la cause est entendue****.*

En revanche, il existe peu de corpus parallèles annotés en MWEs pour entraîner et évaluer la qualité des systèmes de traduction automatique. Pour répondre à ce besoin, Monti *et al.* (2015) ont proposé une méthodologie d'annotation de corpus parallèles en tous types de MWEs. Leur corpus parallèle anglais-italien `MWE-TED` est composé des contenus de conférences *TED* (1,5k phrases, 31k tokens anglais), complété par des traductions automatiques générées par l'outil `MOSES` (Koehn *et al.*, 2007). Le guide d'annotation est basé sur le modèle de PARSEME MWE et sur les test des propriétés des MWEs (Savary *et al.*, 2015; Losnegaard *et al.*, 2016). L'annotation est divisée en trois phases : annotation individuelle, comparaison inter-annotateur et validation finale. Les annotateurs doivent identifier les MWEs dans le texte source et la traduction humaine. Ils évaluent aussi si les traductions sont correctes. Quand la traduction humaine ou automatique est erronée, l'annotateur doit fournir la traduction correcte.

À notre connaissance, il n'existe pas encore de corpus annoté en procédés de traductions. Après l'examen des corpus anglais-français parallèles existants et selon nos critères

11. `https://wit3.fbk.eu/`

12. Les MWEs incluent : idiome, mot composé, terme spécifique d'un domaine, collocation, entité nommée, acronyme, etc.

de choix du corpus, nous avons décidé d'annoter un corpus de *TED Talks* dans notre étude.

4.2 Typologie proposée de procédés de traduction

Inspirés des travaux que nous avons passés en revue dans le chapitre 2 (Vinay et Darbelnet, 1958; Chuquet et Paillard, 1989; Molina et Hurtado Albir, 2002), nous proposons une typologie de procédés pour la paire anglais-français, établie pendant une première phase d'annotation manuelle et d'analyse des phénomènes rencontrés dans le corpus.

La figure 4.1 montre notre typologie, où les nœuds colorés représentent nos catégories d'annotation, et les autres nœuds servent à établir la hiérarchie (*i.e. Non Littéral, Non Aligné, Aucun Type*).

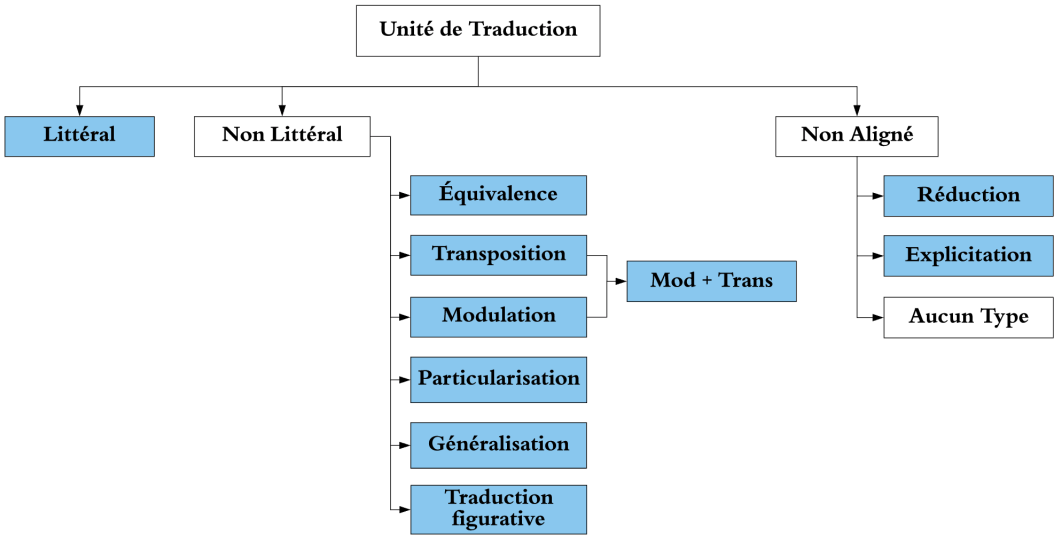

FIGURE 4.1 – Typologie de procédés de traduction pour le couple anglais-français

Par rapport aux typologies proposées dans les travaux précédents, notre typologie présente les différences suivantes :

— la faisabilité de la tâche d'annotation étant prise en compte, notre typologie contient moins de catégories.

— certaines typologies précédentes contiennent des procédés qui concernent les transformations dans les deux sens de traduction. Dans notre corpus, la direction de traduction est anglais → français, ainsi chaque catégorie décrit la transformation que la traduction française a reçue.

— le *calque* (emprunt d'un syntagme étranger avec traduction littérale de ses éléments) est considéré comme un procédé séparé dans les travaux de Vinay et Darbelnet (1958) (*ex. skyscraper → gratte-ciel*), mais ce phénomène est annoté comme une traduction littérale dans notre typologie.

— l'*emprunt* (*ex. t-shirt, parking*) est aussi annoté par la catégorie *Littéral*.

— l'*adaptation* culturelle (*ex. before you could say Jack Robinson → en un clin d'œil*) est annoté en tant que équivalence culturelle.

— dans notre typologie, la *transposition* regroupe des catégories plus fines proposées par Vinay et Darbelnet (1958), par exemple l'*étoffement* et l'*amplification*.

— nous avons ajouté la catégorie combinée *Mod+Trans* pour les traductions libres où les deux genres de transformations sont combinées.

— nous avons ajouté la catégorie *traduction figurative*.

— puisque nous annotons tous les mots dans le corpus, il existe trois cas de figure pour des segments non alignés : *Explicitation*, *Réduction*, et *Aucun type*.

Nous présentons ci-dessous la définition et des exemples typiques pour chaque catégorie.

4.3 Définitions et exemples typiques

Après notre analyse du corpus et les lectures précédentes, nous constatons que l'utilisation des procédés de traduction non littéraux se divise en deux cas :

— Usage obligatoire : parce qu'il n'existe pas de traduction littérale valable. Les traducteurs doivent respecter les conventions d'usage de la langue cible, par exemple :
Birds of a feather flock together. → *Qui se ressemble s'assemble.*

— Usage facultatif et choisi librement par le traducteur, par exemple :
his face wasn't twitching at all → *le visage parfaitement calme*
(une traduction littérale possible : *son visage ne tressaillait pas du tout*)

Nos catégories d'annotation sont divisées en celles pour des segments alignés, c'est-à-dire que des alignements manuels de mots bilingues sont faisables ; et celles pour les segments non alignés, pour lesquels il n'existe aucun alignement de mots pour un segment source ou cible.

Les définitions ci-dessous sont génériques ; nous les avons complétées par des règles dans notre guide d'annotation.

Dans chaque exemple ci-dessous, nous ne nous intéressons qu'à la partie en gras.

4.3.1 Catégories pour les segments alignés

1. Littéral :

   — traduction mot à mot (incluant l'insertion ou la suppression des articles), cela concerne aussi des unités polylexicales :
   certain *kinds of* → ***certains*** *types de*
   hatpin → ***épingle à chapeau***

   — traduction littérale de certains idiomes :
   facts *are stubborn* → ***les faits*** *sont têtus*

   — quand une traduction mot-à-mot n'a pas de sens, la traduction usuelle est considérée comme une traduction littérale :
   in other words *(?en d'autres mots)* → ***en d'autres termes***

   — traduction utilisant un calque :
   the myths of the Inuit elders still ***resonate with*** *meaning*
   → *les mythes des anciens Inuit* ***résonnent*** *encore* ***de*** *sens*

2. Équivalence :

— une traduction mot-à-mot a du sens mais le traducteur a produit une traduction différente, sans changement des catégories grammaticales, ni de point de vue :
sense each other* → *se reconnaître entre eux
(traduction littérale : *se sentir l'un l'autre*)

— traduction non-littérale de certains proverbes, idiomes ou expressions figées :
like a bull in a china shop* → *comme un chien dans un jeu de quilles
on the brink of* → *à deux doigts de
a bird in the hand is worth two in the bush* → *un tiens vaut mieux que deux tu l'auras

— traduction utilisant la mesure utilisée dans la culture cible :
about ***a mile and a half*** *deep → vers* ***2500m*** *de profondeur*

— traduction d'une abréviation par la forme complète :
UN* → *Organisation des Nations Unies

3. Transposition : traduire des mots ou des expressions à l'aide d'autres catégories grammaticales que celles utilisées dans la langue source, sans pour autant modifier le sens de l'énoncé :
astonishingly inquisitive* → *dotée d'une curiosité stupéfiante
patients ***over*** *the age of 40 → les patients* ***ayant dépassé*** *l'âge de 40 ans*
swim across *the river →* ***traverser*** *la rivière* ***à la nage***

4. Modulation : ce procédé consiste à changer le point de vue, ce qui révèle une manière différente de voir les choses pour les locuteurs de la langue cible. Le traducteur peut utiliser ce procédé pour contourner des difficultés de traduction et adopter des traductions naturelles. Nous pouvons constater ce procédé au niveau lexical et syntaxique. Le sens peut être légèrement changé.

— changement de point de vue :
and ***that scar has stayed with him*** *for his entire life*
→ et que, toute sa vie, ***il a souffert de ce traumatisme***
(traduction plus éloignée, changement sémantique et syntaxique)

— changement de voix (actif, passif) :
the statistics ***you hear about*** *→ les statistiques* ***qui nous sont communiquées***

— le sujet devient l'objet :
I had a *really astonishing* ***assignment***
→ ***on m'avait confié une mission*** *étonnante*

— changement entre forme affirmative et négative :
It's difficult*. →* *Ce* ***n'est pas facile****.*

— modulation métonymique :
Buy Coca-Cola ***by the carton****. → Achetez Coca-Cola* ***en gros****.*
(substituer l'abstrait « *en gros* » au concret « *by the carton* »)

— adoption des expressions naturelles dans la langue cible :
unless you ***think of it in the terms that I do***
→ à moins que vous ne ***regardiez la chose comme moi***

— changement léger de sens au niveau lexical :
allowing people to more fully engage with their ***abilities***
→ permettent aux gens de développer pleinement leur ***potentiel***

5. Modulation + Transposition : la traduction présente des changements simultanés de catégories grammaticales, structures syntaxiques et/ou point de vue. Puisqu'il est difficile de distinguer quelle transformation est plus importante, nous avons ajouté cette catégorie combinée.
this is a people ***who cognitively do not distinguish***
→ c'est un peuple ***dont l'état des connaissances ne permet pas de faire la distinction***
this is a completely ***unsustainable*** *pattern*
→ il est absolument ***impossible de continuer sur*** *cette tendance*
6. Particularisation :
   — selon le contexte, le traducteur utilise un mot ou un segment cible dont le sens est plus spécifique ou concret que celui du mot ou segment source :
   they ***have*** *a screen and a wireless radio*
   → ils ***sont équipés d'****un écran et d'une radio sans fil*
   the director ***said*** *→ le directeur* ***déclara***
   language loss → l'extinction du langage
   — préciser le sens d'un mot en contexte :
   if ***you're queasy*** *→ si* ***vous ne supportez pas la vue du sang***
   — traduire un pronom par la/les chose(s) qu'il référence :
   this is where ***it*** *reaches the sea*
   → voilà l'endroit où ***la rivière*** *se jette dans la mer*
7. Généralisation :
   — plusieurs mots ou expressions sources peuvent être traduits en un mot ou une expression cible avec un sens plus général, et le traducteur utilise ce dernier :
   as we ***sit here*** *in Monterey → alors que nous* ***sommes*** *à Monterey*
   — utiliser un pronom pour traduire la/les chose(s) qu'il référence :
   if you're starting with digital information in the computer, ***that digital information*** *has to be really accurate*
   *→ si vous partez des données numérisées sur ordinateur, il faut qu'****elles*** *soient extrêmement précises*
   people in the back or ***people*** *on video*
   → les gens du fond ou ***ceux*** *qui regarderont la vidéo*

   Dans d'autres cas, ce procédé rend le sens plus accessible dans la langue cible :

   — traduction d'un idiome par une expression non figée :
   trial and error → procéder par tâtonnements
   sea change → changement radical
   — suppression d'une métaphore :
   ancient Tairona civilization which once ***carpeted*** *the Caribbean coastal plain*
   → anciennes civilisations tyranniques qui ***occupaient*** *jadis la plaine côtière des Caraïbes*

8. Traduction figurative :
   - introduction d'un idiome pour traduire une expression non figée :
     at any given moment → ***à un instant "t"***
   - introduction d'une métaphore pour traduire des segments non métaphoriques :
     if you ***faint*** *easily → si vous* ***tombez dans les pommes*** *facilement*
     one woman almost ***passed out*** *→ une dame* ***a*** *presque* ***tourné de l'oeil***
   - conservation d'une métaphore à l'aide d'une traduction non littérale :
     the Sun begins to ***bathe*** *the slopes of the landscape*
     → le soleil qui ***inonde*** *les flancs de ce paysage*

4.3.2 Catégories pour les segments non alignés

1. Explicitation : introduction dans la langue cible des clarifications pour des éléments implicites dans la langue source, mais qui émergent du contexte ou de la situation :
   [...] live amongst those who have not forgotten the old ways, who still feel their past in the wind
   → [...] vivre parmi ceux qui n'ont pas oublié les anciennes coutumes, qui ressentent encore leur passé ***souffler*** *dans le vent*
2. Réduction : délibérément ne pas traduire certains mots avec un sens concret, qui auraient pu être traduits :
   and you'll ***suddenly*** *discover what it would be like*
   → et vous découvrirez ce que ce serait
3. Aucun type attribué :
   - mots outils nécessaires dans une langue mais pas dans l'autre :
     the last example I have time to → le dernier exemple ***que*** *j'ai le temps de*
     minus 271 degrees, colder than → moins 271 degrés, ***ce qui est*** *plus froid*
   - segments non traduits mais qui n'influencent pas le sens :
     this is an entire book, ***so this is*** *an example of non-image data*
     → voici un livre entier, un exemple de données qui n'est pas une image
   - segments donnant des informations répétées en contexte :
     when we try to think of biological processes or any ***process***
     → quand on essaye de concevoir des processus, biologique ou autre
   - segments cibles qui ne correspondent à aucun segment source :
     the exciting phase → le moment ***le plus*** *excitant*

4.3.3 Catégories indépendantes des procédés de traduction

Afin de pouvoir annoter tous les phénomènes de traduction dans le corpus, nous avons ajouté ces trois catégories qui ne sont pas liées aux procédés de traduction, mais utiles pour l'annotation. Ces catégories ont été ajoutées dans notre guide d'annotation.

1. Changement lexical (*lexical shift*) : la traduction n'est pas littérale, mais ce sont des changements mineurs au niveau lexical, qui n'implique aucun procédé de traduction.

— changer le temps verbal ou modalité verbale (le verbe ne change pas) :
*give you un update on how that machine **worked***
*→ vous mettre au courant de quelle façon cette machine **fonctionne***
*which **might** seem like an odd thing → qui **peut** paraître bizarre*

— changer la préposition :
*when you do a web search **for** images*
*→ quand on fait une recherche web **sur** des images*

— changer l'article
*if they believe enough there is a measurable effect in **the** body*
*→ s'ils y croient assez fort, il y a un effet mesurable dans **leur** corps*

— changer le sujet
***you** can move them → **on** peut les déplacer*

— changer l'adverbe de position
*there's a hole **there** → il y a un trou **ici***

— échanger entre la forme au singulier et au pluriel
*sugar pills have **a measurable effect** in certain kinds of studies*
*→ des pilules de sucre ont **des effets mesurables** dans certains types d'études*

2. Erreur de traduction : cette catégorie concerne des erreurs de traduction évidentes, par exemple :
*we found two **instances** of natural death*
*→ nous avons découvert deux **circonstances** de mort naturelle*
*is not going to be **remembered** for its wars*
*→ ne sera pas **reconnu** pour ses guerres*
3. Incertain : les annotateurs utilisent cette catégorie quand ils ont un doute et veulent avoir plus de discussions.

4.4 Conclusion

Ayant pour but de reconnaître automatiquement les procédés de traduction, nous avons décidé de les étudier d'abord dans un corpus parallèle anglais-français. Le choix de travailler sur le corpus de `TED Talks` est justifié par ces trois propriétés : la direction de traduction (anglais → français), la diversité des domaines et la diversité des phénomènes de traduction. En nous basant sur des typologies existantes des procédés de traduction et sur notre analyse de corpus, nous avons proposé notre propre typologie adaptée à notre corpus, et nous l'utilisons pour l'annotation manuelle. Des définitions et exemples typiques pour chaque catégorie ont été présentés.

Une perspective de ce travail est l'annotation des corpus d'autres genres, par exemple les corpus littéraires bilingues.

Chapitre 5

Annotation en procédés de traduction

Sommaire

Dans notre travail, nous avons d'abord annoté un corpus parallèle anglais-français, ensuite pour vérifier la généricité de notre démarche, nous avons adapté ces études sur le couple de langues anglais-chinois.

Pour mener à bien cette tâche d'annotation manuelle, plusieurs problèmes se posent, tels que la segmentation en unité de traduction, l'alignement manuel de mots, le développement du guide d'annotation, le choix d'un processus d'annotation pour garantir la qualité et l'extension des études sur le couple anglais-chinois.

Dans ce chapitre, nous expliquons comment nous avons résolu ces problèmes. Nous présentons les informations détaillées sur les corpus annotés, les processus d'annotation et les statistiques calculées sur les corpus annotés.

5.1 Corpus parallèle anglais-français

Le corpus de *TED Talks* que nous avons utilisé a d'abord été mis à disposition pour les campagnes d'évaluation *IWSLT* 2013 et 2014. Nous avons utilisé le corpus d'entraînement de 2014 (160 656 lignes), de développement et de test de 2010 (880 et 1 556 lignes, respectivement). La langue d'origine, c'est-à-dire celle dans laquelle se sont originellement exprimés les orateurs, est l'anglais. Les traductions fournies sont en français, chinois, arabe, espagnol et russe. Ces corpus multilingues ne sont pas tous parallèles, en conséquence, nous avons constitué une intersection de corpus en nous appuyant sur les

phrases source anglaises qu'ils partagent. [1] Nous avons annoté principalement le corpus anglais-français, et une partie du corpus anglais-chinois (la raison sera expliquée dans la section 5.4). Les corpus parallèles bilingues avec des traductions dans d'autres langues (arabe, espagnol et russe) ont été constitués pour servir à d'autres chercheurs.

Pour l'annotation, nous avons combiné le corpus de développement et de test, qui contient un total de 2 436 lignes de phrases parallèles pour chaque paire de langues (nous utilisons la segmentation de phrases fournie par le site *WIT*[3]). Ce corpus contient les transcriptions de 19 conférences. Pour l'annotation, nous avons divisé le corpus en 16 sous-corpus. Chaque sous-corpus contient une ou deux interventions complètes, afin de mieux comprendre le contexte. La table 5.1 récapitule les informations principales du corpus anglais-français.

Langue source	anglais
Langue cible	français
Nombre de conférences	19
Nombre de lignes	2 436
Nombre de tokens	51 930 en anglais
	53 749 en français
Durée totale	environ 4,6 heures

Tableau 5.1 – Informations sur le corpus parallèle anglais-français de *TED Talks* annoté

La tokenisation a été réalisée avec l'outil `Stanford Tokenizer` [2], pour son efficacité et son usage répandu dans la communauté. Afin d'accélérer la phase d'annotation sur les mots traduits littéralement, nous avons effectué un alignement de mots automatique. Pour cela, nous avons choisi l'outil `FastAlign`, parce qu'il est plus efficace en termes de temps que `GIZA++` (Och et Ney, 2003) et les résultats sont comparables en qualité (Dyer *et al.*, 2013). Nous avons utilisé ses paramètres par défaut et l'avons entraîné sur l'intégralité du corpus parallèle (soit 163k paires de phrases et 3M de tokens anglais).

5.2 Annotation manuelle

5.2.1 Outil d'annotation

Nous utilisons l'application Web `Yawat` [3] (Germann, 2008). Cet outil nous permet d'aligner des mots et des segments (continus et discontinus) dans un corpus parallèle, puis d'attribuer des catégories configurables adaptées à notre tâche à des unités bilingues ou monolingues (c'est-à-dire non alignés) (voir figure 5.1).

À des fins d'illustration, considérons l'exemple dans la figure 5.2 :

Les segments en noir sont traduits littéralement avec ces alignements (les frontières sont visibles quand la souris passe sur des mots) :

well → *eh bien*

, → ,

we → *nous*

1. Les frontières de phrases ont été corrigées dans le corpus de test français pour franchir cette étape.
2. `http://nlp.stanford.edu/software/tokenizer.shtml`
3. *Yet Another Word Alignment Tool* ; cet outil est disponible pour la recherche sous la licence GNU Affero General Public License v3.0.

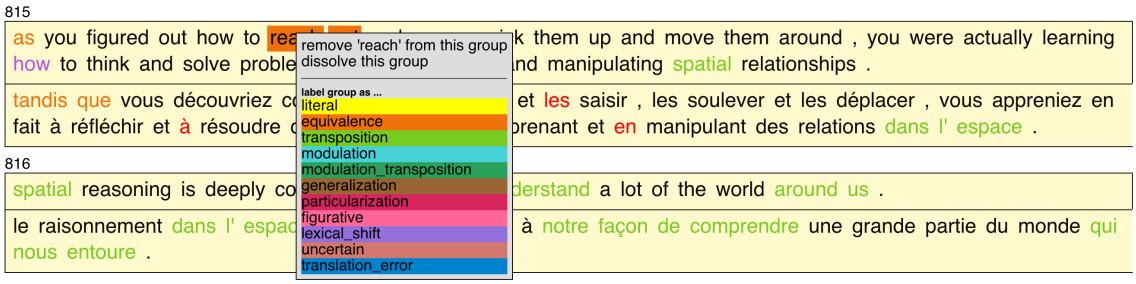

FIGURE 5.1 – Interface de l'outil Yawat pour effectuer l'annotation

use → *employons*
that → *cet*
euphemism → *euphémisme*
which → *qui*
be → *est*

Les mots en rose (« *great* », « *is exposed to* ») sont annotés avec la catégorie « *Réduction* », signifiant qu'ils auraient pu être traduits. L'idiome « *trial and error* » est traduit en une expression non figée (« *procéder par tâtonnements* »), donc la catégorie est *Généralisation*. Enfin l'adjectif « *meaningless* » est traduit en « *dénué de sens* » avec différentes catégories grammaticales, la catégorie est ainsi *Transposition*.

Concernant les alignements de segments (avec plus de deux mots de chaque côté), l'alignement via l'outil `Yawat` sera une correspondance de groupe d'indices, et pas une correspondance d'indice au niveau du mot individuel.

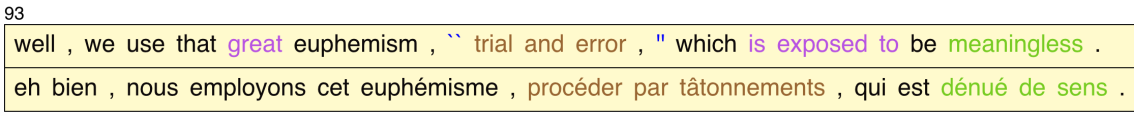

FIGURE 5.2 – Exemple d'annotation pour une paire de phrases dans Yawat

5.2.2 Segmentation en unité de traduction et alignement de mots

Pendant l'annotation de chaque paire de phrases parallèles, nous effectuons d'abord un alignement de mots manuel, ensuite nous attribuons les catégories correspondantes.

Les segmentations en unités de traduction au sein d'une phrase ne sont pas fournies préalablement. En respectant la tokenisation, les annotateurs doivent décider les frontières des segments bilingues à aligner selon le contexte. Nous avons fixé notre définition de l'« *unité de traduction* » en nous basant sur des travaux précédents.

Dans le chapitre 2, nous avons mentionné la définition de l'unité de traduction proposée par Vinay et Darbelnet (1958) : *« le plus petit segment de l'énoncé dont la cohésion des signes est telle qu'ils ne doivent pas être traduits séparément, ex : "prendre son élan", "battre à coups précipités". Les unités de traduction permettent d'effectuer le découpage d'un texte. »* Cette définition est seulement axée sur le texte source. Or dans notre tâche d'annotation réelle, nous trouvons que la frontière d'un segment source est forcément décidée par sa traduction.

Nádvorníková (2017) a signalé qu'avant d'énoncer des jugements critiques concernant la qualité du travail des traducteurs, les utilisateurs de corpus parallèles devraient consulter

le contexte le plus large possible des éléments analysés (au-delà d'une phrase, ou même au niveau du texte entier), pour voir si les non-équivalences observées ne sont pas explicables par des compensations ou par la volonté d'éviter des répétitions. Par exemple, le trait « *familier* » d'une expression peut être effacé dans une expression s'il est repris par une autre expression dans la phrase suivante ou précédente, dans ce cas-là, le caractère global du passage est conservé ; le traducteur peut aussi changer la structure d'une phrase ou modifier/remplacer une expression pour éviter sa répétition. En revanche, dans cette thèse, nous nous limitons à l'annotation des procédés de traduction au sein de chaque paire de phrases. Tout de même, les annotateurs travaillent à chaque fois sur une présentation entière (un sous-corpus), ainsi ils ont accès aux informations contextuelles pour chaque paire de phrases.

Concernant la construction des alignements de mots de référence, Melamed (1998) a été le premier à proposer un guide d'annotation complet pour le projet *Blinker (Bilingual Linker)*, pour annoter 250 paires de versets de la Bible (anglais-français), avec un schéma d'annotation binaire (aligné ou non traduit). Och et Ney (2000) ont utilisé ce guide pour aligner 484 paires de phrases du corpus `Hansard` (actes du parlement Canadien, anglais-français) et ont introduit la distinction entre les alignements sûrs et possibles pour les alignements un-à-un. Ensuite leur schéma a été réutilisé par plusieurs travaux sur l'alignement de mots (Mihalcea et Pedersen, 2003a; Holmqvist et Ahrenberg, 2011).

L'évaluation des alignements de mots incluent des mesures intrinsèques et extrinsèques. La mesure intrinsèque la plus communément utilisée est AER* (*Alignment Error Rate*), proposée par Och et Ney (2000). Cette mesure s'appuie sur un schéma d'annotation particulier pour des alignements de référence, qui distingue les alignements sûrs et possibles. Le score est une F1-mesure où le rappel et la précision sont calculés différemment pour ces deux types d'alignements. Des critiques ont été portées sur AER et sur ce schéma d'annotation, notamment dû à un manque de sémantique claire des alignements « *possibles* », qui ont tendance à être utilisés dans beaucoup trop de situations (traductions non littérales, alignements plusieurs-à-plusieurs, etc.) (Fraser et Marcu, 2007).

Xu et Yvon (2016) ont séparé l'alignement de mots un-à-un et plusieurs-à-plusieurs. Pour que des annotations manuelles soient utiles de façon maximale et pour éviter la sémantique imprécise de la catégorie « *possible* », ils ont distingué ces quatre catégories pour les alignements un-à-un :

— Sûr : dans la plupart des contextes, où ils expriment le même sens (*dog* → *chien*)

— Contextuel : exprimer le même sens seulement dans un contexte spécifique (*tomorrow* → *samedi*)

— Partiel : la paire de mots ne constitue pas un bon alignement par eux-mêmes, mais ils doivent être inclus dans une paire de segments alignés plus longs : *(make) use (of)* → *(se) servir (de)*

— Faux : la paire correspondante ne doit pas être alignée

Ils ont testé ce schéma d'annotation sur les alignements un-à-un produits automatiquement par `MGIZA` (Gao et Vogel, 2008), sur un corpus multilingue de cinq paires de langues (anglais-français, anglais-espagnol, espagnol-français, grec-anglais, grec-français). Les résultats montrent que même si les catégories « *contextuel* » et « *partiel* » ont été moins fréquemment utilisées que « *sûr* » et « *faux* », elles représentent une portion non négligeable et parfois importante. Cette observation confirme leur hypothèse qu'une catégorisation plus fine des alignements « *possibles* » est nécessaire.

Concernant la collection des alignements plusieurs-à-plusieurs, Xu et Yvon (2016) ont proposé un protocole basé sur les divisions récursives d'une paire de phrases parallèles. Des heuristiques sont utilisées pour guider le processus d'annotation, surtout sur le choix des points de division.

Compte tenu de ces recherches précédentes, dans notre annotation des procédés de traduction, l'unité de traduction* est définie comme suit :

Définition ***Unité de traduction*** *: une paire d'unités sous-phrastiques bilingues (mots ou segments) en relation de traduction. L'unité de sens doit être la plus petite possible. Par contre, pour certaines traductions non littérales, il est nécessaire d'étendre la frontière pour inclure certains mots traduits littéralement, pour que le sens soit compréhensible.*

Nous avons établi ces conventions pour décider la frontière :

— Pour des traductions littérales au niveau lexical, annoter l'unité sémantique la plus petite possible :
there is → ***il y a***
effect → ***effet***
did n't have → ***n' avions pas*** (en cas de négation)

— Pour des articles français n'ayant pas de correspondance en anglais, nous attachons ceux-ci avec le nom modifié pour préciser leur appartenance :
play with ***blocks*** → *jouer avec* ***des cubes***

— Grouper la préposition avec le verbe qui déclenche son apparition :
we do n't ***want to*** *encourage people to eat* → *on ne* ***veut*** *pas encourager les gens à manger*

— Pour des traductions non littérales, il est parfois nécessaire d'élargir la frontière de segments en vue de clarifier le sens, même s'il existe des mots traduits littéralement à l'intérieur de ces segments.
spend a large sum of money → ***dépenser massivement*** (« *spend* » et « *dépenser* » sont traduits littéralement)
stamp a letter into it → ***avec une lettre en creux*** (« *a letter* » et « *une lettre* » sont traduits littéralement)
(la catégorie *Transposition* est attribuée à ces deux paires)

Des désaccords peuvent exister entre les annotateurs sur les frontières des unités de traduction, nous expliquerons plus tard comment nous résolvons ce problème.

5.2.3 Guide d'annotation

La typologie des catégories permet de donner une vue d'ensemble sur les catégories (voir figure 4.1). En vue de bien comprendre le contexte, les annotateurs peuvent regarder des vidéos des conférences correspondantes avant d'annoter.

Nous avons établi un guide d'annotation qui est consultable en accès libre (voir l'annexe B). [4] Le guide donne des définitions et des règles spécifiques pour l'ensemble des catégories. Des exemples typiques, des contre-exemples et des exemples difficiles sont fournis systématiquement. Nous avons créé des tableaux qui récapitulent les informations

4. Nous tenons à remercier Cyril Grouin pour ses nombreux conseils qui aident à améliorer le guide. Le guide est disponible en ligne : `https://yumingzhai.github.io/files/Annotation_guide_EN_FR.pdf`

essentielles pour mieux guider les annotateurs dans les étapes de décision. Les annotateurs sont encouragés à consulter des ressources de langue, par exemple les dictionnaires de Cambridge, Larousse, Le Robert, TLFi, etc. Des conventions ont été établies pour annoter les ponctuations, les segments non alignés et les anaphores linguistiques. Le guide contient aussi un tutoriel sur l'utilisation de l'outil `Yawat`.

Afin de garantir la qualité de corpus et de générer un jeu de données propre pour développer notre classifieur automatique, nous avons corrigé les fautes d'orthographe mineures dans le corpus, par exemple *ca* → *ça*, *a quel point* → *à quel point*, *l'endroit ou* → *l'endroit où*, etc. Pour cela, les annotateurs doivent noter l'identifiant de la phrase et les paires qui contiennent les fautes d'orthographe pendant l'annotation, et nous les corrigeons dans le corpus.

Le corpus ayant été préalablement aligné automatiquement par l'outil `FastAlign`, nous avons importé ces alignements produits en vue d'accélérer le processus d'alignement, en particulier pour les mots traduits littéralement. Les annotateurs ont pour consigne de corriger ces alignements si nécessaire.

5.2.4 Étude de contrôle

Nous avons évalué de manière conventionnelle la faisabilité de notre tâche d'annotation en mesurant un accord inter-annotateurs sur un corpus de contrôle. Deux annotateurs ont annoté indépendamment 100 paires de phrases (3 055 tokens anglais et 3 238 tokens français). [5] Puisqu'il existe des désaccords sur la frontière de certains segments, nous avons calculé la valeur du Kappa de Cohen (Cohen, 1960) uniquement pour les segments de mêmes frontières et obtenu la valeur 0,672, qui signifie un accord fort (selon Landis et Koch (1977)). Le nombre de tokens anglais annotés dans des segments de mêmes frontières est de 1 906 pour la catégorie *Littéral* et de 312 pour les autres catégories, ce qui couvre 72,60% des tokens source anglais.

Si nous calculons un accord inter-annotateur de manière plus flexible en incluant des paires avec des segmentations différentes mais compatibles (*i.e.* pas de chevauchement aux frontières) mais avec une annotation commune [6], la valeur de Kappa diminue à 0,617, ce qui dépasse le seuil d'un accord fort (0,61), mais est proche du seuil d'un accord modéré (0,60). Cependant, dans cette configuration, la couverture des tokens anglais augmente à 85,56%. Les tokens restants appartiennent eux à des segments aux frontières incompatibles (voir tableau 5.2).

	κ	%EN tokens
strict	0,672	72,60%
flexible	0,617	85,56%

Tableau 5.2 – L'accord inter-annotateur pour le corpus de contrôle anglais-français

Dans la figure 5.3, nous présentons la table de confusion d'annotation par deux annotateurs sur les segments de mêmes frontières. Sans surprise, la majorité des situations

5. Ils ont annoté ce corpus en janvier 2018. Le guide d'annotation n'était pas encore figé.

6. Par exemple, « *I was asked by* » et « *I was asked by my professor at Havard* » ont tous les deux été annotés par la catégorie *Modulation*, et « *my professor at Havard* » a été annoté en *Littéral* par le premier annotateur. Nous considérons que les deux segmentations sont compatibles et qu'il y a un accord entre les deux annotateurs sur le plus grand segment du fait de la catégorie commune *Modulation*.

d'accord correspondent aux traductions littérales. Il existe tout de même certains désaccords entre *Littéral* et *Équivalence (ex. in this way → de cette façon), Modulation (ex. this entire time → tout ce temps), Particularisation (ex. snuff → tabac)* et *Transposition (ex. their prayers **alone** → **seulement** leurs prières)*. Néanmoins nos catégories *Équivalence* et *Littéral* sont très proches, confusions que nous aurions donc pu considérer comme admissibles dans une mesure plus flexible. *Modulation* présente le plus grand nombre de confusions avec *Littéral* et *Transposition (ex. from the forest floor → tombées par terre)*, ce qui indique qu'il est nécessaire de mieux expliciter les différences entre ces catégories quand nous formons les annotateurs. *Mod+Trans* est une catégorie combinée pour laquelle certains annotateurs ne perçoivent parfois que l'une des deux catégories (*ex. a great distance → de loin*). Il existe très peu de confusion pour *Généralisation (ex. because they're denatured → étant dénaturés)* mais c'est moins le cas pour *Particularisation*. *Non aligné - Explicitation* et *Non aligné - Réduction* présentent très peu de confusion avec les autres catégories, mais sont parfois en compétition avec *Aucun Type*. La catégorie *Figurative*[7] est à l'origine de quelques désaccords (*ex. at the **base** of glaciers → aux **pieds** des glaciers*), qui peuvent notamment s'expliquer par la difficulté d'annotation pour un annotateur non natif de la langue cible.

Annot1\Annot2	Équivalence	Littéral	Modulation	Transposition	Mod+Trans	Généralisation	Particularisation	Explicitation	Réduction	Idiome	Métaphore	Incertain
Équivalence	21	4		5			1		1			
Littéral	27	1857	26	6			10					7
Modulation	4	8	37	7	1	1	3		2			
Transposition	6	7	10	30		1						
Mod+Trans		1	6	2	2	2						
Généralisation		1				17						1
Particularisation	4	13	6	2		1	29					2
Explicitation								10				
Réduction									40			
Idiome										0		
Métaphore	1					1	1				1	
Incertain	1	8	2	2		4			1			4

FIGURE 5.3 – Matrice de confusion d'annotation entre deux annotateurs sur le corpus de contrôle (nombre d'instances)

5.2.5 Processus en plusieurs passes

Le calcul d'une valeur d'accord inter-annotateurs permet certaines interprétations standard sur le corpus de contrôle, et la table de confusion nous aide à identifier des difficultés de la tâche. Afin de converger sur les frontières de segments et sur les attributions de catégories, nous avons adopté un processus d'annotation en plusieurs passes en vue d'obtenir une meilleure qualité d'annotation. Pour chaque sous-corpus, un premier annotateur réalise une première passe pour vérifier les alignements et attribuer l'ensemble des catégories puis un deuxième annotateur prend le relais, ce qui lui permet de modifier les alignements et/ou les catégories s'il existe un désaccord. Chaque fichier d'annotation est sauvegardé à l'issue de chaque passe pour documenter les différences dans l'annotation. Cette alternance peut se répéter jusqu'à la convergence de toutes les annotations. En pratique, nous nous limitons à 3 passes, la troisième étant effectuée par le premier annotateur du corpus. Nous constatons que le nombre de modifications dans la troisième passe décroît au fur et à mesure des sous-corpus annotés, rendant compte d'une adaptation progressive et rapide des annotateurs à la tâche. Ce mode d'annotation, plus coûteux, a toutefois été rendu

7. Ce tableau reflète le fait que nous n'avions pas, au début, regroupé *Idiome* et *Métaphore* dans la catégorie *Figurative*, ce qui est aussi visible dans d'autres figures ou tableaux ci-dessous.

nécessaire par la qualité visée ainsi que par les difficultés inhérentes à la segmentation observées sur le corpus de contrôle.

Les frontières de segments peuvent différer selon les annotateurs. Le processus en plusieurs passes par alternance permet de faire disparaître progressivement ce type de désaccord. Par exemple : *a learning tool for* ***language learners*** → *un outil d'apprentissage pour* ***ceux qui apprennent des langues***, le premier annotateur avait séparé *language* et *learners*, le second les avait ensuite regroupés en attribuant la catégorie *Modulation+Transposition*, ce qui a été finalement approuvé par le premier annotateur dans la troisième passe. Cet autre exemple consiste en une séparation : *I* ***want to start by*** *showing you* → *je* ***vais*** *vous montrer*, les deux annotateurs sont finalement tombés d'accord pour ne pas inclure « *showing* » et « *montrer* » dans la paire de catégorie *Généralisation* : *want to start by* → *vais*.

Pour des procédés *Équivalence*, *Particularisation* et *Figurative*, des annotateurs natifs de la langue cible sont plus à l'aise pour prendre des décisions. Quand un annotateur hésite sur le choix d'une catégorie appropriée, une bonne pratique est de réfléchir à une possible traduction littérale pour identifier des procédés de traduction suivis par le traducteur humain.

5.3 Statistiques sur le corpus annoté

Nous présentons dans la figure 5.4 les statistiques sur les changements effectués pendant notre processus d'annotation en trois passes sur un sous-corpus [8]. Les chiffres dans la dernière colonne de la deuxième figure (passe 2 à passe 3) signifient que, par exemple, le deuxième annotateur était d'accord pour les 37 instances de catégorie *Mod+Trans*, mais que le premier annotateur a corrigé ses premiers choix lors de la troisième passe. Ces deux tableaux montrent que les désaccords sur les frontières et sur les catégories diminuent progressivement grâce à ce processus en plusieurs passes. Des exemples de certains types de changements sont présentés dans le tableau 5.3.

Changement	Exemples
étendre la frontière	*the arctic ice cap* ***is, in a sense,*** → ***on peut voir*** *la calotte glacière arctique* ***comme*** (inclure « *on peut voir* »)
rajouter *Réduction*	***most of*** *the last three years* → *ces 3 dernières années*
Équivalence → *Modulation*	*nice one* → *pas mal*
Modulation → *Transposition*	*increasing rapidly* → *en augmentation rapide*
Incertain → *Généralisation*	*sea change* → *changement de tendance*
Modulation → *Mod+Trans*	*the proposal has been to* → *ils projettent de*

Tableau 5.3 – Exemples de changement qui ont lieu lors d'une deuxième passe

À ce jour nous avons annoté un corpus de contrôle et 9 sous-corpus anglais-français, ce qui représente 1 103 paires de phrases parallèles (24k tokens anglais et 25k tokens

8. Pour les 17 cas de désaccords de *Transposition* avec la même frontière, 11 cas concernent un changement de ponctuation. L'annotation de changement de ponctuation a ensuite été supprimée car cela ne concerne pas les procédés de traduction.

passe 1 à passe 2					
		même frontière		frontière différente	
	nb d'instances	même type	type différent	même type	type différent
Littéral	1443	1411	29	0	3
Équivalence	66	60	5	1	0
Généralisation	17	11	3	2	1
Particularisation	29	23	5	0	1
Modulation	54	50	4	0	0
Transposition	41	17	17	2	5
Mod+Trans	88	58	20	2	8
Idiome	2	1	1	0	0
Métaphore	5	0	5	0	0
Explicitation	3	3	0	0	0
Réduction	4	4	0	0	0
Incertain	17	13	4	0	0
Total	1769	1651	93	7	18

passe 2 à passe 3								
	même frontière			frontière différente				
	total	accord	désaccord	total	AFAT	AFDT	DFDT	correction du premier annotateur
Littéral	**29**	22	7	**3**	2	1	0	3
Équivalence	**5**	3	2	**1**	0	1	0	3
Généralisation	**3**	2	1	**3**	1	1	1	0
Particularisation	**5**	4	1	**1**	1	0	0	0
Modulation	**4**	4	0	**0**	0	0	0	1
Transposition	**17**	17	0	**7**	6	0	1	3
Mod+Trans	**20**	16	4	**10**	8	2	0	37
Idiome	**1**	1	0	**0**	0	0	0	0
Métaphore	**5**	5	0	**0**	0	0	0	0
Incertain	**4**	4	0	**0**	0	0	0	0

FIGURE 5.4 – Statistiques sur les changements de types (nombre d'instances) pendant un processus d'annotation en trois passes sur un sous-corpus. Acronymes utilisés pour la passe 2 à passe 3 : AFAT : accord sur la frontière et le type ; AFDT : accord sur la frontière mais avec différent type ; DFDT : frontière différente et type différent

français). [9]

Le tableau 5.4 donne les statistiques sur le nombre de tokens annotés par langue et par catégorie. Il apparaît que 74,42% des tokens anglais sont annotés avec les catégories *Littéral* et *Équivalence*, et 13,22% sont annotés avec *Modulation, Transposition, Mod+Trans, Généralisation* et *Particularisation*, qui sont les catégories de traduction les plus intéressantes pour la génération de variantes non paraphrastiques par pivot.

5.4 Extension des études au couple anglais-chinois

Ayant réalisé les travaux sur le premier couple de langues, nous voulons tester la généricité de notre typologie de procédés de traduction et du guide d'annotation. Par conséquent, nous avons étendu les études au couple anglais-chinois, qui partage moins de points communs au niveau linguistique et culturel par rapport au couple anglais-français.

La figure 5.5 montre un exemple de traduction anglais-chinois. Les segments « *well* » et « *we use that great euphemism* » sont omis en chinois. L'idiome « *trial and error* » est traduit par une généralisation « *faire diverses expériences et commettre des fautes sans arrêt* ». Le segment « *which is exposed to be* » est traduit par une modulation « 结

9. La version finale de la thèse inclura les statistiques sur les sept sous-corpus restants. Le corpus annoté final contiendra 2 436 paires de phrases.

	anglais	français	% EN tokens
Littéral	23,733	25,288	69,49%
Équivalence	1,685	1,853	4,93%
Transposition	1,141	1,403	3,34%
Modulation	1,247	1,243	3,65%
Mod+Trans	1,171	1,263	3,43%
Généralisation	401	267	1,17%
Particularisation	555	679	1,63%
Figurative	57	59	0,17%
Changement lexical	1,049	1,028	3,07%
Réduction	797	0	2,33%
Explicitation	0	364	0,00%
Incertain	306	304	0,90%
Tous les types	32,213	33,818	94,32%
Aucun Type	1,939	1,770	5,68%
Nb tokens total	34,152	35,588	-

Tableau 5.4 – Statistiques sur les annotations anglais-français (nombre de tokens)

well, we use that great euphemism, "trial and error", which is exposed to be meaningless.

我们	普通人	会	做	各种各样	的	实验
nous	les personnes normales	particule au temps futur	faire	divers	particule d'attribut	expérience
不断	地	犯错误	结果	却	一无所获	
sans arrêt	particule d'adverbe	commettre une faute	par conséquent	cependant	ne rien gagner	

FIGURE 5.5 – Un exemple de traduction anglais-chinois

果(*par conséquent*) 却(*cependant*) ». L'adjectif « *meaningless* » est traduit par un idiome en quatre caractères « 一无所获(*ne rien gagner*) ».

L'anglais et le français sont très similaires en vocabulaire et en grammaire. En revanche, les changements de structures de phrase ont tendance à être plus importants en chinois qu'en français, et la langue chinoise privilégie des phrases courtes et compactes. La figure 5.6 montre un tel exemple. Tandis que la traduction française peut suivre le même ordre de mots que celui de la phrase source anglaise, la traduction chinoise a divisé une sous-phrase en deux, qui met en avant « *les premières expériences* » par la préposition « 在...前 (*avant*) », et la partie « *en commençant par* [...] » a été liée par l'adverbe « 先 (*d'abord*) ».

it took several stages, in fact, starting with a bioethical review before we did the first experiments.

en réalité cela s'est fait en plusieurs étapes, en commençant par un rapport bioéthique avant les premières expériences.

实际上	它	经过	了	好几个	阶段	。
en réalité	il	passer	particule au temps passé	plusieurs	étape	.
在	我们	做	最初	的	试验	前
在. . . 前 (avant)	nous	faire	premier	particule d'attribut	expérience	在. . . 前 (avant)
,	先	进行	了	一	次	生物
,	d'abord	mener	particule au temps passé	un	fois	biologie
伦理学	的	评估	。			
éthique	particule d'attribut	évaluation	.			

FIGURE 5.6 – Différence en structure de phrase entre la traduction française et chinoise

Corpus de *TED Talks* Nous avons constaté que dans notre corpus multilingue pa-

rallèle de *TED Talks*, la qualité de la traduction chinoise n'est pas aussi bonne que la traduction française. Pour illustrer le problème, nous avons annoté trois sous-corpus trilingues. Les statistiques comparatives sont présentées dans le tableau 5.5.

	anglais	français	%EN tokens	anglais	chinois	%EN tokens
Littéral	4,267	4,423	68,13%	3,307	5,311	52,80%
Équivalence	406	514	6,48%	426	629	6,80%
Transposition	142	195	2,27%	166	258	2,65%
Modulation	589	617	9,40%	615	863	9,82%
Mod+Trans	188	225	3,00%	90	134	1,44%
Généralisation	90	65	1,44%	200	208	3,19%
Particularisation	197	256	3,15%	273	661	4,36%
Idiome	4	6	0,06%	10	21	0,16%
Métaphore	10	15	0,16%	6	10	0,10%
Réduction	92	-	1,47%	314	-	5,01%
Explicitation	-	58	-	-	929	-
Incertain	79	79	1,26%	157	277	2,51%
Tous les types	6,064	6,453	96,82%	5,564	9,301	88,84%
Aucun type	199	223	3,18%	699	318	11,16%
Nb tokens total	6,263	6,676	-	6,263	9,619	-

Tableau 5.5 – Statistiques en contraste sur trois sous-corpus parallèles trilingues annotés (anglais-français-chinois) (nombre de tokens)

Nous voyons que moins de tokens anglais sont traduits littéralement en chinois. Pour *Équivalence, Transposition* et *Modulation*, les proportions ne sont pas très différentes, mais les frontières des segments peuvent être différentes. Des traductions chinoises utilisent moins le procédé complexe *Modulation+Transposition*, mais elles ont plus recours aux procédés *Généralisation* et *Particularisation*. En général, la catégorie *Figurative* est moins présente dans les deux langues cibles.

Pourtant, l'introduction des idiomes chinois en quatre caractères pour traduire des expressions non figées est considérée comme une bonne pratique qui permet d'obtenir des textes concis adaptés à la culture chinoise (voir les exemples dans le tableau 5.6). Puisque ces idiomes peuvent être traduits de différentes manières plus ou moins libres, ils peuvent contribuer de façon importante à l'obtention de paraphrases par pivot.

bring our children into the world	生儿育女 *(give birth to and raise children)*
(forest) upon which the people depend	栖身之所 *(shelter)*
pressured the people a little bit about it	刨根问底 *(inquire into the root of the matter)*
died getting old	行将就木 *(getting closer and closer to the coffin)*
with impunity	毫发无损 *(without injury; unhurt)*

Tableau 5.6 – Exemples qui utilisent des idiomes chinois pour traduire, trouvés dans notre corpus de *TED Talks* anglais-chinois

Les catégories *Réduction* et *Explicitation* montrent des différences claires avec les traductions en français. Nous avons observé que les traductions chinoises des sous-titres expriment souvent seulement les informations les plus importantes et laissent d'autres phrases de contenu non traduites. Concernant la catégorie *Explicitation*, cela inclut l'ajout

des classificateurs (mot de mesure) nécessaires en chinois ; des mots pleins (par exemple une phrase nominale en tant que sujet) pour garder la phrase grammaticale, au lieu de traduire littéralement mot à mot ; et quelques instances d'anaphore résomptive (Charolles, 2002), qui consiste à ajouter un segment ou une phrase qui résume l'information présente dans la phrase précédente. Il existe aussi plus d'instances de la catégorie *Incertain*, ensemble avec *Réduction*, ces phénomènes reflètent que la qualité des traductions chinoises ne sont pas aussi bonnes que les traductions françaises. Beaucoup plus de mots anglais n'ont aucune catégorie attribuée en raison de différences importantes en grammaire. En même temps, les traductions chinoises sont plus concises que les transcriptions de l'anglais oral, ce qui conduit à l'omission de beaucoup de mots anglais de transition.

Cette moins bonne qualité est causée par des raisons multiples : manque de connaissances du domaine spécifique d'une conférence ; manque d'édition finale pour corriger des erreurs évidentes, etc. De cela il résulte notamment des mots anglais laissés non traduits et des traductions erronées. En outre, des traductions extrêmement libres posent même de réelles difficultés pour l'alignement de mots manuel.

L'accord inter-annotateur sur un corpus de contrôle anglais-chinois de *TED Talks* (100 lignes, 3 055 tokens anglais et 4 195 caractères chinois [10]) est montré dans le tableau 5.7. Les deux valeurs de kappa signifient un accord modéré. [11]

	κ	%EN tokens
strict	0,61	52,76%
flexible	0,60	74,10%

Tableau 5.7 – L'accord inter-annotateur pour le corpus de contrôle anglais-chinois

Compléter avec d'autres corpus anglais-chinois Compte tenu de la qualité non satisfaisante du corpus anglais-chinois de *TED Talks*, nous avons cherché d'autres corpus disponibles pour les compléter. Il faut que la direction de traduction soit au mieux anglais → chinois et que le corpus contienne plusieurs genres différents (pas seulement le genre du discours préparé). En plus, si l'alignement de mots est déjà fait, nous pourrons concentrer nos efforts sur d'autres tâches.

Nous listons ci-dessous les corpus anglais-chinois que nous avons identifiés :

— `UM-corpus` (Tian *et al.*, 2014) : ce corpus a été construit pour servir à la traduction automatique. La direction de traduction est anglais → chinois. Il contient 2,2M de phrases parallèles, et est divisé en huit genres (*matériel pour l'éducation, loi, microblog, journal, science, parlé, sous-titre, thèse*) avec une distribution à peu près équilibrée. L'alignement au niveau de la phrase a été manuellement corrigé. En revanche, il reste des erreurs, par exemple les cas où un grand segment n'est pas traduit dans une phrase. Ainsi il faut annoter ce corpus en filtrant les paires incomplètes ou erronées. La segmentation de mots en chinois et l'alignement de mots ne sont pas fournis. Le corpus est en libre accès sous la licence *Creative Commons Non-Commercial 4.0*. Sur le droit de rediffuser les annotations sur ce corpus, nous n'avons pas encore eu la réponse des auteurs.

10. Le côté anglais est le même que celui du corpus de contrôle anglais-français.
11. Deux annotateurs chinois ont annoté ce corpus indépendamment en avril 2018. Le guide d'annotation n'était pas encore figé.

— `Tsinghua-corpus` (Liu et Sun, 2015) : c'est un jeu de donnée de développement et de test pour évaluer l'outil d'alignement automatique de mots des auteurs. Il contient 40 715 paires de phrases qui ont été alignées manuellement au niveau du mot. L'alignement de la phrase est beaucoup plus propre que celui du `UM-corpus` et la segmentation de mots en chinois est déjà faite. Par contre, selon l'auteur, la direction de traduction est parfois inverse (*i.e.* chinois → anglais). Le pourcentage de chaque genre (journal, sous-titre, etc.) est inconnu, mais il est sûr que le genre journalistique occupe une majeure partie. Le corpus est en libre accès, et nous pouvons rediffuser le corpus annoté.

— `UP-corpus` (Chang et Bai, 2003) : ce corpus a été construit principalement pour entraîner les systèmes de traduction automatique. L'alignement au niveau de la phrase a été manuellement vérifié, et le corpus contient une grande variété de genres. Nous avons demandé une partie du corpus dans les domaines de la littérature, l'art et la science, qui est fourni gratuitement pour la recherche après la signature d'une convention de transfert. En revanche, nous n'avons pas le droit de rediffuser le corpus annoté.

— `UnitedNations-corpus` (Ziemski *et al.*, 2016) [12] : ce corpus en libre accès contient des comptes rendus officiels et des documents parlementaires des Nations Unies. Nous prenons un échantillon de ce corpus pour annoter les procédés de traduction dans le genre textuel *politique*.

— Pour le genre textuel *Thèse*, après notre examen, la partie contenue dans `UM-corpus` a une qualité non satisfaisante pour l'annotation. Ainsi, nous avons construit notre propre échantillon de corpus, en collectant des résumés bilingues depuis ces revues en ligne : *Chinese Linguistics* [13], *Chinese Journal of Software* [14], *Chinese Journal of Computers* [15]. La direction de traduction est majoritairement du chinois vers l'anglais. Seuls les résumés bilingues proposant le même niveau de contenu ont été conservés.

La plateforme *Linguistic Data Consortium* (LDC) [16] propose seulement des corpus avec la direction de traduction chinois → anglais. Quelques corpus sont alignés au niveau du mot, mais ils sont payants. La plateforme CLARIN [17] fournit quelques corpus anglais-chinois dont les genres sont déjà couverts par les corpus que nous avons mentionnés.

En résumé, pour le couple anglais-chinois, nous avons annoté un corpus parallèle qui contient ces onze genres différents : *matériel pour l'éducation, loi, microblog, journal, science, parlé, sous-titre, thèse, politique, littérature*, et *art*. Le corpus anglais-français contenant 2 436 paires de phrases, nous prenons un échantillon de 2 200 paires de phrases anglais-chinois pour l'annotation, qui contient 200 paires de phrases pour chaque genre. [18]

Pré-traitement du corpus Nous avons utilisé l'outil `THULAC` (Li et Sun, 2009) pour la segmentation de mots en chinois. L'alignement de mot automatique est fait par l'outil

12. `https://cms.unov.org/UNCorpus/en/DownloadOverview`
13. `http://www.cuhk.edu.hk/journal/jcl/`
14. `http://www.jos.org.cn/josen/ch/index.aspx`
15. `http://english.ict.cas.cn/`
16. `https://www.ldc.upenn.edu/`
17. `https://www.clarin.eu/resource-families/parallel-corpora`
18. En vue d'annoter des paires de phrases de bonne qualité, pour le genre du journal, nous avons utilisé `Tsinghua-corpus` au lieu de `UM-corpus`. Pour les autres genres, l'échantillon de 200 paires de phrases est extrait dans le corpus correspondant.

`TsinghuaAligner` (Liu et Sun, 2015). Ces deux outils obtiennent des résultats au niveau de l'état-de-l'art sur les tâches correspondantes.

Le corpus chinois segmenté automatiquement contient des erreurs qui peuvent conduire à des alignements et des attributions de catégorie incorrects. Certains mots chinois ont donc besoin d'une re-segmentation manuelle avant l'annotation, afin de mieux correspondre aux segments anglais.

Par exemple :
only is → 仅仅是
a été corrigé en :
only is → 仅仅 *(only)* 是*(is)*
Les annotateurs doivent noter ces cas de re-segmentation nécessaires, ainsi que les fautes d'orthographes, afin que nous les corrigions dans le corpus.

Collaboratrices Nous avons travaillé avec deux étudiantes chinoises sur cette partie : Lufei Liu et Xinyi Zhong, titulaires des masters en interprétation anglais-français-chinois et en traduction anglais-chinois, respectivement. Xinyi nous a d'abord aidés à améliorer le guide au niveau de la structure et des consignes, pour améliorer la clarté du guide pour les annotateurs. Lufei a ensuite travaillé avec nous en tant que stagiaire pendant trois mois. Elle a annoté une partie du corpus et a enrichi le guide d'annotation.

Guide d'annotation Nous avons adapté et enrichi le guide d'annotation en nous basant sur le guide pour le couple de langues anglais-français. [19] La structure du guide reste la même et nous avons remplacé des exemples anglais-français par des exemples anglais-chinois trouvés dans le corpus. Après avoir annoté plus de 300 paires de phrases représentatives de différents genres, nous avons conservé la même typologie de procédés de traduction que celle du couple anglais-français. Ceci montre que notre typologie est assez générale, car elle permet d'annoter ce couple de langues qui sont bien moins similaires.

En revanche, les nombres de règles précises pour certains procédés sont différents. Par exemple, nous avons dû ajouter plusieurs règles pour la particule chinoise 《 的 》 ("de"), qui présente différentes fonctions selon le contexte ; et des règles sous la catégorie *Explicitation*, qui correspondent à des phénomènes linguistiques spécifiques en chinois.

Parmi les couples que nous avons étudiés, il semble que certains phénomènes n'existent que pour le couple anglais-français. Par exemple, la double transposition nommée « *chassé-croisé* » qui consiste à un changement de catégories grammaticales et une permutation syntaxique des éléments qui constituent le sens :
so they ***speared*** *the five missionaries* ***to death***
→ *ils ont donc* ***abattu*** *les cinq missionnaires* ***à coups de lance***
(*speared* → *à coups de lance, to death* → *abattu*)

Statistiques de l'annotation Lufei a annoté 316 paires de phrases pendant 3 mois dans ce nouveau corpus que nous avons constitué. Le tableau 5.8 montre les statistiques (tous les genres de corpus confondus) avant déduplication.

Après déduplication, nous obtenons 2 221 instances de catégorie *Littéral*, et 465 instances de catégorie *Non Littéral* (qui contient les instances de *Équivalence* jusqu'à *Changement lexical* dans le tableau 5.8). La direction de traduction est de l'anglais vers le

19. Le guide est disponible en ligne : `https://yumingzhai.github.io/files/Annotation_guide_EN_ZH.pdf`

Catégorie	nb d'instances	nb de tokens anglais	nb de caractères chinois
Littéral	3 410	4 440	6 604
Équivalence	169	298	420
Transposition	167	197	280
Modulation	73	199	245
Mod+Trans	13	14	29
Généralisation	25	52	55
Particularisation	72	104	221
Figurative	10	16	30
Changement lexical	27	44	54
Réduction	464	507	-
Explicitation	710	-	1 013
Erreur de traduction	12	23	28
Incertain	13	22	33
Total	5 165	5 916	9 012

Tableau 5.8 – Statistiques sur les annotations anglais-chinois (nombre d'instances et nombre de tokens)

chinois. Seul le genre *Thèse* dans notre corpus est traduit totalement du chinois vers l'anglais, nous étudierons ces résultats d'annotation à part.

5.5 Perspectives

Avec ce nouveau jeu de données sur le couple anglais-chinois, nous pouvons mener des expériences de classification binaire, en essayant de transférer des expériences pour le couple anglais-français que nous présenterons dans le prochain chapitre. Les instances de catégorie *Réduction* et *Explicitation* fournissent des exemples intéressants pour mener une analyse linguistique explicitant les différences entre ces deux langues.

Pendant l'annotation sur les traductions en chinois, nous avons rencontré des exemples typiques de changement de structure de phrase, qui relèvent d'un changement stylistique. Par exemple, dans la figure 5.7, la phrase nominale anglaise « *the export of [...] which [...]* » est beaucoup trop longue pour être traduite littéralement en chinois. Par conséquent, le traducteur a mis cette partie au début de la phrase, pour rendre la traduction plus facile à comprendre. Ce phénomène est pour l'instant hors du cadre de notre étude qui se focalise sur les phénomènes de traduction au niveau sous-phrastique.

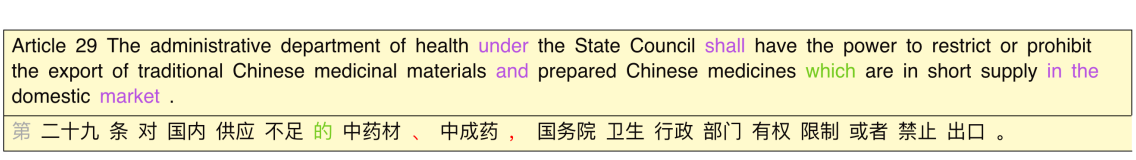

FIGURE 5.7 – Exemple de changement de structure de la phrase. Voici une traduction littérale de la phrase chinoise en français : « *Article 29 Pour des matières médicinales et des médicaments préparés traditionnels chinois qui sont en pénurie dans le marché intérieur, le département administratif de la santé sous l'égide du conseil d'État a le droit de restreindre ou interdire l'exportation.* »

Les annotateurs (quelle que soit la langue cible) rencontrent parfois des paires de

phrases où la traduction est très éloignée de la phrase source au niveau de la forme. Le noyau sémantique est plus ou moins conservé, mais la traduction semble parfois être une erreur. Cela montre encore une fois que cette tâche d'annotation est difficile.

Pour permettre aux annotateurs d'indiquer leur confiance sur la qualité de traduction, nous envisageons d'utiliser un symbole (*ex.* $\$\$\$$) à la fin de chaque phrase source. Les annotateurs peuvent l'annoter d'une autre couleur s'ils sont incertains de la qualité de traduction, comme ce qui a été fait dans le travail de Xu et Yvon (2016). Cela nous permettra d'écarter ces phrases quand nous construirons le jeu de données d'apprentissage, ou de leur attribuer un poids moins important.

5.6 Conclusion

Dans ce chapitre, nous avons présenté les détails du corpus anglais-français de *TED Talks* utilisé pour l'annotation, suivi par différents aspects liés à l'annotation : l'outil, la segmentation en unité de traduction et l'alignement de mots, le guide d'annotation, l'étude de contrôle et le processus à trois passes adopté. Nous avons montré les statistiques calculées sur le corpus annoté jusqu'à présent. Les exemples annotés serviront comme jeu de données pour entraîner un classifieur automatique, que nous présenterons dans le chapitre suivant.

Ayant travaillé sur ce premier couple de langues, nous avons étendu les études sur le couple anglais-chinois, pour lequel nous avons constitué un corpus qui contient onze genres différents. Pendant l'annotation, nous avons adapté et enrichi le guide d'annotation, mais nous avons pu garder la même typologie de procédés de traduction. Nos collaboratrices chinoises finalisent actuellement l'annotation. Nous pourrons transférer des expériences menées sur le couple anglais-français pour ce nouveau couple, et calculer un nouvel accord inter-annotateur selon le guide mis à jour.

Chapitre 6

Reconnaissance des procédés de traduction

Sommaire

Dans ce chapitre, nous proposons une classification automatique des procédés de traduction en nous basant sur des exemples annotés manuellement dans un corpus parallèle anglais-français de *TED Talks*. L'annotation de ce corpus a été décrite dans le chapitre précédent. Même si le jeu de données est petit, les résultats expérimentaux valident notre hypothèse de travail : il est possible de reconnaître les procédés de traduction. Les expériences montrent les directions à suivre dans les travaux futurs.

Une partie des travaux réalisés dans ce chapitre ont été publiés dans ces deux articles : Zhai *et al.* (2019c) et Zhai *et al.* (2019a).

6.1 Travaux précédents

Notre étude se concentre sur la classification des procédés de traduction au niveau sous-phrastique, en nous basant sur des traductions humaines annotées manuellement.

Cette classification pourrait être utilisée pour évaluer la qualité de la traduction automatique au niveau sous-phrastique à grain fin, en comparant avec la traduction humaine. Notre but à long terme serait d'utiliser cette classification pour avoir un meilleur contrôle sémantique lors de l'extraction de paraphrases sous-phrastiques à partir des corpus parallèles bilingues. Cette recherche pourrait aussi être utile pour aider les apprenants de langues étrangères pour qu'ils soient conscients des procédés de traduction utilisés dans

les corpus parallèles. Cet aspect orienté éducation n'a pas encore été développé dans les études précédentes. Nous détaillerons ces points dans le chapitre 7.

Nous allons d'abord détailler quelques travaux concernant les traductions humaines et surtout la traduction non littérale.

Pour trouver une façon de décrire pourquoi la traduction est difficile et non déterministe, Carl et Schaeffer (2017) ont proposé une définition hypothétique de la « *traduction littérale absolue* », qui désigne les cas où la traduction est identique à la langue source aux niveaux syntaxique et sémantique. Ils ont développé un cadre informatique pour mesurer la « *non littéralité* » des traductions réelles. Des preuves théoriques et empiriques montrent que la traduction littérale est plus facile et rapide à produire. Plus la traduction dévie des critères de « littéralité », plus il sera difficile et coûteux en temps de produire une telle traduction depuis le début ou pendant la post-édition de la traduction automatique.

Dans l'étude de Carl et Schaeffer, une approche empirique pilotée par le corpus est présentée, pour mesurer la similarité cross-lingue syntaxique et sémantique. Leur corpus multilingue TPR-DB* (*Translation Process Research DataBase*) contient de nombreuses traductions alternatives. Les indicateurs de « *littéralité* » syntaxique et sémantique renvoient à la variation de l'ordre des mots, ainsi que celle des choix lexicaux. Ils se focalisent sur l'effort en termes de temps utilisé, avec l'hypothèse que des temps de production plus courts impliquent moins d'effort cognitif. D'après eux, il semble que la variation syntaxique en traduction soit corrélée à la variation du choix lexical et réciproquement.

Tan et Bond (2011) ont constaté que dans le corpus multilingue parallèle de NTU (*Nanyang Technological University*), l'introduction des idiomes chinois pour traduire un segment anglais non idiomatique est souvent rencontrée. Pour identifier des idiomes dans la traduction chinoise, Ho *et al.* (2014) ont créé un lexique de 4 000 idiomes en quatre caractères pour étiqueter toutes les occurrences des idiomes dans le corpus de NTU en quatre genres. L'extension de la liste se réalise via une révision et correction manuelle. Leur but final est d'intégrer la liste dans le Wordnet chinois.

Poirier (2014) a proposé un algorithme semi-automatique pour la détection des erreurs et divergences de traduction (*translation shift*) au niveau du contenu. Dans son travail, les divergences de traduction* sont imposées par les normes de la langue cible, qui permettent d'exprimer la même idée avec un nombre différent de mots pleins et une construction différente. Nous pouvons constater que si la construction de la phrase source avait été copiée telle quelle, cela apporterait des expressions mal formées ou non fluides dans la langue cible.

Cette méthode nécessite un couple de phrases déjà alignées en entrée, et les mots outils sont enlevés selon une liste préalablement établie pour chaque langue. L'algorithme va comparer la quantité de mots pleins restants dans les deux langues pour voir si les nombres sont égaux. Cette étape est suivie par une localisation manuelle des erreurs ou divergences de traduction, par une association manuelle des unités de sens (mot ou segment). Cette étude peut indiquer aux traducteurs des erreurs ou divergences potentielles dans la traduction, qui doivent être ensuite vérifiées plus attentivement.

La tâche QE (*Quality Estimation**) consiste à prédire la qualité de la traduction automatique quand l'évaluation automatique ou humaine n'est pas possible. Typiquement elle se passe lors de l'exécution du système de traduction. L'estimation de la qualité se fait sur des unités de traduction variées (mot, segment, phrase, document) sans s'appuyer sur des traductions de référence (Specia *et al.*, 2010). Les applications potentielles de QE sont multiples : prédire l'effort de post-édition pour des traducteurs professionnels (Specia, 2011) ; rendre les utilisateurs finaux conscients de la qualité de traduction ; sélectionner la

meilleure traduction parmi des options venant de multiples systèmes, etc.

Concernant le QE au niveau de la phrase, Blatz *et al.* (2004) ont d'abord considéré ce problème d'estimation comme une tâche binaire. Specia *et al.* (2009) ont proposé une tâche de régression pour estimer des scores continus, dont les résultats sont plus appropriés pour l'estimation du temps nécessaire pour la post-édition. Dans la première tâche partagée sur QE, Callison-Burch *et al.* (2012) ont proposé deux sous-tâches : classer les résultats générés par un système de traduction statistique basé sur des segments, et leur attribuer des scores. Des analyses en détail du jeu de données et des résultats ont été menées, entre autres, par Wisniewski *et al.* (2013). La tâche partagée plus récente sur QE inclut des traductions générées par des systèmes de traduction automatique neuronaux, et les niveaux de granularité sont *mot, segment, phrase et document* (Specia *et al.*, 2018).

Pour des prédictions au niveau de la phrase, dans la tâche partagée en 2019 par exemple, les systèmes doivent donner des scores aux phrases selon des efforts de post-édition, via un pourcentage des éditions nécessaires (HTER (*Human-targeted Translation Edit Rate*)). Au niveau du mot, les participants doivent détecter les erreurs de traduction pour chaque token et des tokens qui n'ont pas été traduits. Au niveau sous-phrastique (ou du mot), nous voyons que l'estimation de la qualité de traduction non littérale n'est pas encore prise en compte.

Récemment, différents modèles ont été proposés pour détecter automatiquement des divergences de traduction dans des corpus parallèles. Le but est de filtrer automatiquement des couples de phrases qui ne sont pas parallèles au niveau du sens (erreur de traduction, problème d'alignement de phrase, omission de segment source, explicitation dans la traduction, etc.), afin d'améliorer la qualité du corpus d'entraînement pour des systèmes de traduction automatique. Carpuat *et al.* (2017) ont introduit un détecteur de divergence cross-lingue basé sur SVM, en utilisant des traits en alignement de mots et en longueur de phrase. Vyas *et al.* (2018) ont proposé une approche basée sur des réseaux neuronaux profonds, dont l'entraînement ne demande pas d'annotation manuelle. D'une façon non supervisée, Pham *et al.* (2018) ont généré des plongements phrastiques en fonction de la similarité entre les mots. Ils mesurent l'équivalence sémantique entre les phrases afin de guider le filtrage.

Une tâche sémantique proche de la nôtre concerne la reconnaissance d'implication textuelle (Dagan *et al.*, 2005).

Définition ***TE* (Textual Entailment)*** *: l'implication textuelle désigne la relation entre deux phrases en langue naturelle (une prémisse P et une hypothèse H) qui existerait si un humain lisant P pourrait inférer que H est probablement vrai.*

Les traits exploités pour le RTE (*Recognizing Textual Entailment*) monolingue ont été utilisés pour améliorer la mesure d'évaluation en traduction automatique (Padó *et al.*, 2009a,b). Par la modélisation des correspondances et non-correspondances entre la traduction automatique et la référence humaine, ils distinguent les variantes de traduction qui préservent le sens des mauvaises traductions.

Mehdad *et al.* (2010) ont proposé l'extension du cadre de la RTE monolingue vers une tâche cross-lingue (*Cross-Lingual Textual Entailment (CLTE)*). Ils ont organisé deux tâches partagées sur l'utilisation de CLTE pour la synchronisation automatique des contenus multilingues (Negri *et al.*, 2012, 2013). Sur des paires de phrases cross-lingues, les participants doivent identifier la relation d'implication multi-directionnelle ("en avant", "en arrière", "bi-directionnel", "pas d'implication"). Étant donné deux documents sur le même sujet écrits dans deux langues différentes (*ex.* les pages de Wikipédia), l'application de cette tâche de synchronisation consiste à détecter automatiquement et résoudre des

différences dans les informations qu'ils contiennent, en vue de produire des documents alignés et mutuellement enrichis.

Plus récemment, Upadhyay *et al.* (2018) ont proposé des approches non supervisées pour identifier l'hyperonymie cross-lingue, une relation d'implication asymétrique à grain fin. Leur approche distributionnelle utilise un modèle de plongement de mots cross-lingue appris sur un petit dictionnaire bilingue et une variété de contextes syntaxiques monolingues extraits de corpus analysé en dépendance. En revanche, leur jeu de données cross-lingue construit par le *crowdsourcing* est composé de paires de mots sans contexte.

De notre côté, l'étude se passe au niveau sous-phrastique, à savoir des mots ou des segments. Nous choisissons cette granularité au lieu de la phrase entière parce que les segments sont plus réutilisables et plus facilement identifiables dans les corpus. Notre classification automatique s'appuie sur des données annotées manuellement, qui sont différentes de celles parfois synthétiques dans les travaux précédents.

Contrairement aux efforts qui ont lieu au niveau phrastique et qui effectuent une décision binaire (bonne ou mauvaise traduction), nous nous intéressons à un niveau plus fin (sous-phrastique), et aux bonnes traductions mais effectuées en utilisant différents procédés de traduction (surtout ceux non littéraux). Nous menons une classification binaire et aussi en multi-classes. Certains procédés apportent des *glissements* sémantiques ou syntaxiques, mais ce ne sont pas des divergences de traduction qui influencent la qualité de façon négative.

6.2 Jeu de données

Statistiques Le tableau 6.1 présente le nombre d'instances par catégorie pour notre expérience de classification. La catégorie *modulation+transposition* a été annotée dans le corpus, parce qu'elle représente un phénomène important où les deux procédés sont combinés, et nous ne pouvons pas dire que l'un est plus important que l'autre. En revanche, à ce stade, il n'existe que 53 instances de cette catégorie. En conséquence, nous combinons *transposition* et *modulation+transposition* dans une catégorie *contient_transposition* pour l'expérience, où la catégorie *modulation* est considérée comme neutre. La catégorie *traduction_figurative* est sous-représentée dans notre corpus, nous décidons de l'ignorer pour l'expérience. Le jeu de données n'étant pas équilibré, nous évaluerons les classifieurs sous différentes configurations.

littéral	3771	littéral (3771)
équivalence	289	
contient_transposition (transposition + mod_trans)	342	
modulation	195	non_littéral (1127)
généralisation	86	
particularisation	215	

Tableau 6.1 – Nombre d'instances par catégorie

Pré-traitement Pour le corpus anglais et français, la tokenisation a été faite avec `Stanford Tokenizer` [1]. Nous l'avons aussi utilisé pour pré-traiter le corpus avant l'annotation manuelle. Nous mettons tous les mots en minuscules. La lemmatisation en anglais est faite avec `Stanford CoreNLP` (Manning *et al.*, 2014), celle en français avec

1. `http://nlp.stanford.edu/software/tokenizer.shtml`

`Tree Tagger` (Schmid, 1995) [2]. Pendant la lemmatisation, la tokenisation initiale générée par `Stanford Tokenizer` a été conservée pour faciliter l'extraction de traits dans la prochaine étape.

Objectif de classification Nous menons des expériences dans un scénario simplifié, où nous connaissons déjà les frontières des couples bilingues, et nous ne prédisons que le procédé de traduction. Par exemple, étant donné le couple *deceptive* → *une illusion*, le but est de prédire son étiquette *contient_transposition*.

6.3 Des traits indépendants du contexte

Puisque le nombre d'instances pour la validation croisée est assez limité, nous avons travaillé principalement sur l'ingénierie des traits, et entraîné différents classifieurs d'apprentissage statistique avec la boîte à outils `Scikit-Learn` (Pedregosa *et al.*, 2011). Nous avons aussi testé des architectures en réseau neuronaux (voir section 6.4). [3]

Nous avons d'abord travaillé sur les traits qui sont indépendants du contexte. En analysant les caractéristiques des données, nous avons proposé des traits dont l'implémentation diffère en complexité et en ressources utilisées. Cinq groupes de traits sont décrits ci-dessous pour le couple anglais-français : différences en formes de surface, en analyse morpho-syntaxique et en analyse syntaxique ; des traits qui utilisent des ressources externes, et des informations en alignement de mots automatique.

Les jeux d'étiquettes des deux langues pour l'analyse morpho-syntaxique, l'analyse syntaxique en constituant et en dépendance ont été convertis en trois jeux unifiés et compacts (Petrov *et al.*, 2012; Leung *et al.*, 2016).

1. Surface

- Nous comptons le nombre de tokens en anglais (l_e) et en français (l_f). Ensuite nous calculons le ratio de ces nombres (l_e/l_f, l_f/l_e) et la distance de Levenshtein (Levenshtein, 1966) en caractères entre les segments. Le français et l'anglais étant proches en orthographe pour certains mots apparentés (Hammer et Monod, 1976), une paire de mots ou segments en traduction littérale pourrait avoir une petite distance Levenshtein.

2. Analyse morpho-syntaxique (PoS)

- L'analyse est faite avec `Stanford CoreNLP` pour les deux langues. Pour chaque langue, le nombre d'occurrence de chaque étiquette est compté dans un vecteur (il y a au total 17 étiquettes de PoS universelles). Ces deux vecteurs sont utilisés comme traits (voir tableau 6.2). Nous calculons ensuite la similarité cosinus entre ces deux vecteurs. Cette similarité est calculée sur tous les mots ainsi que sur les mots pleins uniquement. [4]
- En nous basant sur des instances de la catégorie *contient_transposition*, nous avons manuellement construit une cinquantaine de patrons de changement de séquence de

2. Puisque la lemmatisation en français n'est pas encore possible avec `Stanford CoreNLP`.
3. Le jeu de données et le code sont disponibles ici : `https://github.com/YumingZHAI/ctp`.
4. Les étiquettes de mots pleins contiennent : ADJ, ADV, NOUN, PROPN, VERB. Si un segment ne contient aucun mot plein, nous utilisons le segment original.

	ADJ	DET	NOUN	...	INTJ
anglais	1	0	0	0	0
français	0	1	2	0	0

Tableau 6.2 – Vecteurs des nombres d'occurrences pour chaque étiquette de PoS

PoS (Hunston et Francis, 2000). Nous vérifions ensuite si une instance correspond à un patron dans la liste.

À ce stade, 48% d'instances de la catégorie *contient_transposition* correspondent à ces patrons. Par exemple :

methodologically → *de façon méthodologique* (ADV → ADP NOUN ADJ)

vision impairment → *l'altération visuelle* (NOUN NOUN → DET NOUN ADJ)

are increasing rapidly → *est en augmentation rapide*
(VERB VERB ADV → VERB ADP NOUN ADJ)

3. Analyse syntaxique

- L'analyse syntaxique en constituant est faite avec `Bonsai` (Candito *et al.*, 2010) pour le français, et `Stanford CoreNLP` pour l'anglais [5]. C'est un trait binaire et nous avons trois cas de figures :

  a) si c'est un couple de mots en traduction, nous comparons les étiquettes de PoS. Si elles sont identiques, le trait vaut 0, sinon 1.

  b) si c'est un couple de segments en traduction, nous comparons leur étiquette du nœud non terminal ;

  c) si c'est un mot traduit par un segment ou vice versa, nous comparons la catégorie des étiquettes (*ex.* si un adjectif est traduit par un syntagme verbal, le trait sera 1).

- L'analyse syntaxique en dépendance est faite avec `Stanford CoreNLP` pour les deux langues afin de partager le même jeu d'étiquettes. À l'intérieur des segments, nous comptons le nombre d'occurrences de chaque relation de dépendance (voir figure 6.1 et tableau 6.3).

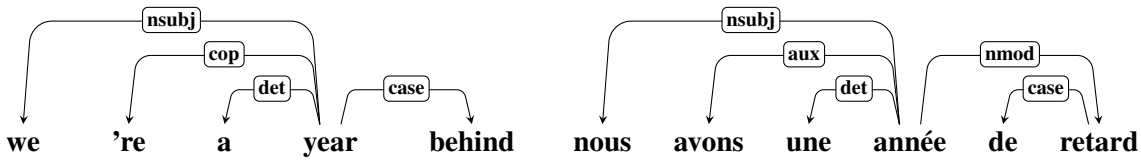

FIGURE 6.1 – Analyse en dépendance à l'intérieur du segment

	amod	det	nmod	case	...	nsubj
anglais	0	1	0	1	...	1
français	0	1	1	1	...	1

Tableau 6.3 – Comptage des étiquettes de relation de dépendance

5. Concernant l'analyse en constituant en français, `Bonsai` est beaucoup plus efficace que `Stanford CoreNLP` et a moins d'erreurs évidentes au niveau du nœud non terminal.

Dans la figure 6.2, le mot « *adapt* » est traduit par « *adaptation* », et à l'extérieur des segments, « *ability to* » et « *capacité d'* » sont en lien de dépendance avec les mots en question. Parmi les mots externes liés en dépendance, nous gardons seulement les mots alignés manuellement par les annotateurs. Ainsi nous comptons les relations de dépendance comme illustré dans la figure.

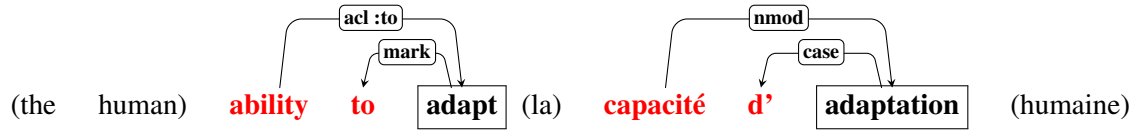

FIGURE 6.2 – Analyse en dépendance à l'extérieur du segment

4. Ressource externe

Nous exploitons ici la ressource multilingue ConceptNet [6], qui est un réseau sémantique en accès libre, conçu pour permettre aux machines d'accéder aux sens des mots. Cette ressource est constituée des connaissances obtenues par *crowdsourcing*, de la lexicographie, des jeux sérieux et des données ouvertes liées (*Linked Open Data*). Le système basé sur ConceptNet a remporté la première place dans la tâche "Similarité sémantique lexicale multilingue et cross-lingue" de SemEval2017 (Camacho-Collados *et al.*, 2017; Speer et Lowry-Duda, 2017).

- Pour calculer la similarité cosinus entre les plongements bilingues, nous utilisons ConceptNet Numberbatch (Speer *et al.*, 2017). Les plongements de cette ressource ont été construits en utilisant des données de ConceptNet, de Word2Vec (Mikolov *et al.*, 2013) et de Glove (Pennington *et al.*, 2014). Les représentations de l'espace vectoriel ont été affinées à l'aide d'informations relationnelles dans les lexiques sémantiques, avec une variante de *retrofitting* (Faruqui *et al.*, 2015). Cela permet que des mots liés dans des lexiques aient des représentations vectorielles similaires.

  Certaines expressions polylexicales ont leur propre plongement dans cette ressource. Si ce n'est pas le cas, nous calculons la moyenne des plongements uniquement sur les mots pleins. Les mêmes traits ont été calculés sur les segments lemmatisés.

- La ressource ConceptNet fournit aussi des assertions sous forme de triplet : un couple de mots ou expressions liés par une relation. [7] En nous basant sur ces assertions, nous avons un trait en entier qui vérifie si un couple anglais-français est :

  a) directement lié ;

  b) indirectement lié par un autre segment français (*complete* ← *complet / entier* → *total*) ;

  c) simplement pas lié. [8]

  Trois formes ont été testées : forme originale, forme lemmatisée et forme lemmatisée filtrée. [9] Par exemple :

  forme originale : *increasing rapidly* → *en augmentation rapide*

6. http://conceptnet.io/
7. https://github.com/commonsense/conceptnet5/wiki/Downloads
8. Les assertions anglais-français et français-français sont utilisées dans ce travail.
9. Nous filtrons les mots selon une liste manuelle, qui contient les verbes légers, déterminants, pronoms, etc.

forme lemmatisée : *increase rapidly* → *en augmentation rapide*

forme lemmatisée filtrée : *increase rapidly* → *augmentation rapide*

- Sur la forme lemmatisée filtrée, nous calculons le nombre de tokens bilingues qui sont liés indirectement, puis le divisons par le nombre total des tokens dans les deux langues. Par exemple « *deceptive* » et « *illusion* » ne sont pas directement liés dans la ressource, mais tous les deux sont liés à « *illusoire* ». Ainsi ces deux mots sont liés indirectement. Ce genre de situation apparaît souvent dans une paire traduite par le procédé *Transposition*.

5. Alignement de mots automatique

Pour ce groupe de traits, nous avons exploité le tableau de probabilité de traduction lexicale, fourni par l'outil d'alignement de mots `Berkeley Word Aligner` (Liang *et al.*, 2006). Cet outil statistique a été entraîné sur un corpus parallèle anglais-français combiné de notre corpus *TED Talks* entier (163 092 lignes) et d'une partie du corpus `Paracrawl`* [10], pour un total de 1.8M de couples de phrases et 41M de tokens anglais.

- L'entropie des distributions de probabilités de traduction lexicale est calculée selon l'équation ci-dessous (Gray, 1990; Carl et Schaeffer, 2017). Ces valeurs sont directement fournies par `Berkeley Word Aligner` :

$$H(X) = \sum_i P(x_i)I(x_i) = -\sum_i P(x_i)log_e P(x_i) \tag{6.1}$$

  Nous calculons l'entropie moyenne sur les mots pleins. Une entropie plus grande indique que les mots possèdent des sens plus généraux ou qu'ils sont polysémiques. Le même trait est calculé sur les mots pleins lemmatisés.

- Nous calculons la pondération lexicale bidirectionnelle sur les mots pleins, selon l'équation proposée par Koehn *et al.* (2003). Puisque nous avons aligné les traductions non littérales bloc par bloc (c'est-à-dire à un niveau plus large que celui des mots), nous avons supposé un alignement de mots *n-m* entre les deux segments (e et f), à savoir chaque mot source est aligné avec tous les mots cibles. Et l'ensemble des alignements constitue A :

$$lex(e|f, A) = \prod_{i=1}^{length(e)} \frac{1}{|\{j|(i,j) \in A\}|} \sum_{\forall(i,j)\in A} w(e_i|f_j) \tag{6.2}$$

  Pour calculer la pondération lexicale directe, chacun des mots anglais e_i est généré par des mots étrangers alignés f_j avec la probabilité de traduction lexicale $w(e_i|f_j)$. Et de même pour la pondération lexicale inverse $lex(f|e, A)$. Les mêmes traits ont été calculés sur les mots pleins lemmatisés. Plus la valeur est grande, plus la confiance dans l'alignement au niveau du mot entre les deux segments est forte.

- Nous calculons la somme de différence de probabilités de traduction lexicale entre la traduction la plus probable (selon le tableau de probabilité) et la traduction humaine dans le corpus. Pour chaque mot source, nous prenons le mot cible dans la traduction humaine ayant la plus grande probabilité.

  Par exemple pour la paire *« alternatives → solutions de remplacement »*, la traduction la plus littérale est « *alternatives* » avec une probabilité de 0,4. Or dans

10. `https://paracrawl.eu/index.html`

la traduction humaine, le mot « *solutions* » possède la plus grande probabilité, mais qui est seulement 0,07. Nous additionnons ainsi cette différence de probabilité pour chaque mot source. Selon cette méthode, nous comptons aussi les mots cibles n'ayant pas la plus grande probabilité comme des mots non alignés, afin de calculer un ratio de ces mots considérés comme non alignés sur le nombre total de tokens de chaque côté. Ces traits ont été calculés dans les deux directions de traduction.

6.3.1 Résultats expérimentaux et analyse

Plan expérimental Le nombre d'instances de *non_littéral* (1127) est seulement un tiers de celui de *littéral* (3771). Compte tenu de cet écart, nous avons évalué les classifieurs sous plusieurs configurations :

(a) six classes (*littéral, équivalence, généralisation, particularisation, modulation, contient_transposition*), où *littéral* contient soit toutes les instances, soit 200 instances pour avoir une distribution approximativement équilibrée ; [11]

(b) deux classes (*littéral* et *non_littéral*), avec trois répartitions (3 :1, 2 :1, 1 :1) ;

(c) cinq classes : seulement les catégories non littérales.

Les classifieurs suivants ont été entraînés : *RandomForest, Multi-Layer Perceptron, Logistic Regression, Support Vector Machine, K-nearest Neighbors, Decision Tree, Bernoulli Naive Bayes, Multinomial Naive Bayes* et *Gaussian Naive Bayes*. Pour chaque configuration, nous avons optimisé les hyperparamètres de ces classifieurs [12]. L'évaluation est menée par une validation croisée à cinq [13] plis (en utilisant *StratifiedKFold*, où les plis ont été construits en préservant le pourcentage des instances de chaque classe). Les mesures utilisées contiennent l'exactitude (*accuracy*) moyenne, la F-mesure micro-moyenne et macro-moyenne (Tsoumakas *et al.*, 2011). Les résultats sous différentes configurations sont présentés dans le tableau 6.4, où le classifieur *Dummy* est utilisé comme une baseline. Il prédit toujours la classe majoritaire. Pour toutes les configurations, le classifieur *RandomForest* obtient toujours la meilleure performance [14].

Discussion sur les résultats Nous effectuons d'abord une classification directe en six classes. Les résultats de notre classifieur dépassent largement ceux du classifieur *Dummy*. En revanche, la difficulté de la tâche en multi-classes est aussi reflétée dans la distribution approximativement équilibrée (le nombre de classes est élevé mais le nombre d'instances par classe est limité). Ainsi étudions une classification binaire (littéral *versus* non littéral) et indépendemment une classification entre les cinq classes non littérales.

Pour la classification binaire, les deux meilleurs classifieurs sont *RandomForest* et *Multilayer Perceptron*. En plus, *RandomForest* est meilleur que les deux assemblés par la méthode *hard voting* ou *soft voting*. Sous la distribution équilibrée, l'exactitude la plus élevée atteint 87,09%. De la distribution naturelle (3 :1) à la distribution équilibrée (1 :1), la F-mesure moyenne pour la classe *non_littéral* augmente de 0,78 à 0,88. Une analyse des erreurs sur la distribution 3 :1 montre que parmi les 290 instances *non_littéral* classifiées en *littéral*, 117 sont de classe *équivalence*. Cela indique que ces deux classes sont difficiles à distinguer pour le classifieur (*ex. search history* → *historique de recherche* : annoté comme *littéral*, classifié comme *équivalence*).

11. Des instances de *littéral* ont été extraites au hasard pour les configurations a et b.
12. Pour trouver les meilleurs hyperparamètres, 10% de données sont séparées comme test, et une validation croisée à trois plis est exécutée sur 90% de données d'entraînement.
13. Le nombre de 5 est choisi de façon empirique.
14. Les hyperparamètres détaillés sont donnés ensemble avec le code.

Distribution de classes	Classifieur	Exactitude moyenne	Micro-F1	Macro-F1
Six classes				
six classes, avec 3771 *littéral*	Dummy	76,99%	0,77	0,14
	RandomForest	**83,10%** ± 0,35%	**0,83**	**0,44**
six classes, avec 200 *littéral*	Dummy	25,77%	0,26	0,07
	RandomForest	57,04% ± 1,47%	0,57	0,52
Deux classes				
littéral (3) : *non_littéral* (1)	Dummy	76,99%	0,77	0,43
	RandomForest	**90,16%** ± 0,98%	**0,90**	**0,86**
littéral (2) : *non_littéral* (1)	Dummy	66,67%	0,67	0,40
	RandomForest	88,85% ± 0,71%	0,89	0,88
littéral (1) : *non_littéral* (1)	Dummy	50,00%	0,50	0,33
	RandomForest	87,09% ± 2,50%	0,87	0,87
Cinq classes				
cinq classes non_littéral	Dummy	30,35%	0,30	0,09
	RandomForest	55,10% ± 1,45%	0,55	0,47

Tableau 6.4 – Résultats sous différentes configurations, utilisant tous les traits. Les écart-types ont été calculés sur chaque pli de la validation croisée

En utilisant tous les traits, l'exactitude la plus élevée pour la classification entre les cinq classes non littérales est de 55,10%. Des F-mesures moyennes sur les cinq plis pour chaque classe sont présentées dans le tableau 6.5. La catégorie *généralisation* contient beaucoup moins d'instances que les autres catégories, qui nécessite une augmentation ; il existe beaucoup de confusion entre *modulation* et les autres catégories, ce qui suggère une amélioration du guide d'annotation ; la confusion existe aussi entre *équivalence* et *contient_transposition* (*ex. all the people in the world → la population mondiale*).

équivalence	généralisation	particularisation	modulation	contient-transposition
0,51 ± 0,02	0,25 ± 0,09	0,56 ± 0,05	0,36 ± 0,08	0,68 ± 0,02

Tableau 6.5 – F-mesures moyennes sur les cinq plis pour chaque procédé non littéral

Le tableau 6.6 montre les performances des autres classifieurs que nous avons entraînés, sur le jeu de données équilibré en deux classes. Sous les autres configurations, *RandomForest* est aussi toujours le classifieur le plus performant.

Contribution de traits Avec le meilleur classifieur *RandomForest*, nous avons effectué une étude sur la contribution de traits : un par un et aussi en groupe. Pour la classification binaire (distribution équilibrée, voir tableau 6.7), le trait « *pondération lexicale bidirectionnelle* » contribue le plus, permettant l'obtention d'une F1 moyenne de 0,78 pour la classe *littéral* et de 0,80 pour la classe *non_littéral* par lui seul. Concernant les groupes de traits, le groupe « *alignement de mots* » est celui qui contribue le plus. Nous observons que la combinaison de tous les traits produit quasiment les meilleurs résultats. Les meilleurs résultats sont obtenus en supprimant les traits « *analyse syntaxique en constituant* », « *analyse en dépendance à l'extérieur des segments* » et « *l'entropie de probabilités de traduction lexicale* ».

Pour la classification en cinq classes (voir tableau 6.8), la combinaison de tous les traits génère les meilleurs résultats, où les groupes « *analyse morpho-syntaxique* » et

Algorithme	Exactitude Moyenne	F1 (Littéral)	F1 (Non littéral)
RandomForest	**87,09%**	**0,87**	**0,87**
MLP	85,01%	0,85	0,85
rbfSVC	85,14%	0,84	0,86
LogisticRegression	84,78%	0,85	0,85
KNN	83,41%	0,83	0,83
BernoulliNB	81,14%	0,80	0,82
DecisionTree	79,95%	0,79	0,81
MultinomialNB	80,83%	0,81	0,81
GaussianNB	64,51%	0,73	0,50

Tableau 6.6 – Classification binaire (distribution équilibrée), utilisant tous les traits

« *analyse syntaxique* » contribuent plus que les autres groupes.

Trait	F1 moyenne (littéral)	F1 moyenne (non littéral)
distance_Levenshtein	0,75	0,75
(ratio)_longueur_token	0,74	0,71
pos_vecteur_comptage	**0,78**	0,75
posCosinus_tous_les_mots	0,69	0,69
posCosinus_mots_pleins	0,68	0,64
pos_changement_patron	0,70	0,34
analyse_constituant	0,66	0,50
analyse_dépendance_interne	0,76	0,73
analyse_dépendance_externe	0,61	0,62
ConceptNet_Embedding	0,73	0,73
ConceptNet_lien	0,70	0,78
ConceptNet_pourcentage_indirect	0,70	0,62
différence_probabilité_traduction	0,77	0,78
entropie_traduction	0,60	0,62
pondération_lexicale	**0,78**	**0,80**
surface	0,72	0,70
analyse_PoS	0,78	0,75
analyse_syntaxique	0,76	0,76
ressource_ConceptNet	0,77	0,78
alignement_de_mots	0,84	0,85
tous les traits	**0,87**	0,87
tous les traits - les traits en vert	**0,87**	**0,88**

Tableau 6.7 – Contribution de traits pour la classification binaire (distribution équilibrée)

Combinaison de classes Dans la classification binaire, notre analyse d'erreur montre que la distinction entre *littéral* et *équivalence* est difficile ; dans la classification en multi-classes, la plus grande confusion existe entre *équivalence* et *contient_transposition*. Par conséquent, nous avons mené trois autres expériences de classification binaire (voir tableau 6.9), où dans toutes les configurations chaque classe contient 549 instances pour rendre les résultats comparables :

(a) *littéral* (L) versus *non_littéral* (NL)

(b) *littéral* combiné avec *équivalence* (E), versus les autres classes

(c) *littéral* combiné avec *équivalence* (E) et *transposition* (T), versus les autres classes

Trait	Micro-F1	Macro-F1	F1 (E)	F1 (G)	F1 (P)	F1 (M)	F1 (T)
distance_Levenshtein	0,36	0,26	0,32	0,00	0,31	0,18	0,49
(ratio)_longueur_token	0,39	0,34	0,36	0,21	0,39	0,26	0,49
pos_vecteur_comptage	0,52	0,44	**0,51**	0,16	0,54	0,34	0,65
posCosinus_tous_les_mots	0,39	0,29	0,25	0,00	0,45	0,23	0,54
posCosinus_mots_pleins	0,39	0,25	0,06	0,00	0,42	0,19	0,58
pos_changement_patron	0,37	0,20	0,43	0,00	0,00	0,00	0,56
analyse_constituant	0,36	0,19	0,00	0,00	0,41	0,00	0,56
analyse_dépendance_interne	0,45	0,39	0,37	0,21	0,47	0,31	0,59
analyse_dépendance_externe	0,35	0,26	0,27	0,05	0,33	0,15	0,51
ConceptNet_Embedding	0,32	0,27	0,28	0,02	0,38	0,25	0,40
ConceptNet_lien	0,32	0,18	0,13	0,00	0,30	0,00	0,46
ConceptNet_pourcentage_indirect	0,29	0,16	0,22	0,00	0,00	0,18	0,42
différence_probabilité_traduction	0,38	0,32	0,35	0,13	0,38	0,26	0,50
entropie_traduction	0,35	0,27	0,36	0,00	0,39	0,15	0,44
pondération_lexicale	0,32	0,24	0,36	0,02	0,28	0,15	0,41
surface	0,38	0,34	0,36	0,23	0,37	0,27	0,47
analyse_PoS	0,51	0,43	0,48	0,15	0,52	0,34	0,64
analyse_syntaxique	0,48	0,40	0,38	0,18	0,52	0,30	0,63
ressource_ConceptNet	0,34	0,28	0,24	0,02	0,46	0,23	0,44
alignement_de_mots	0,45	0,38	0,44	0,15	0,51	0,26	0,54
pos + surface + syntaxique	0,54	0,47	0,49	**0,28**	0,54	**0,38**	0,67
tous - ConceptNet	0,54	0,47	**0,51**	0,27	0,54	0,37	0,67
tous les traits	**0,55**	0,47	0,50	0,25	**0,55**	0,37	**0,67**

Tableau 6.8 – Contribution de traits pour la classification entre les cinq classes non littéral. E (équivalence), G (généralisation), P (particularisation), M (modulation), T (contient_transposition)

La troisième configuration est plus intéressante, parce que le groupe de procédés *LET* ne cause pas de changement de sens, alors que le groupe *non-LET* le peut. Après l'inclusion de *transposition* (changer les classes grammaticales sans en changer le sens), les résultats deviennent meilleurs que ceux obtenus lorsque l'on regroupe seulement *littéral* et *équivalence*, puisque nous évitons la confusion entre *équivalence* et *transposition*. Cette observation nous motive à développer des classifieurs en cascade dans de futurs travaux, en considérant d'abord de distinguer les traductions littérales mot-à-mot, ou celles qui ne causent pas de changement de sens, et ensuite mener une classification à grain fin parmi les autres catégories.

Configuration	Exactitude moyenne	F-mesure moyenne (class1)	F-mesure moyenne (class2)
Dummy	48,63%	0,49	0,49
L vs NL	**85,24%**	0,84	0,86
LE vs non-LE	75,32%	0,74	0,77
LET vs non-LET	**79,42%**	0,78	0,81

Tableau 6.9 – Résultats de classification après le regroupement de classes, chaque classe contient 549 instances

6.4 Classifieurs en réseaux neuronaux et résultats

En plus des classifieurs statistiques avec l'ingénierie des traits, nous avons également développé des classifieurs à base de réseaux neuronaux pour comparer les perfor-

mances. [15]

Les segments en anglais et en français sont encodés par des unités de GRU *(Gated Recurrent Unit)* bidirectionnelles (taille 10). [16] Les sorties des réseaux récurrents en avant (dit *forward* dans la suite) et en arrière (*backward*) sont concaténées pour former les représentations de segment source et cible (taille 20).

Après la couche d'encodage, nous avons essayé deux architectures différentes :

1) Nous construisons une matrice d'alignement pour une paire de segments, utilisant le produit scalaire de leur représentation. Cette démarche s'inspire des travaux de Legrand *et al.* (2016) et de Pham *et al.* (2018). Un classifieur CNN (*Convolutional Neural Network*) est ensuite appliqué à cette matrice d'alignement. Il est composé d'une couche de convolution, suivie par une couche de « *pooling* ». Puisque la taille des matrices d'alignement varie d'une paire de segments à une autre, un pooling adaptatif est utilisé (He *et al.*, 2015). La sortie de la couche pooling est envoyée à une couche « *fully connected* », qui est suivie par une couche linéaire qui génère la sortie finale (voir figure 6.3).

2) Pour la sortie source et cible de la couche d'encodage, une moyenne est calculée au fil de chaque étape de temps, ce qui génère deux vecteurs de dimension fixe. Ils sont ensuite concaténés (taille 40) et envoyés à un classifieur MLP (*Multi-Layer Perceptron*). La couche cachée de MLP a 10 nœuds cachés et la fonction d'activation est *tanh* (voir figure 6.4).

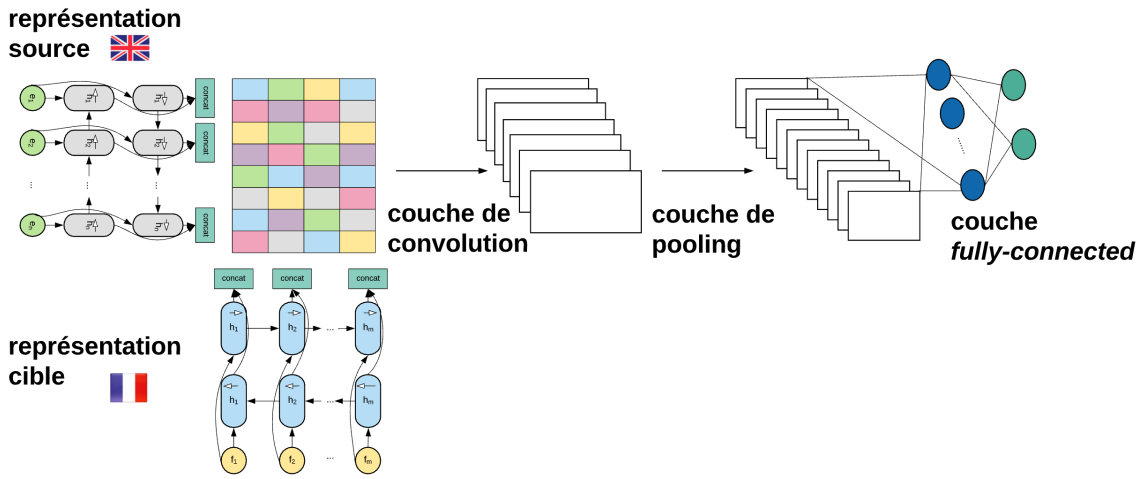

FIGURE 6.3 – Première architecture : matrice d'alignement de mots + classifieur CNN

Les phrases sont souvent courtes, surtout pour les instances de traductions mot-à-mot. En vue de construire une matrice d'alignement de mots plus robuste et d'éviter le problème de mots hors vocabulaire, nous avons choisi des plongements de caractères. Comme le tableau 6.10 le montre, nous avons essayé des plongements de caractères initialisés aléatoirement (taille 10), et aussi entraîné nos propres plongements lexicaux avec le modèle *skipgram* de `FastText` (Bojanowski *et al.*, 2017) sur notre corpus de *TED Talks* (le corpus anglais et français contiennent chacun environ 3M de tokens). La dimension

15. Le travail présenté dans la section 6.4 a été réalisé avec la collaboration de Pooyan Safari, doctorant dans le groupe TLP au LIMSI.

16. Nous avons aussi testé l'encodage par LSTM (*Long-Short Term Memory*), mais les résultats ne sont pas différents de façon significative. Par exemple dans le tableau 6.10, l'architecture qui utilise les plongements de FastText et le classifieur MLP donne une exactitude moyenne de 71,47%. Chung *et al.* (2014) ont montré que GRU et LSTM sont comparables en performance. Puisque GRU a besoin d'entraîner moins de paramètres que LSTM, nous l'avons utilisé pour les expériences par la suite.

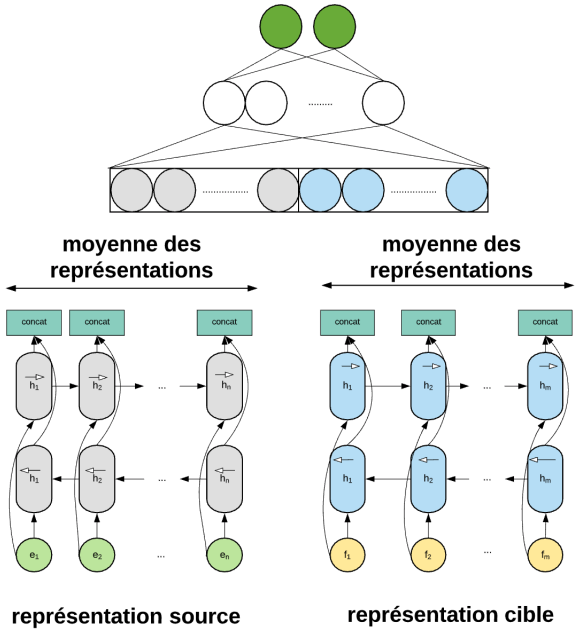

FIGURE 6.4 – Deuxième architecture : représentation en moyenne + classifieur MLP

de ces plongements est de 100. Pour les paramètres de `FastText`, la taille minimum de n-gram est de 3, le maximum est de 6.

Tous les modèles ont été entraînés avec 200 *epochs* [17], avec un taux d'apprentissage de 0.0001. Nous avons utilisé *Adam* pour l'optimisation. La taille de *minibatch* est de 20, et le *dropout* a été appliqué à toutes les couches sauf la couche de plongement et de sortie.

Les tableaux 6.10 et 6.11 montrent les résultats des classifieurs en réseaux neuronaux de bout en bout, pour la classification binaire (distribution équilibrée) et la classification parmi les cinq classes non littéral. À part les pré-traitements mentionnés au début, pour des classifieurs neuronaux, nous avons aussi normalisé des formes en contraction en mots complets (*ex. 're → are*), et des chiffres en forme de lettre (*ex. 42 → quatre deux* (pas *quarante-deux*)). L'architecture utilisant des plongements lexicaux et MLP obtient de meilleurs résultats, et est plus efficace au niveau du temps que les deux autres architectures. Pourtant, le jeu de données actuel est trop petit pour que les classifieurs neuronaux produisent des résultats satisfaisants (les résultats obtenus sont plus bas que ceux obtenus par le classifieur *RandomForest*, à savoir 87,09% et 55,10%).

Architecture	Exactitude	F1 (L)	F1 (NL)
Plongement de caractère initialisé aléatoirement			
CNN	59,99%	0,60	0,60
MLP	71,16%	0,71	0,71
Plongement de mot pré-entraîné (FastText)			
MLP	**71,25%**	0,71	0,71

Tableau 6.10 – Classification binaire (distribution équilibrée)

17. Un *epoch* désigne une itération de l'algorithme d'apprentissage sur l'ensemble de données.

Architecture	Exactitude	Micro-F1	Macro-F1
Plongement de caractère initialisé aléatoirement			
CNN	34,08%	0,34	0,20
MLP	40,74%	0,41	0,34
Plongement de mot pré-entraîné (FastText)			
MLP	**43,22%**	0,43	0,34

Tableau 6.11 – Classification en multi-classes (cinq classes non littéral)

6.5 Classification sensible au contexte

Jusqu'à présent, nous avons obtenu des résultats encourageants sur la classification binaire (*littéral* versus *non_littéral*). Dans la suite de ce travail, nous nous focalisons sur la classification des procédés non littéraux. Dans notre travail, l'information contextuelle n'a pas été prise en compte jusqu'à maintenant. En revanche, certaines traductions non littérales ne peuvent être interprétées qu'en contexte. Nous avons voulu exploiter des indices utiles fournis par le contexte pour reconnaître le procédé utilisé.

En vue de représenter l'information contextuelle de façon appropriée, nous avons d'abord travaillé sur une tâche monolingue comparable à la nôtre. Nous avons vérifié la reproductibilité de deux systèmes de classification sur un jeu de données anglais annoté en relation sémantique à grain fin et en contexte (Shwartz et Dagan, 2016; Vyas et Carpuat, 2017). Avec de nouveaux traits proposés venant des représentations d'ELMo (*Embeddings from Language Models*) (Peters *et al.*, 2018) et de Context2Vec (Melamud *et al.*, 2016), nous avons légèrement amélioré des résultats de l'état-de-l'art sur cette tâche monolingue.

Par conséquent, nous étudions la pertinence de l'information contextuelle pour notre tâche cross-lingue, via l'adaptation des traits sensibles au contexte dont nous avons prouvé l'utilité dans la tâche monolingue. Basé sur notre jeu de données annoté manuellement, nous comparons ce nouveau système avec le système précédent qui a seulement utilisé des traits indépendants du contexte. Nous menons l'analyse linguistique pour expliquer les problèmes rencontrés et proposons de futurs travaux pour l'amélioration des résultats.

6.5.1 Inférence lexicale monolingue sensible au contexte

Ayant des difficultés pour trouver des travaux précédents similaires à notre tâche, nous avons d'abord étudié une tâche comparable sur l'inférence lexicale monolingue, en vue de développer des méthodes de l'état-de-l'art sur la représentation du contexte.

Définition de la tâche

MacCartney et Manning (2007) ont proposé une typologie de relations sémantiques à grain fin (*ex.* équivalence, exclusion, implication textuelle, etc.). En vue d'entraîner un classifieur qui peut prédire automatiquement ces relations sur des paires de segments dans la ressource de paraphrases PPDB, un sous-ensemble de paires de segments sans contexte a été annoté manuellement, et le jeu de données est nommé PPDB-fine-human (Pavlick *et al.*, 2015a) (nous avons présenté ce travail dans le chapitre 3).

Basé sur ce jeu de données, Shwartz et Dagan (2016) ont proposé le premier jeu de données d'inférence lexicale *en contexte* (Context-PPDB-fine-human), par une réannotation de relations sémantiques à grain fin dans deux contextes de phrase différents. Étant donné une paire de mots et leur contexte différent (w_x, C_x, w_y, C_y), les relations

sémantiques annotées sont *Équivalence, Implication* [18]*, Exclusion Mutuelle, Lié autrement* et *Indépendant*. Le jeu de données d'apprentissage contient 2 550 paires de mots en contexte, et le jeu de test contient 890 paires.

Systèmes existants

Shwartz et Dagan (2016) ont utilisé tous les traits disponibles et indépendants du contexte dans PPDB 2.0 (Pavlick *et al.*, 2015b), qui contiennent par exemple la probabilité prédite pour chaque relation sémantique ; les similarités distributionnelles ; le score prédit de paraphrase ; le score de classement dans PPDB 2.0, etc. Pour que le système soit sensible au contexte, ils ont utilisé des plongements lexicaux de Glove (*Global Vectors for word representation*) (Pennington *et al.*, 2014) pour représenter les mots, et calculé les traits suivants pour représenter les similarités les plus importantes entre les mots et les contextes. [19]

$$\{\max_{w \in C_y} \overrightarrow{w_x} \cdot \overrightarrow{w}, \max_{w \in C_x} \overrightarrow{w_y} \cdot \overrightarrow{w}, \max_{w_x \in C_x, w_y \in C_y} \overrightarrow{w_x} \cdot \overrightarrow{w_y}\} \tag{6.3}$$

Vyas et Carpuat (2017) ont exploité les représentations contextualisées pour représenter les sens des mots en contexte. Ils ont amélioré le système de Shwartz et Dagan (2016) avec des représentations masquées, basées sur les plongements lexicaux de Glove. En suivant le travail de Tang *et al.* (2014), pour chaque phrase de contexte, ils génèrent d'abord une matrice contenant dans chaque ligne des plongements lexicaux de Glove des mots dans cette phrase. Ensuite pour chaque colonne, la valeur maximum, minimum et moyenne est calculée, qui produit trois vecteurs de dimension d ($\overrightarrow{C}_{max}, \overrightarrow{C}_{min}, \overrightarrow{C}_{mean}$).

Les nouvelles représentations masquées sensibles au contexte sont le produit élément par élément du plongement de Glove $\overrightarrow{w_x}$ (indépendant du contexte) avec chacun de ces vecteurs de dimension d. Ils répètent le même calcul pour $\overrightarrow{w_y}$. Ils utilisent seulement 6 parmi 43 traits dans la ressource PPDB, à savoir les probabilités prédites pour chacune relation sémantique dans PPDB.

Pour ces deux systèmes, nous avons demandé le code source aux deux premiers auteurs pour assurer la reproductibilité et être clair sur les détails techniques. Pendant l'inspection de leur code, nous avons trouvé un bug lors de la lecture du jeu de données. En fait, afin de garder seulement une catégorie pour *Implication*, ils ont inversé des paires de mots pour des instances de classe *implication_en_arrière*, pour ne garder que la classe de l'implication en avant. En revanche, les phrases de contexte n'ont pas été inversées. Cette erreur conduit à des similarités mots/contextes plus grandes pour ces instances. Ce problème a été confirmé avec les auteurs, et les résultats après correction sont montrés dans le tableau 6.12. [20]

Ajouter des traits dérivés d'ELMo et de Context2Vec

Pour la tâche monolingue sur le jeu de données Context-PPDB-fine-human, notre contribution est de fournir de nouveaux traits sensibles au contexte dérivés de représentations en utilisant ELMo (Peters *et al.*, 2018) et Context2Vec (Melamud *et al.*, 2016). Différent des plongements lexicaux de Glove qui sont indépendants du contexte, ELMo attribue à chaque token dans une phrase des représentations dépendantes au contexte qui

18. L'implication en avant et en arrière sont combinées.
19. Ils ont aussi calculé les produits entre chaque deux traits parmi ces trois.
20. Les résultats déclarés dans leur article était de 0,67 et 0,72, respectivement. La différence est due à une performance plus basse sur la catégorie *Implication*.

sont une fonction de la phrase entière. Les représentations sont dérivées d'un LSTM bidirectionnel qui est entraîné avec un objectif de modèle de langue sur un corpus de référence contenant 1 milliard de mots (Chelba *et al.*, 2013). Des améliorations nettes ont été montrées après l'application d'ELMo dans une grande variété de tâches en TAL (Sheikhshabbafghi *et al.*, 2018; Peters *et al.*, 2018; Che *et al.*, 2018). Concernant Context2Vec, avec un objectif que le contexte puisse prédire le mot manquant via un modèle log-linéaire, le modèle neuronal projette le mot en question et sa phrase de contexte dans un même espace en utilisant un LSTM bidirectionnel.

Nous utilisons la librairie `allennlp` (Gardner *et al.*, 2018) pour générer les plongements lexicaux d'ELMo pour chaque phrase, et récupérons le plongement du mot cible selon sa position dans la phrase. Le modèle pré-entraîné le plus grand a été utilisé, et nous avons essayé la dernière couche ainsi qu'une moyenne de toutes les couches comme les représentations vectorielles d'ELMo. Pour Context2Vec, nous utilisons le modèle anglais pré-entraîné sur le corpus ukWaC* (Ferraresi *et al.*, 2008). Pour générer des représentations sur une phrase de contexte, le terme en question a été remplacé par un blanc.

Nous présentons ci-dessous les traits que nous proposons :

1) Nous remplaçons les vecteurs de Glove par ceux d'ELMo. Après avoir filtré les mots outils, nous calculons les traits de la similarité mot/contexte selon la proposition de Shwartz et Dagan (2016).

2) Suivant le travail de Vyas et Carpuat (2017), nous générons des représentations masquées des mots, basé sur Context2Vec :

$$\{\overrightarrow{w}_{x,c2v} \odot \overrightarrow{C}_{x,c2v}, \overrightarrow{w}_{y,c2v} \odot \overrightarrow{C}_{y,c2v}\} \tag{6.4}$$

3) Dans l'idée de comparer la similarité entre un mot cible et le contexte d'une autre phrase, nous calculons le produit scalaire entre leurs plongements, ainsi que le produit scalaire entre les plongements de deux phrases de contexte, qui sont tous générés par Context2Vec :

$$\{\overrightarrow{w}_{x,c2v} \cdot \overrightarrow{C}_{y,c2v}, \overrightarrow{w}_{y,c2v} \cdot \overrightarrow{C}_{x,c2v}, \overrightarrow{C}_{x,c2v} \cdot \overrightarrow{C}_{y,c2v}\} \tag{6.5}$$

4) Comme le trait précédent, mais le plongement lexical du mot cible généré par Context2Vec est remplacé par le plongement d'ELMo.

$$\{\overrightarrow{w}_{x,elmo} \cdot \overrightarrow{C}_{y,c2v}, \overrightarrow{w}_{y,elmo} \cdot \overrightarrow{C}_{x,c2v}\} \tag{6.6}$$

Les résultats et traits utilisés par ces trois systèmes sont résumés dans le tableau 6.12. Dans notre expérience, les vecteurs de la dernière couche d'ELMo génèrent des résultats légèrement meilleurs que les vecteurs issus d'une moyenne de trois couches. Tous les systèmes utilisent le classifieur *Logistic Regression* avec les hyper-paramètres par défaut. Nous obtenons un score légèrement meilleur que Vyas et Carpuat (2017). Il resterait cependant encore des traits efficaces à tester pour tenter d'améliorer la performance.

6.5.2 Classification des procédés de traduction sensible au contexte

Ayant validé que nos méthodes de représentation contextuelle génèrent des résultats au niveau de l'état-de-l'art, nous revenons à notre tâche de classification des procédés de traduction sensible au contexte. Nous montrerons comment nous adaptons et transférons des traits pertinents d'une tâche monolingue à une tâche cross-lingue.

Système	Traits	F1 moyenne pondérée
Shwartz et Dagan (2016)	43 traits dans PPDB, 6 traits de similarité mot/contexte utilisant les plongements de Glove (300d). Les mots outils ont été filtrés.	0,600
Vyas et Carpuat (2017)	6 traits dans PPDB, 6 traits de similarité mot/contexte utilisant les plongements de Glove (50d), les représentations masquées en basant sur GloVe. Les mots outils n'ont pas été filtrés.	0,672
Notre méthode	Traits utilisés par Vyas et Carpuat (2017), plus quatre traits que nous avons proposés.	0,684

Tableau 6.12 – Comparaison de systèmes sur le jeu de données Context-PPDB-fine-human

Définition de la tâche

Nous utilisons toujours le jeu de données annoté manuellement sur le corpus de *TED Talks* (référé comme TED-translation-process dans la suite). Les phrases de contexte bilingues sont gardées pour extraire des traits sensibles au contexte. Dans cette tâche, nous gardons les catégories *Équivalence, Transposition, Généralisation, Particularisation, Modulation* et *Modulation + Transposition*. Un exemple pour chaque catégorie est présenté dans le tableau 6.13, le contexte de phrase est ajouté pour illustrer l'importance du contexte pour interpréter certaines traductions.

Nous combinons ces procédés de traduction à grain fin en trois catégories selon le degré de glissement de sens : *Équivalence* (équivalence, transposition), *Implication textuelle* (généralisation, particularisation), et *Lié en thématique* (modulation, modulation + transposition). En vue d'utiliser le modèle de Context2Vec pour générer des représentations sur les phrases de contexte, nous ne gardons pas de segments discontinus. Le nombre d'instances pour chaque catégorie est présenté dans le tableau 6.14. La problématique de recherche principale est d'étudier si l'ajout des traits sensibles au contexte améliore la performance de classification.

Des traits indépendants du contexte

Nous réutilisons des traits décrits dans la section 6.3. Concernant des plongements lexicaux bilingues, nous avons aussi mené des expériences avec MUSE (*Multilingual Unsupervised and Supervised Embeddings*)[21], dont les plongements multilingues ont été entraînés sur le corpus de Wikipédia en utilisant `fastText` (Bojanowski *et al.*, 2017), et alignés de façon supervisée dans un même espace vectoriel (Conneau *et al.*, 2017). Une différence importante par rapport aux plongements multilingues de ConceptNet Numberbatch que nous avons utilisé auparavant, est qu'il n'existe pas de plongements pour des expressions polylexicales dans MUSE.

Des traits sensibles au contexte

Il existe plusieurs différences entre notre tâche cross-lingue et la tâche monolingue proposée par Shwartz et Dagan (2016). Les phrases de contexte sont totalement différentes dans leur jeu de données, mais elles sont en relation de traduction dans le nôtre. Par conséquent, les similarités les plus importantes entre les mots et les contextes ne sont pas utiles dans notre cas (voir les traits 6.3 proposés par Shwartz et Dagan (2016)). Nos couples de traduction à reconnaître contiennent des mots et des segments, alors que le jeu

21. `https://github.com/facebookresearch/MUSE`

Couple anglais-français	Procédé de traduction	Contexte
every one of us → chaque personne	équivalence	There are now three people on the planet for **every one of us** that existed in 1946 ; within 40 years, there'll be four. Il y a maintenant sur cette planète trois personnes pour **chaque personne** qui vivait en 1946 ; d'ici 40 ans, il y en aura quatre.
stamp a letter into it → avec une lettre en creux	transposition	But if you change the form that you give the placebo in, like you make a smaller pill, and **stamp a letter into it**, it is actually measurably more effective. Si vous changez la forme du placebo, par exemple un comprimé plus petit, **avec une lettre en creux**, c'est en fait plus efficace de façon mesurable.
rapid in its melting → touchée	généralisation	And west Antarctica cropped up on top some under-sea islands, is particularly **rapid in its melting**. La région ouest de l'Antarctique, juchée sur des îles sous-marines, est particulièrement **touchée**.
reaches → se jette dans	particularisation	If you want to know how sea level rises from land-base ice melting this is where it **reaches** the sea. Si vous voulez voir comment le niveau de la mer monte à cause de la fonte des glaces terrestres voilà l'endroit où la rivière **se jette dans** la mer.
drive → alimenter	modulation	We're looking to see if we can take captured CO_2, which can easily be piped to sites, convert that CO_2 back into fuel to **drive** this process. Nous essayons de voir si nous pouvons récupérer du CO_2 capté, qu'on peut facilement transporter sur place par tuyaux, et reconvertir ce CO_2 en carburant pour **alimenter** ce processus.
became the basis for → ont inspiré	modulation + transposition	Steve's columns **became the basis for** a book, which was turned into a movie. Les chroniques de Steve **ont inspiré** un livre, qui a été adapté à l'écran.

Tableau 6.13 – Exemples de procédé de traduction non littéral dans le jeu de données TED-translation-process

de données anglais contient seulement des couples de mots. Nous devons également réaliser quelques adaptations pour transférer des traits sensibles au contexte prouvés utiles

Catégorie	Nombre d'instances
Équivalence	710
Implication textuelle	384
Lié en thématique	305

Tableau 6.14 – Nombre d'instances de chaque catégorie dans TED-translation-process

dans la tâche monolingue.

Premièrement, pour pouvoir utiliser des représentations d'ELMo et de Context2Vec dans cette tâche cross-lingue, nous avons utilisé le modèle pré-entraîné pour le français fourni par Che *et al.* (2018)[22]. Nous avons aussi entraîné un modèle pour le français de Context2Vec sur un corpus interne issu des journaux en ligne[23] (32.9M de phrases, 0.8 milliard de tokens, avec un GPU).

Ensuite, nous avons proposé ces traits sensibles au contexte :

1) Des plongements lexicaux d'ELMo anglais et français[24], ou la moyenne des plongements pour représenter des segments.[25]

2) La similarité cosinus entre des plongements d'ELMo bilingues.

3) Les plongements de phrases de contexte anglais et français, où le mot ou le segment cible est remplacé par un blanc, généré par Context2Vec.

4) Des représentations de mot masquées basées sur Context2Vec et MUSE, suivant le travail de Vyas et Carpuat (2017).

Afin de garder seulement une direction pour la catégorie *Implication textuelle*, pour des instances de catégorie *particularisation*, nous inversons l'ordre des traits anglais et français (des traits indépendants du contexte et aussi sensibles au contexte)[26], en suivant la même pratique dans la tâche monolingue d'inférence lexicale.

6.5.3 Résultats expérimentaux et discussion

La boîte à outils `Scikit-learn` est utilisée pour entraîner différents classifieurs avec des hyper-paramètres par défaut. Nous évaluons des classifieurs par une validation croisée à trois plis (utilisant *StratifiedKFold*). Le classifieur *Dummy* est utilisé comme une baseline. Il prédit toujours la catégorie majoritaire.

Comme présenté dans le tableau 6.15, en comparant avec la seule utilisation des traits indépendants du contexte, l'ajout des traits sensibles au contexte améliore la performance pour trois classifieurs sauf *RandomForest*, qui semble avoir des difficultés face à des milliers de traits. Les classifieurs *Logistic Regression* et *Multi-Layer Perceptron* obtiennent la F-mesure moyenne pondérée la plus élevée (0,74) et l'amélioration est la plus évidente pour *Multi-Layer Perceptron*. Nous enlevons les traits liés à Context2Vec parce que pendant l'étude de contribution de traits, ils n'améliorent pas des résultats de façon significative, et apportent plutôt du bruit pour la classification.

22. `https://github.com/HIT-SCIR/ELMoForManyLangs`
23. Ce corpus interne a été compilé et pré-traité par Xavier Tannier.
24. Nous avons mené des expériences avec une moyenne de trois couches et la dernière couche d'ELMo. Les résultats ne sont pas différents de façon significative. Nous présentons des résultats obtenus par une moyenne des couches dans la prochaine section.
25. La moyenne des plongements donne des meilleurs résultats que la somme des plongements.
26. Cela concerne ces traits indépendants du contexte : les traits de surface, les vecteurs qui comptent les occurrences des étiquettes de PoS, l'entropie moyenne de traduction lexicale, différence de probabilité de traduction, pondération lexicale et analyse en dépendance interne.

Classifieur	Exactitude moyenne	F1 moyenne pondérée
Dummy	50.75%	0.34
Seulement des traits indépendants du contexte (132 traits)		
Logistic Regression	69.76%	0.68
LinearSVM	68.55%	0.67
Multi-Layer Perceptron	67.12%	0.66
RandomForest	68.98%	0.67
Seulement des traits sensibles au contexte (6249 traits)		
Logistic Regression	69.84%	0.67
LinearSVM	67.83%	0.67
Multi-Layer Perceptron	66.98%	0.67
RandomForest	68.05%	0.64
Des traits sensibles au contexte sauf des traits liés à Context2Vec (3849 traits)		
Logistic Regression	69.19%	0.64
LinearSVM	70.26%	0.66
Multi-Layer Perceptron	66.98%	0.67
RandomForest	66.91%	0.64
Tous les traits (6381 traits)		
Logistic Regression	74.48%	0.74
LinearSVM	71.91%	0.71
Multi-Layer Perceptron	72.91%	0.72
RandomForest	69.62%	0.66
Tous les traits sauf des traits liés à Context2Vec (3981 traits)		
Logistic Regression	74.77%	**0.74**
LinearSVM	72.27%	0.72
Multi-Layer Perceptron	**74.91%**	**0.74**
RandomForest	68.19%	0.65

Tableau 6.15 – Classification des procédés de traduction en trois classes : Équivalence (710), Implication textuelle (384) et Lié en thématique (305)

En général, l'utilisation du procédé de traduction *équivalence* ou *transposition* est indépendante du contexte, alors que les quatre autres procédés sont plus dépendants du contexte (voir les exemples dans le tableau 6.13). Pour cette raison, nous avons mené une classification binaire entre *équivalence* (710 instances) et la somme de deux autres catégories (689 instances). Ainsi le jeu de données est plus équilibré. Les résultats correspondants sont présentés dans le tableau 6.16. Après avoir ajouté des traits sensibles au contexte, la meilleure performance est obtenue par le classifieur *Logistic Regression* (la F-mesure moyenne pondérée est de 0,79). L'amélioration est plus petite que la classification en trois classes, parce que des traits sensibles au contexte seuls obtiennent des performance plus basses que des traits indépendants du contexte dans cette configuration.

Ces deux expériences montrent la pertinence d'ajouter l'information contextuelle en plus de traits indépendants du contexte. D'un autre côté, leur interprétation n'est pas triviale. Nous menons une analyse qualitative par la suite.

Analyse des performances Les F-mesures moyennes pour chaque catégorie sont présentées dans les tableaux 6.17 et 6.18, obtenues par le meilleur classifieur *Logistic Regression*, utilisant seulement des traits indépendants du contexte ou tous les traits sauf ceux liés à Context2Vec. Nous pouvons voir qu'après l'ajout des traits sensibles au contexte,

Classifieur	Exactitude moyenne	F1 moyenne pondérée
Dummy	50.75%	0.34
Seulement des traits indépendants du contexte (132 traits)		
Logistic Regression	76.27%	0.76
LinearSVM	75.05%	0.75
Multi-Layer Perceptron	75.77%	0.76
RandomForest	74.70%	0.75
Seulement des traits sensibles au contexte (6249 traits)		
Logistic Regression	71.26%	0.71
LinearSVM	70.77%	0.71
Multi-Layer Perceptron	71.55%	0.71
RandomForest	69.26%	0.68
Des traits sensibles au contexte sauf des traits liés à Context2Vec (3849 traits)		
Logistic Regression	72.62%	0.72
LinearSVM	72.34%	0.72
Multi-Layer Perceptron	70.98%	0.71
RandomForest	69.62%	0.69
Tous les traits (6381 traits)		
Logistic Regression	78.77%	0.79
LinearSVM	77.27%	0.77
Multi-Layer Perceptron	76.27%	0.76
RandomForest	70.34%	0.70
Tous les traits sauf des traits liés à Context2Vec (3981 traits)		
Logistic Regression	**79.27%**	**0.79**
LinearSVM	77.84%	0.78
Multi-Layer Perceptron	78.06%	0.78
RandomForest	70.26%	0.70

Tableau 6.16 – Classification des procédés de traduction en deux classes : Équivalence (710), versus la somme de deux autres classes (689)

des F-mesures augmentent pour chaque catégorie.

Catégorie	traits indépendants du contexte	+ traits sensibles au contexte
Équivalence	0.78	0.80
Implication textuelle	0.73	0.84
Lié en thématique	0.39	0.48

Tableau 6.17 – F-mesures en moyenne par classe pour la classification en trois classes

Catégorie	traits indépendants du contexte	+ traits sensibles au contexte
Équivalence	0.77	0.80
Non-équivalence	0.76	0.79

Tableau 6.18 – F-mesures en moyenne par classe pour la classification binaire

Nous avons ensuite comparé les résultats de classification binaire. Par exemple, après l'ajout de traits sensibles au contexte, cette instance est correctement classifiée comme une traduction de non équivalence :

Last month scientists reported the entire continent ***is now in negative ice balance****.*
→ Le mois dernier, des scientifiques ont annoncé que le continent ***perd désormais de la glace.***

En même temps, il existe toujours des erreurs de classification graves. Par exemple, cette instance a été classifiée dans la catégorie *Équivalence* :
there was a time, when I was ***a boy****, [...] → quand j'étais* ***enfant*** *[...]*
Cela suggère que nous avons besoin de mieux capturer l'hyperonymie cross-lingue, par exemple en suivant le travail de Upadhyay *et al.* (2018).

Analyse du système Concernant l'ingénierie des traits, il existe d'autres traits indépendants du contexte que nous pouvons explorer. Par exemple, exploiter un tableau de traduction de phrase ; extraire le nombre de sens d'un mot dans une ressource externe, etc. Il y a aussi un problème pour nos traits sensibles au contexte : les modèles bilingues d'ELMo et de Context2Vec ne sont pas entraînés sur des corpus parallèles, et des représentations bilingues n'ont pas été alignées dans un même espace vectoriel.

Les résultats montrent que des plongements contextualisés d'ELMo et des représentations de mot masquées basées sur MUSE contribuent à l'amélioration. En revanche, des traits liés à Context2Vec ne sont pas aussi utiles que nous le pensions. Nous supposons que c'est partiellement à cause de la taille actuelle du jeu de données et de la méthodologie d'annotation. Prenons ce couple de phrases comme exemple :
So, this may sound like genomic alchemy, but we can, by moving the software of DNA around, ***change things quite dramatically****.*
→ Cela peut ressembler à de l'alchimie génétique, mais nous pouvons réellement, en transférant l'ADN logiciel ici et là, ***faire des changements radicaux****.*
Pour l'instant, nous alignons la partie en gras comme un couple de segments en entier. En revanche, les phrases de contexte autour des segments sont quasiment traduites de façon littérale. Après avoir remplacé la partie alignée par un blanc pour générer des plongements de contexte par Context2Vec, nous supposons que des représentations contextuelles bilingues partagent trop de points en commun pour que le classifieur apprenne des indices utiles.

Afin de mieux exploiter le modèle de Context2Vec, nous proposons de diviser les alignements en : *change → changements, dramatically → radicaux*, pour que leurs contextes immédiats soient différents, qui sont traduits de façon non littérale.

Analyse du jeu de données Puisque nous nous focalisons sur des traductions non littérales, le jeu de données contient seulement 252 paires lexicales (traduction mot à mot) et 1 147 paires où il existe plus d'un token dans un côté de langue. Pour chaque procédé de traduction, nous prenons aléatoirement 100 instances pour analyser le pourcentage de traductions qui ne peuvent être interprétées qu'en contexte. Le résultat montre que le procédé *particularisation* contient le plus de traductions dépendantes du contexte.[27] Par exemple :
In the last five years we've ***added*** *70 million tons of CO_2 every 24 hours – 25 million tons every day to the oceans.*
→ Au cours des 5 dernières années nous avons ***rejeté*** *dans l'atmosphère 70 millions de tonnes de CO_2 chaque 24 heures – 25 millions de tonnes dans les océans tous les jours.*

27. Voici le nombre des traductions dépendantes du contexte par catégorie : 32/100 équivalence, 9/100 transposition, 22/toutes les 129 instances de généralisation, 59/102 particularisation, 35/100 modulation, 14/toutes les 61 instances de modulation_transposition.

Nous avons combiné les procédés *généralisation* et *particularisation* en une seule catégorie *Implication_Textuelle* pendant l'ingénierie des traits, et nous voyons que l'amélioration apportée par des traits sensibles au contexte est la plus grande sur cette catégorie (de 0,73 à 0,84, voir tableau 6.17).

L'analyse d'erreur des résultats de classification révèle aussi la nécessité de revoir les annotations manuelles. Cet exemple illustre ce point :
This is the official dogma, the one that we all take to be true, and it's ***all*** *false. It is not true.*
→ *C'est le dogme officiel, celui que nous croyons vrai, et qui est* ***complètement*** *faux.*
Le système l'a classifié comme une traduction d'équivalence, mais l'annotateur l'a annoté comme non équivalent. La décision du système est en fait correcte. Le résultat de classification pour la catégorie *Lié en thématique* est bien plus bas que les deux autres catégories (voir tableau 6.17). L'analyse qualitative montre aussi les situations où la décision du classifieur est plus justifiée pour cette catégorie.

Par conséquent, nous avons besoin de réviser les annotations avec une analyse plus fine des résultats de classification, pour aider à éviter les inconsistances de l'annotation humaine et améliorer le guide d'annotation de façon itérative.

6.6 Perspectives

Dans les travaux futurs, nous avons ces perspectives. À court terme :

1) Développer un classifieur en deux étapes qui effectue d'abord une classification binaire entre la traduction littérale et celle non littérale ; ensuite mener une classification en multi-classes sur les instances classifiées en tant que non littérales lors de la première étape ; ceci permettrait d'analyser les erreurs à chaque niveau.

2) Combiner les traits en plongements lexicaux avec les traits linguistiques, et utiliser les architectures à base de réseaux neuronaux comme classifieur.

3) Finaliser le transfert des expériences menées pour le couple anglais-français sur le couple anglais-chinois.

Et à plus long terme :

1) Étendre la granularité du niveau sous-phrastique au niveau phrastique : pour une phrase source, prédire si sa traduction contient des segments traduits non littéralement. Puisque plusieurs procédés de traduction peuvent se cumuler dans une phrase entière, la prédiction sera seulement binaire. Cette classification est utile pour constituer un jeu de test qui contient des traductions de référence non littérales afin d'évaluer la traduction automatique. Elle pourrait aussi être utilisée pour fournir des phrases d'exemple aux apprenants de langues étrangères.

2) Prédire la complexité de traduction : pour une phrase source, prédire s'il vaut mieux utiliser des procédés de traduction non littéraux que la traduction littérale. Cette prédiction peut guider un système de traduction automatique pour être mieux adapté aux difficultés de traduction. Elle pourrait aussi être utilisée pour constituer des phrases de test pour entraîner les traducteurs humains.

3) Détecter automatiquement la frontière de traduction non littérale : pour l'instant nos expériences s'appuient sur des instances dont les frontières sont annotées manuellement. Il sera important de le faire automatiquement.

6.7 Conclusion

Dans ce chapitre, nous avons présenté notre étude sur la classification automatique des procédés de traduction. Le jeu de données vient de notre annotation manuelle sur le corpus parallèle anglais-français de *TED Talks* (décrit dans le chapitre 5).

Nous avons présenté trois expériences : 1) l'ingénierie des traits indépendants du contexte, suivi par l'utilisation des classifieurs statistiques 2) l'utilisation des classifieurs neuronaux en prenant seulement des représentations vectorielles en entrée 3) l'ajout des traits sensibles au contexte pour étudier l'importance de l'information contextuelle dans la tâche de classification.

Bien que notre jeu de données soit petit, notre travail valide notre hypothèse de travail et ouvre des pistes de recherche. Nous continuons l'annotation pour fournir plus d'exemples. Les résultats obtenus pour la classification binaire sont encourageants, et nous devons renforcer la classification en multi-classes pour les différents procédés de traduction non littéraux. Notre but à long terme étant d'avoir un meilleur contrôle sémantique pendant l'extraction de paraphrases à partir des corpus parallèles bilingues, il est important de reconnaître au moins ces trois catégories : *équivalence, implication textuelle, lié en thématique*.

Chapitre 7

Validation externe

Sommaire

Dans ce chapitre, nous proposons d'explorer deux voies pour valider l'apport de notre étude sur la reconnaissance des procédés de traduction. L'une concerne la contribution à certaines tâches en TAL (la construction de ressource de paraphrases, l'évaluation de l'alignement automatique de mots et l'évaluation de la qualité de la traduction automatique), pour lesquelles nous proposons des pistes de recherche. L'autre concerne la conception d'un outil pour favoriser l'apprentissage du français langue étrangère.

7.1 Contribution à certaines tâches en TAL

7.1.1 Aide à la construction de ressource de paraphrases

Dans le chapitre 3, nous avons passé en revue des travaux sur la méthode par pivot pour extraire des paraphrases à partir des corpus parallèles bilingues (Bannard et Callison-Burch, 2005; Callison-Burch, 2008; Ganitkevitch *et al.*, 2011, 2012), ainsi que ceux dans la continuité des précédents sur la construction et l'amélioration de la ressource de paraphrases PPDB (Ganitkevitch *et al.*, 2013; Ganitkevitch et Callison-Burch, 2014; Pavlick *et al.*, 2015b,a).

Après nos études sur les procédés de traduction, nous proposons les problématiques de recherche suivantes, qui ne sont pas encore abordées dans les travaux cités :

(a) Les travaux précédents ont regroupé différentes langues pivot comme une langue unique. Des paraphrases candidates obtenues sont-elles différentes selon la langue pivot utilisée, surtout si la langue source et la langue pivot sont moins similaires en linguistique et en culture ? Si oui, de quelle manière sont-elles différentes ?

(b) Les paraphrases dans PPDB sont extraites en adaptant la grammaire hors contexte synchrone (Ganitkevitch *et al.*, 2011), ainsi chaque paire de paraphrases partagent une même catégorie syntaxique. Cela garantit une meilleure grammaticalité de la phrase après la substitution.

En complément de cette ressource, si nous ne considérons que la préservation du sens original, des paraphrases ayant différentes catégories syntaxiques, ou même différentes structures syntaxiques sont également utiles comme ressource. Par exemple, dans le cadre d'aide à la rédaction pour des apprenants de langues étrangères, suggérer des paraphrases ayant différentes catégories syntaxiques les encourage à adapter la rédaction de leur phrase.

(c) Des procédés de traduction, au moins à gros grain (littéral versus non littéral), n'ont pas été pris en compte jusqu'à présent. L'équivalence de traduction approximative a été d'abord assumée, et les paraphrases candidates ont été classées avec des méthodes de plus en plus avancées, en utilisant la probabilité de paraphrase et de nombreux autres traits. Concernant la probabilité de paraphrase, des traductions littérales sont forcément privilégiées grâce à leur haute fréquence. Les traductions non littérales sont probablement moins exploitées, pour leur fréquence moins haute et la difficulté dans leur alignement automatique.

D'un autre côté, le travail de Pavlick *et al.* (2015a) montre bien que différentes relations sémantiques existent à part l'équivalence stricte dans PPDB. Selon leur estimation, dans son ensemble de paraphrases le plus grand (taille XXXL), la relation *Équivalence* n'occupe qu'une très petite partie (voir figure 3.4). À part les problèmes de faux alignements de mots et de qualité de corpus parallèles, est-ce que différents procédés de traduction influencent l'équivalence entre le segment source et la paraphrase obtenue ?

Afin d'étudier ces questions, nous proposons d'adapter le travail de Kok et Brockett (2010), qui ont introduit un modèle sous le nom de *HTP (Hitting Time Paraphraser)*. Ce système à base de graphes est utilisé pour étendre la méthode par pivot. En se reposant sur des parcours aléatoires (*random walk*) et sur le temps d'atteinte (*hitting time*), le système parcourt des chemins de longueur supérieure à 2 (contrairement à la méthode classique) en utilisant l'information entre les nœuds qui représentent des segments (dans la langue d'origine et dans des langues étrangères). Des connaissances monolingues sous forme de nœuds spéciaux sont également prises en compte.

Nous visons à étiqueter les chemins de parcours de ce système (à savoir les liens entre les segments sous-phrastiques) par le procédé de traduction classifié automatiquement, afin de guider le parcours. Par exemple, privilégier plutôt la traduction littérale si nous voulons obtenir des paraphrases strictes, ou plutôt la traduction non littérale avec un glissement de sens pour extraire des réécritures plus variées.

Nous proposons d'étudier les problématiques présentées précédemment au moyen de plusieurs expériences :

— Comparer les paraphrases candidates obtenues selon différentes langues pivot utilisées.

— Extraire des paraphrases avec des catégories syntaxiques différentes (par exemple, via des traductions pivots non littérales qui changent la structure syntaxique des segments source), afin de compléter les résultats de la méthode classique.

— Étudier l'influence des traductions pivots non littérales sur les types de réécritures obtenues.

— Avoir plus d'interprétabilité et plus de contrôle sémantique dans cette adaptation de la méthode par pivot classique.

L'intégration des procédés de traduction est plus complexe et indirecte dans cette direction de recherche, pourtant il est important de les étudier. Cependant la quantité de travail pour mener à bien cette recherche constitue un projet à part entière, et ne peut être réalisé dans cette thèse. Néanmoins, ce travail ouvre cette perspective.

7.1.2 Évaluation de l'alignement automatique de mots

Des jeux de données sur l'alignement de mots construits manuellement sont précieux pour le développement des techniques d'alignement automatique. D'un côté, ils peuvent être utilisés comme exemples de supervision pour ces méthodes. De l'autre, cela permet d'évaluer directement la qualité des outils d'alignement et garantit l'investigation des patrons d'erreur (Xu et Yvon, 2016).

L'évaluation des alignements de mots inclut des mesures intrinsèques et extrinsèques. La mesure intrinsèque la plus communément utilisée est AER (*Alignment Error Rate*), proposée par Och et Ney (2000). Cette mesure s'appuie sur un schéma d'annotation particulier pour des alignements de référence, qui distingue les alignements *sûrs* et *possibles*. Étant donné un ensemble d'alignements A, et l'ensemble d'alignements de référence G, chaque ensemble contient deux sous-ensembles A_S, A_P et G_S, G_P, qui correspondent à des alignements sûrs et possibles. Les mesures suivantes (précision, rappel, F-mesure) sont définies, où T est le type d'alignement, qui peut prendre la valeur de S ou P :

$$P_T = \frac{|A_T \cap G_T|}{|A_T|} \tag{7.1}$$

$$R_T = \frac{|A_T \cap G_T|}{|G_T|} \tag{7.2}$$

$$F_T = \frac{2P_T R_T}{P_T + R_T} \tag{7.3}$$

$$AER = 1 - \frac{|A_P \cap G_S| + |A_P \cap G_P|}{|A_P| + |G_S|} \tag{7.4}$$

Pour mieux comprendre les questions dans ce domaine, des corpus parallèles ont été alignés au niveau du mot, pour des paires de langues telles que allemand-anglais, français-anglais, roumain-anglais, etc. Ces corpus ont ensuite servi comme données de base aux tâches partagées pour évaluer les systèmes d'alignement (Mihalcea et Pedersen, 2003b; Martin *et al.*, 2005).

L'évaluation de la qualité d'alignement de mots dépend de l'application en aval, par exemple la traduction automatique. Langlais *et al.* (1998) sont parmi les premiers qui ont discuté la relation entre la qualité d'alignement et la performance en traduction automatique. Vilar *et al.* (2006) ont mis en évidence la non correspondance entre AER et la

performance en traduction automatique. Lopez et Resnik (2006) ont montré l'impact de la qualité de l'alignement de mots sur des modèles de traduction basés sur des segments. Dans la même direction de recherche, Ayan et Dorr (2006) ont de plus mis l'accent sur l'interaction entre la performance en traduction et différentes méthodes d'extraction de segments.

De nouvelles mesures basées au niveau du segment ont été proposées pour surmonter la faiblesse de AER (Carl et Fissaha, 2003). Une mesure alternative WAA* (*Word Alignment Agreement*) a été proposée et ensuite améliorée par Davis *et al.* (2007). En donnant moins de poids aux points d'alignement qui connectent plusieurs mots alignés, la corrélation avec la performance en traduction a été améliorée en utilisant WAA.

Des critiques ont été portées sur AER et sur ce schéma d'annotation, notamment dû à un manque de sémantique claire des alignements « possibles », qui ont tendance à être utilisés dans beaucoup trop de situations (traductions non littérales, alignements plusieurs-à-plusieurs, etc.) (Fraser et Marcu, 2007). Les expériences et observations de Xu et Yvon (2016) confirment qu'une catégorisation plus fine des alignements que la distinction entre les alignements *sûrs* et *possibles* est nécessaire.

Les corpus parallèles que nous avons annotés pourront servir comme données de test pour évaluer les systèmes d'alignement automatique de mots, puisque nous avons effectué l'alignement manuel de mots ou de segments, et que nous avons catégorisé finement les différents procédés de traduction non littéraux. Cela permet une évaluation et une analyse d'erreurs à grain fin. De nouvelles données peuvent être obtenues en suivant notre guide d'annotation. La corrélation entre notre schéma d'évaluation et la performance des systèmes dans une tâche extrinsèque est à étudier dans des travaux futurs.

7.1.3 Évaluation de la traduction automatique

Pendant l'annotation du corpus, nous constatons que certaines traductions non littérales de bonne qualité sont pourtant très éloignées en surface d'une traduction littérale. Nous nous demandons comment l'évaluation de la traduction automatique devrait fonctionner sur ces traductions de référence non littérales.

D'abord, passons en revue plusieurs principales mesures d'évaluation existantes.

Quand des humains évaluent les résultats de traduction automatique, plusieurs aspects sont pris en compte, tels que l'adéquation, la fidélité et la fluidité. En revanche, ces approches manuelles sont pour la plupart coûteuses en temps (White *et al.*, 1994; Hovy, 1999; Reeder, 2001), et cela ralentit le développement en traduction automatique.

Face au besoin d'avoir une mesure d'évaluation automatique qui est rapide, peu coûteuse, indépendante de la langue et qui se corrèle bien avec l'évaluation humaine, Papineni *et al.* (2002) ont proposé la mesure BLEU* (*BiLingual Evaluation Understudy*).

Pour évaluer la traduction candidate d'une phrase source, la mesure BLEU peut prendre en compte de multiples traductions de référence humaine, en vue d'autoriser des différences légitimes en choix de mots et en ordre de mots. Elle consiste à calculer une moyenne géométrique de précision en n-grammes (version modifiée), entre la traduction candidate et des traduction de référence. Ces correspondances en n-grammes sont indépendantes des positions dans la phrase. Une pénalité de concision (*brevity penalty*) est proposée pour pénaliser les traductions de longueur trop courte par rapport aux références. Plus le score BLEU est élevé, plus la traduction candidate est considérée comme étant de meilleure qualité. BLEU peut mesurer l'adéquation (traduction correcte de sens) et la fluidité (la grammaticalité) des traductions, et les expériences montrent que le classement des

systèmes par BLEU corrèle bien avec le classement humain.

Après la proposition de BLEU, plusieurs autres mesures ont été inventées pour maximiser la corrélation entre les mesures automatiques et l'évaluation humaine au niveau phrastique ou sous-phrastique.

Doddington (2002) a proposé la mesure NIST [1], qui est une variante de BLEU, où le gain d'information de chaque n-gramme est pris en compte. L'idée est de donner plus de poids si une correspondance trouvée concerne des n-grammes plus rares.

Banerjee et Lavie (2005) ont proposé METEOR (*Metric for Evaluation of Translation with Explicit ORdering*). Cette mesure aligne une traduction candidate à une ou plusieurs traductions de référence. Des alignements sont basés sur des correspondances exactes, sur les mots racinisés et sur les synonymes et les paraphrases. Les scores de METEOR incluent aussi une pénalité de fragmentation, qui mesure à quel point les unigrammes correspondants sont bien générés dans l'ordre.

Snover *et al.* (2006) ont présenté TER (*Translation Edit Rate*), qui mesure le nombre d'éditions qu'un humain aurait besoin d'effectuer pour modifier une traduction candidate, pour qu'elle corresponde exactement à une traduction de référence. Les auteurs ont aussi défini HTER (*Human-targeted TER*), où les annotateurs humains génèrent une autre traduction de référence qui est la plus proche de la traduction candidate, et le calcul du nombre d'éditions à effectuer est basé sur cette nouvelle référence. Snover *et al.* (2009) ont introduit TER-plus comme une extension de TER, qui résout certains points faibles de TER en utilisant les paraphrases, le racinisation, les synonymes, etc.

Des différences entre une traduction automatique et une traduction humaine particulière n'indiquent pas toujours une mauvaise qualité de traduction automatique. Récemment, Fomicheva *et al.* (2016) ont proposé une mesure spécifique pour mieux intégrer le critère de la fluidité dans l'évaluation. Cet aspect détermine à quel point une traduction respecte des régularités linguistiques de la langue cible, et il constitue un indicateur fort de la qualité générale de traduction.

Par rapport à des techniques statistiques, l'arrivée des techniques de traduction neuronale a conduit à des améliorations profondes de la qualité de traduction automatique (Kalchbrenner et Blunsom, 2013; Cho *et al.*, 2014; Sutskever *et al.*, 2014; Wu *et al.*, 2016; Vaswani *et al.*, 2017), surtout pour des paires de langues qui sont relativement proches, par exemple anglais-français, anglais-espagnol. Alors que ces techniques continuent à s'améliorer, des mesures d'évaluation existantes perdent leur efficacité de façon inévitable. Un autre défi posé par les systèmes de NMT concerne leur opacité. En effet, en utilisant des systèmes statistiques, il est plus clair d'expliquer quels phénomènes sont mal gérés et la raison correspondante, mais il est plus difficile d'y répondre pour NMT.

En vue de compléter les mesures d'évaluation de la traduction automatique, qui ne rendent qu'imparfaitement compte de la performance des systèmes, Isabelle *et al.* (2017) ont présenté 108 paires de phrases (anglais $\rightarrow$ français), conçues manuellement pour évaluer les systèmes de traduction automatique sur du matériel linguistique "difficile". Cette méthodologie est complémentaire de la pratique standard de sélectionner aléatoirement un jeu de test dans un nouveau texte.

Le but est d'évaluer la capacité d'un système à réduire l'écart causé par des divergences structurales entre langues. Les phrases sources anglaises ont été choisies de sorte que leur équivalent français le plus proche est divergent en structure de la phrase source d'une manière cruciale (Vinay et Darbelnet, 1958; Dorr, 1994). Chaque paire se focalise

1. Ce nom vient de l'institut qui a proposé cette mesure : *the US National Institute of Standards and Technology*.

sur un phénomène linguistique particulier, ce qui rend facile la collection des évaluations manuelles fiables via des questions directes "oui/non" (voir figure 7.1). Ils classifient les phénomènes de divergence en trois types : morpho-syntaxique, lexico-syntaxique et purement syntaxique. Les auteurs considèrent leur approche complémentaire aux évaluations de qualité sur des phrases complexes où différents phénomènes coexistent.

Src	The repeated calls from his mother **should** have alerted us.
Ref	Les appels répétés de sa mère **auraient** dû nous alerter.
Sys	Les appels répétés de sa mère devraient nous avoir alertés.
Is the subject-verb agreement correct (y/n)? **Yes**	

FIGURE 7.1 – Un exemple de question défi dans le jeu de test conçu par Isabelle *et al.* (2017) pour évaluer des systèmes de traduction automatique. Pour une phrase source, les auteurs comparent la référence avec la traduction produite par un système automatique. La question sur le défi attend une réponse binaire

Isabelle *et al.* (2017) ont mené des expériences avec des systèmes de traduction statistique basés sur des segments (PBSMT) et des systèmes NMT. Les résultats montrent que NMT est systématiquement meilleur que PBMT, mêmes si les différences en score BLEU sont petites. Une classification linguistique fine sur des succès et des échecs des systèmes NMT est fournie selon différentes catégories de défi. Les auteurs arrivent à déterminer les difficultés sur lesquelles butent encore les systèmes NMT les plus récents (même avec des données d'apprentissage massives). Ceci inclut notamment des problèmes de la généralisation incomplète, la traduction des idiomes communs et des idiomes syntaxiquement flexibles, ou la traduction des verbes de mouvement *(ex. swim across the river → traverser la rivière à la nage)*, etc.

Dans la même direction de recherche, Isabelle et Kuhn (2018) ont conçu 506 paires des phrases de défi pour la direction de traduction français → anglais, en ajoutant le test au niveau purement lexical. Ils ont comparé les performances entre *Google Translate* [2] et *DeepL* [3], avec deux dates différentes pour chacun.

Les corpus parallèles que nous avons annotés pourront être utilisés afin de constituer un ensemble de défis pour les systèmes de traduction automatique, ce qui permet d'évaluer les systèmes dans des situations où les traducteurs humains recourent à différents procédés de traduction.

Nous pouvons prévoir les cas d'évaluation suivants :

— La traduction non littérale est obligatoire, parce qu'une traduction littérale n'existe pas.
swim across *the river* → ***traverser*** *la rivière* ***à la nage***

— Certains idiomes et proverbes ont leur correspondance figée (une ou plusieurs) dans la langue cible, souvent via une traduction non littérale.
like a bull in a china shop → comme un éléphant dans un magasin de porcelaine
like a bull in a china shop → comme un chien dans un jeu de quilles

2. `https://translate.google.com/`
3. `https://www.deepl.com/fr/translator`

— La traduction non littérale est recommandée, sinon ce n'est pas naturel dans la langue cible.

unless you ***think of it in the terms that I do***
→ *à moins que vous ne* ***regardiez la chose comme moi***

— La traduction non littérale est facultative, ce qui peut tester la limite de la capacité des systèmes automatiques sur le glissement des expressions, sans générer une erreur de traduction.

and ***that scar has stayed with him*** *for his entire life*
→ *et que, toute sa vie,* ***il a souffert de ce traumatisme***

Dans le but de proposer notre évaluation complémentaire aux mesures couramment utilisées, nous devons suivre les travaux de Isabelle *et al.* (2017) sur la conception méticuleuse des défis.

Après avoir passé en revue les travaux existants dans chaque domaine, nous avons proposé les directions de recherche pour trois tâches en TAL, afin d'étudier l'apport de notre recherche sur les procédés de traduction.

Dans la section suivante, nous présentons un autre aspect possible de validation, qui concerne l'aide à l'apprentissage des langues étrangères.

7.2 Conception d'un logiciel d'aide à la compréhension écrite pour les apprenants de FLE

Une partie de travail décrit dans cette section a été publié dans cet article : Zhai *et al.* (2019b).

7.2.1 Problématique de recherche

L'apprentissage des langues étrangères est important pour les étudiants, surtout quand ils veulent poursuivre des études dans un pays étranger. En effet, les études et l'intégration dans la société étrangère nécessitent un niveau de langue intermédiaire voire avancé. Prenons l'exemple des étudiants chinois qui étudient en France. L'anglais est la première langue étrangère qu'une majorité d'entre eux apprennent depuis l'école primaire. Au cours de leur apprentissage du français comme une autre langue étrangère, avoir recours à sa langue maternelle (chinois) ou à une autre langue plus proche du français (anglais) est une pratique courante. La traduction est ainsi la méthode la plus utilisée pour aider les apprenants à mieux assimiler et comparer les connaissances sur la langue.

Une autre stratégie pour apprendre des langues étrangères est la reformulation (Martinot, 2012). Une telle compétence aide les apprenants à élargir leur vocabulaire et leur répertoire d'expressions, à prendre l'habitude de réfléchir dans la langue étrangère au lieu de la langue maternelle. Un texte écrit par un apprenant puis reformulé par un natif donne à l'apprenant de fortes motivations pour analyser les façons natives d'exprimer les mêmes idées (Cohen, 1982, 1983; Sulistyo et Heriyawati, 2017). Les reformulations paraphrastiques (Rossari, 1994; Eshkol-Taravella et Grabar, 2014) aident les apprenants à varier leurs expressions, ce qui est une capacité importante dans la production pour réaliser les transformations lexicales, syntaxiques et même discursives, en vue de simplifier ou rendre leurs énoncés plus complexes (Chachu, 2017; Chen *et al.*, 2013). La capacité de reformulation est aussi utile pour progresser en compréhension écrite. Transformer les mots du

texte en ses propres termes est crucial pour les apprenants, afin de vérifier leur compréhension et la lier avec leurs connaissances préalables en vue de faire des inférences (Kletzien, 2009).

Compte tenu de ces deux aspects sur l'apprentissage d'une langue étrangère (traduction et reformulation paraphrastique), nous proposons les problématiques de recherche suivantes concernant l'apprentissage du français pour des étudiants chinois (*a priori*) :

— Quelles sont les influences positives ou négatives de la maîtrise de l'anglais sur l'apprentissage du français ? Notre hypothèse est que ces influences varient selon le niveau de maîtrise de l'anglais.

— Comment concevoir un outil qui combine les informations de traduction et de paraphrase pour aider les étudiants pendant leur apprentissage ?

Afin de positionner notre étude, nous passons en revue ci-dessous des travaux précédents dans le domaine de la didactique des langues étrangères en trois aspects : le rôle de la langue maternelle, de la traduction et de la reformulation. Nous désignerons la langue maternelle par $L1$, et une langue étrangère en cours d'apprentissage par $L2$.

7.2.2 Travaux antérieurs en didactique

Langue maternelle

Pour les locuteurs natifs, les régularités d'une langue sont acquises par imprégnation au cours des interactions quotidiennes avec la famille et plus tard à l'école. En revanche, les apprenants des langues étrangères ont besoin des processus d'apprentissage plus explicites (Ellis, 2008; Paradis, 2009).

Le rôle de la L1 dans l'apprentissage d'une nouvelle L2 a évolué dans l'histoire en France. Irénée Carré, ancien inspecteur général de l'enseignement primaire, a créé la méthode directe (ou « maternelle ») d'enseignement d'une langue en 1888. L'enseignement du français via cette méthode demande à l'instituteur de procéder comme la mère de famille qui, pour apprendre à parler à son bébé, va directement de l'objet au mot, et du simple au complexe. Le recours à une traduction en langue régionale doit être proscrit. L'enseignant est, en outre, invité à constituer dans les locaux de la classe une collection d'objets usuels que les élèves doivent apprendre à nommer en français. La méthode se veut concrète et pratique. [4]

Michel Bréal, ancien inspecteur de l'enseignement supérieur, est au contraire favorable à une prise en compte des langues premières des élèves dans une approche contrastive des langues. Dès 1965, des classes spécifiques ont été créées pour donner une formation de français langue étrangère aux enfants de travailleurs migrants. C'est une démarche interculturelle où « l'enfant peut se construire en regard de sa langue-culture maternelle ». Il a fallu attendre la fin des années 1970 pour qu'une reconnaissance du rôle de la L1 dans l'apprentissage d'une L2 soit établie.

Tout apprentissage des langues en effet « repose, consciemment ou non, sur une comparaison entre le ou les systèmes langagiers préexistants et la langue à apprendre. Apprendre une autre langue, c'est toujours calquer le système à atteindre sur son système d'origine, quel que soit le niveau linguistique (son, syntaxe, lexique, etc.) », ainsi rappelle Auger (2010). Au cours de l'apprentissage surgissent des phénomènes d'interférences, nommés aussi « transferts négatifs ». Weinreich (1953) est le premier à avoir caté-

4. https://fr.wikipedia.org/wiki/Irénée_Carré

gorisé ces phénomènes qui apparaissent lorsque deux langues entrent en contact. Ce processus inévitable, à travers les « emprunts » inconscients d'ordre phonologique, lexical, syntaxique ou encore sémantique, participe à la construction d'une « interlangue ». Celle-ci est toujours évolutive, tant que l'apprenant poursuit son apprentissage de la langue cible, dont il se rapprochera par ailleurs peu à peu pour parvenir un jour à la maîtrise d'une langue dite « standard ». [5]

Nation (2003) a identifié les tâches dans un cours de langue où utiliser la L1 a de la valeur. Quand un enseignant estime qu'une tâche en L2 dépasserait la capacité des apprenants, par exemple une discussion sur un sujet avant la rédaction, le faire en L1 aiderait à surmonter certains obstacles. Utiliser la traduction en L1 pour apprendre des mots inconnus est efficace. Parce que la définition en L1 est souvent claire, courte et familière pour les apprenants. Au contraire, un apprenant doit maîtriser un vocabulaire suffisamment grand (au moins 2 000 mots pour l'anglais), pour pouvoir consulter efficacement des dictionnaires monolingues. Si L1 et L2 sont dans la même famille de langues (par exemple l'anglais et le suédois), la L1 peut apporter encore plus d'aide sur l'expansion du vocabulaire.

Nation a mis en évidence que la L1 fournit une façon familière et efficace de comprendre rapidement la signification et le contenu des matériels utilisés en L2. La L1 doit être vue comme un outil utile comme les images, les objets réels, la démonstration, etc. et doit être utilisée en cas de besoin sans être sur-utilisée, afin de maximiser l'utilisation de la L2 en classe.

Traduction

Utiliser la traduction comme une méthode d'enseignement de langue fait toujours l'objet d'études, et demeure un sujet fréquemment discuté parmi les linguistes, les méthodologues et les enseignants.

Shiyab et Abdullateef (2001) ont montré que des études analytiques et descriptives, ainsi que les observations des enseignants ont révélé la validité d'utiliser la traduction comme un outil dans l'enseignement d'une L2.

Mahmoud (2006) a passé en revue le contexte historique du rôle de la L1 dans l'enseignement d'une L2. Il a présenté comment la L1 est utilisée actuellement dans les classes de L2. L'auteur soutient spécifiquement l'exercice de la traduction écrite en L1 après avoir lu un texte en L2, pour évaluer la compréhension écrite des apprenants. Parce que pendant cet exercice, l'apprenant se concentre sur le texte entier et se focalise sur la compréhension. Leur traduction peut représenter leur capacité de compréhension et leur développement d'interlangue.

Vermes (2010) a présenté les avantages et inconvénients d'utiliser la traduction dans l'enseignement d'une L2. L'auteur pense que la traduction peut être utilisée dans une façon constructive, mais cela conduit à des questions concernant plusieurs aspects : quand, comment, dans quelles circonstances et pour quels buts.

L'enquête menée par Dagilienė (2012) suggère que la traduction est un outil pédagogique efficace dans des cours d'anglais. La traduction intégrée dans les activités quotidiennes de classe s'avère utile pour le progrès des étudiants sur diverses compétences de langue. La traduction améliore aussi la compréhension des étudiants sur les structures de

5. https://cursus.edu/articles/35674/transferts-linguistiques-interferences-dans-lapprentissage

deux langues. Pourtant la traduction ne doit pas être utilisée excessivement, elle doit être intégrée dans l'enseignement pour des étudiants appropriés aux bons moments.

Reformulation

À part la traduction et l'utilisation de la langue maternelle, le rôle de la reformulation en L2 dans la didactique des langues a également été étudié.

La correction des dissertations étant une tâche importante et chronophage, les enseignants de L2 peuvent se contenter de corriger les problèmes les plus flagrants au niveau lexical, grammatical ou rhétorique. Ces corrections peuvent également être incohérentes. Dans ce cas-là, les étudiants n'obtiennent que des retours partiels. Ceci freine d'une certaine manière leur progrès dans la rédaction, pour que le résultat ressemble le plus possible à une version produite par un locuteur natif.

En vue d'aider les étudiants à mieux maîtriser la production écrite (améliorer le style et la clarté des idées), Cohen (1983) a utilisé la reformulation comme un complément d'analyse d'erreurs, et a appliqué cette méthode dans une classe de L2.

Sur une dissertation originale, l'enseignant apporte d'abord des corrections de surface pour que le texte respecte les règles de la langue étrangère utilisée. Une version révisée est produite par les apprenants suite à cette étape. Ensuite, une reformulation de toute la dissertation est effectuée par un locuteur natif, qui doit reconsidérer le texte pour qu'il reflète un style natif, tout en préservant le sens. Cette méthode de reformulation donne aux apprenants de fortes motivations à analyser la rédaction native pour développer les mêmes idées. Les apprenants ont ainsi l'opportunité de voir leurs points faibles dans la rédaction, et déterminer des zones à développer dans le futur, par exemple : choix de vocabulaire, choix et ordre des structures syntaxiques, marqueur de cohésion, fonctions de discours, etc.

L'auteur souligne que la reformulation pourrait avoir l'impact le plus fort sur les élèves au niveau avancé, qui essaient vraiment de perfectionner leur compétence de rédaction en langue seconde. En tout cas, ce genre d'exercice a le potentiel pour susciter un intérêt réel à la rédaction native parmi des apprenants non natifs.

Kletzien (2009) a mis en évidence que la capacité de paraphraser aide les apprenants à surveiller leur compréhension, et les encourage à accéder à ce qu'ils connaissent déjà sur un sujet. L'exercice de paraphrase met l'accent sur la compréhension, qui est le but de la lecture, et aide à tisser des liens entre les connaissances déjà acquises et ce qui est en train d'être lu.

Selon Martinot (2012), si un jeune enfant natif pratique spontanément la reformulation des énoncés produits par ses parents, son entourage ou lui-même, sans en avoir la moindre conscience, l'apprenant de langues étrangères a au contraire intérêt à prendre conscience des différentes procédures reformulatoires qu'il peut appliquer aux énoncés de la L2. La réactivation de ces procédures entraîne l'apprenant à réfléchir à ses stratégies langagières et l'invite à comparer systématiquement ses productions aux productions des natifs. C'est ce travail de va-et-vient conscientisé qui permet d'optimiser les stratégies d'apprentissage de chaque apprenant.

Après avoir revu les travaux précédents dans ces trois domaines, nous décidons de nous focaliser sur l'aide au développement de la compétence en reformulation pour des apprenants de langues étrangères. Nous considérons aussi les contributions que la traduction peut apporter dans ce processus, dans la langue maternelle ou dans une autre langue

étrangère que les apprenants maîtrisent. Nous expliquons la motivation de notre travail dans la section suivante.

7.2.3 Motivation de travail

Dans cette section, nous présentons notre motivation et nos hypothèses de travail sur la conception d'un outil pour aider la compréhension écrite.

Prenons un exemple où un étudiant qui apprend le chinois comme langue étrangère veut comprendre le sens de cette phrase :

从汗牛充栋的古籍中，找出所需要的资料相当不容易。

Voici les traductions générées par *Google Translate* en français et en anglais :

- Parmi les livres anciens enthousiastes, il n'est pas facile de trouver les informations nécessaires.

- From the ancient books that are full of enthusiasm, it is not easy to find the information needed.

La deuxième partie de la phrase est compréhensible, mais utiliser le mot « *enthousiaste* » pour décrire des livres anciens ne fait pas de sens pour l'apprenant.

En fait,《汗牛充栋》est un idiome, qui se traduit littéralement comme « *faire transpirer le bœuf qui les transporte ou remplir une maison jusqu'aux chevrons* », et le sens figuré est « *avoir un très grand nombre de livres* ». Ainsi la phrase peut être traduite comme suit :

Parmi un très grand nombre de livres anciens, il n'est pas facile de trouver les informations nécessaires.

Nous voyons que c'est une situation typique qui bloquerait la compréhension écrite pour les apprenants de langues étrangères. Revenons à l'apprentissage du français, selon Yilmaz Güngör (2015), pour les apprenants débutants, une syntaxe simple, un contexte contenant des informations socioculturelles familières pour les apprenants, et l'absence des mots inconnus sont des éléments qui facilitent la compréhension des textes.

Les difficultés en compréhension écrite évoluent selon le niveau de langue des apprenants. Par exemple, les débutants peuvent confondre des mots homographes, sans pouvoir distinguer leur catégorie grammaticale et leur usage en contexte.

Qu'en penses-tu ? (Le mot « *en* » est un pronom au lieu d'une préposition.)

(Une reformulation qui explicite l'usage du pronom est plus facile à comprendre : « *Que penses-tu de cela ?* »)

Je n'ai envie d'aller nulle part. (Le mot « *nulle* » pourrait être compris dans le sens de « *sans aucune valeur* » au lieu de « *aucun* ».)

Pour les apprenants dans un niveau intermédiaire même avancé, nous avons résumé les phénomènes suivants qui posent des difficultés de compréhension, en nous basant sur l'article de Yilmaz Güngör (2015) :

1) Des mots inconnus des domaines généraux ; des mots déjà vus mais ils apparaissent dans un contexte donné avec un sens inconnu ou figuré :

*Dans la publicité, le journalisme, la mode, les métiers artistiques et de la culture en général, « être **branché** » est important.*

(Ici « *branché* » veut dire « *à la mode* ».)

2) Des termes spécifiques d'un certain domaine :

*Aubert, **évêque** d'Avranches, installa sur le site une communauté de douze **chanoines** pour servir le **sanctuaire** et accueillir les **pèlerins**.*

3) Des nuances et subtilités de langue ; différents registres de langue :

Par exemple, la nuance entre « *persuader* » et « *convaincre* » est assez subtile, mais elle existe cependant. À première vue, ces deux verbes ont des significations tellement proches qu'ils sont considérés comme des synonymes. Pourtant, la persuasion fait appel au pathétique, à l'instinct et au ressenti, alors que la conviction est un phénomène qui joue sur l'argumentation logique. L'emploi de « *convaincre* » devrait donc être réservé aux opinions qui se basent sur un argumentaire étayé, par exemple, sur des faits scientifiques ; alors que « *persuader* » s'emploierait davantage avec des ressentis et des émotions. [6]

4) Des expressions figées, des idiomes :
Quand je suis allé à la réunion, il y avait ***trois pelés et un tondu*** *...*
(Cette expression veut dire « *presque personne* ».)

5) Des structures syntaxiques complexes, des phrases longues avec une logique complexe. Pour l'instant, les difficultés sur la structure syntaxique dépassent le cadre de notre travail.

Compte tenu des travaux précédents en didactique des langues étrangères et des difficultés des apprenants, nous proposons ici un cadre possible d'évaluation de nos recherches, qui consiste à développer un outil pour aider les apprenants (*a priori* chinois) de Français Langue Étrangère (FLE*) en compréhension écrite, via la proposition des réécritures des mots ou des suites de mots en contexte.

La production écrite étant plus compliquée, nous travaillons d'abord sur l'aide à la compréhension écrite. Nous ciblons les apprenants chinois pour tester notre hypothèse que l'apprentissage d'une deuxième langue étrangère (français) peut bénéficier de la maîtrise d'une première langue étrangère similaire (anglais), parce qu'une majorité des étudiants chinois apprend l'anglais depuis l'école primaire.

Étant donné ce but précis, les réécritures proposées doivent être contrôlées et de bonne qualité. La génération de réécritures suivra l'approche d'extraction des paraphrases via la méthode par pivot dans des corpus parallèles bilingues. Nous intégrons la reconnaissance automatique des procédés de traduction dans ce processus, pour avoir un meilleur contrôle sémantique pendant la recherche de réécritures, aussi pour faciliter la compréhension des justifications du système et la comparaison des connaissances dans des langues différentes.

Notre conception de l'outil s'appuie sur l'hypothèse de travail que pendant l'extraction de réécritures via la méthode par pivot, la reconnaissance de procédés de traduction nous permet de classer les réécritures et les distribuer dans ces trois classes :

— réécritures avec une équivalence sémantique stricte, à savoir des paraphrases, et la phrase après la substitution reste grammaticale.

— réécritures en relation d'implication (plus général ou plus spécifique), si elles existent.

— réécritures qui sont reliées avec le mot ou la suite de mots d'une certaine manière (par exemple, ils appartiennent au même champ sémantique ; il faut adapter la structure de la phrase pour la substitution).

Avant de concevoir cet outil, pour avoir une meilleure idée sur les difficultés et les besoins des apprenants chinois, nous avons mené une expérience préliminaire sur la compréhension écrite avec des étudiants chinois. Nous présentons cette expérience dans la prochaine section.

6. https://www.correctrice-web.fr/nuances-subtilites-francaises/

7.2.4 Expérience préliminaire

Plan expérimental Nous avons préparé deux textes, du niveau A2 et B2, respectivement. Les questions de compréhension portent sur des mots ou expressions potentiellement difficiles pour les étudiants. Les tests se sont passés sur le plateforme *Moodle*, dont les captures d'écran de l'interface sont présentées dans l'annexe A.

Les textes ont été préparés selon la grille pour l'auto-évaluation du niveau A2 et B2 du CECR [7]. Pour la compréhension écrite, la capacité correspondante est décrite comme suit :

> **niveau A2** : « Je peux lire des textes courts très simples. Je peux trouver une information particulière prévisible dans des documents courants comme les publicités, les prospectus, les menus et les horaires et je peux comprendre des lettres personnelles courtes et simples. »
>
> **niveau B2** : « Je peux lire des articles et des rapports sur des questions contemporaines dans lesquels les auteurs adoptent une attitude particulière ou un certain point de vue. Je peux comprendre un texte littéraire contemporain en prose. »

Le texte du niveau A2 est issu d'un site de ressource pour l'éducation [8], dont le sujet est sur le réchauffement climatique, et nous l'avons modifié légèrement. Celui du niveau B2 a été écrit par nous à partir d'une liste de mots et segments. Nous testons la compréhension des étudiants sur des mots ou expressions potentiellement difficiles dans les textes. (Les textes et questions sont présents dans l'annexe A.) Pour chaque groupe d'étudiants (du niveau A2 et B2), le test est divisé en trois phases, où chaque fois une version différente du texte est présentée aux étudiants :

1) version originale (sans aucune information complémentaire)

2) une version où les étudiants peuvent passer la souris sur le texte pour regarder les paraphrases des mots ou expressions en gras

3) en plus de ces paraphrases en français, les traductions anglaises sont aussi ajoutées

À chaque fois, les étudiants répondent au même ensemble de questions. Voici les principes que nous avons respectés pour la conception des questions :

1) Les questions concernent les mots ou expressions potentiellement difficiles, mais pas la logique du texte. Les étudiants doivent comprendre les mots ou expressions pour pouvoir répondre aux questions.

2) Les paraphrases et les traductions anglaises fournies ne doivent pas contenir exactement les mêmes chaînes de caractères contenues dans les réponses.

3) Les options de réponse doivent être assez proches pour être des bons distracteurs. Les distracteurs ne doivent pas aider à induire la bonne réponse de façon évidente.

Les participants sont des étudiants chinois adultes. Nous avons contacté quelques étudiants chinois en France, et aussi trois enseignantes de FLE dans ces trois établissements chinois : Université des langues étrangères de Dalian, Alliance Française de Dalian et Collège de technologie professionnelle de la ville de Ningbo. Nous avons échangé des idées sur la préparation des textes et des questions avec ces enseignantes. Parmi les 44 étudiants qui ont été sollicités pour participer aux tests, il y a eu finalement 15 étudiants

7. https://www.coe.int/fr/web/common-european-framework-reference-languages/

8. http://fr.hellokids.com/c_16133/lire-et-apprendre/reportages-pour-enfant/les-sciences/le-developpement-durable-explique-aux-enfants/le-rechauffement-climatique

du niveau A2 et 11 étudiants du niveau B2 qui ont fini les tests. Et 18 étudiants n'ont pas passé les tests malgré plusieurs relances.

Avant de commencer l'expérience, nous avons fait passer les tests à trois étudiants chinois (du niveau A1-A2, B1-B2 et B2-C1, respectivement). Selon leur retour, nous avons ajusté les questions posées, les paraphrases et traductions fournies comme information complémentaire. Finalement nous avons fixé le temps limité pour chaque test : 30 minutes pour le premier test, et 15 minutes pour chacun des deux tests suivants. Il a été interdit d'utiliser les dictionnaires (peu importe le format) pendant les tests. Après avoir fini les tests, les étudiants ont dû répondre à un questionnaire, qui est aussi consultable dans l'annexe A.

Ajout de paraphrases et traductions au texte Dans cette expérience, une partie des paraphrases et traductions fournies pour aider les étudiants sont extraites automatiquement d'un corpus parallèle bilingue. Pour cela, nous avons adapté un système de traduction automatique statistique, développé par Gong *et al.* (2013) [9]. Le corpus utilisé est combiné des corpus de `TED Talks` [10] et de `Tatoeba` [11]. Après avoir enlevé les doublons, ce corpus anglais-français contient 397k lignes de phrases parallèles. L'alignement automatique de mots a été effectué par l'outil `FastAlign` (Dyer *et al.*, 2013) avec les paramètres par défaut. Pendant cet alignement, la langue source est le français, et celle cible est l'anglais, afin d'utiliser l'anglais comme une langue « pivot » pour générer des paraphrases françaises.

Étant donné une entrée (un mot ou une expression, *ex.* « de par le monde »), ce système peut réaliser ces trois tâches essentielles :

1) afficher toutes les phrases françaises où l'entrée apparaît, ainsi que la phrase de traduction en anglais. Ce concordancier bilingue peut faciliter l'examen manuel du corpus parallèle.

2) extraire toutes les traductions possibles anglaises pour cette entrée (*ex.* « *over the world, around the globe* », etc.).

3) à partir de chaque traduction intermédiaire en anglais, extraire leur traduction en français, à savoir des paraphrases candidates (*ex.* « à travers le monde, aux quatre coins du monde » etc.).

Nous avons sélectionné manuellement des paraphrases et traductions adéquates pour les ajouter au texte comme information complémentaire. Quand ce système ne peut pas fournir ces informations, à cause de la taille limitée du corpus ou de la difficulté d'alignement de mots, nous avons eu recours à la ressource en ligne *Linguee*. Par exemple dans cette phrase :

Quand il n'écrivait pas, il racontait des récits ***édifiants*** *sur les prouesses accomplies par des personnages dans sa tête.*

Le mot « édifiant » n'existe pas dans notre corpus, mais grâce à *Linguee*, nous avons pu trouver ses traductions anglaises (dans ce contexte) « *amazing* » ou « *stunning* », ce qui peut être retraduit en français comme « étonnant », « incroyable », « époustouflant », etc.

Voici un autre exemple :

On y voyait un ivrogne, ayant son calepin à la main, errer avec ***dans son sillage***

9. Le premier auteur Li GONG a développé ce système en java pendant sa thèse au LIMSI.
10. Nous avons utilisé deux corpus anglais-français de `TED Talks`, qui ont été publiés pour la campagne d'évaluation IWSLT en 2013 et 2016 (`https://wit3.fbk.eu/`).
11. C'est un site de collection de phrases et de traductions (`https://tatoeba.org/eng`).

certains de ses compagnons.

Notre système a proposé « *in its wake* » et « *in his spirit* » comme traduction, or la dernière n'est pas appropriée dans ce contexte. En plus, la paraphrase obtenue reste identique (toujours « dans son sillage »). Sur *Linguee*, nous avons pu prendre « *in the tracks of* » comme une traduction supplémentaire, et ainsi « sur ses traces » comme une paraphrase possible.

Il existe aussi des cas où nous ne pouvons ajouter que des traductions anglaises, parce qu'il n'existe pas de paraphrase (même avec l'aide de *Linguee*), par exemple « pétrole » ou « charbon ».

Enfin les paraphrases ont été classées selon leur niveau de difficulté, avec l'aide du site FLELex [12] (Francois *et al.*, 2014), qui est le premier lexique classé pour le FLE qui indique les fréquences de mots par niveau de difficulté (selon l'échelle du CECR).

Hypothèses de travail

1) Les questions ont été conçues pour tester la compréhension des étudiants sur les mots ou expressions potentiellement difficiles pour leur niveau actuel. Dans le premier test où les étudiants lisent le texte original, ils feront des fausses inférences quand ils répondront aux questions.

2) Pour la deuxième et troisième version qui contiennent des informations complémentaires, les étudiants comprendront mieux le texte et auront des meilleures performances.

3) Par rapport au chinois, l'anglais est plus similaire au français, et ces étudiants chinois apprennent l'anglais depuis l'école primaire. La compréhension du texte français bénéficiera-t-elle de leur maîtrise de l'anglais ? La troisième version du texte qui contient des traductions anglaises en plus est-elle nécessaire ? Nous supposons que les performances vont varier selon leur niveau différent en anglais.

Résultats du groupe A2 Les résultats par participant sont montrés dans la figure 7.2. Pour trois étudiants (N° 3, 5, 13), leur note au premier test est meilleure que celle aux deux tests suivants. Du deuxième test au troisième test, quatre étudiants ont répondu correctement à moins de questions (N° 1, 2, 3, 6). Certains étudiants ont eu les mêmes notes pour les trois tests (N° 7, 9, 10, 12), ou les notes très similaires (N° 8, 11, 14). Pour deux étudiants (N° 4, 15), des informations complémentaires semblent apporter le plus d'aide.

La figure 7.3 montre le nombre d'étudiants ayant la bonne réponse pour chaque question. Nous voyons que la deuxième question est la plus difficile. Elle porte sur cette phrase :

[...] il fera bientôt beaucoup trop chaud sur la Terre pour que certaines espèces puissent ***survivre***.

La question est :

Quand il fera bientôt beaucoup plus chaud, certains animaux vont :

- disparaître

- souffrir

- changer de lieu de vie

Les paraphrases donnés sont « continuer à vivre » et « toujours en vie », et la traduction anglaise est « *survive* ». La plupart des étudiants ont choisi « changer de lieu de vie » or la bonne réponse est « disparaître ». Nous supposons que la difficulté dans cette phrase ne

12. `http://cental.uclouvain.be/flelex/`

concerne non seulement le mot « survivre », mais aussi l'expression « trop + adj. + pour que ... puisse ... ».

Nous constatons que la traduction anglaise fournie en plus des paraphrases aide le plus pour la quatrième question, qui concerne cette phrase :

[...] réduire notre consommation d'énergie émettant des gaz à effet de serre au profit d'autres énergies moins ***nuisibles*** *pour la planète.*

Et la question est :

Les nouvelles énergies :

- sont plus profitables

- sont plus faciles à utiliser

- causent moins de mal

Les paraphrases données sont « destructrice » et « dommageable », et les traductions sont « *detrimental* » et « *damaging* ». Avec des traductions anglaises, plusieurs étudiants ont pu corriger leur réponse au troisième test.

Nous avons analysé leurs réponses à notre questionnaire (voir annexe A). En général, le niveau en anglais n'est pas très élevé pour plusieurs étudiants, surtout ceux qui ont eu des moins bons résultats. Ceux qui ont répondu correctement à plus de questions ont dit qu'ils apprennent l'anglais et le français en les comparant, tandis que d'autres étudiants ne veulent pas mélanger l'apprentissage des deux langues. La majorité des étudiants ont confirmé que la maîtrise de l'anglais favorise l'apprentissage du français, surtout pour le vocabulaire et la grammaire. Dix personnes (sur quinze au total) ont indiqué qu'un outil qui propose des paraphrases en contexte pourrait les aider dans la lecture, et neuf personnes sont d'accord que la proposition des traductions anglaises est utile dans un tel outil, qui est mieux si un contexte de phrase est donné en même temps. L'étudiant N° 15 a signalé qu'il existe beaucoup d'applications pour aider à apprendre l'anglais, mais c'est moins le cas pour le français.

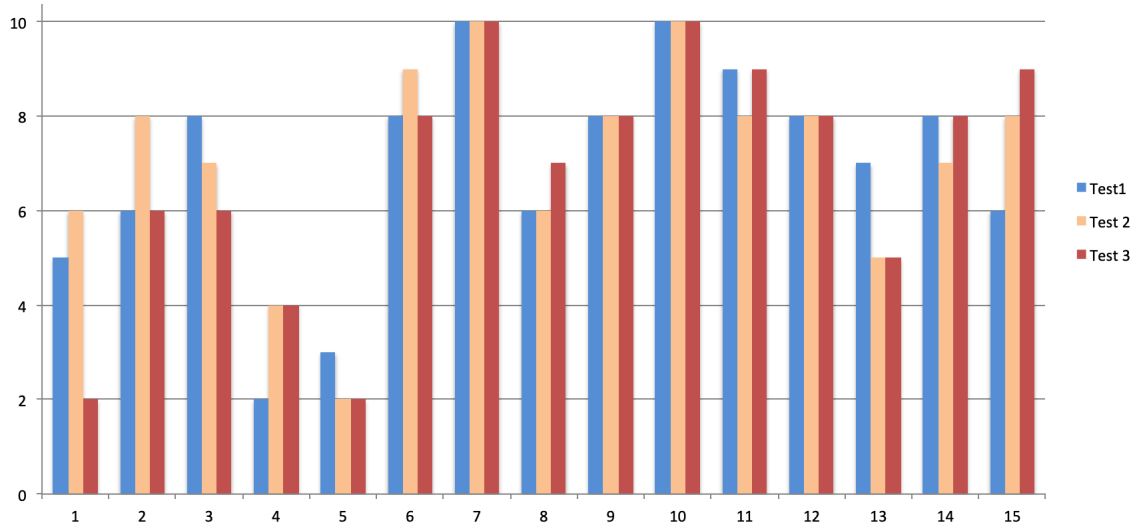

FIGURE 7.2 – Résultats en nombre de bonnes réponses par participant : test niveau A2, 15 participants (axe X), 11 questions (axe Y)

Résultats du groupe B2 Les résultats par participant et par question sont montrés dans les figures 7.4 et 7.5. À part l'étudiant N° 8, tous les autres étudiants ont eu des meilleures notes avec des informations complémentaires par rapport au premier test. Pour trois étudiants (N° 1, 2, 8), l'ajout des traductions anglaises apporte de l'aide supplémen-

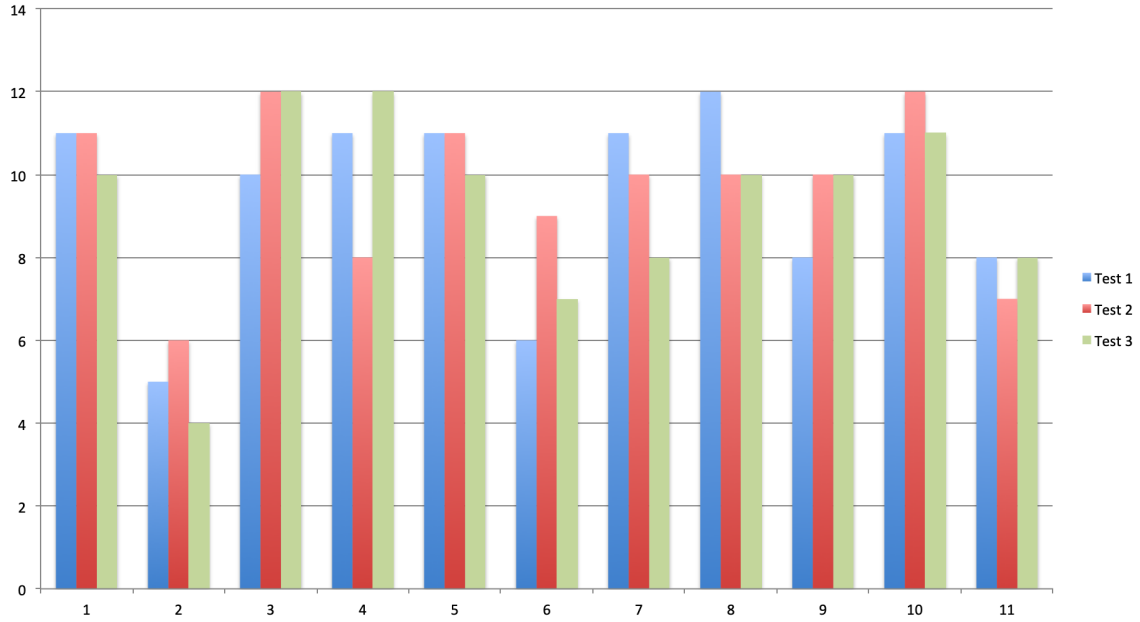

FIGURE 7.3 – Le nombre de personnes qui ont correctement répondu par question : test niveau A2, 11 questions (axe X), 15 participants (axe Y)

taire en plus des paraphrases. En général, l'évolution des performances du groupe B2 est plus homogène que celle du groupe A2.

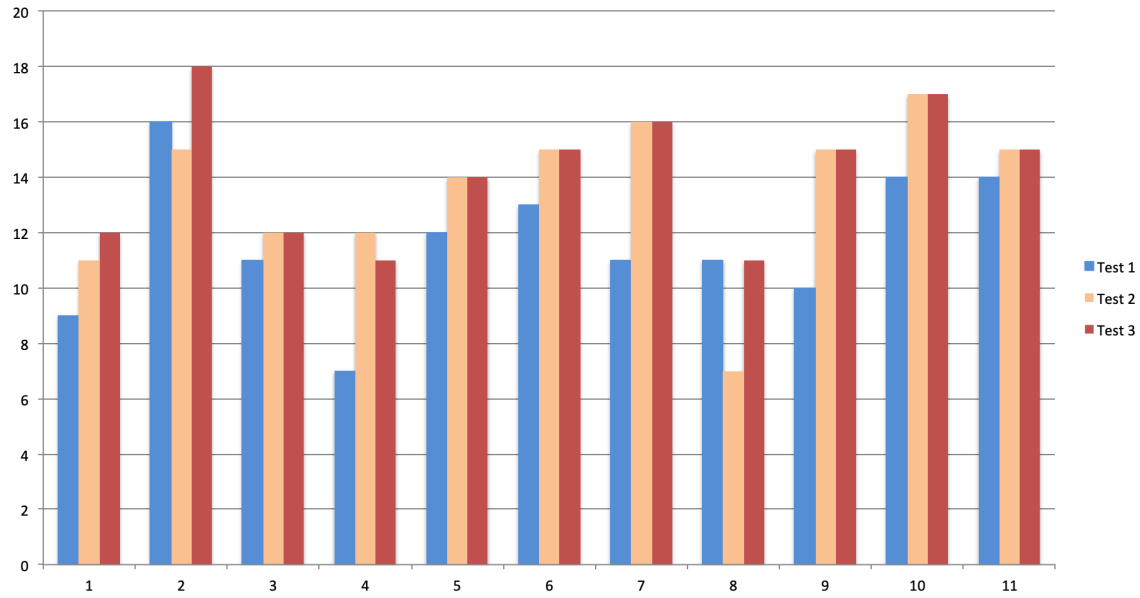

FIGURE 7.4 – Résultats en nombre de bonnes réponses par participant : test niveau B2, 11 participants (axe X), 19 questions (axe Y)

La question la plus difficile est la question 10, peu importe la version du texte, seuls deux étudiants ont eu la bonne réponse. Elle concerne cette phrase :

On y voyait un ivrogne errer avec son calepin à la main, avec ***dans son sillage*** *certains de ses compagnons.*

La question est :

Où voyait-on les compagnons de cet homme ?

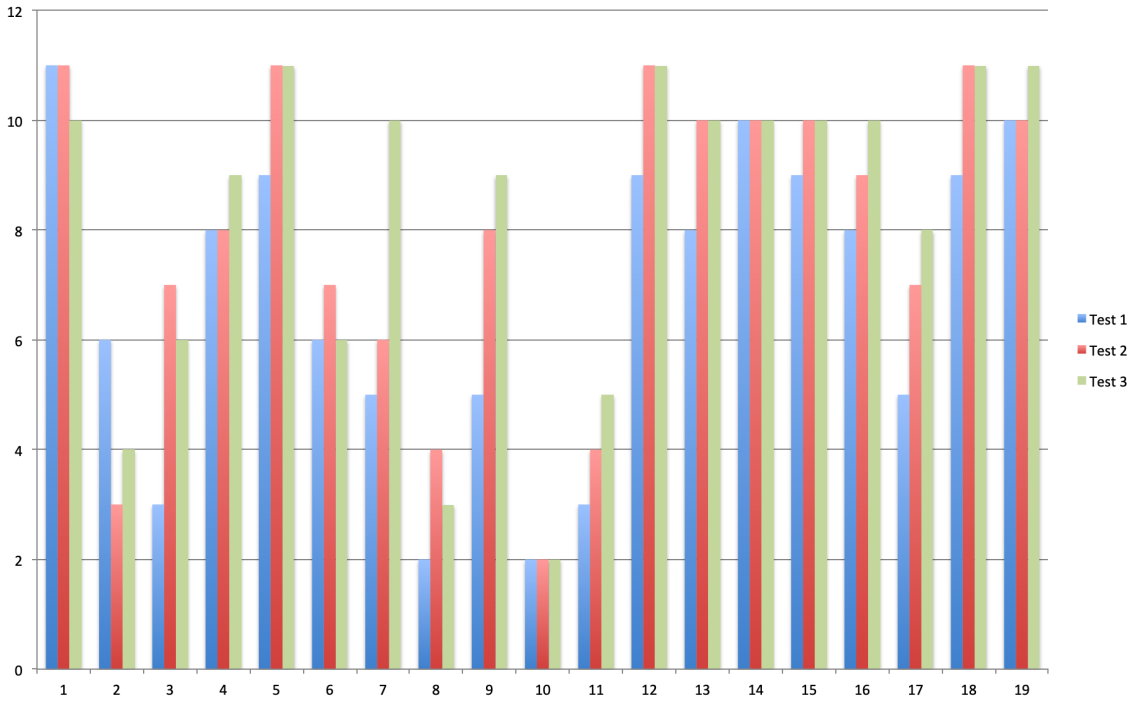

FIGURE 7.5 – Le nombre de personnes qui ont correctement répondu par question : test niveau B2, 19 questions (axe X), 11 participants (axe Y)

- autour de lui
- derrière lui
- devant lui
- ce n'est pas dit

Cela montre que les paraphrases et les traductions ne sont pas suffisamment claires pour aider la compréhension. C'est aussi dû aux principes sur la conception de cette expérience, où les informations supplémentaires ne doivent pas rendre la réponse évidente.

Dans la figure 7.5, nous constatons qu'à part la deuxième question, des informations complémentaires aident la compréhension.

La deuxième question porte sur cette phrase :

*Ainsi, j'ai pu voir à la **lisière** d'une ville, dans un pâté de maisons isolé [...]*

La question est :

Où se déroule l'histoire ?
- au centre-ville
- à la campagne
- aucune des deux réponses

Nous avons fourni « frontière » et « bordure » comme paraphrase, et « *edge* » comme traduction. La bonne réponse est « aucune des deux » mais « à la campagne » a été choisie par la majorité des étudiants au troisième test. Seuls quatre étudiants ont capté la nuance entre la « lisière d'une ville » et « la campagne », sachant que cette notion existe aussi en chinois.

Parmi ces 11 étudiants du niveau B2, trois sont en quatrième année de licence de la langue française en Chine. Les autres poursuivent leurs études en France, dont certains ont eu des notes respectables en TOEIC (*Test of English for International Communication*). Quatre étudiants (N° 2, 6, 7, 8) préfèrent ne pas mélanger l'anglais et le français pendant l'apprentissage, et les autres apprennent ces deux langues en les comparant. Par exemple,

l'étudiant N° 2 a répondu qu'au début l'anglais l'aide beaucoup à apprendre le français, mais maintenant les interférences entre les deux le perturbent. Tous les autres confirment que la maîtrise de l'anglais favorise l'apprentissage du français. À part les similarités sur le vocabulaire et la grammaire, ils ont aussi mentionné l'étymologie et les expressions figées. Excepté deux étudiants qui préfèrent toujours utiliser un dictionnaire, les autres expriment leur envie d'avoir un outil qui propose des paraphrases en contexte pour élargir leur vocabulaire. L'ajout des traductions anglaises dans un tel outil est apprécié par huit étudiants, et les trois autres veulent toujours apprendre dans la même langue.

Nous avons recueilli leurs conseils sur la conception de l'outil, par exemple, les paraphrases plus difficiles que le mot d'origine doivent être présentées avec plus d'information, sinon l'aide fournie est limitée ; l'utilisation des définitions plus simples au lieu des paraphrases peut être considérée ; c'est mieux si un contexte d'utilisation des mots peut être fourni en même temps. L'importance de proposer des paraphrases en contexte après une désambiguïsation a aussi été mentionnée.

Bilan et discussion Avant de concevoir et d'implémenter notre outil d'aide à la compréhension écrite pour les apprenants de FLE, nous avons mené une expérience avec la participation des étudiants chinois adultes. L'anglais est leur première langue étrangère depuis l'école primaire et le français est leur deuxième langue étrangère. Nous avons décrit les détails sur la préparation des textes, des questions et du questionnaire. Pour les deux groupes d'étudiants, selon leur niveau différent en anglais et en français, les notes de trois tests varient de façon différente. Nos hypothèses sur l'évolution de leur performance dans les trois tests sont validées, et de façon plus évidente pour le groupe B2. Selon les figures 7.3 et 7.5, l'hypothèse que les traductions anglaises fournissent de l'aide supplémentaire n'est pas encore complètement validée. Nous pensons qu'une expérience de plus large envergure, avec plus de textes et de participants, est nécessaire pour mieux étudier cette problématique.

Bien que cette expérience soit limitée, nous constatons que des étudiants avec un meilleur niveau en français et en anglais ont tendance à plus bénéficier des paraphrases et des traductions. Pour des débutants, nous pensons qu'il est probablement nécessaire de fournir l'aide sous d'autres formes pour être plus adapté à leur niveau.

Dans un premier temps, à travers cette expérience, la majorité des étudiants a confirmé l'utilité d'un outil pouvant proposer des paraphrases françaises en contexte, ce qui est une motivation forte pour notre futur travail. Pendant les tests, les paraphrases et traductions ont été simplement ajoutées dans une liste, et nous n'avons pas encore ajouté l'information sur les procédés de traduction dans l'interface.

Nous sommes aussi conscients de ces aspects à prendre en compte pendant l'implémentation et l'évaluation d'un tel outil :

— quel type de réécriture conviendrait mieux à quel type d'apprenants ? (paraphrase, réécriture de relation d'implication textuelle par rapport à l'entrée cherchée, réécriture liée d'une certaine manière à l'entrée cherchée)

— le stockage des traces d'utilisation des apprenants pour mener l'analyse de l'apprentissage

— l'adaptation du système face aux demandes différentes des apprenants

— l'évaluation longitudinale de l'outil selon les différents niveaux des apprenants (débutant, intermédiaire, avancé)

7.2.5 Conception de l'outil

L'expérience préliminaire avec des étudiants chinois confirme notre hypothèse de travail et nous permet de concevoir un outil qui répond à leur besoin. Nous proposons notre conception par la suite.

Outils existants pour l'apprentissage Face aux définitions parfois obscures des dictionnaires, les apprenants peuvent être découragés en tentant d'identifier le sens exact selon le contexte par eux-mêmes. *DeepL*, le concurrent de *Google Translate*, permet de varier les expressions sous-phrastiques dans la langue cible. *Linguee*, un concordancier bilingue qui souligne la traduction correspondante (mais elle pourrait être erronée), montre l'usage des segments en contexte. *Writefull* [13] exploite les informations de fréquence à partir de larges bases de données textuelles, permettant aux utilisateurs de vérifier l'usage de segments en production écrite grâce à ses multiples fonctions. *Rewordify* [14] simplifie automatiquement un texte anglais en enlevant les ambiguïtés, pour faciliter la lecture aux apprenants débutants ou intermédiaires.

Notre conception Nous visons à proposer des réécritures françaises en contexte en nous basant sur la méthode par pivot. Plusieurs techniques en TAL seront utilisées : la désambiguïsation sémantique, la traduction automatique, l'alignement automatique de mots, etc. Nous allons garder les traces des traductions pivots anglaises et chinoises pour les montrer aux apprenants chinois. Nous utiliserons aussi la reconnaissance des procédés de traduction au niveau sous-phrastique pour contrôler le processus d'extraction de réécritures.

Notre outil sera développé sous forme d'une application web. Chaque apprenant peut s'inscrire sur le site pour sauvegarder son historique de requêtes. Nous avons besoin des informations telles que sa langue maternelle, les autres langues étrangères maîtrisées et le niveau atteint, etc. pour créer son profil. L'interface la plus importante est montrée dans la figure 7.6. L'apprenant charge un texte dans le bloc 1 (au moins une phrase complète), et à chaque fois il sélectionne un mot ou une suite de mots qui lui pose des difficultés de compréhension. Le système récupère la phrase entière où cette entrée cherchée apparaît, afin de proposer une liste de candidats selon le contexte, affichés dans le bloc 2.

Les réécritures seront classées et contenues dans trois couleurs différentes :

1) réécritures avec une équivalence sémantique stricte, à savoir des paraphrases, et la phrase après la substitution reste grammaticale

2) réécritures en relation d'implication (plus générale ou plus spécifique), si elles existent

3) réécritures qui sont reliées avec le mot ou la suite de mots d'une certaine manière (*ex.* ils appartiennent au même champ sémantique ; il faut adapter la structure de la phrase pour la substitution).

L'apprenant peut cliquer sur chaque candidat proposé, et le bloc 3 s'affichera pour justifier la réécriture. Ici l'entrée cherchée est « d'affilée », et nous montrons le processus pour trouver le candidat « successivement » et « durer », respectivement. Pour chaque paire de traduction trouvée, nous appliquons une reconnaissance automatique pour indiquer si elle est littérale ou non littérale. C'est le chemin avec chaque étape typée jusqu'au candidat qui permet d'interpréter et de classer les résultats. Par exemple, la locution adverbiale anglaise « *on end* » est traduite par un verbe français « durer » dans la dernière paire

13. `https://writefullapp.com/`
14. `https://rewordify.com/`

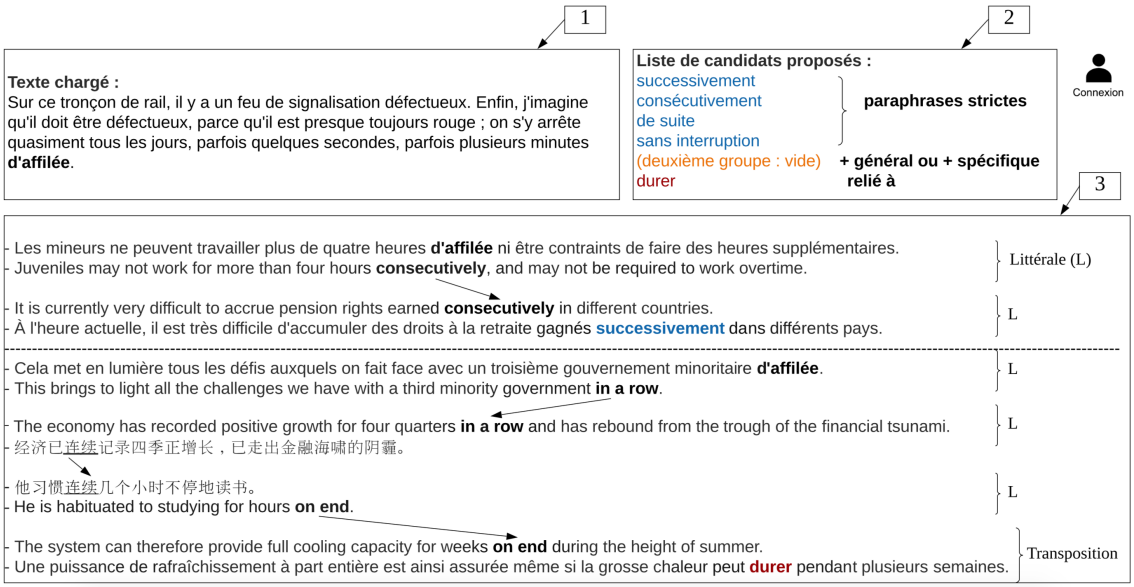

FIGURE 7.6 – Interface du prototype : réécrire en contexte pour mieux comprendre

de phrases. La classe grammaticale change avec le procédé de traduction *Transposition*, ainsi ce candidat n'est pas une paraphrase stricte en contexte. L'apprenant peut décider de n'afficher que les paires de traduction littérale ou non littérale.

Nous garderons les traces d'utilisation des apprenants pour étudier la façon dont ils apprennent, voir quel procédé de traduction est plus utile pour aider la compréhension, etc. À long terme, nous souhaitons que cet outil puisse aussi aider les apprenants à maîtriser la traduction selon les différents procédés de traduction. Notre site web proposera aussi d'autres fonctionnalités :

1) évaluation qualitative par les utilisateurs sur l'utilité des procédés de traduction identifiés

2) visualisation de nos annotations de procédés de traduction par catégorie dans les corpus parallèles

3) forum d'échange entre les utilisateurs

7.2.6 Développement du prototype

Nous décrivons ci-dessous les étapes de développement du prototype de notre outil décrit dans la section précédente. La figure 7.7 présente un résumé du flux de travail.

1. Rassembler plusieurs corpus bilingues parallèles (anglais-français, anglais-chinois, français-chinois), et effectuer des pré-traitements nécessaires (ex. analyse morpho-syntaxique, lemmatisation, etc.).
   — Nous pouvons construire un tableau des suffixes (Manber et Myers, 1993) pour la recherche rapide de mots ou de segments dans le corpus.
2. Constituer un jeu de test qui contient des mots ou expressions polylexicales français à réécrire en contexte.
   — Chacun doit avoir de multiples occurrences dans le corpus pour permettre la recherche de réécritures via la méthode par pivot.
   — Chacun doit apparaître dans une phrase complète sans ambiguïté, pour que nous puissions proposer les réécritures selon le contexte.

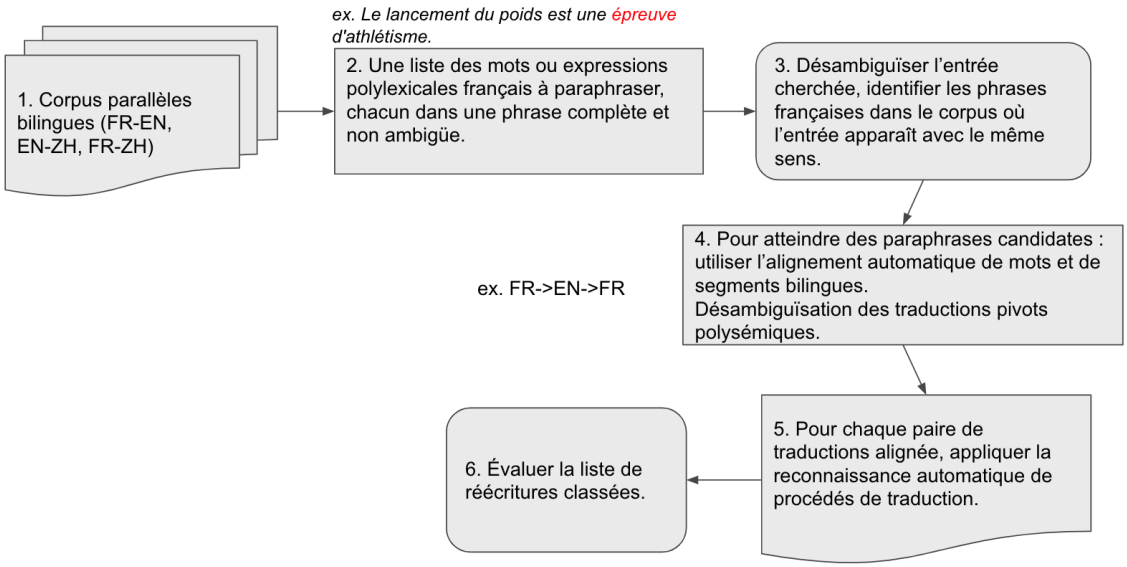

FIGURE 7.7 – Flux de travail pour développer le prototype

3. Désambiguïser l'entrée cherchée, identifier les phrases françaises dans le corpus où l'entrée apparaît avec le même sens. Par exemple, pour cette phrase donnée : « *Le lancement du poids est une* ***épreuve*** *d'athlétisme.* », l'entrée cherchée « épreuve » désigne une compétition sportive.
4. Exploiter l'alignement automatique de mots ou de segments bilingues effectué dans le corpus.
   — À partir des phrases françaises sélectionnées dans l'étape 3 : effectuer un enchaînement d'alignements intermédiaires via les langues de pivot, pour atteindre des paraphrases candidates (ex. français → anglais → français).
   — Désambiguïsation des pivots polysémiques (*a priori* des mots) : il est important de pouvoir continuer à pivoter en préservant le sens de départ.
5. Pour chaque paire alignée, appliquer la reconnaissance automatique des procédés de traduction.
6. Évaluer la liste de réécritures classées. Pour le classement, nous pouvons réutiliser des traits proposés dans des travaux précédents, par exemple ceux de Pavlick *et al.* (2015b), en ajoutant les informations sur les procédés de traduction prédits.

À ce stade, nous disposons de plusieurs corpus parallèles bilingues et des programmes pour effectuer des pré-traitements. La constitution du jeu de test reste à réaliser, et nous travaillerons sur la désambiguïsation lexicale avec des collaborateurs. Pour l'alignement automatique de mots, nous allons tester différents outils, tels que `GIZA++`, `FastAlign`, `Berkeley word aligner`, etc. La reconnaissance automatique des procédés de traduction est au cœur de notre contribution, qui nécessite encore des améliorations pour une meilleure performance. Le développement du prototype de notre outil est complexe, et nous devons rassembler ces différentes composantes pour arriver à l'évaluation finale.

7.3 Conclusion

Nous avons présenté deux cadres de validation de nos études. Le premier concerne trois tâches en TAL : la construction de ressource de paraphrases, l'évaluation de l'ali-

gnement automatique de mots et l'évaluation de la qualité de la traduction automatique. Nous avons passé en revue des travaux existants afin de développer à l'aide des procédés de traduction nos pistes de recherche.

Le second cadre de validation concerne l'aide à l'apprentissage des langues étrangères. Étant donné l'importance de la capacité en reformulation pour l'apprentissage des langues étrangères ainsi que les difficultés des apprenants, nous avons conçu un outil d'aide à la compréhension écrite pour les apprenants de FLE. Diverses techniques (traduction automatique, alignement automatique de mots, désambiguïsation sémantique, reconnaissance de procédés de traduction) peuvent être utilisées pour le développement de cet outil. Une expérience avec des étudiants chinois confirme notre hypothèse de travail et nous a permis de développer cet outil.

Les travaux menés précédemment sur les procédés de traduction constituent les étapes préliminaires, qui vont permettre de mener à bien les travaux dans ces deux directions de recherche.

Troisième partie

Conclusions et Perspectives

Chapitre 8

Conclusion et perspectives

Sommaire

8.1 Bilan

Dans l'introduction de cette thèse, après avoir présenté notre thématique de recherche, nous avons posé une hypothèse de travail générale : il est possible de reconnaître automatiquement différents procédés de traduction au niveau sous-phrastique.

Pour la confirmer, nous avons réalisé ce qui est décrit comme un cycle de travail complet en apprentissage automatique dans l'ouvrage « *Natural Language Annotation for Machine Learning* » de Pustejovsky et Stubbs (2012). Cela consiste à annoter manuellement le jeu de données, entraîner le modèle, tester le modèle, évaluer le modèle, réviser le modèle et les données. En effet, afin de répondre à notre hypothèse posée, nous avons besoin d'un jeu de données pour entraîner un classifieur supervisé. Puisque des données de ce type n'existaient pas, nous avons dû annoter manuellement un corpus parallèle en procédés de traduction.

Afin de fixer une typologie de procédés de traduction pour annoter un corpus parallèle, nous avons d'abord passé en revue et comparé les travaux précédents qui ont proposé différentes typologies (chapitre 2). Ensuite, nous avons justifié le choix du corpus parallèle anglais-français de `TED Talks` pour appliquer nos annotations. En nous basant sur les typologies existantes et sur notre analyse du corpus annoté, nous avons proposé notre propre typologie utilisée pour l'annotation. Les définitions et les exemples typiques pour chaque catégorie ont été présentées dans le chapitre 4.

L'annotation manuelle en détail est présentée dans le chapitre 5 : les caractéristiques du corpus anglais-français, l'outil d'annotation, les choix sur la segmentation en unité de traduction et sur l'alignement de mots. Le guide d'annotation complet est consultable dans l'annexe B. L'étude de contrôle montre que l'accord inter-annotateurs (0,67) est significatif mais il dépasse peu le seuil d'un accord fort (0,61). Face à la difficulté de cette tâche, nous adoptons un processus d'annotation à trois passes pour garantir la qualité de l'annotation. Les statistiques calculées sur le corpus annoté sont présentées, pour les changements effectués pendant le processus d'annotation en trois passes, et pour le nombre de tokens annotés par langue et par catégorie.

L'annotation manuelle nous fournit un jeu de données pour entraîner le classifieur automatique. Dans le chapitre 6, nous avons d'abord positionné notre nouvelle tâche de reconnaissance par rapport aux travaux similaires. Après la présentation des données, nous avons décrit nos différentes expériences. Ces expériences ont été menées dans un scénario simplifié, où nous fournissons au classifieur des paires de segments bilingues avec la frontière donnée, et le but du classifieur est de prédire le procédé utilisé. Nous avons présenté trois expériences : 1) l'ingénierie des traits indépendants du contexte, suivi par l'utilisation des classifieurs statistiques 2) l'utilisation des classifieurs neuronaux en prenant seulement des représentations vectorielles en entrée 3) l'ajout des traits sensibles à la phrase de contexte pour étudier l'importance de l'information contextuelle à la tâche de reconnaissance.

Les résultats obtenus par des classifieurs statistiques sont meilleurs par rapport aux classifieurs neuronaux. Nous avons aussi montré que l'ajout des traits sensibles à la phrase de contexte est pertinent pour améliorer la reconnaissance automatique. Bien que notre jeu de données soit petit, cette étude valide notre hypothèse de travail et ouvre des pistes de recherche. Les résultats obtenus pour la classification binaire sont encourageants, et nous devons renforcer la classification en multi-classes pour les différents procédés de traduction non littéraux.

Ayant travaillé sur le couple de langues anglais-français qui sont des langues proches, nous avons étendu les études sur le couple de langues anglais-chinois (chapitre 5). Le but est de tester la généricité de notre typologie de procédés de traduction et du guide d'annotation sur un couple de langues qui partagent beaucoup moins de points communs au niveau linguistique et culturel.

Pour ce nouveau couple de langues, nous avons constitué un corpus qui contient onze genres différents. Pendant l'annotation, nous avons adapté et enrichi le guide d'annotation, qui est aussi consultable dans l'annexe B. Après la première phase d'annotation, nous avons pu garder la même typologie de procédés de traduction, ce qui justifie d'étudier le transfert des expériences menées pour le couple anglais-français au couple anglais-chinois.

Les deux guides d'annotation sont librement disponibles. Les corpus et leurs annotations seront disponibles à des fins de recherche dès que nous aurons obtenu l'autorisation des organismes concernés.

Revenons aux deux questions qui ont motivé nos travaux : l'aide à la construction de ressources de paraphrases, et l'aide à la compréhension écrite pour les apprenants de Français Langue Étrangère.

Dans le chapitre 3, nous avons passé en revue diverses définitions et typologies de la paraphrase en linguistique et en TAL. Ensuite, nous avons présenté en détail des travaux sur l'extraction de paraphrases via la méthode par pivot. Cette méthode s'appuie sur des équivalences de traduction et sur des techniques de traduction automatique. Dans le chapitre 7, nous avons proposé de nouvelles pistes de recherche concernant cette problématique. Nous pouvons intégrer la reconnaissance des procédés de traduction pour répondre aux deux questions posées : 1) Des paraphrases candidates obtenues sont-elles différentes selon la langue pivot utilisée, surtout si la langue source et la langue pivot sont moins similaires en linguistique et en culture ? Si oui, de quelle manière sont-elles différentes ? 2) À part les problèmes de faux alignements et de qualité de corpus parallèles, est-ce que différents procédés de traduction influencent l'équivalence entre le segment source et la paraphrase obtenue ?

Étant donné l'importance de la capacité en reformulation pour l'apprentissage des

langues étrangères, ainsi que les difficultés des apprenants, nous avons conçu un outil d'aide à la compréhension écrite pour les apprenants de FLE. Pour adapter la conception de l'outil, nous avons mené une expérience avec des étudiants chinois sur la compréhension écrite. L'analyse des résultats valide notre hypothèse de travail et nous permet d'affiner la conception de cet outil.

Les procédés de traduction représentent un sujet linguistique fondamental. Pour cette raison, après avoir passé en revue des travaux existants, nous proposons que nos corpus annotés soient utilisés dans une mesure d'évaluation complémentaire pour ces deux tâches : l'évaluation de l'alignement automatique de mots et l'évaluation de la qualité de la traduction automatique. Pour l'alignement de mots, notre corpus annoté avec différents procédés de traduction permet une évaluation et des analyses d'erreurs plus fines que le schéma d'alignement couramment utilisé, qui contient seulement des alignements *sûrs* et *possibles*. Pour la traduction automatique, notre corpus annoté peut être utilisé comme un ensemble de défis, pour évaluer les systèmes dans des situations où les traducteurs humains recourent à différents procédés de traduction non littéraux.

Les travaux menés sur les procédés de traduction constituent les étapes préliminaires, qui vont permettre de mener à bien les travaux dans ces deux directions de recherche. À ce stade, la performance de la reconnaissance des procédés de traduction au niveau sous-phrastique que nous obtenons est bien meilleure qu'une décision aléatoire. En revanche, pour être intégrée dans les tâches que nous avons décrites, la reconnaissance doit être améliorée de façon importante, surtout pour les procédés non littéraux.

8.2 Perspectives

Nous présentons ci-dessous les perspectives que nous considérons importantes à poursuivre dans les travaux futurs.

À court terme Concernant l'annotation des corpus anglais-français, nous devons annoter des corpus d'autres genres que le discours préparé, par exemple, les corpus littéraires contiennent probablement plus de procédés de traduction non littéraux.

Par rapport à la reconnaissance, nous devons tester d'autres traits, par exemple, les probabilités de traduction entre segments bilingues, la liste de ressources sur les idiomes et les expression figées, le nombre total de sens des mots, etc. Il nous reste à étudier les erreurs d'un classifieur en cascade, qui effectue d'abord une classification binaire entre la traduction littérale et celle non littérale, et qui mène ensuite une classification en multi-classes sur les instances étiquetées en tant que non littérales lors de la première étape. Nous devons aussi combiner les traits en plongements lexicaux avec les traits linguistiques, et utiliser les architectures à base de réseaux neuronaux comme classifieur.

Nous transférerons des expériences effectuées pour le couple anglais-français au couple anglais-chinois, afin de tester l'efficacité des traits déjà utilisés sur ce nouveau couple de langues. Les instances non alignées de catégorie *Réduction* et *Explicitation* fournissent des exemples intéressants pour mener une analyse linguistique sur les différences entre ces deux langues distantes.

Nous demanderons aux annotateurs d'indiquer leur degré de confiance sur la qualité de traduction des paires de phrases. Quand nous construisons le jeu de données d'apprentissage, cela nous permettra d'écarter les phrases qui sont proches des erreurs de traduction, ou de leur attribuer un poids moins important.

À moyen terme Nous souhaitons mettre en place des méthodes d'apprentissage

semi-supervisé, ou de supervision à distance, dans le but d'alléger le besoin de l'annotation manuelle sur des exemples faciles, et de laisser les annotateurs se concentrer sur des exemples difficiles.

Pendant l'annotation sur les traductions en chinois, nous rencontrons des exemples typiques sur le changement de la structure de phrase, qui relèvent d'un changement stylistique. Par exemple, dans la figure 8.1, le syntagme nominal anglais « *the export of [...] which [...]* » est beaucoup trop long pour être traduit littéralement en chinois. Par conséquent, le traducteur a mis cette partie au début de la phrase, pour rendre la traduction plus facile à comprendre. Ce phénomène est pour l'instant hors du cadre de notre étude, et nous devons réfléchir aux façons de les traiter dans les travaux futurs.

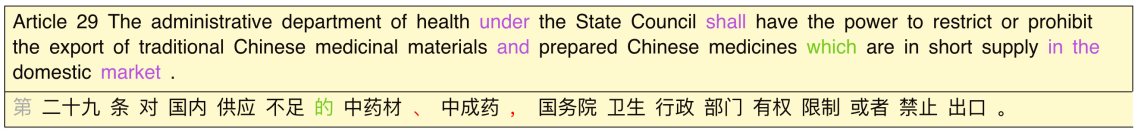

FIGURE 8.1 – Exemple de changement de structure de la phrase dans la traduction chinoise. Voici une traduction littérale de la phrase chinoise en français : « *Article 29 Pour des matières médicinales et des médicaments préparés traditionnels chinois qui sont en pénurie dans le marché intérieur, le département administratif de la santé sous l'égide du conseil d'État a le droit de restreindre ou interdire l'exportation.* »

À long terme La granularité d'étude peut être étendue du niveau sous-phrastique au niveau phrastique. Pour une phrase source, le but est de prédire si sa traduction contient des segments traduits de façon non littérale. Puisque plusieurs procédés de traduction peuvent être utilisés ensemble dans une phrase entière, la prédiction sera seulement binaire. Cette classification est utile pour constituer un jeu de test qui contient des traductions de référence non littérales, pour évaluer la qualité de la traduction automatique. Elle pourrait aussi être utilisée pour fournir des phrases d'exemple aux apprenants de langues étrangères.

La prédiction de la complexité de traduction peut être considérée. Pour une phrase source, le but est de prédire s'il vaut mieux utiliser des procédés de traduction non littéraux que la traduction littérale. Cette prédiction peut guider un système de traduction automatique pour être mieux adapté aux difficultés de traduction. Elle pourrait aussi être utilisée pour constituer des phrases de test pour entraîner les traducteurs humains.

Pour l'instant, nos expériences s'appuient sur des instances dont les frontières sont annotées manuellement. Il sera important de pouvoir détecter ces frontières et aligner les segments traduits non littéralement de façon automatique.

Le développement de l'outil que nous avons conçu pour aider les apprenants de FLE sur la compréhension écrite constitue aussi un cadre de travail à long terme, ainsi que l'évaluation de cet outil auprès des apprenants et des enseignants.

Au final, nous posons deux nouvelles problématiques : 1) En utilisant la méthode par pivot pour extraire des réécritures dans des corpus parallèles bilingues, est-il possible de catégoriser la relation sémantique entre un segment source et sa réécriture grâce à la catégorisation de procédés de traduction ? 2) Peut-on adapter la traduction automatique à l'utilisateur selon les procédés de traductions qui lui conviennent ?

Liste des tableaux

Table des figures

Index

Bibliographie

AGIRRE, E., GONZALEZ-AGIRRE, A., LOPEZ-GAZPIO, I., MARITXALAR, M., RIGAU, G. et URIA, L. (2016). Semeval-2016 task 2 : Interpretable semantic textual similarity. *In Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 512–524. (cité à la page : 53)

AHO, A. V. et ULLMAN, J. D. (1972). *The Theory of Parsing, Translation, and Compiling*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. (cité à la page : 40)

AHRENBERG, L. (2017). Comparing machine translation and human translation : A case study. *In Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria. Association for Computational Linguistics, Shoumen, Bulgaria. (cité aux pages : 24, 25, 27, and 143)

ANDROUTSOPOULOS, I. et MALAKASIOTIS, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187. (cité aux pages : 47, 51, and 52)

APIDIANAKI, M. (2008). *Automatic sense acquisition for Word Sense Disambiguation and lexical selection in translation*. Theses, Université Paris-Diderot - Paris VII. (cité à la page : 9)

AUGER, N. (2010). *Élèves nouvellement arrivés en France : réalités et perspectives pratiques en classe*. Archives contemporaines. (cité à la page : 118)

AYAN, N. F. et DORR, B. J. (2006). Going beyond AER : An extensive analysis of word alignments and their impact on MT. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 9–16. (cité à la page : 114)

BAHDANAU, D., CHO, K. et BENGIO, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*. (cité à la page : 26)

BAKER, M. (1993). *In other words : A coursebook on translation*. (cité à la page : 21)

BALLARD, M. (2006). À propos des procédés de traduction. *Palimpsestes. Revue de traduction*, (Hors série):113–130. (cité aux pages : 16, 17, 24, and 54)

BANERJEE, S. et LAVIE, A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. *In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72. (cité à la page : 115)

BANNARD, C. et CALLISON-BURCH, C. (2005). Paraphrasing with bilingual parallel corpora. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604. Association for Computational Linguistics. (cité aux pages : 35, 36, 37, 38, 39, 43, 44, and 111)

BARREIRO, A. (2009). *Make it simple with paraphrases : Automated paraphrasing for authoring aids and machine translation*. Thèse de doctorat, Universidade do Porto. (cité à la page : 53)

BARRÓN-CEDEÑO, A., VILA, M., MARTÍ, M. et ROSSO, P. (2013). Plagiarism meets paraphrasing : Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947. (cité aux pages : 32, 53)

BARZILAY, R. (2003). *Information Fusion for Multidocument Summarization : Paraphrasing and Generation*. Thèse de doctorat, New York, NY, USA. AAI3088294. (cité aux pages : 31, 32, 47, and 53)

BARZILAY, R. et LEE, L. (2003). Learning to Paraphrase : An Unsupervised Approach Using Multiple-Sequence Alignment. *In* HEARST, M. A. et OSTENDORF, M., éditeurs : *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics. (cité à la page : 51)

BARZILAY, R. et MCKEOWN, K. (2001). Extracting Paraphrases from a Parallel Corpus. *In Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference, July 9-11, 2001, Toulouse, France.*, pages 50–57. (cité aux pages : 30, 33, and 34)

BARZILAY, R. et MCKEOWN, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328. (cité à la page : 53)

BARZILAY, R., MCKEOWN, K. R. et ELHADAD, M. (1999). Information fusion in the context of multi-document summarization. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 550–557, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 53)

BHAGAT, R. (2009). *Learning Paraphrases from Text*. Thèse de doctorat, Los Angeles, CA, USA. AAI3368694. (cité aux pages : 30, 32)

BHAGAT, R. et HOVY, E. H. (2013). What Is a Paraphrase ? *Computational Linguistics*, 39(3):463–472. (cité à la page : 47)

BHAGAT, R. et RAVICHANDRAN, D. (2008). Large scale acquisition of paraphrases for learning surface patterns. *In Proceedings of ACL-08 : HLT*, pages 674–682. (cité à la page : 53)

BJERVA, J., BOS, J., van der GOOT, R. et NISSIM, M. (2014). The meaning factory : Formal semantics for recognizing textual entailment and determining semantic similarity. *In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 642–646, Dublin, Ireland. Association for Computational Linguistics. (cité à la page : 49)

BLATZ, J., FITZGERALD, E., FOSTER, G., GANDRABUR, S., GOUTTE, C., KULESZA, A., SANCHIS, A. et UEFFING, N. (2004). Confidence estimation for machine translation. *In Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 87)

BOJANOWSKI, P., GRAVE, E., JOULIN, A. et MIKOLOV, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. (cité aux pages : 97, 102)

BOUAMOR, H. (2012). *Etude de la paraphrase sous-phrastique en traitement automatique des langues. (A study of sub-sentential paraphrases in Natural Language Processing)*. Thèse de doctorat, University of Paris-Sud, Orsay, France. (cité aux pages : 30, 31, and 33)

BRADFORD, R. B. (2010). Machine translation using vector space representations. US Patent 7,765,098. (cité à la page : 9)

BRANTS, T. et FRANZ, A. (2006). Web 1T 5-gram version 1. (cité aux pages : 42, 44)

CALLISON-BURCH, C. (2007). *Paraphrasing and Translation*. Thèse de doctorat, University of Edinburgh, Edinburgh, Scotland. (cité aux pages : 30, 47)

CALLISON-BURCH, C. (2008). Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. *In 2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 196–205. (cité aux pages : 37, 38, 39, 40, 43, 46, 111, and 143)

CALLISON-BURCH, C., KOEHN, P., MONZ, C., POST, M., SORICUT, R. et SPECIA, L. (2012). Findings of the 2012 workshop on statistical machine translation. *In Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 10–51, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 87)

CALLISON-BURCH, C., KOEHN, P. et OSBORNE, M. (2006a). Improved statistical machine translation using paraphrases. *In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics. (cité à la page : 53)

CALLISON-BURCH, C., OSBORNE, M. et KOEHN, P. (2006b). Re-evaluation the role of Bleu in machine translation research. *In 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics. (cité à la page : 53)

CAMACHO-COLLADOS, J., PILEHVAR, M. T., COLLIER, N. et NAVIGLI, R. (2017). Semeval-2017 task 2 : Multilingual and cross-lingual semantic word similarity. *In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26. Association for Computational Linguistics. (cité à la page : 91)

CANDITO, M., NIVRE, J., DENIS, P. et ANGUIANO, E. H. (2010). Benchmarking of statistical dependency parsers for french. *In Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, pages 108–116. Association for Computational Linguistics, Chinese Information Processing Society of China. (cité à la page : 90)

CARL, M. et FISSAHA, S. (2003). Phrase-based evaluation of word-to-word alignments. *In Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts : Data Driven Machine Translation and Beyond*, pages 31–35. (cité à la page : 114)

CARL, M. et SCHAEFFER, M. J. (2017). Why translation is difficult : A corpus-based study of non-literality in post-editing and from-scratch translation. *HERMES-Journal of Language and Communication in Business*, (56):43–57. (cité aux pages : 86, 92)

CARPUAT, M., VYAS, Y. et NIU, X. (2017). Detecting cross-lingual semantic divergence for neural machine translation. *In Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79. Association for Computational Linguistics. (cité à la page : 87)

CATFORD, J. C. (1965). *A linguistic theory of translation : An essay in applied linguistics*. Oxford University Press. (cité à la page : 11)

CETTOLO, M., GIRARDI, C. et FEDERICO, M. (2012). Wit^3 : Web inventory of transcribed and translated talks. *In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy. (cité à la page : 61)

CHACHU, S. (2017). The intermediary role of reformulation in learning French as a foreign language : The case of students at the University of Ghana. *SHS Web of Conferences*, 38:00007. (cité aux pages : 4, 117)

CHAN, T. P., CALLISON-BURCH, C. et VAN DURME, B. (2011). Reranking bilingually extracted paraphrases using monolingual distributional similarity. *In Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–42, Edinburgh, UK. Association for Computational Linguistics. (cité aux pages : 41, 42)

CHANG, B. et BAI, X. (2003). The Markup Guidelines for the Chinese-English Parallel Corpus of Peking University. *Journal of Chinese Language and Computing*, 13(2):195–214. (cité à la page : 81)

CHAROLLES, M. (2002). *La référence et les expressions référentielles en français*. Ophrys. (cité à la page : 80)

CHE, W., LIU, Y., WANG, Y., ZHENG, B. et LIU, T. (2018). Towards better UD parsing : Deep contextualized word embeddings, ensemble, and treebank concatenation. *In Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics. (cité aux pages : 101, 104)

CHELBA, C., MIKOLOV, T., SCHUSTER, M., GE, Q., BRANTS, T. et KOEHN, P. (2013). One billion word benchmark for measuring progress in statistical language modeling. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. (cité à la page : 101)

CHEN, M., TANG, Q., WISEMAN, S. et GIMPEL, K. (2019). Controllable paraphrase generation with a syntactic exemplar. *In 57th Annual Meeting of the Association for Computational Linguistics*. (cité à la page : 52)

CHEN, M.-H., HUANG, S.-T., CHANG, J. et LIOU, H.-C. (2013). Developing a corpus-based paraphrase tool to improve EFL learners' writing skills. *Computer Assisted Language Learning*. (cité aux pages : 4, 117)

CHEN, Q., KWONG, O. Y. et ZHU, J. (2018). Detecting free translation in parallel corpora from attention scores. *In Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics. (cité aux pages : 24, 25, 26, and 27)

CHIANG, D. (2005). A hierarchical phrase-based model for statistical machine translation. *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 263–270, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 40)

CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. et BENGIO, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*. (cité à la page : 115)

CHUNG, J., GULCEHRE, C., CHO, K. et BENGIO, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv :1412.3555*. (cité à la page : 97)

CHUQUET, H. et PAILLARD, M. (1989). *Approche linguistique des problèmes de traduction anglais-français*. Ophrys. (cité aux pages : 2, 10, 11, 12, and 62)

CIVERA, J. et JUAN, A. (2007). Domain adaptation in statistical machine translation with mixture modelling. *In Proceedings of the Second Workshop on Statistical Machine Translation*, pages 177–180. Association for Computational Linguistics. (cité à la page : 9)

CLARKE, D. (2009). Context-theoretic semantics for natural language : An overview. *In Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 112–119, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 49)

CLARKE, J. et LAPATA, M. (2008). Global inference for sentence compression an integer linear programming approach. *Journal of Artificial Intelligence Research*, 31(1):399–429. (cité à la page : 42)

COHEN, A. D. (1982). Writing like a native : The process of reformulation. (cité à la page : 117)

COHEN, A. D. (1983). Reformulating Second-Language Compositions : A Potential Source of Input for the Learner. (cité aux pages : 117, 120)

COHEN, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46. (cité à la page : 74)

CONNEAU, A., LAMPLE, G., RANZATO, M., DENOYER, L. et JÉGOU, H. (2017). Word translation without parallel data. *arXiv preprint arXiv :1710.04087*. (cité à la page : 102)

CRISTEA, T. (2001). *Structures signifiantes et relations sémantiques en français contemporain*. Editura Fundaţiei România de Mâine. (cité à la page : 31)

DAGAN, I., GLICKMAN, O. et MAGNINI, B. (2005). The PASCAL recognising textual entailment challenge. *In* CANDELA, J. Q., DAGAN, I., MAGNINI, B. et D'ALCHÉ-BUC, F., éditeurs : *Machine Learning Challenges Workshop*, volume 3944 de *Lecture Notes in Computer Science*, pages 177–190. Springer. (cité à la page : 87)

DAGILIENĖ, I. (2012). Translation as a learning method in english language teaching. *Kalbų studijos*, (21):124–129. (cité à la page : 119)

DAUNAY, B. (2004). Réécriture et paraphrase. *Le francais aujourd'hui*, (1):25–32. (cité à la page : 29)

DAVIS, P. C., XIE, Z. et SMALL, K. (2007). All links are not the same : evaluating word alignments for statistical machine translation. *Machine Translation Summit XI, European Associaton for Machine Translation, Copenhagen, Denmark*, pages 119–126. (cité à la page : 114)

DELÉGER, L. et ZWEIGENBAUM, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. *In Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora (BUCC)*, pages 2–10, Singapore. Association for Computational Linguistics. (cité à la page : 53)

DELISLE, J. (1993). *La traduction raisonnée : manuel d'initiation à la traduction professionnelle de l'anglais vers le français : méthode par objectifs d'apprentissage*. Numéro vol. 2 de Collection Pédagogie de la traduction. Presses de l'Université d'Ottawa. (cité aux pages : 14, 15, and 16)

DENG, D. et XUE, N. (2017). Translation divergences in chinese–english machine translation : An empirical investigation. *Computational Linguistics*, 43(3):521–565. (cité aux pages : 22, 23, and 141)

DENKOWSKI, M., AL-HAJ, H. et LAVIE, A. (2010). Turker-assisted paraphrasing for English-Arabic machine translation. *In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 66–70, Los Angeles. Association for Computational Linguistics. (cité aux pages : 44, 45)

DENKOWSKI, M. et LAVIE, A. (2010). METEOR-NEXT and the METEOR paraphrase tables : Improved evaluation support for five target languages. *In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden. Association for Computational Linguistics. (cité à la page : 44)

DODDINGTON, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *In Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc. (cité à la page : 115)

DOLAN, B., QUIRK, C. et BROCKETT, C. (2004). Unsupervised Construction of Large Paraphrase Corpora : Exploiting Massively Parallel News Sources. *In Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité aux pages : 33, 34)

DOLAN, W. B. et BROCKETT, C. (2005). Automatically constructing a corpus of sentential paraphrases. *In Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. (cité à la page : 35)

DONG, L., MALLINSON, J., REDDY, S. et LAPATA, M. (2017). Learning to paraphrase for question answering. *In* PALMER, M., HWA, R. et RIEDEL, S., éditeurs : *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 875–886. Association for Computational Linguistics. (cité à la page : 53)

DORR, B. J. (1994). Machine Translation Divergences : A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597–633. (cité aux pages : 22, 115)

DRAS, M. (1997). Representing Paraphrases Using Synchronous TAGs. *In 35th Annual Meeting of the Association for Computational Linguistics*. (cité à la page : 33)

DRAS, M. (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Thèse de doctorat, Macquarie University NSW 2109 Australia. (cité aux pages : 30, 31, 33, and 34)

DUCLAYE, F., YVON, F. et COLLIN, O. (2003). Learning paraphrases to improve a question-answering system. *In Proceedings of the EACL Workshop on Natural Language Processing for Question Answering Systems*, pages 35–41. (cité à la page : 53)

DYER, C., CHAHUNEAU, V. et SMITH, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. *In* VANDERWENDE, L., III, H. D. et KIRCHHOFF, K., éditeurs : *Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648. The Association for Computational Linguistics. (cité aux pages : 70, 124)

EISELE, A. et CHEN, Y. (2010). MultiUN : A Multilingual Corpus from United Nation Documents. *In* CALZOLARI, N., CHOUKRI, K., MAEGAARD, B., MARIANI, J., ODIJK, J., PIPERIDIS, S., ROSNER, M. et TAPIAS, D., éditeurs : *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association. (cité à la page : 60)

ELHADAD, N. et SUTARIA, K. (2007). Mining a lexicon of technical terms and lay equivalents. *In Proceedings of the Workshop on BioNLP 2007 : Biological, Translational, and Clinical Language Processing*, pages 49–56. Association for Computational Linguistics. (cité à la page : 53)

ELLIS, N. C. (2008). Implicit and explicit knowledge about language. *Encyclopedia of language and education*, pages 1878–1890. (cité à la page : 118)

ESHKOL-TARAVELLA, I. et GRABAR, N. (2014). Repérage et analyse de la reformulation paraphrastique dans les corpus oraux. *In TALN2014*, pages 304–315. (cité à la page : 117)

FARUQUI, M., DODGE, J., JAUHAR, S. K., DYER, C., HOVY, E. H. et SMITH, N. A. (2015). Retrofitting word vectors to semantic lexicons. *In* MIHALCEA, R., CHAI, J. Y. et SARKAR, A., éditeurs : *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 1606–1615. The Association for Computational Linguistics. (cité aux pages : 53, 91)

FEDERICO, M., BENTIVOGLI, L., PAUL, M. et STÜKER, S. (2011). Overview of the IWSLT 2011 Evaluation Campaign. *In International Workshop on Spoken Language Translation (IWSLT)*. (cité à la page : 61)

FERNÁNDEZ GUERRA, A. (2012). Translating culture : Problems, strategies and practical realities. *SIC-Journal of Literature, Cultural and Literary Translation. Broj 1, Godina 3 : Art and Subversion.*, 3. (cité à la page : 10)

FERRARESI, A., ZANCHETTA, E., BARONI, M. et BERNARDINI, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. (cité à la page : 101)

FOMICHEVA, M., BEL, N., SPECIA, L., da CUNHA, I. et MALINOVSKIY, A. (2016). CobaltF : A fluent metric for MT evaluation. *In Proceedings of the First Conference on Machine Translation : Volume 2, Shared Task Papers*, pages 483–490, Berlin, Germany. Association for Computational Linguistics. (cité à la page : 115)

FRANCOIS, T., GALA, N., WATRIN, P. et FAIRON, C. (2014). FLELex : a graded Lexical Resource for French Foreign Learners. *In* CHAIR), N. C. C., CHOUKRI, K., DECLERCK, T., LOFTSSON, H., MAEGAARD, B., MARIANI, J., MORENO, A., ODIJK, J. et PIPERIDIS, S., éditeurs : *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA). (cité à la page : 125)

FRASER, A. et MARCU, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303. (cité aux pages : 72, 114)

FROELIGER, N. (2008). Le problème de la nuance en traduction pragmatique. *Traduire. Revue française de la traduction*, (218):77–93. (cité à la page : 2)

FUCHS, C. (1982). *La paraphrase*. Presses universitaires de France. (cité aux pages : 30, 31)

FUCHS, C. (1994). *Paraphrase et énonciation*. Editions Ophrys. (cité aux pages : 30, 32)

FUJITA, A. (2005). *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Thèse de doctorat, Ph. D. thesis, Nara Institute of Science and Technology. (cité aux pages : 30, 31)

GANITKEVITCH, J. et CALLISON-BURCH, C. (2014). The multilingual paraphrase database. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014.*, pages 4276–4283. (cité aux pages : 3, 45, 46, and 111)

GANITKEVITCH, J., CALLISON-BURCH, C., NAPOLES, C. et DURME, B. V. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1168–1179. ACL. (cité aux pages : 40, 41, 42, 43, 44, 111, and 112)

GANITKEVITCH, J., DURME, B. V. et CALLISON-BURCH, C. (2012). Monolingual distributional similarity for text-to-text generation. *In* AGIRRE, E., BOS, J. et DIAB, M. T., éditeurs : *Proceedings of the First Joint Conference on Lexical and Computational Semantics, *SEM 2012, June 7-8, 2012, Montréal, Canada.*, pages 256–264. Association for Computational Linguistics. (cité aux pages : 42, 43, and 111)

GANITKEVITCH, J., VAN DURME, B. et CALLISON-BURCH, C. (2013). PPDB : The paraphrase database. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 758–764. (cité aux pages : 3, 43, 44, and 111)

GAO, Q. et VOGEL, S. (2008). Parallel implementations of word alignment tool. *In Software engineering, testing, and quality assurance for natural language processing*, pages 49–57. (cité à la page : 72)

GARDNER, M., GRUS, J., NEUMANN, M., TAFJORD, O., DASIGI, P., LIU, N. F., PETERS, M., SCHMITZ, M. et ZETTLEMOYER, L. (2018). AllenNLP : A deep semantic natural language processing platform. *In Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics. (cité à la page : 101)

GERMANN, U. (2008). Yawat : Yet Another Word Alignment Tool. *In ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Demo Papers*, pages 20–23. The Association for Computer Linguistics. (cité à la page : 70)

GIBOVÁ, K. (2012). *Translation Procedures in the Non-literary and Literary Text Compared*. Books on Demand. (cité à la page : 10)

GONG, L., MAX, A. et YVON, F. (2013). Improving bilingual sub-sentential alignment by sampling-based transpotting. *In Proceedings of IWSLT*, Heidelberg, Germany. (cité à la page : 124)

GRAY, R. M. (1990). *Entropy and Information Theory*. Springer-Verlag, Berlin, Heidelberg. (cité à la page : 92)

HAMMER, P. et MONOD, M. (1976). *English-French Cognate Dictionary*. (cité à la page : 89)

HARRIS, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162. (cité à la page : 33)

HARRIS, Z. S. (1957). Co-occurrence and transformation in linguistic structure. *Language*, 33(3):283–340. (cité à la page : 30)

HASSAN, S., CSOMAI, A., BANEA, C., SINHA, R. et MIHALCEA, R. (2007). Unt : Subfinder : Combining knowledge sources for automatic lexical substitution. *In Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410–413. (cité à la page : 51)

HE, K., ZHANG, X., REN, S. et SUN, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916. (cité à la page : 97)

HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In 14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, August 23-28, 1992*, pages 539–545. (cité à la page : 33)

HO, W. Y., KNG, C., WANG, S. et BOND, F. (2014). Identifying idioms in Chinese translations. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 716–721, Reykjavik, Iceland. European Languages Resources Association (ELRA). (cité à la page : 86)

HOLMQVIST, M. et AHRENBERG, L. (2011). A gold standard for English-Swedish word alignment. *In Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 106–113, Riga, Latvia. Northern European Association for Language Technology (NEALT). (cité à la page : 72)

HONECK, R. P. (1971). A study of paraphrases. *Journal of Verbal Learning and Verbal Behavior*, 10(4):367–381. (cité à la page : 30)

HOVY, E. (1999). Toward finely differentiated evaluation metrics for machine translation. *Proceedings of the EAGLES Workshop on Standards and Evaluation Pisa, Italy, 1999*. (cité à la page : 114)

HUNSTON, S. et FRANCIS, G. (2000). *Pattern grammar : A corpus-driven approach to the lexical grammar of English*, volume 4. John Benjamins Publishing. (cité à la page : 90)

HUTCHINS, J. (1995). «The Whisky Was Invisible» or Persistent Myths of MT. *MT News International*. (cité à la page : 9)

IBRAHIM, A., KATZ, B. et LIN, J. (2003). Extracting Structural Paraphrases from Aligned Monolingual Corpora. *In Proceedings of the Second International Workshop on Paraphrasing, IWP@ACL 2003, Sapporo, Japan, July 11, 2003*. (cité aux pages : 33, 34)

ISABELLE, P., CHERRY, C. et FOSTER, G. F. (2017). A Challenge Set Approach to Evaluating Machine Translation. *In* PALMER, M., HWA, R. et RIEDEL, S., éditeurs :

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 2476–2486. Association for Computational Linguistics. (cité aux pages : 61, 115, 116, 117, and 144)

ISABELLE, P. et KUHN, R. (2018). A challenge set for french -> english machine translation. *CoRR*, abs/1806.02725. (cité à la page : 116)

IYYER, M., WIETING, J., GIMPEL, K. et ZETTLEMOYER, L. (2018). Adversarial Example Generation with Syntactically Controlled Paraphrase Networks. *In* WALKER, M. A., JI, H. et STENT, A., éditeurs : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1875–1885. Association for Computational Linguistics. (cité à la page : 52)

KALCHBRENNER, N. et BLUNSOM, P. (2013). Recurrent continuous translation models. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709. (cité à la page : 115)

KAUCHAK, D. et BARZILAY, R. (2006). Paraphrasing for automatic evaluation. *In Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 455–462, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 53)

KINGSBURY, P. et PALMER, M. (2002). From TreeBank to PropBank. *In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA). (cité à la page : 44)

KLETZIEN, S. B. (2009). Paraphrasing : An effective comprehension strategy. *The Reading Teacher*, 63(1):73–77. (cité aux pages : 4, 118, and 120)

KOEHN, P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. *In Conference Proceedings : the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT. (cité aux pages : 37, 60)

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical MT. *In Proc. ACL :Systems Demos*, pages 177–180, Prague, Czech Republic. (cité à la page : 61)

KOEHN, P., OCH, F. J. et MARCU, D. (2003). Statistical phrase-based translation. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics. (cité aux pages : 36, 40, 45, 53, and 92)

KOK, S. et BROCKETT, C. (2010). Hitting the Right Paraphrases in Good Time. *In Human Language Technologies : Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 145–153. (cité aux pages : 43, 112)

KOZLOWSKI, R., MCCOY, K. F. et VIJAY-SHANKER, K. (2003). Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. *In Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 1–8. Association for Computational Linguistics. (cité aux pages : 31, 33)

LANDIS, J. R. et KOCH, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174. (cité à la page : 74)

LANGLAIS, P., SIMARD, M. et VÉRONIS, J. (1998). Methods and practical issues in evaluating alignment techniques. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98/COLING '98, pages 711–717, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 113)

LEGRAND, J., AULI, M. et COLLOBERT, R. (2016). Neural Network-based Word Alignment through Score Aggregation. *In Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 66–73. The Association for Computer Linguistics. (cité à la page : 97)

LEMAIRE, C. (2017). *Traductologie et traduction outillée : du traducteur spécialisé professionnel à l'expert métier en entreprise*. Thèse de doctorat. (cité aux pages : 2, 54)

LEUNG, H., POIRET, R., WONG, T., CHEN, X., GERDES, K. et LEE, J. (2016). Developing universal dependencies for mandarin chinese. *In* HASIDA, K., WONG, K., CALZORARI, N. et CHOI, K., éditeurs : *Proceedings of the 12th Workshop on Asian Language Resources, ALR@COLING 2016, Osaka, Japan, December 12, 2016*, pages 20–29. The COLING 2016 Organizing Committee. (cité à la page : 89)

LEVENSHTEIN, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710. (cité à la page : 89)

LI, X., STRASSEL, S., GRIMES, S., ISMAEL, S., MAAMOURI, M., BIES, A. et XUE, N. (2012). Parallel aligned treebanks at LDC : New challenges interfacing existing infrastructures. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1848–1855, Istanbul, Turkey. European Languages Resources Association (ELRA). (cité à la page : 22)

LI, Z., CALLISON-BURCH, C., DYER, C., GANITKEVITCH, J., IRVINE, A., KHUDANPUR, S., SCHWARTZ, L., THORNTON, W., WANG, Z., WEESE, J. et ZAIDAN, O. (2010). Joshua 2.0 : A toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. *In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 133–137, Uppsala, Sweden. Association for Computational Linguistics. (cité à la page : 41)

LI, Z., JIANG, X., SHANG, L. et LI, H. (2018). Paraphrase generation with deep reinforcement learning. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. Association for Computational Linguistics. (cité à la page : 52)

LI, Z. et SUN, M. (2009). Punctuation As Implicit Annotations for Chinese Word Segmentation. *Computational Linguistics*, 35(4):505–512. (cité à la page : 81)

LIAN, S. (2006). *A Course Book on English-Chinese Translation*. Beijing Higher Education Press. (cité à la page : 21)

LIANG, P., TASKAR, B. et KLEIN, D. (2006). Alignment by agreement. *In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111. Association for Computational Linguistics. (cité à la page : 92)

LIN, D. (1998). Automatic retrieval and clustering of similar words. *In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98/COLING '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 49)

LIN, D., CHURCH, K., JI, H., SEKINE, S., YAROWSKY, D., BERGSMA, S., PATIL, K., PITLER, E., LATHBURY, R., RAO, V., DALWANI, K. et NARSALE, S. (2010). New tools for web-scale n-grams. *In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Languages Resources Association (ELRA). (cité aux pages : 42, 44)

LIN, D. et PANTEL, P. (2001). DIRT – Discovery of Inference Rules from Text. *In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 323–328, New York, NY, USA. ACM. (cité aux pages : 33, 34, and 47)

LISON, P. et TIEDEMANN, J. (2016). Opensubtitles2016 : Extracting large parallel corpora from movie and tv subtitles. (cité à la page : 60)

LIU, Y. et SUN, M. (2015). Contrastive unsupervised word alignment with non-local features. *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2295–2301. AAAI Press. (cité aux pages : 81, 82)

LOPEZ, A. et RESNIK, P. (2006). Word-based alignment, phrase-based translation : What's the link. *In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 90–99. (cité à la page : 114)

LOSNEGAARD, G. S., SANGATI, F., PARRA ESCARTÍN, C., SAVARY, A., BARGMANN, S. et MONTI, J. (2016). PARSEME Survey on MWE Resources. *In 9th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2299–2306, Portorož, Slovenia. (cité à la page : 61)

LU, W. et FANG, H. (2012). Reconsidering Peter Newmark's Theory on Literal Translation. *Theory & Practice in Language Studies*, 2(4). (cité aux pages : 18, 19)

LUONG, M.-T., PHAM, H. et MANNING, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv :1508.04025*. (cité à la page : 26)

MA, S., SUN, X., LI, W., LI, S., LI, W. et REN, X. (2018). Query and output : Generating words by querying distributed word representations for paraphrase generation. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*,

pages 196–206, New Orleans, Louisiana. Association for Computational Linguistics. (cité à la page : 52)

MACCARTNEY, B. (2009). *Natural language inference*. Citeseer. (cité à la page : 48)

MACCARTNEY, B. et MANNING, C. D. (2007). Natural logic for textual inference. *In Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague. Association for Computational Linguistics. (cité à la page : 99)

MADNANI, N., AYAN, N. F., RESNIK, P. et DORR, B. J. (2007). Using paraphrases for parameter tuning in statistical machine translation. *In Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 120–127, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 53)

MADNANI, N. et DORR, B. J. (2010). Generating phrasal and sentential paraphrases : A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387. (cité aux pages : 25, 30, and 52)

MAHMOUD, A. (2006). Translation and foreign language reading comprehension : A neglected didactic procedure. *In English Teaching Forum*, volume 44, page 28. ERIC. (cité à la page : 119)

MALAKASIOTIS, P. (2011). Paraphrase and textual entailment recognition and generation. *PhD thesis, Department of Informatics, Athens University of Economics and Business*. (cité à la page : 30)

MALLINSON, J., SENNRICH, R. et LAPATA, M. (2017). Paraphrasing revisited with neural machine translation. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, volume 1, pages 881–893. (cité à la page : 43)

MANBER, U. et MYERS, G. (1993). Suffix arrays : a new method for on-line string searches. *siam Journal on Computing*, 22(5):935–948. (cité à la page : 131)

MANNING, C. D., SURDEANU, M., BAUER, J., FINKEL, J., BETHARD, S. J. et MCCLOSKY, D. (2014). The Stanford CoreNLP natural language processing toolkit. *In Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. (cité à la page : 88)

MARCUS, M. P., MARCINKIEWICZ, M. A. et SANTORINI, B. (1993). Building a large annotated corpus of english : The penn treebank. *Comput. Linguist.*, 19(2):313–330. (cité aux pages : 44, 61)

MARELLI, M., BENTIVOGLI, L., BARONI, M., BERNARDI, R., MENINI, S. et ZAMPARELLI, R. (2014). SemEval-2014 task 1 : Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics. (cité à la page : 48)

MARGOT, J. C. (1979). *Traduire sans trahir : la théorie de la traduction et son application aux textes bibliques*. Age d'homme. (cité aux pages : 14, 15)

MARTIN, J., MIHALCEA, R. et PEDERSEN, T. (2005). Word alignment for languages with scarce resources. *In Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 65–74. (cité à la page : 113)

MARTIN, R. (1976). Inférence, antonymie et paraphrase éléments pour une théorie sémantique. (cité aux pages : 30, 31)

MARTINOT, C. (2012). De la reformulation en langue naturelle, vers son exploitation pédagogique en langue étrangère : pour une optimisation des stratégies d'apprentissage. (cité aux pages : 4, 117, and 120)

MARTON, Y., CALLISON-BURCH, C. et RESNIK, P. (2009). Improved statistical machine translation using monolingually-derived paraphrases. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Singapore. Association for Computational Linguistics. (cité à la page : 53)

MAX, A. (2008). Génération de reformulations locales par pivot pour l'aide à la révision. *In Actes de TALN*. (cité à la page : 51)

MAX, A. (2010). Example-based paraphrasing for improved phrase-based statistical machine translation. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 656–666. (cité à la page : 31)

MCKEOWN, K. R. (1979). Paraphrasing using given and new information in a question-answer system. *In 17th Annual Meeting of the Association for Computational Linguistics*, pages 67–72. (cité aux pages : 33, 53)

MCKEOWN, K. R. (1983). Paraphrasing questions using given and new information. *Computational Linguistics*, 9(1):1–10. (cité à la page : 51)

MEHDAD, Y., NEGRI, M. et FEDERICO, M. (2010). Towards cross-lingual textual entailment. *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324, Los Angeles, California. Association for Computational Linguistics. (cité à la page : 87)

MELAMED, I. D. (1998). Manual annotation of translational equivalence : The blinker project. *Technical Report IRCS TR*. (cité à la page : 72)

MELAMUD, O., GOLDBERGER, J. et DAGAN, I. (2016). context2vec : Learning generic context embedding with bidirectional LSTM. *In* GOLDBERG, Y. et RIEZLER, S., éditeurs : *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 51–61. ACL. (cité aux pages : 99, 100)

MEL'CUK, I. (1988). Paraphrase et lexique dans la théorie linguistique sens-texte. *Lexique*, (6):13–54. (cité à la page : 30)

MEL'CUK, I. (1992). Paraphrase et lexique : la théorie sens-texte et le dictionnaire explicatif et combinatoire. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques III. Les Presses de l'Université de Montréal*, pages 9–59. (cité à la page : 30)

METEER, M. et SHAKED, V. (1988). Strategies for effective paraphrasing. *In Proceedings of the 12th Conference on Computational Linguistics - Volume 2*, COLING '88, pages 431–436, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 33)

MIHALCEA, R. et PEDERSEN, T. (2003a). An evaluation exercise for word alignment. *In Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts : data driven machine translation and beyond*, pages 1–10. (cité à la page : 72)

MIHALCEA, R. et PEDERSEN, T. (2003b). An evaluation exercise for word alignment. *In Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts : data driven machine translation and beyond*, pages 1–10. (cité à la page : 113)

MIHÁLTZ, M. (2005). Towards a hybrid approach to word-sense disambiguation in machine translation. Citeseer. (cité à la page : 9)

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S. et DEAN, J. (2013). Distributed representations of words and phrases and their compositionality. *In* BURGES, C. J. C., BOTTOU, L., GHAHRAMANI, Z. et WEINBERGER, K. Q., éditeurs : *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119. (cité à la page : 91)

MILIĆEVIĆ, J. (2007). *La paraphrase : Modélisation de la paraphrase langagière*, volume 80. Peter Lang. (cité à la page : 30)

MILLER, G. A. (1995). Wordnet : A lexical database for english. *Commun. ACM*, 38(11): 39–41. (cité à la page : 37)

MIZUKAMI, M., NEUBIG, G., SAKTI, S., TODA, T. et NAKAMURA, S. (2014). Building a free, general-domain paraphrase database for Japanese. *In 2014 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, pages 1–4. (cité aux pages : 44, 45)

MOLDOVAN, D. et RUS, V. (2001). Logic form transformation of wordnet and its applicability to question answering. *In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 402–409. (cité à la page : 53)

MOLINA, L. et HURTADO ALBIR, A. (2002). Translation Techniques Revisited : A Dynamic and Functionalist Approach. *Meta*, 47(4):498–512. (cité aux pages : 2, 10, 14, 16, 17, 62, and 141)

MONTI, J., SANGATI, F. et ARCAN, M. (2015). TED-MWE : a bilingual parallel corpus with MWE annotation. Towards a methodology for annotating MWEs in parallel multilingual corpora. *In Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it*, pages 193–197, Trento. (cité à la page : 61)

MUNDAY, J. (2016). *Introducing Translation Studies : Theories and Applications*. Taylor & Francis. (cité à la page : 18)

MURAKI, K. (1982). On a semantic model for multi-lingual paraphrasing. *In Proceedings of the 9th Conference on Computational Linguistics - Volume 1*, COLING '82, pages 239–244, Czechoslovakia. Academia Praha. (cité à la page : 33)

NÁDVORNÍKOVÁ, O. (2017). Pièges méthodologiques des corpus parallèles et comment les éviter. *Corela. Cognition, représentation, langage*, (HS-21). (cité à la page : 71)

NAPOLES, C., GORMLEY, M. et VAN DURME, B. (2012). Annotated gigaword. *In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX '12, pages 95–100, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité aux pages : 44, 49)

NATION, P. (2003). The role of the first language in foreign language learning. *Asian EFL*, 5(2):1–8. (cité à la page : 119)

NEGRI, M., MARCHETTI, A., MEHDAD, Y., BENTIVOGLI, L. et GIAMPICCOLO, D. (2012). Semeval-2012 task 8 : Cross-lingual textual entailment for content synchronization. *In *SEM 2012 : The First Joint Conference on Lexical and Computational Semantics – Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 399–407, Montréal, Canada. Association for Computational Linguistics. (cité à la page : 87)

NEGRI, M., MARCHETTI, A., MEHDAD, Y., BENTIVOGLI, L. et GIAMPICCOLO, D. (2013). Semeval-2013 task 8 : Cross-lingual textual entailment for content synchronization. *In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 25–33, Atlanta, Georgia, USA. Association for Computational Linguistics. (cité à la page : 87)

NEUBIG, G., SUDOH, K., ODA, Y., DUH, K., TSUKADA, H. et NAGATA, M. (2014). The NAIST-NTT TED Talk Treebank. *In 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA. (cité à la page : 61)

NEWMARK, P. (1981). *Approaches to Translation (Language Teaching Methodology Senes)*. Oxford : Pergamon Press. (cité aux pages : 2, 10)

NEWMARK, P. (1988). *A textbook of translation*, volume 66. Prentice Hall New York. (cité aux pages : 10, 14, 15, and 18)

NEWMARK, P. (2001). *Approaches to translation*. Shanghai Foreign Language Education Press. (cité à la page : 21)

NIDA, E. A. (1964). *Toward a Science of Translating : With Special Reference to Principles and Procedures Involved in Bible Translating*. E.J. Brill. (cité aux pages : 14, 18)

NIDA, E. A. et TABER, C. R. (1969). *The Theory and Practice of Translation*. Brill. (cité à la page : 14)

NORD, C. (1997). *Translating as a Purposeful Activity : Functionalist Approaches Explained*. Translation theories explained. St. Jerome Pub. (cité à la page : 18)

OCH, F. J. (2003). Minimum error rate training in statistical machine translation. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 40)

OCH, F. J. et NEY, H. (2000). A comparison of alignment models for statistical machine translation. *In Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 1086–1090, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité aux pages : 72, 113)

OCH, F. J. et NEY, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51. (cité aux pages : 36, 37, and 70)

OCH, F. J. et NEY, H. (2004). The alignment template approach to statistical machine translation. *Comput. Linguist.*, 30(4):417–449. (cité aux pages : 36, 38, and 40)

ĐORĐEVIĆ, J. (2017). Translation techniques revisited : the applicability of existing solutions in non-literary translation. *FACTA UNIVERSITATIS-Linguistics and Literature*, 15(1):35–47. (cité aux pages : 18, 19, and 141)

PADÓ, S., GALLEY, M., JURAFSKY, D. et MANNING, C. (2009a). Robust machine translation evaluation with entailment features. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 1 - Volume 1*, ACL '09, pages 297–305, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 87)

PADÓ, S., GALLEY, M., JURAFSKY, D. et MANNING, C. D. (2009b). Textual entailment features for machine translation evaluation. *In Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 37–41, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 87)

PANG, B., KNIGHT, K. et MARCU, D. (2003). Syntax-based Alignment of Multiple Translations : Extracting Paraphrases and Generating New Sentences. *In* HEARST, M. A. et OSTENDORF, M., éditeurs : *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*. The Association for Computational Linguistics. (cité aux pages : 33, 34, and 51)

PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics. (cité aux pages : 25, 40, and 114)

PARADIS, M. (2009). *Declarative and procedural determinants of second languages*, volume 40. John Benjamins Publishing. (cité à la page : 118)

PAVLICK, E., BOS, J., NISSIM, M., BELLER, C., VAN DURME, B. et CALLISON-BURCH, C. (2015a). Adding semantics to data-driven paraphrasing. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, volume 1, pages 1512–1522. (cité aux pages : 3, 46, 47, 48, 49, 50, 54, 99, 111, 112, and 143)

PAVLICK, E. et NENKOVA, A. (2015). Inducing lexical style properties for paraphrase and genre differentiation. *In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 218–224, Denver, Colorado. Association for Computational Linguistics. (cité à la page : 46)

PAVLICK, E., RASTOGI, P., GANITKEVITCH, J., DURME, B. V. et CALLISON-BURCH, C. (2015b). PPDB 2.0 : Better Paraphrase Ranking, Fine-Grained Entailment Relations, Word Embeddings, and Style Classification. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2 : Short Papers*, pages 425–430. (cité aux pages : 3, 46, 50, 54, 100, 111, and 132)

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. et DUCHESNAY, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. (cité à la page : 89)

PENNINGTON, J., SOCHER, R. et MANNING, C. (2014). Glove : Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. (cité aux pages : 91, 100)

PETERS, M. E., NEUMANN, M., IYYER, M., GARDNER, M., CLARK, C., LEE, K. et ZETTLEMOYER, L. (2018). Deep contextualized word representations. *In* WALKER, M. A., JI, H. et STENT, A., éditeurs : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics. (cité aux pages : 99, 100, and 101)

PETROV, S., BARRETT, L., THIBAUX, R. et KLEIN, D. (2006). Learning accurate, compact, and interpretable tree annotation. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia. Association for Computational Linguistics. (cité à la page : 45)

PETROV, S., DAS, D. et MCDONALD, R. T. (2012). A universal part-of-speech tagset. *In* CALZOLARI, N., CHOUKRI, K., DECLERCK, T., DOGAN, M. U., MAEGAARD, B., MARIANI, J., ODIJK, J. et PIPERIDIS, S., éditeurs : *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2089–2096. European Language Resources Association (ELRA). (cité à la page : 89)

PHAM, M. Q., CREGO, J., SENELLART, J. et YVON, F. (2018). Fixing translation divergences in parallel corpora for neural mt. *In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2967–2973. (cité aux pages : 87, 97)

POIRIER, É. A. (2014). A method for automatic detection and manual localization of content-based translation errors and shifts. *Journal of Innovation in Digital Ecosystems*, 1(1-2):38–46. (cité à la page : 86)

PRAKASH, A., HASAN, S. A., LEE, K., DATLA, V. V., QADIR, A., LIU, J. et FARRI, O. (2016). Neural paraphrase generation with stacked residual LSTM networks. *In* CALZOLARI, N., MATSUMOTO, Y. et PRASAD, R., éditeurs : *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2923–2934. ACL. (cité à la page : 52)

PUSTEJOVSKY, J. et STUBBS, A. (2012). *Natural Language Annotation for Machine Learning : A guide to corpus-building for applications*. " O'Reilly Media, Inc.". (cité à la page : 137)

QUIRK, C., BROCKETT, C. et DOLAN, W. (2004). Monolingual machine translation for paraphrase generation. *In Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 142–149. (cité aux pages : 30, 51)

RASTOGI, P., DURME, B. V. et ARORA, R. (2015). Multiview LSA : representation learning via generalized CCA. *In* MIHALCEA, R., CHAI, J. Y. et SARKAR, A., éditeurs : *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 556–566. The Association for Computational Linguistics. (cité aux pages : 46, 53)

RAVICHANDRAN, D. et HOVY, E. H. (2002). Learning Surface Text Patterns for a Question Answering System. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 41–47. ACL. (cité à la page : 53)

REEDER, F. (2001). Additional mt-eval references. *International Standards for Language Engineering, Evaluation Working Group*. (cité à la page : 114)

ROSSARI, C. (1994). *Les opérations de reformulation : analyse du processus et des marques dans une perspective contrastive français-italien*. Sciences pour la communication. Peter Lang. (cité aux pages : 4, 117)

RUVIDIC, I. (2006). *La traduction de la poésie chinoise classique : des pièges théoriques aux obstacles de la pratique*. Thèse de doctorat, Paris 7. (cité à la page : 2)

SAVARY, A., SAILER, M., PARMENTIER, Y., ROSNER, M., ROSÉN, V., PRZEPIÓRKOWSKI, A., KRSTEV, C., VINCZE, V., WÓJTOWICZ, B., LOSNEGAARD, G. S., PARRA ESCARTÍN, C., WASZCZUK, J., CONSTANT, M., OSENOVA, P. et SANGATI, F. (2015). PARSEME – PARSing and Multiword Expressions within a European multilingual network. *In 7th Language & Technology Conference : Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland. Available at `https://hal.archives-ouvertes.fr/hal-01223349/document`. (cité à la page : 61)

SCHMID, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50. (cité à la page : 89)

SENNRICH, R., FIRAT, O., CHO, K., BIRCH, A., HADDOW, B., HITSCHLER, J., JUNCZYS-DOWMUNT, M., LÄUBLI, S., MICELI BARONE, A. V., MOKRY, J. et NADEJDE, M. (2017). Nematus : a toolkit for neural machine translation. *In Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics. (cité à la page : 26)

SHEIKHSHABBAFGHI, G., BIROL, I. et SARKAR, A. (2018). In-domain context-aware token embeddings improve biomedical named entity recognition. *In Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 160–164, Brussels, Belgium. Association for Computational Linguistics. (cité à la page : 101)

SHEMTOV, H. (1996). Generation of paraphrases from ambiguous logical forms. *In COLING 1996 Volume 2 : The 16th International Conference on Computational Linguistics*. (cité à la page : 33)

SHIEBER, S. M. et SCHABES, Y. (1990). Generation and synchronous tree-adjoining grammars. *In Proceedings of the Fifth International Workshop on Natural Language Generation*. (cité à la page : 33)

SHINYAMA, Y., SEKINE, S. et SUDO, K. (2002). Automatic paraphrase acquisition from news articles. *In Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 313–318, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. (cité à la page : 32)

SHIYAB, S. et ABDULLATEEF, M. (2001). Translation and foreign language teaching. *Journal of King Saud University Language & Translation*, 13:1–9. (cité à la page : 119)

SHOJAEI, A. (2012). Translation of idioms and fixed expressions : strategies and difficulties. *Theory and Practice in Language Studies*, 2(6):1220–1229. (cité à la page : 10)

SHWARTZ, V. et DAGAN, I. (2016). Adding Context to Semantic Data-Driven Paraphrasing. *In Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics, *SEM@ACL 2016, Berlin, Germany, 11-12 August 2016*, pages 108–113. The *SEM 2016 Organizing Committee. (cité aux pages : 99, 100, 101, and 102)

SNELL-HORNBY, M. (1995). *Translation studies : An integrated approach (Revised Edition)*. John Benjamins Publishing (First edition in 1988). (cité à la page : 19)

SNOVER, M., DORR, B., SCHWARTZ, R., MICCIULLA, L. et MAKHOUL, J. (2006). A study of translation edit rate with targeted human annotation. *In Proceedings of association for machine translation in the Americas*, volume 200. (cité aux pages : 25, 115)

SNOVER, M. G., MADNANI, N., DORR, B. et SCHWARTZ, R. (2009). TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3):117–127. (cité aux pages : 53, 115)

SNOW, R., JURAFSKY, D. et NG, A. Y. (2004). Learning syntactic patterns for automatic hypernym discovery. *In Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1297–1304. (cité aux pages : 33, 49)

SPECIA, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. *In Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80. (cité à la page : 86)

SPECIA, L., BLAIN, F., LOGACHEVA, V., ASTUDILLO, R. F. et MARTINS, A. F. T. (2018). Findings of the WMT 2018 shared task on quality estimation. *In* BOJAR, O., CHATTERJEE, R., FEDERMANN, C., FISHEL, M., GRAHAM, Y., HADDOW, B., HUCK, M., JIMENO-YEPES, A., KOEHN, P., MONZ, C., NEGRI, M., NÉVÉOL, A., NEVES, M. L., POST, M., SPECIA, L., TURCHI, M. et VERSPOOR, K., éditeurs : *Proceedings of the Third Conference on Machine Translation : Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 689–709. Association for Computational Linguistics. (cité à la page : 87)

SPECIA, L., CANCEDDA, N., DYMETMAN, M., TURCHI, M. et CRISTIANINI, N. (2009). Estimating the sentence-level quality of machine translation systems. *In EAMT09*, pages 28–37. (cité à la page : 87)

SPECIA, L., RAJ, D. et TURCHI, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50. (cité à la page : 86)

SPEER, R., CHIN, J. et HAVASI, C. (2017). Conceptnet 5.5 : An open multilingual graph of general knowledge. *In Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451. (cité à la page : 91)

SPEER, R. et LOWRY-DUDA, J. (2017). Conceptnet at semeval-2017 task 2 : Extending word embeddings with multilingual relational knowledge. *In* BETHARD, S., CARPUAT, M., APIDIANAKI, M., MOHAMMAD, S. M., CER, D. M. et JURGENS, D., éditeurs : *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 85–89. Association for Computational Linguistics. (cité à la page : 91)

STEEDMAN, M. et BALDRIDGE, J. (2011). Combinatory categorial grammar. *Non-Transformational Syntax : Formal and Explicit Models of Grammar*, pages 181–224. (cité à la page : 38)

SULISTYO, T. et HERIYAWATI, D. (2017). Reformulation, text modeling, and the development of EFL academic writing. *Journal on English as a Foreign Language*, 7:1. (cité à la page : 117)

SULTAN, M. A., BETHARD, S. et SUMNER, T. (2014). Back to basics for monolingual alignment : Exploiting word similarity and contextual evidence. *TACL*, 2:219–230. (cité à la page : 53)

SUTSKEVER, I., VINYALS, O. et LE, Q. V. (2014). Sequence to sequence learning with neural networks. *In Advances in neural information processing systems*, pages 3104–3112. (cité aux pages : 52, 115)

SZPEKTOR, I. et DAGAN, I. (2008). Learning entailment rules for unary templates. *In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 849–856, Manchester, UK. Coling 2008 Organizing Committee. (cité à la page : 49)

SZPEKTOR, I., SHNARCH, E. et DAGAN, I. (2007). Instance-based evaluation of entailment rule acquisition. *In ACL*. (cité à la page : 39)

TAN, L. et BOND, F. (2011). Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 362–371, Singapore. Institute of Digital Enhancement of Cognitive Processing, Waseda University. (cité à la page : 86)

TANG, D., WEI, F., YANG, N., ZHOU, M., LIU, T. et QIN, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics. (cité à la page : 100)

TIAN, L., WONG, D. F., CHAO, L. S., QUARESMA, P., OLIVEIRA, F., LU, Y., LI, S., WANG, Y. et WANG, L. (2014). UM-corpus : A large English-Chinese parallel corpus for statistical machine translation. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842, Reykjavik, Iceland. European Language Resources Association (ELRA). (cité à la page : 80)

TIEDEMANN, J. (2012). Parallel data, tools and interfaces in OPUS. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Languages Resources Association (ELRA). (cité à la page : 60)

TSOUMAKAS, G., KATAKIS, I. et VLAHAVAS, I. (2011). Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7): 1079–1089. (cité à la page : 93)

UPADHYAY, S., VYAS, Y., CARPUAT, M. et ROTH, D. (2018). Robust cross-lingual hypernymy detection using dependency context. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Association for Computational Linguistics. (cité aux pages : 88, 107)

VAN BENTHEM, J. (1991). Language in action. *Journal of philosophical logic*, 20(3): 225–263. (cité à la page : 47)

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. et POLOSUKHIN, I. (2017). Attention is all you need. *In Advances in neural information processing systems*, pages 5998–6008. (cité à la page : 115)

VÁZQUEZ-AYORA, G. (1977). *Introducción a la traductología : curso básico de traducción*. Georgetown University School of languages and linguistics. Georgetown University Press. (cité aux pages : 14, 15, and 16)

VENUTI, L. (2004). *The Translator's Invisibility*. Shanghai Foreign Language Education Press. (cité à la page : 21)

VERMEER, H. (1987). *A Frame Work for a General Theory of Translation*. Heidelberg University Press. (cité à la page : 18)

VERMES, A. (2010). Translation in foreign language teaching : a brief overview of pros and cons. *Eger Journal of English Studies*, 10(1):83–93. (cité à la page : 119)

VILA, M., MARTÍ, M. A. et RODRÍGUEZ, H. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90. (cité à la page : 31)

VILA, M., MARTÍ, M. A. et RODRÍGUEZ, H. (2014). Is this a paraphrase ? What kind ? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(01):205. (cité à la page : 32)

VILAR, D., POPOVIĆ, M. et NEY, H. (2006). Aer : Do we need to "improve" our alignments ? *In International Workshop on Spoken Language Translation (IWSLT) 2006*. (cité à la page : 113)

VINAY, J.-P. et DARBELNET, J. (1958). *Stylistique comparée du français et de l'anglais : méthode de traduction*. Bibliothèque de stylistique comparée. Didier. (cité aux pages : 2, 10, 11, 12, 13, 14, 15, 16, 17, 18, 62, 71, 115, and 141)

VYAS, Y. et CARPUAT, M. (2017). Detecting asymmetric semantic relations in context : A case-study on hypernymy detection. *In Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 33–43. (cité aux pages : 99, 100, 101, 102, and 104)

VYAS, Y., NIU, X. et CARPUAT, M. (2018). Identifying Semantic Divergences in Parallel Text without Annotations. *In* WALKER, M. A., JI, H. et STENT, A., éditeurs : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1503–1515. Association for Computational Linguistics. (cité à la page : 87)

WANG, L. (2017). On Strategies of Non-equivalence in English-Chinese Translation. *Theory and Practice in Language Studies*, 7(12):1295–1299. (cité à la page : 20)

WEEDS, J., WEIR, D. et MCCARTHY, D. (2004). Characterising measures of lexical distributional similarity. *In Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 49)

WEESE, J., GANITKEVITCH, J., CALLISON-BURCH, C., POST, M. et LOPEZ, A. (2011). Joshua 3.0 : Syntax-based machine translation with the thrax grammar extractor. *In Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 478–484, Edinburgh, Scotland. Association for Computational Linguistics. (cité à la page : 45)

WEINREICH, U. (1953). Languages in contact. *The Hague : Mouton*. (cité à la page : 118)

WHITE, J., O'CONNELL, T. et O'MARA, F. (1994). The arpa mt evaluation methodologies : evolution, lessons, and future approaches. *In Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*, pages 193–205. (cité à la page : 114)

WISNIEWSKI, G., SINGH, A. K. et YVON, F. (2013). Quality estimation for machine translation : Some lessons learned. *Machine translation*, 27(3-4):213–238. (cité à la page : 87)

WU, Y., SCHUSTER, M., CHEN, Z., LE, Q. V., NOROUZI, M., MACHEREY, W., KRIKUN, M., CAO, Y., GAO, Q., MACHEREY, K., KLINGNER, J., SHAH, A., JOHNSON, M., LIU, X., KAISER, L., GOUWS, S., KATO, Y., KUDO, T., KAZAWA, H., STEVENS, K., KURIAN, G., PATIL, N., WANG, W., YOUNG, C., SMITH, J., RIESA, J., RUDNICK, A., VINYALS, O., CORRADO, G., HUGHES, M. et DEAN, J. (2016). Google's neural machine translation system : Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144. (cité aux pages : 53, 115)

XU, W., RITTER, A., DOLAN, B., GRISHMAN, R. et CHERRY, C. (2012). Paraphrasing for style. *In Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee. (cité à la page : 53)

XU, Y. et YVON, F. (2016). Novel elicitation and annotation schemes for sentential and sub-sentential alignments of bitexts. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 628–635, Portorož, Slovenia. European Language Resources Association (ELRA). (cité aux pages : 72, 73, 84, 113, and 114)

YE, H. (2014). On foreignization of cultural elements in the translation of classical chinese poetry. *Theory & Practice in Language Studies*, 4(6). (cité à la page : 10)

YILMAZ GÜNGÖR, Z. (2015). La compréhension des textes en français langue étrangère : quelles difficultés ? *Journal of International Social Research*, 8(40). (cité à la page : 121)

YU, M. et DREDZE, M. (2014). Improving lexical embeddings with semantic knowledge. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 545–550, Baltimore, Maryland. Association for Computational Linguistics. (cité à la page : 53)

ZHAI, Y. (2018). Construction d'un corpus multilingue annoté en relations de traduction. *In Rencontre Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 85–99, Rennes, France. (cité à la page : 59)

ZHAI, Y., ILLOUZ, G. et VILNAT, A. (2019a). Classification automatique des procédés de traduction. *In 26th Conférence sur le Traitement Automatique des Langues Naturelles*, pages 205–214. (cité à la page : 85)

ZHAI, Y., ILLOUZ, G. et VILNAT, A. (2019b). Conception d'un outil d'aide à la compréhension écrite pour les apprenants de français langue étrangère. *In 9th Conférence sur les Environnements Informatiques pour l'Apprentissage Humain*, pages 379–382. (cité à la page : 117)

ZHAI, Y., MAX, A. et VILNAT, A. (2018). Construction of a Multilingual Corpus Annotated with Translation Relations. *In First Workshop on Linguistic Resources for Natural Language Processing*, pages 102–111. (cité à la page : 59)

ZHAI, Y., SAFARI, P., ILLOUZ, G., ALLAUZEN, A. et VILNAT, A. (2019c). Towards Recognizing Phrase Translation Processes : Experiments on English-French. *CoRR*, abs/1904.12213. (cité à la page : 85)

ZHANG, Y. et YAMAMOTO, K. (2002). Paraphrasing of chinese utterances. *In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics. (cité à la page : 44)

ZHANG, Y. et YAMAMOTO, K. (2005). Paraphrasing spoken chinese using a paraphrase corpus. *Nat. Lang. Eng.*, 11(4):417–434. (cité à la page : 44)

ZHAO, S., LAN, X., LIU, T. et LI, S. (2009). Application-driven statistical paraphrase generation. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 834–842, Suntec, Singapore. Association for Computational Linguistics. (cité aux pages : 30, 51)

ZHAO, S., NIU, C., ZHOU, M., LIU, T. et LI, S. (2008a). Combining multiple resources to improve SMT-based paraphrasing model. *In* MCKEOWN, K., MOORE, J. D., TEUFEL, S., ALLAN, J. et FURUI, S., éditeurs : *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 1021–1029. The Association for Computer Linguistics. (cité aux pages : 40, 51)

ZHAO, S., WANG, H., LIU, T. et LI, S. (2008b). Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. *In* MCKEOWN, K., MOORE, J. D., TEUFEL, S., ALLAN, J. et FURUI, S., éditeurs : *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 780–788. The Association for Computer Linguistics. (cité aux pages : 30, 40, and 51)

ZHOU, L., LIN, C.-Y., MUNTEANU, D. S. et HOVY, E. (2006). ParaEval : Using paraphrases to evaluate summaries automatically. *In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 447–454, New York City, USA. Association for Computational Linguistics. (cité à la page : 53)

ZIEMSKI, M., JUNCZYS-DOWMUNT, M. et POULIQUEN, B. (2016). The united nations parallel corpus v1.0. *In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA). (cité aux pages : 60, 81)

ZOLLMANN, A. et VENUGOPAL, A. (2006). Syntax augmented machine translation via chart parsing. *In Proceedings on the Workshop on Statistical Machine Translation*, pages 138–141, New York City. Association for Computational Linguistics. (cité à la page : 45)

Annexes

Annexe A

Expériences en compréhension écrite avec des étudiants chinois

Dans la section 7.2.4, nous avons décrit notre expérience sur la compréhension écrite menée avec la participation des étudiants chinois. Cette expérience a été réalisée sur le site de Moodle du LIMSI. La figure A.1 montre les consignes de l'expérience qui sont identiques pour les deux groupes d'étudiants (niveau A2 et B2). Pendant le troisième test, le texte est présenté avec des traductions en anglais ainsi que des paraphrases pour les segments soulignés, accessibles quand les étudiants posent leur souris sur ces segments (figure A.2, niveau B2). [1] L'interface pour les choix de réponses est présentée dans la figure A.3.

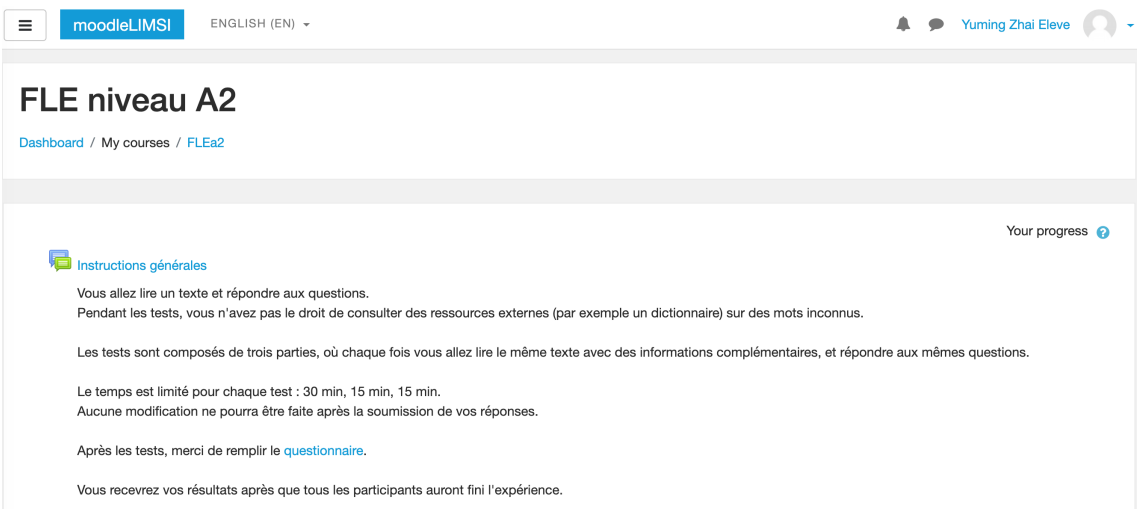

FIGURE A.1 – Consignes pour les participants

Nous présentons ci-dessous les textes, les questions et le questionnaire utilisés dans cette expérience.

Le texte du niveau A2 est issu d'un site de ressource pour l'éducation [2], et nous l'avons modifié légèrement. Celui du niveau B2 a été écrit par nous à partir d'une liste de mots

1. Lors du premier test, les étudiants lisent le texte original. Le deuxième test présente le texte avec seulement des paraphrases en tant qu'informations complémentaires.

2. `http://fr.hellokids.com/c_16133/lire-et-apprendre/reportages-pour-enfant/les-sciences/le-developpement-durable-explique-aux-enfants/le-rechauffement-climatique`

Version 3 : texte avec des paraphrases françaises et des traductions anglaises

De par le monde, il existe des tas d'endroits où l'on peut observer des phénomènes étonnants. Ainsi, j'a*over the world* **é de maisons** isolé, des **chaumières** qui semblaient **dé***from around the world* sonnes qui semblaient **inaptes au travail**. On y voyait un **iv**autour du monde vec **dans son sillage** certains de ses compagnons. Quand il autour du globe **ts** sur les **prouesses** accomplies par des personnages da dans le monde entier à ses amis. Ils **lui en voulaient** de les **effrayer** ainsi. **Il** à travers le monde insi dans un **perpétuel** ennui. Pour leur **sauver la mise**, il au aux quatre coins du monde **bras le corps** pour **remédier à** cette triste situation. Ils en étaient **tout bonnement** incapables, si quelqu'un ne venait pas **mettre au point** avec eux un moyen de s'en sortir. Mais **prôner** n'importe quel travail n'est pas une solution ! Un jour, un homme les entraîna pour travailler pour lui contre **une poignée de** pièces. Mais cette solution trop facile était un piège : cet homme était un voleur et il les incitait à jouer les **fouineurs** pour dévoiler des secrets. Ayant **bravé les interdits**, ces malheureux se sont trouvés en garde à vue. En plus, **dans la foulée**, leurs demeures ont été endommagées par **une pluie battante**. Mais **bon sang** ! Quelle tragédie !

FIGURE A.2 – Lors du troisième test, le texte est présenté avec des traductions anglaises et des paraphrases pour des segments soulignés

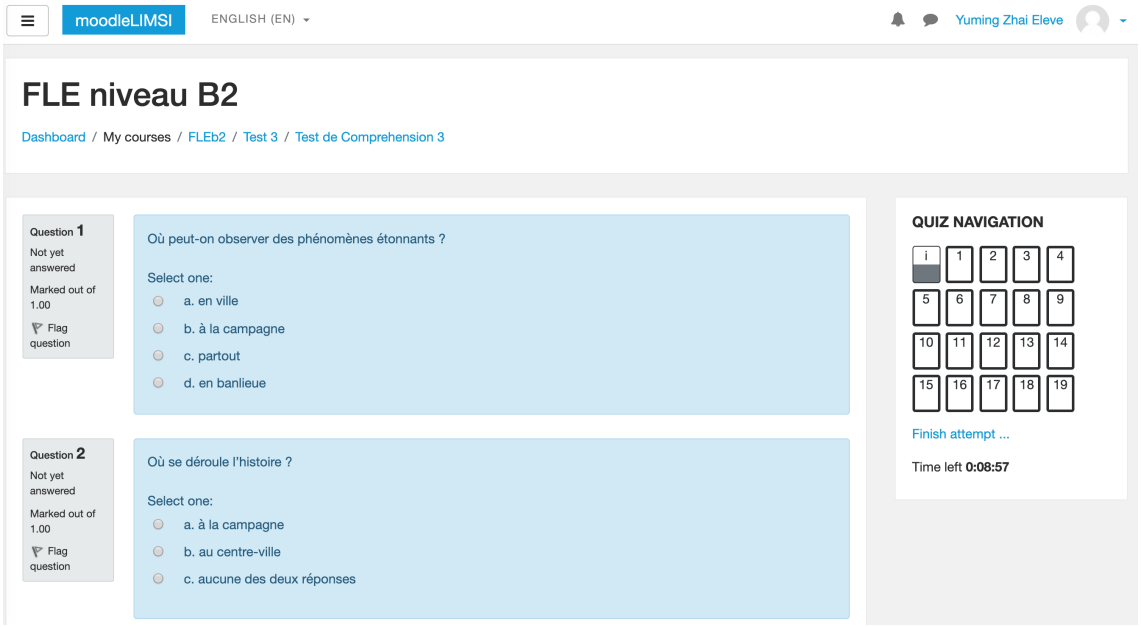

FIGURE A.3 – L'interface pour les choix de réponses

et segments. Nous testons la compréhension des étudiants sur des mots ou expressions en gras dans les textes. La bonne réponse pour chaque question est soulignée.

Le texte du niveau A2 :

LE RÉCHAUFFEMENT CLIMATIQUE

La température moyenne **ne cesse d'**augmenter sur la planète. Au rythme de sa progression, il fera bientôt beaucoup trop chaud sur la Terre pour que certaines **espèces** puissent **survivre**. Le réchauffement cause aussi la **fonte** des glaces et la montée des océans, ce qui multiplie le risque de catastrophes naturelles (tsunamis, inondations ...).

LES GAZ À EFFET DE SERRE

Pour bien comprendre ce qu'il se passe, il faut déjà comprendre ce qu'est l'effet de serre. La planète est en fait entourée d'une couche de gaz qui permet de retenir la chaleur du soleil. Elle permet de réchauffer la surface de la Terre. On les appelle les **gaz à effet de serre**. Si ce phénomène n'existait pas, il ne ferait que -18°C sur Terre ! D'un autre côté que se passerait-il d'après toi si la quantité de ces gaz augmentait fortement ? Et bien il ferait encore plus chaud sur la Terre ! Le problème du réchauffement climatique est que justement le volume des gaz à effet de serre est en trop forte augmentation.

POURQUOI ?

Ce réchauffement **est dû à** l'homme et à ses activités. Les sources d'énergie que nous utilisons (**pétrole**, **charbon**, gaz ...) **émettent** des gaz à effet de serre lorsqu'elles brûlent. Et nous en émettons beaucoup trop ! Les conséquences seront très graves d'ici 50 à 100 ans, entraînant un fort déséquilibre sur la planète.

LES SOLUTIONS

Pour **pallier** le réchauffement de la planète, il n'y a qu'un seul moyen : réduire notre consommation d'énergie émettant des gaz à effet de serre au profit d'autres énergies moins **nuisibles** pour la planète. On peut utiliser des technologies qui ne produisent pas de gaz, comme par exemple, des barrages hydrauliques, des panneaux solaires pour se chauffer, les **éoliennes**, *etc.*

LES BONS GESTES AU QUOTIDIEN

Nous pouvons tous faire un effort au quotidien pour **lutter contre** le réchauffement de la planète. Et si tout le monde **s'y met**, le résultat peut être phénoménal !

Voici 10 petits gestes que tu peux faire toi aussi :

- Ne pas laisser de lumière ni d'appareil électrique allumés inutilement.
- Baisser le chauffage.
- Ne pas **gaspiller** l'eau en laissant couler le robinet.
- N'utiliser l'eau chaude qu'en cas de réel besoin.
- Ne pas gaspiller le papier. Écris bien sur toute la feuille de papier avant d'en utiliser une autre.
- Acheter des produits qui respectent l'environnement.
- **Trier** les déchets.
- Ne pas jeter les **piles**, les **ampoules** et les médicaments avec les autres déchets.
- Pour un petit trajet, se déplacer à pied ou à vélo plutôt qu'en voiture.
- Préférer le train à l'avion pour les voyages si cela est possible.

Les paraphrases et traductions proposées sont montrées dans la figure A.4.

Les questions :

1. La température moyenne augmente :
   de temps en temps
   régulièrement

Original	Paraphrase	Traduction
ne cesse d'	sans cesse continuellement en permanence constamment	constantly keep
survivre	continuer à vivre toujours en vie résister à	survive
dû à	pour raison de pour cause de en raison de à cause de suite à provoqué par provient du	because of due to caused by
pallier	diminuer réduire mitiger atténuer résoudre redresser remédier à	reduce address mitigate alleviate redress
gaspiller	gâcher faire perdre	waste

Original	Paraphrase	Traduction
lutter contre	battre combattre résister à attaquer défendre contre	fight resist tackle
s'y met	commencer à battre s'engager à s'attacher à	set about sth. get on with sth. get into sth.
nuisibles	destructrice dommageable	detrimental damaging
trier	-	sort
piles	batteries	battery
ampoules	-	bulbs
pétrole	-	oil
charbon	-	coal
émettent	-	emit
espèces	-	species
fonte	-	melting
gaz à effet de serre	-	greenhouse gases
éoliennes	-	aeolian

FIGURE A.4 – Les paraphrases et traductions proposées pour le niveau A2

2. Quand il fera bientôt beaucoup plus chaud, certains animaux vont :
   disparaître
   souffrir
   changer de lieu de vie
3. Pourquoi avons-nous le problème du réchauffement climatique ?
   à cause des catastrophes naturelles
   à cause des habitudes humaines
4. Les nouvelles énergies :
   sont plus profitables
   sont plus faciles à utiliser
   causent moins de mal
5. Qu'est-ce que les barrages hydrauliques et les éoliennes utilisent comme ressource ?
   l'eau et le soleil
   l'eau et le vent
   le soleil et le vent
6. Les 10 petits gestes sont donnés pour :
   mieux vivre avec le réchauffement de la planète
   résister au réchauffement de la planète
   manifester contre le réchauffement de la planète
7. Comment aurons-nous un résultat satisfaisant ?
   si tout le monde se met d'accord
   si tout le monde suit ces conseils
   si tout le monde travaille bien

8. Quel est le sens contraire de "gaspiller" ?
   enregistrer
   économiser
   garder
9. Que devons-nous faire pour les déchets ?
   les jeter à un endroit précis
   les séparer selon des règles données
   les réduire au maximum
10. Quelle est la caractéristique des piles ?
    avoir des capacités différentes
    être lumineuses
11. Une ampoule peut :
    stocker de l'énergie
    consommer de l'énergie
    produire de l'énergie

Le texte du niveau B2 :

De par le monde, il existe un tas d'endroits où on peut observer des phénomènes étonnants. Ainsi, j'ai pu voir à la **lisière** d'une ville, dans un **pâté de maisons** isolé, des **chaumières** qui semblaient **désaffectées**, où pourtant habitaient des personnes qui semblaient **inaptes au travail**. On y voyait un **ivrogne errer** avec son **calepin** à la main, avec **dans son sillage** certains de ses **compagnons**. Quand il n'écrivait pas, il racontait des récits **édifiants** sur les **prouesses** accomplies par certains personnages, ce qui faisait **grincer des dents** à ses amis. Ils **lui en voulaient** de les **effrayer** ainsi. **Il s'avère que** tous ces malheureux vivaient ainsi dans un **perpétuel** ennui. Pour leur **sauver la mise**, il aurait fallu qu'ils **prennent leur problème à bras le corps** pour **remédier à** cette triste situation. Ils en étaient **tout bonnement** incapables, si quelqu'un ne venait pas **mettre au point** avec eux un moyen de s'en sortir. Mais **prôner** n'importe quel travail n'est pas une solution ! Un jour, un homme les entraîna pour travailler pour lui contre **une poignée de** pièces. Mais cette solution trop facile était un piège : cet homme était un voleur et il les incitait à jouer les **fouineurs** pour dévoiler des secrets. Ayant **bravé les interdits**, ces malheureux se sont trouvés en garde à vue. En plus, **dans la foulée**, leurs demeures ont été endommagées par **une pluie battante**. Mais **bon sang** ! Que ce genre de tragédie ne se reproduise pas ailleurs.

Les paraphrases et traductions proposées sont montrées dans les figures A.5 et A.6.

Les questions :

1. Où peut-on observer des phénomènes étonnants ?
   à la campagne
   en ville
   partout
   en banlieue
2. Où se déroule l'histoire ?
   au centre-ville

à la campagne
aucune des deux réponses

3. Quelle est la caractéristique des maisons ?
   elles sont bien alignées
   elles sont en groupe
   elles sont séparées les unes des autres
4. Où habitaient ces personnes ?
   dans de grandes maisons
   dans de petites maisons
   dans de grands appartements
   dans de petits appartements
5. Leur lieu d'habitation semble :
   propre
   vieux
   abandonné
6. Ces gens ont-ils actuellement un emploi ?
   probablement non
   probablement oui
   certainement oui
   certainement non
7. Quelle est la caractéristique principale de la personne qu'on voyait avec un calepin à la main ?
   un clochard
   un alcoolique
   un chômeur
8. Qu'est-ce qu'il faisait ?
   parlait avec ses amis
   courait derrière des gens
   se promenait sans destination précise
9. A quoi peut servir un calepin ?
   à caler un pin
   à dessiner
   à payer son repas
10. Où voyait-on les compagnons de cet homme ?
    autour de lui
    derrière lui
    devant lui
    ce n'est pas dit
11. Qu'est-ce qu'il racontait à ses amis ?
    des aventures intéressantes
    des récits instructifs
    des histoires stupéfiantes
    aucune des trois réponses

12. Quelle est la réaction de ses amis ?
    ils sont contents
    ils ont peur
    ils sont tristes
13. Depuis combien de temps ces gens s'ennuyaient ?
    depuis une heure
    souvent
    tout le temps
14. Qu'est-ce que ces malheureux auraient dû faire ?
    proposer des solutions pour améliorer leur vie
    faire du sport pour avoir une meilleure santé
    épargner de l'argent
15. Pourquoi ont-ils besoin d'une aide extérieure ?
    pour ne plus vivre de cette façon
    pour sortir de là où ils habitaient
    pour ne plus avoir peur
16. Le voleur les paie combien pour le travail ?
    beaucoup
    assez peu
    plutôt bien
17. Pourquoi ont-ils eu un problème à cause de ce travail ?
    ils ont volé quelque chose
    à cause d'une contravention à la loi
    ils ont endommagé une maison
    ce n'est pas dit
18. Ils pouvaient encore habiter chez eux ?
    oui, aucun soucis
    non, ce n'était plus possible
19. Comment était la pluie dans l'histoire ?
    pendant peu de temps
    moyenne
    très forte

Questionnaire :

1. Quel est votre niveau en français ? (par exemple, votre note d'examens tels que TCF, TEF, DELF, DALF, TFS4, TFS8, la note moyenne des examens universitaires du français, etc.)
2. Quel est votre niveau en anglais ? (par exemple, votre note d'examens tels que IELTS, TOEFL, TOEIC, CET4, CET6, la note moyenne des examens universitaires de l'anglais, etc.)

3. Avec quelle version de texte vous sentez-vous le plus à l'aise pour répondre aux questions ?
   Version 1
   Version 2
   Version 3
4. Quelle version de texte avez-vous le mieux compris ?
   Version 1
   Version 2
   Version 3
5. La version 3 vous semble-t-elle nécessaire pour comprendre tout le texte (sans considérer les questions) ?
   oui
   non
6. Essayez-vous de séparer l'apprentissage de l'anglais de celui du français ?
   oui, je ne veux pas mélanger les deux
   non, j'apprends en comparant les deux
7. Vos connaissances en anglais vous aident-elles pour l'apprentissage du français ? Si oui, sur quels aspects ? Si non, pourquoi ? (Si vous voulez, vous pouvez répondre en anglais.)
8. Pour aider à l'apprentissage de la lecture de textes, souhaitez-vous utiliser un outil qui propose des paraphrases françaises des mots ou expressions figées en contexte ?
   oui, ce sera très utile pour élargir mon vocabulaire
   non, je préfère consulter un dictionnaire bilingue (français-chinois, français-anglais) ou monolingue français
9. Avez-vous d'autres attentes sur un tel outil ? (Si vous voulez, vous pouvez répondre en anglais.)
10. L'apprentissage des paraphrases françaises via des traductions anglaises vous semble-t-il utile ?
    oui, je peux apprendre plus vite le français grâce aux connaissances en anglais
    oui, mais ce sera mieux si les propositions sont données avec un contexte de phrase
    non, je préfère apprendre dans la même langue

Original	Paraphrase	Traduction
de par le monde	autour du monde autour du globe dans le monde entier à travers le monde aux quatre coins du monde	over the world from around the world
lisière	frontière bordure	edge
pâté de maisons	bloc de maisons quartier	block of houses
chaumières	maison de campagne chalet	cottage
désaffectées	inhabité cesser d'être utilisées hors service non utilisé hors d'usage	unused disused deserted
inaptes au travail	incapable de trouver un emploi inemployable impossible à employer	unfit unemployable incompetent
ivrogne	soûl enivré	drunkard
calepin	cahier carnet de notes	notebook
errer	flâner déambuler vagabonder	roam wander
dans son sillage	sur ses traces	in the tracks of in the wake of
édifiants	étonnant incroyable époustouflant	amazing stunning
tout bonnement	tout simplement	just simply
mettre au point	développer élaborer concevoir mettre en place	develop work out devise come up with set up
prôner	défendre préconiser	advocate extol
une poignée de	très peu quelques-uns	a small number of a handful of a couple of a fistful of a few
fouineurs	ceux qui mettent leur nez partout les gens curieux	busybody snoop

FIGURE A.5 – Les paraphrases et traductions proposées pour le niveau B2

Original	Paraphrase	Traduction
prouesses	réussite succès réalisation accomplissement exploit	achievement feat accomplishment
grincer des dents	enrager	gnash one's teeth grind one's teeth
lui en voulaient	se mettre en colère contre lui être furieux contre lui être mécontent de lui lui en tenir rigueur s'indigner de lui	hate him blame him hold it against him be angry with him resent him
effrayer	faire peur terrifier apeurer	frighten scare spook
Il s'avère que	en fait en réalité effectivement il se trouve que	it turns out in fact
perpétuel	incessant permanent constant	constant perpetual
sauver la mise	sauver la situation régler la situation	save the day save them
prennent leurs problèmes à bras le corps	régler le problème résoudre le problème lutter contre le problème	deal with engage in combat with come to grips with
remédier à	régler résoudre traiter apporter une réponse à faire face à	fix address
bravé les interdits	ne pas respecter les règlements enfreindre les règles contrevenir aux règles en vigueur	break the rules commit the forbidden
dans la foulée	tout de suite après presque aussitôt après sans la moindre perte de temps sans répit	immediately thereafter without ever skipping a beat
une pluie battante	pluie violente pluie torrentielle pleut à verse pleuvoir à boire debout	heavy rain pouring rain driving rain
bon sang	mon dieu jésus christ	(my) god (oh my) gosh for goodness sake for god's sake what the hell

FIGURE A.6 – Les paraphrases et traductions proposées pour le niveau B2

Annexe B

Guides d'annotation pour les couples anglais-français et anglais-chinois

Annotation Guidelines of Translation Techniques for English-French

Yuming Zhai, Gabriel Illouz, Anne Vilnat
LIMSI, CNRS
Université Paris-Sud, Université Paris-Saclay
Orsay, France
zhai@limsi.fr

Contents

1 Introduction

1.1 Task description

This document is a guide for annotating translation techniques in a parallel corpus. For example in our work, the corpus is composed of transcriptions and human translations of TED Talks[1].

The total corpus contains 19 talks and 2 436 lines of parallel sentences (the sentence alignment is already done). These topics are covered: technology, psychology, culture, science, biology, etc. Each sentence pair contains on average 21 English tokens on source side and 22 French tokens on target side.

1.2 Translation techniques

Consider this example (all examples in this guide are shown with tokenization) :

that image reminded me of something . → cette image m' a rappelé quelque chose .

For this pair of sentences, we could conduct word alignments as follows, where all the source segments have been translated literally.

that → cette

image → image

reminded of → a rappelé

me → m'

something → quelque chose

. → .

Translation techniques constitute an important subject of study for translators and linguists (Vinay and Darbelnet, 1958; Newmark, 1981; Newmark, 1988; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002; Gibová, 2012), which distinguish literal translation from other translation techniques on word or phrase level. Based on the above cited work in translation techniques, by annotating and analyzing our English-French parallel corpus, we have proposed a typology of translation techniques (see figure 1) in order to have a global view of these categories. The tables 1 and 2 provide a recapitulation of definition and important rules for each translation technique.

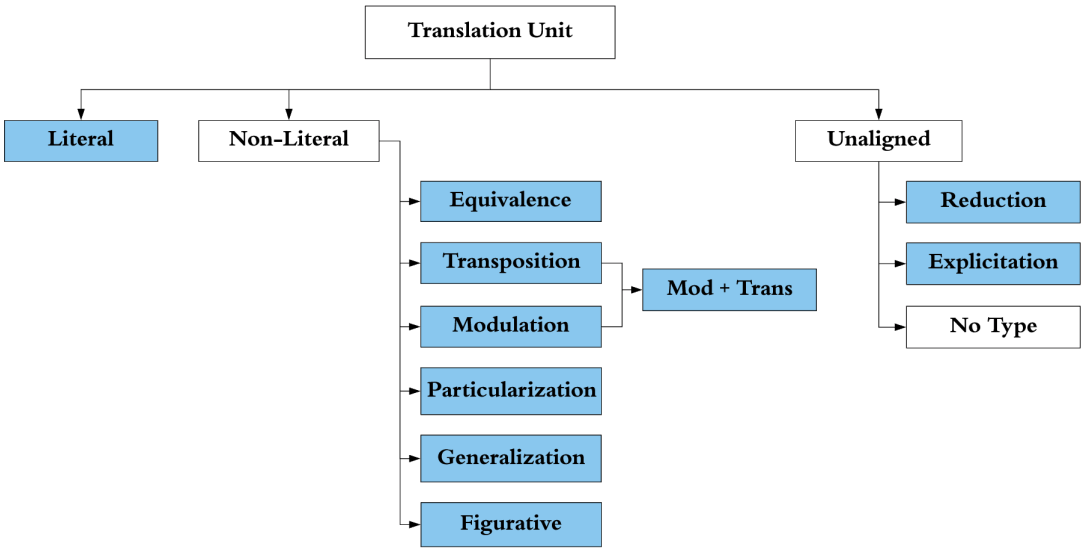

Figure 1: Typology of translation techniques

[1] https://www.ted.com/

Translation Technique	Definition and important rules
Aligned segments	
Literal (6.1.1)	Word-for-word translation; - possible literal translation of idioms; - concerns lexical units in multiword form; - corresponding expression when absolute literal translation does not make sense. *certain **kinds** of → certains **types** de*
Equivalence (6.1.2)	A word-for-word translation makes sense but the translator has expressed differently; - non-literal translation of proverbs, idioms, or fixed expressions; - no change in meaning and point of view. ***sense each other → se reconnaître entre eux***
Transposition (6.1.3)	Change grammatical categories without changing the meaning. *and , if anything , at a far **greater rate** . → et peut-être bien **plus rapidement** .*
Modulation (6.1.4)	Change the point of view; - can occur both on lexical level and in syntactic structures; - metonymical and grammatical modulation; - could bring slight meaning changes. *the statistics **you hear about** → les statistiques **qui nous sont communiquées***
Mod+Trans (6.1.5)	Combine the transformations of *Modulation* and *Transposition*. *this is a completely **unsustainable** pattern . → il est absolument **impossible de continuer sur** cette tendance .*
Particularization (6.1.6)	The translation has a more concrete or particular meaning; - specify the meaning of a word in context; - translate a pronoun by the thing(s) it references. *they **have** a screen and a wireless radio → ils **sont équipés d'** un écran et d' une radio sans fil*
Generalization (6.1.7)	The translator used a target word or segment whose meaning is more general than that of the source word or segment; - the translation of an idiom by a non-fixed expression; - the removal of a metaphorical image; - use pronoun to translate the thing(s) that it references. *as we **sit here** today in Monterey → alors que nous **sommes** à Monterey aujourd'hui*
Figurative translation (6.1.8)	- introduce an idiom to translate a non-fixed expression, or a metaphorical expression to translate non-metaphor; ***at any given moment → à un instant " t "*** - keep the same metaphorical image by using a non-literal translation; *the Sun begins to **bathe** the slopes of the landscape → le soleil qui **inonde** les flancs de ce paysage*
Lexical shift (6.1.9)	Change verbal tense or verbal modality, preposition, determiner, subject, position adverb, between singular and plural forms. *when you do a web search **for** images → quand on fait une recherche web **sur** des images*
Translation error (6.1.10)	Obvious translation errors. *is not going to be **remembered** for its wars → ne sera pas **reconnu** pour ses guerres*
Uncertain (6.1.11)	Not sure about which category to assign, need more discussion.

Table 1: Definition and important rules for aligned segments

Translation Technique	Definition and important rules
Unaligned segments	
Explicitation (6.2.1)	- resumptive anaphora; - introduce clarifications that is implicit in source text. *[...] live amongst those who have not forgotten the old ways, who still feel their past in the wind* *→ [...] vivre parmi ceux qui n'ont pas oublié les anciennes coutumes, qui ressentent encore leur passé* ***souffler*** *dans le vent*
Reduction (6.2.2)	Deliberately remove certain words with concrete meaning that could be translated. *look* ***carefully*** *at the area of the eastern Pacific → regardez le secteur oriental de l' océan pacifique*
No Type (6.2.3)	- function words only necessary in one language; - segments not translated but they do not impact the meaning; - segments giving repeated information in context; - translated segments which do not correspond to any source segment. *minus 271 degrees , colder than → moins 271 degrés ,* ***ce qui est*** *plus froid que*

Table 2: Definition and important rules for unaligned segments

2 General instructions

2.1 How to deal with minor misspelling and tokenization errors?

In order to guarantee the corpus quality and to generate a clean data set for developing our automatic classifier, we correct minor spelling errors in the corpus, for example *ca → ça*, *a quel point → à quel point*, *l' endroit ou → l' endroit où*, etc.

During the annotation, please note down the sentence ID and the misspelled pair. The same for the word tokenization errors that you have found (*e.g. lorsqu'on → lorsqu' on*).

2.2 Three conditions of annotation

Annotators can watch corresponding videos of TED Talks before annotating to better understand the context of each sub-corpus, the links are provided in another file.

During the annotation, there will be three possible configurations:

- Raw text without any manual annotation: you should conduct segmentation of translation units, correct existing automatic word alignments, attribute categories of translation technique.
- Text which has already been annotated once: in order to guarantee the quality of our work, you need to verify the annotation. You can modify the alignment and category attribution if there is a disagreement. You can also modify the phrase boundary.
- Text which has been annotated twice, and it was you who conducted the first pass: please try to reach consensus with the other annotator on the remaining differences between you two. We will provide you with the differences between the two versions to accelerate the reviewing.

2.3 External resources

Annotators are encouraged to use language resources such as *Cambridge Dictionary, Larousse, Le Robert, TLFi*, etc. to consult word senses, widely used literal translations, the translation of multiword expressions and so on.

For example, in *Linguee*[2], we can see the relation between *« faint »* and *« tomber dans les pommes »*, this can also help to decide the boundary:

[2]https://www.linguee.com/

*if you **faint** easily → si vous **tombez dans les pommes** facilement* (Category *Figurative* (subsection 6.1.8))

3 Decision helper

In order to facilitate the annotation task, please see figure 2 to help you to make decisions on the most confusing categories. Please note that this table recapitulates the most distinguishing aspects for each category, and doesn't include all the definitions and rules presented below.

	word-for-word	non-literal translation of proverbs, idioms, fixed expressions	change point of view	change syntactic structure	change meaning	change grammatical category	more specific meaning	more general meaning
Literal	x							
Equivalence		x						
Transposition						x		
Modulation			perhaps	perhaps	perhaps			
Mod+Trans			perhaps	perhaps	perhaps	x		
Particularization					x		x	
Generalization					x			x
Lexical shift								

Figure 2: A table to help annotators to make decisions on the most confusing categories

During the annotation, there exist different categories concerning idioms and fixed expressions in the source or target language, the table 3 recapitulates them. Below is their definition.

Idiom: a phrase or an expression that has a figurative, or sometimes literal, meaning. The figurative meaning is based on the whole rather than on the individual words in it. Idiomatic expressions are strongly cultural and have different meanings derived from the cultures they come from. For example: *a piece of cake, every cloud has silver lining*.

Fixed expression: a standard form of expression that has taken on a more specific meaning than the expression itself. It is used as a part of a sentence, and is the standard way of expressing a concept or idea. Unlike idioms, they are generally transparent in meaning. For example: *as a matter of fact, all of a sudden, to whom it may concern*.

Translation phenomenon	Category attributed
literal translation of an idiom	*Literal* (see subsection 6.1.1)
idiom → equivalent idiom	*Equivalence* (see subsection 6.1.2)
idiom → non-fixed expression	*Generalization* (see subsection 6.1.7)
non-idiom → idiom	*Figurative* (see subsection 6.1.8)
non-fixed expression → fixed expression	*Equivalence* (see subsection 6.1.2)

Table 3: Different categories concerning idioms and fixed expressions

4 Annotation tool

We use the web application Yawat (Germann, 2008) for our annotation, if you don't know how to use this tool, please read the section 7.

5 Annotation conventions

5.1 How to decide the segment boundary?

In principle The segment boundary is not provided to annotators and it should be fixed by respecting the given tokenization, while excluding the part not involved (follow the bold part in this annotation guide):

have generated ***sufficient*** *interest → ont suscité* ***suffisamment d'*** *intérêt*

For simple literal lexical translation We annotate the smallest semantic unit as we can, for example given this pair:

there is a measurable effect → il y a un effet mesurable

We should segment and align like this:

there is → il y a
a → un
measurable → mesurable
effect → effet

Negation

Don 't *believe everything she says . →* ***Ne*** *crois* ***pas*** *tout ce qu' elle te dira .*
since it ***did n't happen*** *here → comme ça* ***ne se passe pas*** *ici*
we ***did n't have*** *polio in this country yesterday → nous* ***n' avions pas*** *la polio dans ce pays hier*
are n't *you* ***afraid*** *→ vous* ***n' avez pas peur de***

Articles and prepositions

- In the following examples, we annotate the article together with the noun, because the French article corresponds to an empty position in English:
  1. *it 's really* ***text*** *→ c' est vraiment* ***du texte***
  2. *in other words ,* ***sugar pills*** *have a measurable effect*
     → en d' autres termes , ***des pilules de sucre*** *ont des effets mesurables*
  3. *you can just say a few names and* ***people*** *will understand*
     → il suffit de citer quelques noms et ***les gens*** *comprennent*
- Here we align *to you* with *vous*, which are both indirect objects of the verb *show* and *montrer*.
  I 'll show it ***to you*** *→ je vais* ***vous*** *le montrer*
- Regroup the preposition with the verb that triggers its appearance:
  we do n't ***want to*** *encourage people to eat → on ne* ***veut*** *pas encourager les gens à manger*
- In this example, *lift off* is aligned to *soulevant*. Adding the preposition *off* here changes the meaning of the verb, because *lift* alone means *lever*. Then *the table* is aligned to *de la table*, because *de* corresponds to an empty position in English (the notion of *from* is implicitly implied in English).

  and he can bring new characters into the scene , just by ***lifting*** *the Siftables* ***off the table*** *that have that character shown on them .*
  → il peut amener de nouveaux personnages dans la scène , simplement en ***soulevant de la table*** *les Siftables présentant ce personnage .*

For non-literal translations Sometimes it is necessary to enlarge the boundary of the segments, in order to clarify the meaning, even though there are words which could be annotated as *Literal* inside this segment. For example :

1. *and the great indicator of that , of course , is* ***language loss***
   → et l' indicateur le plus fiable est bien sûr ***l' extinction du langage*** (Particularization, subsection 6.1.6)
2. ***spend a large sum of money → dépenser massivement*** (*Transposition*)
   (We keep this group to clarify the very general meaning of *« massivement »*.)
3. ***stamp a letter into it → avec une lettre en creux*** (*Transposition*)
4. *les Buddhists* ***still pursue*** *the breath of the Dharma*
   → les Bouddhistes ***continuent à rechercher*** *le souffle du Dharma* (*Transposition*)
   (*still + verb → continuer à + verb* is a pattern, which should be annotated as a group.)

Composition of categories In the above cases, the literal part is considered as neutral, when combined with another category, the latter becomes the translation technique for the entire segment. For example:

stamp a letter into it → avec une lettre en creux

(*Transposition*, where *a letter* and *une lettre* is a pair of literal translation)

Thus, in our work, the translation unit could be a word, a phrase or a **short** sentence (do not include the final punctuation):

Nice one . → ***Pas mal*** . (*Equivalence*, subsection 6.1.2)

How long have you been *here ?* → ***Quand êtes -vous arrivés*** *ici?* (*Modulation*)

5.2 How to align punctuation?

See table 4 for a recapitulation.

Punctuation	Alignment
if the punctuation doesn't change	
Final punctuation (period, exclamation mark, question mark), comma, colon, semicolon, quotation marks, angle quotes, brackets, ellipsis, etc.	Align them out of the segments.
Apostrophe, dash, hyphen	Respect the given tokenization. If you disagree with them, please contact us.
if the punctuation changes	
For example, the translation replaces a double dash by a comma.	Annotate as *Lexical shift* (6.1.9).

Table 4: How to align punctuation

5.3 Mutually exclusive categories

There exist difficult borderline examples, but we do not allow multiple categorization in this task, which means that a pair of segments receive always one category listed in our typology (see figure 1).

For borderline examples, after discussion, annotators should agree on a category which better reflects the technique used by the translator.

5.4 How to annotate unaligned segments?

For the categories of *Unaligned - Explicitation* and *Unaligned - Reduction* (see definition in table 2), please annotate separately each span of reduced or added segment (separated by aligned segments), do not make a whole group.

For example, in figure 3, please annotate the four instances separately: *facing, from you, is just going to, out*.

in a moment when my hand moves from facing you to being away from you , this finger right here , my index finger is just going to shift from where it is , to a position pointing out like this .

au moment quand ma main s' écarte de vous pour s' éloigner , ce doigt juste ici , mon index , passe de là où il est à une position pointant comme ceci .

Figure 3: How to annotate unaligned reduced segments

5.5 How to deal with linguistic anaphora?

In our work we do not manually resolve the linguistic anaphora, for example, we leave the « ma » unaligned in the following pair:

it is also ***my*** *great love and fascination* → *c' est aussi* ***mon*** *grand amour et ma fascination*

Another two examples are shown in figure 4 and 5, see the figure caption for how to annotate.

In figure 6, we annotate the repeated preposition « de » after the conjunction word « et » as *Explicitation* (see subsection 6.2.1).

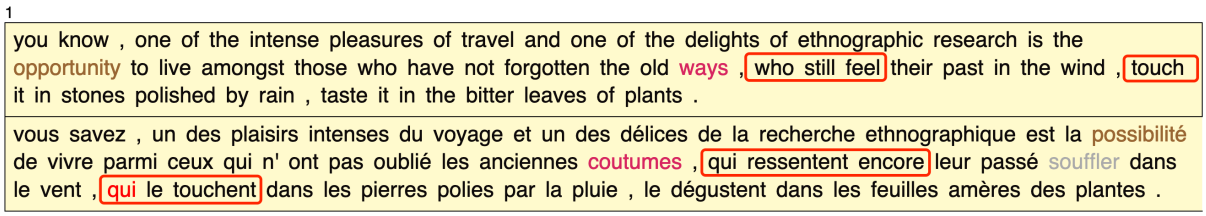

Figure 4: On target side, leave the second *qui* unaligned

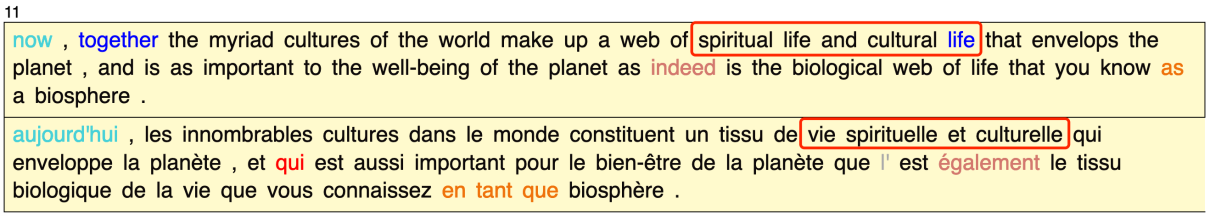

Figure 5: On source side, leave the second *life* unaligned

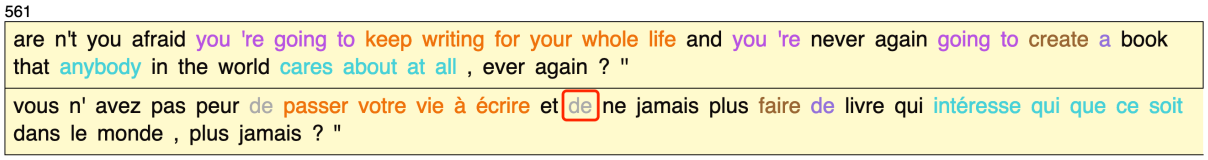

Figure 6: Annotate the repeated preposition after the conjunction as *Explicitation*

6 Annotation in practice

We have already annotated 17 924 English tokens, below we will show the percentage of literally translated tokens, and the percentage of each other category among the non-literal cases.

6.1 Annotation of aligned segments

Literal translations

6.1.1 Literal

Percentage: 73.80%.

Definition Word-for-word translation (including insertion or deletion of articles), or possible literal translation of some idioms:

certain ***kinds*** *of* → *certains* ***types*** *de*

we ***all*** *share the same* → *nous partageons* ***tous*** *les mêmes*

facts *are stubborn* → ***les faits*** *sont têtus* (literal translation of an idiom, align word by word)

Rule 1 The literal translation also concerns lexical units but in multiword form:

1. *But we* ***wonder*** *whether it has gone too far .*
   → *Nous* ***nous demandons*** *cependant si cela n'est pas allé trop loin .*
2. ***there are*** → ***il y a***
3. ***this is*** → ***voici***
4. ***hatpin*** → ***épingle à chapeau***
5. ***the largest*** → ***le plus grand***
6. ***ago*** → ***il y a***
7. ***magic trick*** → ***tour de magie***

8. ***NASA geospatial image* → *image géospatiale de la NASA***
   (annotate as a group, useful to teach learners the composition rules in French when adding the preposition ***de***)
9. ***all of you* → *vous tous***
   (annotate as a group: all of + pron → pron + tous)

Rule 2 When an absolute word-for-word translation does not make sense in the target language, the corresponding expression is deemed to be literal:
look forward to *(?regarder en avant)*[3] → ***avoir hâte de***
in other words *(?en d'autres mots)* → ***en d' autres termes***
I give you my word *. (?Je vous donne mon mot .)* → ***Je vous donne ma parole*** *.*
experience a hallucination *(?expérimenter une hallucination)* → ***avoir une hallucination***

Rule 3 The change of numbers between the spelled out form and the Arabic form is annotated as *Literal*:
go from ***six and a half*** *to nine billion people → passer de* ***6,5*** *à 9 milliards d' êtres humains*

Rule 4 Translating by using calque or anglicism is annotated as *Literal*:
the myths of the Inuit elders still ***resonate with*** *meaning → les mythes des anciens Inuit* ***résonnent*** *encore* ***de*** *sens*
(« résonner de sens » is not a natural expression here)
but one of them is ***distinctly*** *worse than the other → mais l' une d' entre elles est* ***distinctement*** *pire que l' autre*

Rule 5 Fixed modulation is annotated as *Literal*:
a ***life jacket*** *→ un* ***gilet de sauvetage***
(The means « sauvetage » substituted for the result « life », from a linguistic point of view, the category should be *Modulation*. However, since there is no other possible translation for « life jacket », and it's a recorded pair in bilingual dictionaries, we annotate this case by *Literal*.)

For each translation technique, we show some counterexamples and borderline examples:
Counterexamples:
that looks kind of neat → c' est pas mal
(The translation is not word-for-word, and the translation uses a negative form. The category should be *Modulation*.)

Borderline examples:
Sort of . → En quelque sorte .
(Borderline with *Equivalence*, but it should be annotated as *Literal*. If it's translated by « Peut-être. » or « Presque. », then it should be annotated as *Equivalence*.)

Non-literal translations

6.1.2 Equivalence

Percentage: 17.79% (of non-literal translations)
Definition There is no change of point of view like in *Modulation* (subsection 6.1.4). A word-for-word translation makes sense but the translator has expressed differently. However, if there exist changes of grammatical categories, the pair should be annotated with *Transposition* (see examples in subsection 6.1.3).
if you 'll pardon *the pun →* ***si vous me passez*** *ce calembour*
sense each other → se reconnaître entre eux

[3] The question mark means this translation is word-for-word but actually is incorrect.

more than that* → *plus encore
right now *he 's over Ohio →* ***là*** *il survole l' Ohio*
are n't you afraid you 're never going to ***be able to top that*** *?*
→ vous n' avez pas peur de ne jamais ***réussir à faire mieux*** *?*

Rule 1 Non-literal translation of proverbs, idioms, or fixed expressions is annotated as *Equivalence*:
Birds of a feather flock together .* → *Qui se ressemble s' assemble .
like a bull in a china shop* → *comme un chien dans un jeu de quilles
on the brink of* → *à deux doigts de

Rule 2 Equivalence in context is annotated as *Equivalence*:
and we started talking about music , from Bach to Beethoven and Brahms , Bruckner , all the B 's , from Bartók , ***all the way up to*** *Esa-Pekka Salonen .*
→ et nous avons commencé à parler de musique , de Bach à Beethoven , de Brahms , Bruckner , tous les B , de Bartók , ***jusqu'à*** *Esa-Pekka Salonen .*
(need world knowledge, chronological sense)

Rule 3 Changing the measure into the one used in the target culture is annotated as *Equivalence*:
about ***a mile and a half*** *deep → vers* ***2500 m*** *de profondeur*
is ***20 inches*** *→ est de* ***50 centimètres***

Rule 4 Translate a non-fixed expression by a fixed expression:
now they are ***metaphorically*** *in the womb of the great mother*
→ ils sont maintenant dans le ventre de la grande mère , ***métaphoriquement parlant***
(The word "métaphoriquement" alone would be a literal translation, but "-ment + parlant" is a fixed expression, but it doesn't have a figurative meaning.)

Rule 5 Translate an abbreviation into a full version or vice versa:
UN → Organisation des Nations unies
IPCC → Groupe d' experts intergouvernemental sur l' évolution du climat

Counterexamples:
magic trick → tour de magie
(Word-for-word translation, the category is *Literal*.)
at no time → à aucun moment
(We can not say *à aucun temps* in French. This is a literal translation (see Rule 2 of *Literal*).)

Borderline examples:
which is → soit
(Borderline with *Literal*. Since it is not the most literal translation, it is annotated as *Equivalence*.)
that 's something the world needs right now → c' est quelque chose dont le monde a besoin maintenant
(Borderline with *Literal*, annotated as *Equivalence*.)

6.1.3 Transposition

Percentage: 14.81%
Definition Translating words or expressions by using other grammatical categories than the ones used in the source language, without altering the meaning of the utterance.
The change of grammatical categories could occur on a complete syntagm, for example:
people ***are suspicious*** *→ les gens* ***se méfient***
or locally on a term in a syntagm, which globally doesn't change the category, for example:
I said it ***as a joke*** *→ J' ai dit ça* ***pour plaisanter***

Below are some typical examples found in our corpus:

adv -> conjunction
*there are **only** three fingers down here → il **n'** y a **que** trois doigts ici*
*it takes 142 pages **just** to print this genetic code*
*→ ça prendrait 142 pages **rien que** pour imprimer ce code génétique*
verb -> prep
*" what is life ? " is something that **I think** many biologists have been trying to understand*
*→ " qu' est -ce que la vie ? " est **selon moi** ce que beaucoup de biologistes ont cherché à comprendre*
verb -> noun
*unless **something changes** , they 're already dead*
*→ à moins qu' **un changement ait lieu** , elles sont déjà mortes*
*these two morphologically unrelated plants that **when combined** in this way*
*→ ces deux plantes sans aucun lien morphologique qui **lorsque mises en synergie** de cette façon*
noun -> verb
*and **that is the idea** that the world in which we live*
*→ et **cela veut dire** que le monde dans lequel nous vivons*
*the **computer vision** algorithms have registered these images*
*→ les algorithmes de **vision informatisée** ont enregistré ces images*
noun -> adv
*and , if anything , at a far **greater rate** . → et peut-être bien **plus rapidement** .*
adj -> adv
*have generated **sufficient** interest → ont suscité **suffisamment d'** intérêt*
adj -> verb
*even those of us **sympathetic with** the plight of indigenous people*
*→ même ceux d' entre nous qui **compatissons avec** les difficultés du peuple indigène*
adj -> noun
*we would say to be **friendly gestures** , → on pourrait qualifier de **témoignage d' amitié** ,*
prep -> verb
*patients **over** the age of 40 → les malades **ayant dépassé** l' âge de 40 ans*
*how we understand a lot of the world **around us** .*
*→ notre façon de comprendre une grande partie du monde **qui nous entoure** .*
*these were calling cards **from** the devil → c' était des cartes **venant du** diable*

The « chassé-croisé » is a double transposition involving both a change of grammatical category and a syntactic permutation of the elements that constitute the meaning.

1. *all that we can be as an **astonishingly inquisitive** species .*
   *→ tout ce que nous pouvons être en tant qu' espèce **dotée d' une curiosité stupéfiante** .*
2. *he **strode into** the house → il **entra à grands pas** dans la maison*
3. *so they **speared** the five missionaries **to death***
   *→ ils ont donc **abattu** les cinq missionnaires **à coups de lance***

Counterexamples:
where they get mixed . → où elles se mélangent .
(The passive voice (« get + past participle » expresses a passive meaning) has been changed to active voice, the category should be *Modulation*.)

Borderline examples:
all of these peoples teach us that there are other ways of being
→ tous ces peuples nous enseignent qu' il y a d' autres façons d' être
(Here, "all" is a pronoun and "tous" is an adjective, but it is borderline with *Literal*. It is annotated as *Literal* now.)

they go , " well , he 's certainly not dumb enough to stab himself through the skin to entertain us for a few minutes .

→ ils font : " il n' est pas assez bête pour se planter des choses à travers la peau pour nous amuser quelques minutes ."

(Without preposition, "quelques temps" is not a direct object as in English. It is annotated as *Transposition* now.)

6.1.4 Modulation

Percentage: 16.03%

Definition This translation technique consists of a change in the point of view that enables us to convey the same phenomenon in two languages in a different way.

- can be encountered both in lexis as well as in syntactic structures
- reveal a specific way of seeing things for the speakers of the target language
- circumvent translation difficulties
- the translator desires to achieve expression naturalness in the target language
- could bring meaning changes between source and target text

According to Chuquet and Paillard (1989), there are mainly two types of modulation:

- metonymical modulations (the cause substituted for the effect, the container for the content, the part for the whole, etc.)
- grammatical modulations (change between affirmative form and negative form; between injunction and interrogation; between passive voice and active voice; the subject becomes the object, etc.)

Examples:

Metonymical modulations:

1. *Buy Coca-Cola* ***by the carton*** *.* → *Achetez Coca-Cola* ***en gros*** *.* (the abstract « en gros » substituted for the concrete « by the carton »)
2. ***global warming pollution*** → ***pollution à effet de serre*** (the means « à effet de serre » substituted for the result « global warming »)
3. ***He shut the door in my face*** *.* → ***Il me claqua la porte au nez*** *.* (the part « nez » substituted for the whole « face »)
4. Geographical modulation:
   In French, *Holland* is often used to refer to the *Netherlands*, while the former is just the name of a region in the Netherlands.
   By metonomy, we often refer to *United Kingdom* by *England* or *Great Britain*.

Change the point of view:

5. ***How long have you been*** *here ?* → ***Quand êtes -vous arrivés*** *ici ?*
6. *their hunters could smell animal urine at 40 paces and tell you* ***what species left it behind*** *.*
   → *leurs chasseurs pouvaient sentir l' urine animale à 40 pas et vous dire* ***de quelle espèce elle provenait*** *.*
7. ***what could be more lonely than*** *to be enveloped in silence*
   → ***comment ne pas se sentir seuls*** *, enveloppés dans le silence*
8. *and* ***that scar has stayed with him*** *for his entire life*
   → *et que , toute sa vie ,* ***il a souffert de ce traumatisme***
9. *the beta-carbolines* ***found within*** *that liana* → *les béta-carbolines* ***dont est composée*** *la liane*

Passive voice and active voice:

10. *where they* ***get mixed*** *.* → *où elles* ***se mélangent*** *.*
11. *the statistics* ***you hear about*** → *les statistiques* ***qui nous sont communiquées***

Affirmative form to negative form, the negation of the opposite:

12. *It **'s difficult** . → Ce **n' est pas facile** .*

The subject becomes the object:

13. ***He was knee-deep in water** . → **L' eau lui arrivait jusqu'aux genoux** .*
14. ***I had a** really astonishing **assignment** → **on m' avait confié une mission** étonnante*

Syntax changes (no meaning change):

15. *something **has happened** → **il m' est arrivé** quelque chose*
16. *and you can go backwards **, you can go** forwards ; you can not stay where you are .*
    *→ vous pouvez aller en arrière **ou** en avant ; vous ne pouvez pas rester où vous êtes .*
17. ***those are two very different entities** → **ces deux entités sont très différentes***

Circumvent translation difficulties, achieve expression naturalness:

18. *unless you **think of it in the terms that I do** , → à moins que vous ne **regardiez la chose comme moi** :*
19. *which has **caused me to** have to **recalibrate** my whole relationship with this work*
    *→ qui a **transformé** ma relation à ce travail*

Slight meaning change in lexical level:

20. *at no time can anything **travel** , → à aucun moment quoi que ce soit ne peut **passer** .*
21. *allowing people to more fully engage with their **abilities** → permettent aux gens de développer pleinement leur **potentiel***
22. *I do n't **expect** that 's going to change → je ne **crois** pas que cela va changer*
23. *remember the **central** revelation of anthropology*
    *→ se rappeler de la révélation **essentielle** de l' anthropologie*

Counterexamples:
and saw the front page was like that → et voyait la page d' accueil comme ça
(The point of view does not change, the category should be *Equivalence*.)

Borderline examples:
we 're looking at → on a devant les yeux
(Borderline with *Mod+Trans*, annotated as *Modulation* now.)

6.1.5 Mod+Trans

Percentage: 4%

This category combines the transformations of *Modulation* and *Transposition*, because the translation includes both changes in grammatical category and in syntactic structure or point of view. This often results in many-to-many alignment:

1. *this is a people **who cognitively do not distinguish** the color blue from the color green*
   *→ c' est un peuple **dont l' état des connaissances ne permet pas de faire la distinction** entre la couleur bleue et verte*
2. *this is a completely **unsustainable** pattern .*
   *→ il est absolument **impossible de continuer sur** cette tendance .*
3. *you can learn **a great deal** about deception → vous pouvez apprendre **pas mal de choses** sur l' illusion*
4. ***it 's going to be a stretch** to do it for 9 . → **ça sera d' autant plus difficile** de le faire pour 9 .*
5. *so , our pace of digitizing life has been increasing **at an exponential pace***
   *→ bien que le rythme de notre numérisation de la vie se soit accru **de manière exponentielle***
6. *an enzyme **found naturally** in the human gut → une enzyme **se trouvant de façon naturelle** dans l' intestin de l' homme*

(*found* → *se trouvant*: passive voice to active voice)

7. ***The houses were all dark*** . → ***Pas une maison n' avait de lumière*** .
8. *are n't you afraid that you 're going to* ***work your whole life at this craft*** *and nothing 's ever going to come of it and you 're going to die on a scrap heap of broken dreams with* ***your mouth filled with*** *bitter ash of failure ?*
   → *tu n' as pas peur de* ***passer ta vie à écrire*** *sans résultat et de mourir sur un tas de rêves brisés avec ,* ***dans la bouche*** *, le goût amer des cendres de l' échec ?*
9. ***you are taken by*** *the spirit* → *l' esprit* ***prend le contrôle***

Counterexamples:
that was done entirely computationally → *qui a été faite entièrement par calcul informatique*
(There is only *Transposition* used in this translation.)

Borderline examples:
brilliant , Juilliard-trained musician → *brillant musicien issu de Juilliard*
(Annotated as *Mod+Trans*.)

6.1.6 Particularization

Percentage: 7.71%

Definition The source word could be translated into several target words with more specific meaning, and the translator has chosen one of them according to the context.

they ***have*** *a screen and a wireless radio* → *ils* ***sont équipés d'*** *un écran et d' une radio sans fil*
this plant ***had in it*** *some very powerful tryptamines*
→ *cette plante* ***contenait*** *de très puissantes tryptamines*
the ***great*** *indicator of that* → *l' indicateur* ***le plus fiable*** *est*
because of *someone 's perception of it* → ***grâce à*** *la perception de quelqu'un*

Longer boundary Annotators could enlarge the boundary to help disambiguate the meaning of the segment:

1. *and the great indicator of that , of course , is* ***language loss***
   → *et l' indicateur le plus fiable est bien sûr* ***l' extinction du langage***
2. *live amongst those who have not forgotten the old* ***ways***
   → *vivre parmi ceux qui n' ont pas oublié les anciennes* ***coutumes***
3. *I 'll* ***get my sleeve back*** . → *je* ***retrousse ma manche*** .

Rule 1 Specify the meaning of a word in context is annotated as *Particularization*:
if ***you 're queasy*** → *si* ***vous ne supportez pas la vue du sang***

Rule 2 Translate a pronoun by the thing(s) it references is annotated as *Particularization*:
if you want to know how sea level rises from land-base ice melting this is where ***it*** *reaches the sea .*
→ *si vous voulez voir comment le niveau de la mer monte à cause de la fonte des glaces terrestres voilà l' endroit où* ***la rivière*** *se jette dans la mer .*

Counterexamples:
the weight of the marshmallow causes the entire structure to buckle and to collapse → *le poids du marshmallow fait que toute la structure se déforme et s' écroule*
(This is an instance of *Generalization*.)

Borderline examples:
you get to choose which one you want to use → *on peut choisir celle qu' on veut utiliser*

(It is not a very obvious example, it's annotated as *Particularization* now.)

6.1.7 Generalization

Percentage: 3.85%

Definition This translation technique is the opposite of *Particularization*. Several source words or expressions could be translated into a more general target word or expression, and the translator used the latter to translate. Or some semantic properties have been lost via the generalized translation.

1. *the most optimistic scenario* ***in the realm of*** *cultural diversity*
   → *le scénario le plus optimiste* ***au sein de*** *la diversité culturelle*
2. *and* ***because you are possessed*** *, you are taken by the spirit*
   → *et* ***étant possédés*** *, l' esprit prend le contrôle*
   (remove the causality)
3. *as we* ***sit here*** *today in Monterey → alors que nous* ***sommes*** *à Monterey aujourd'hui*
4. *it 's held so* ***tightly*** *→ il tient si* ***bien***
   (the notion of "tightly" is lost)

Rule 1 The translation of an idiom by a non-fixed expression is annotated as *Generalization*:
we use that great euphemism , " ***trial and error*** *" , which is exposed to be meaningless .*
→ nous employons cet euphémisme , ***procéder par tâtonnements*** *, qui est dénué de sens .*

Rule 2 The removal of a metaphorical image is annotated as *Generalization*:
ancient Tairona ***civilization which once carpeted*** *the Caribbean coastal plain of Colombia*
→ anciennes ***civilisations*** *tyranniques* ***qui occupaient jadis*** *la plaine côtière des Caraïbes de Colombie*

Rule 3 Use pronoun to translate the thing(s) that it references:
people in the back or ***people*** *on video → les gens du fond ou* ***ceux*** *qui regarderont la vidéo*
if you 're starting with digital information in the computer , ***that digital information*** *has to be really accurate*
→ si vous partez des données numérisées sur ordinateur , il faut qu' ***elles*** *soient extrêmement précises*

Counterexamples:
it does n't matter how much information we 're looking at → ce quelle que soit la quantité d' informations que l' on visionne
(This is an instance of *Particularization*.)

Borderline examples:
I 'm going to grab hold of my wrist → je vais tenir mon poignet
(It is not a very obvious example, it's annotated as *Generalization* now.)
let 's pretend right here we have a machine → supposons que nous ayons ici une machine
(Boderline with *Equivalence*.)
what 's interesting is the metaphor that defines the relationship between the individual and the natural world .
→ la métaphore est intéressante , définissant la relation entre l' individu et le monde naturel .
(Borderline with *Modulation*, syntactic change without changing the meaning.)

6.1.8 Figurative translation

Percentage: 0.57%

Rule 1 Introduce an idiom to translate a non-fixed expression, or a metaphorical expression to translate non-metaphor:
the number of pixels on your screen ***at any given moment***
→ le nombre de pixels sur votre écran ***à un instant " t "***
if you ***faint*** *easily → si vous* ***tombez dans les pommes*** *facilement*
and one woman almost ***passed out*** *→ et une dame* ***a*** *presque* ***tourné de l' oeil***

Rule 2 Keep the same metaphorical image by using a non-literal translation:
the Sun begins to ***bathe*** *the slopes of the landscape → le soleil qui* ***inonde*** *les flancs de ce paysage*
(Here we don't annotate it as *Equivalence*, because the non-literal translation not only keeps the same meaning, but also keeps the metaphorical image. We use the category *Figurative* to quantify this phenomenon.)

Counterexamples:

Borderline examples:
we can do a lot of things → on peut faire un tas de choses
(Borderline with *Équivalence*.)

6.1.9 Lexical shift

Percentage: 10.11%
Definition The translation is not literal, but there is no change in meaning. They are minor lexical level changes, which do not involve any translation technique.

change verbal tense or verbal modality (without changing the verb)
give you un update on how that machine ***worked***
→ vous mettre au courant de quelle façon cette machine ***fonctionne***
which ***might*** *seem like an odd thing → qui* ***peut*** *paraître bizarre*

change preposition
when you do a web search ***for*** *images → quand on fait une recherche web* ***sur*** *des images*

change determiner
if they believe enough there is a measurable effect in ***the*** *body*
→ s' ils y croient assez fort , il y a un effet mesurable dans ***leur*** *corps*

change subject
you *can move them →* ***on*** *peut les déplacer*
it 's clear that ***we*** *can make food → il est clair que l'* ***on*** *peut faire de la nourriture*

change position adverb
there 's a hole ***there*** *→ il y a un trou* ***ici***

change between singular and plural
sugar pills have ***a measurable effect*** *in certain kinds of studies*
→ des pilules de sucre ont ***des effets mesurables*** *dans certains types d' études*
(the boundary has been enlarged to keep the segment as a whole group)

change punctuation
I went through all the types of batteries that get made −− for cars ,
→ J' ai étudié tous les types de batteries qu' on fabrique ***,*** *pour les voitures ,*

Counterexamples:

I 've argued that we → je pense que nous
(The verb changes, not only the verbal tense, so it is a *Modulation (meaning changes in lexical level)*.)
at no time can anything travel , → à aucun moment quoi que ce soit ne peut passer .
(Same as above, this is a case of *Modulation (meaning changes in lexical level)*.)

6.1.10 Translation error

Percentage: 0.80%

Definition This category concerns obvious translation errors.

1. *the world in which we live does not exist in some absolute sense , but is just one **model** of reality*
   *→ le monde dans lequel nous vivons n' existe pas dans un sens absolu , mais est uniquement un **exemple** de réalité*
2. *to this day , they remain ruled by a ritual **priesthood** but the training for the priesthood is rather extraordinary .*
   *→ à ce jour , ils sont dirigés par un **clergé** rituel cependant , leur formation au clergé est plutôt extraordinaire .*
3. *there 's no one home anymore to **experience** a hallucination*
   *→ il n' y a plus personne à la maison pour **faire** une hallucination*
4. *we found two **instances** of natural death → nous avons découvert deux **circonstances** de mort naturelle*
5. *is not going to be **remembered** for its wars → ne sera pas **reconnu** pour ses guerres*

6.1.11 Uncertain

Percentage: 3.05%

Please attribute this category when you do not know which label to assign: because it is a difficult borderline example, or because there is something not clear in this annotation guide, and you want more discussion about this pair of segments.

- *this young man 's father had been **ascribed** to the Panchen Lama .*
  *→ le père de ce jeune homme avait été **affecté** au Panchen Lama .*

6.2 Annotation of unaligned segments

Definition Certain segments are left unaligned, there exist these three cases:

6.2.1 Unaligned - Explicitation

Definition Translations that could not be aligned to any source segment. (As a result, the annotation percentage is 0% at source side.)

Rule 1 Resumptive anaphora (Charolles, 2002): add a phrase or sentence summarizing the preceding information (which could be present in the previous sentence), to help understanding the present sentence.

Rule 2 Introduce in the target language clarifications that remain implicit in the source language but emerge from the situation:

[...] live amongst those who have not forgotten the old ways, who still feel their past in the wind
*→ [...] vivre parmi ceux qui n'ont pas oublié les anciennes coutumes, qui ressentent encore leur passé **souffler** dans le vent*

Rule 3 Specify that the sentence comes from which source (*e.g.* video, another speaker, etc.).
*the sun is rising . → **vidéo :** le soleil se lève .*

Rule 4 Adding obligatory words to keep the sentence grammatically correct:
*writing books is my profession → écrire des livres , **c'** est mon métier*

Rule 5 Annotate an added preposition or a repeated preposition after the conjunction as *Explicitation*:
are n't you afraid you 're going to keep writing for your whole life and you 're never [...]
*→ vous n' avez pas peur **de** passer votre vie à écrire et **de** ne jamais plus [...]*

6.2.2 Unaligned - Reduction

Percentage: 6.83%

Definition Deliberately remove certain words with concrete meaning that could be translated:

1. *and you 'll **suddenly** discover what it would be like → et vous découvrirez ce que ce serait*
2. *they say we , who are the younger brothers , are **the ones** responsible for*
   → ils disent , nous , les jeunes frères , sommes responsables de
3. *look **carefully** at the area of the eastern Pacific → regardez le secteur oriental de l' océan pacifique*
4. *we 've never needed **progress in** science more than we need it right now*
   → nous n' avons jamais eu autant besoin de la science que maintenant
5. *We **do** want to help you . → Nous voulons t' aider .*

6.2.3 Unaligned - No Type

Percentage: 14.47% (on source side).

Rule 1 Function words necessary in one language but not in the other:

1. *the last example I have time to → le dernier exemple **que** j' ai le temps de*
2. *a sequence that you can assemble → une séquence que **l'** on peut assembler*
3. *minus 271 degrees , colder than → moins 271 degrés , **ce qui est** plus froid que*

Rule 2 Segments not translated but which do not impact the meaning:

1. *he 's going to land in a couple of hours , **he 's going to** rent a car , and **he 's going to** come to Long Beach*
   → il va atterrir dans quelques heures , louer une voiture et arriver à Long Beach
   (The stylistic choice in English has been omitted in French)
2. ***and** everyone in this room has to get into it . → chacun de nous ici doit y monter .*

Rule 3 Target segments which do not correspond to any source segment:

Rule 4 Reformulation of the speaker: in figure 7, we don't annotate the reformulation which are incomplete segments.

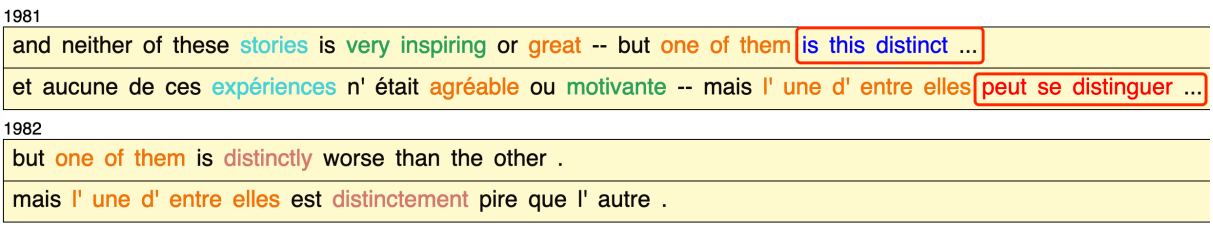

Figure 7: Don't annotate the incomplete segment of reformulation

7 Tutorial for using Yawat

In order to conduct all the operations available in Yawat, please use Firefox (version 67 or before)[4] as navigator instead of Chrome or Safari.

Url : `https://pakin.limsi.fr/cgi-bin/cgi/yawat.cgi`

Every annotator has an account created by the administrator. After logging in, according to the current task, please find the corresponding sub-corpus in different directories (*e.g.* pass1, pass2, pass3). Then click the corpus name to begin the annotation. (We also prepare a small test corpus for you to be familiar with the functions in Yawat.)

Here are some instructions about using this tool:

1. You can change the layout to show pairs of sentences in two lines or in two columns, by clicking on the button in the top right corner.
2. The black color is by default which means the category *Literal*. (But this does not mean that the boundary and alignment are all correct. Annotators should verify the alignment of all words.)
   At source side: blue means *unaligned and no type*.
   At target side: red means *unaligned and no type*.
3. Automatic word alignments have been imported, annotators should correct them when necessary. For example, in figure 8, there was just *open* and *ouvre* aligned automatically, and we want to add *up* to *open*.
   First, left click on either *open* or *ouvre*, the pair will be in a state of "being chosen", then left click *up*, we can see a new discontinuous boundary is created. Finally right click on any of these three words, the new alignment will be confirmed.

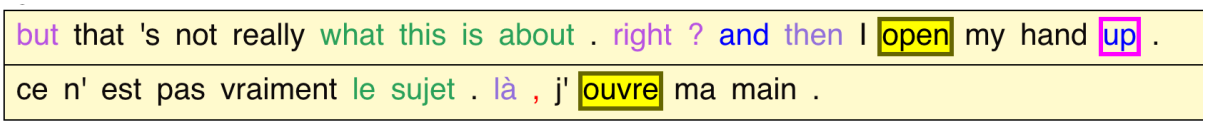

Figure 8: How to correct alignment in Yawat

4. If we want to remove a word from a segment, right click on it, and click *remove ... from this group*. If it is a lexical pair like the example in figure 9, when removing one word from one side, the link between them is broken in consequence. The corresponding word will change into the *unaligned and no type* state a moment later, you can hover your mouse on it to confirm it.

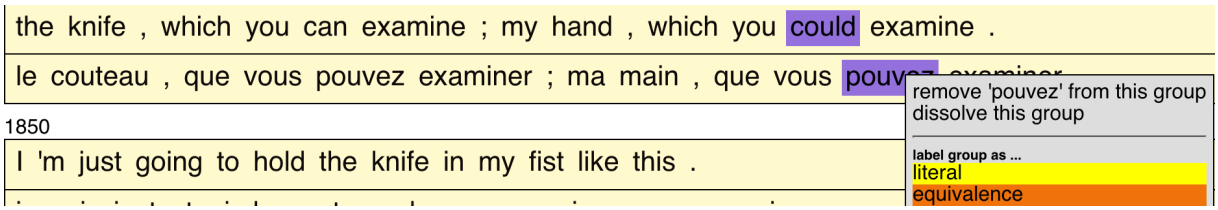

Figure 9: How to remove a word from a segment

5. To facilitate the alignment of a pair of long segments, for example in figure 10, first left click *nothing* and *il*, then right click either of them to create this pair.
   Next, left click this pair (by clicking either word inside) and add the rest of the segment by left clicking the other words, and finally right click on any word of them to confirm this long segment: *nothing goes up or down my sleeve* → *il n' y a rien dans mes manches*.
6. In order to annotate unaligned segments, *i.e. Explicitation* and *Reduction*, left click words to fix the boundary then right click the segment to choose the category from the menu (figure 11).

[4] If you need to use higher version of Firefox for other applications, you can continue to use Yawat by downloading the Firefox ESR version here: `https://www.mozilla.org/en-US/firefox/organizations/`.

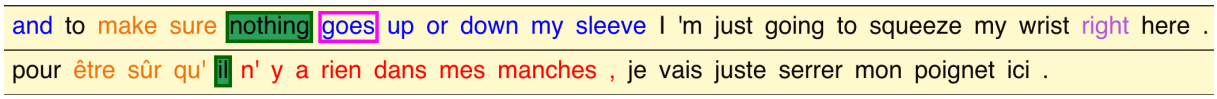

Figure 10: How to align long segments in Yawat

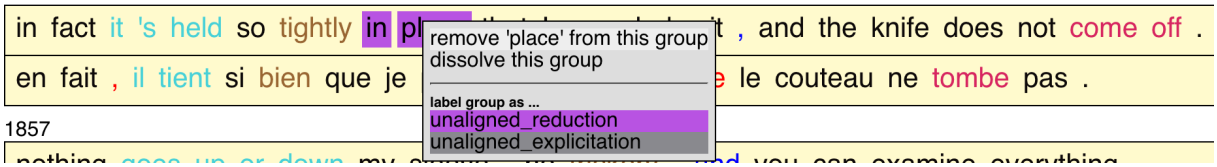

Figure 11: How to annotate unaligned segments

7. Once the boundary and the alignment are fixed, we can attribute a category. Right click any word in the pair, and choose a category from the menu (figure 12).

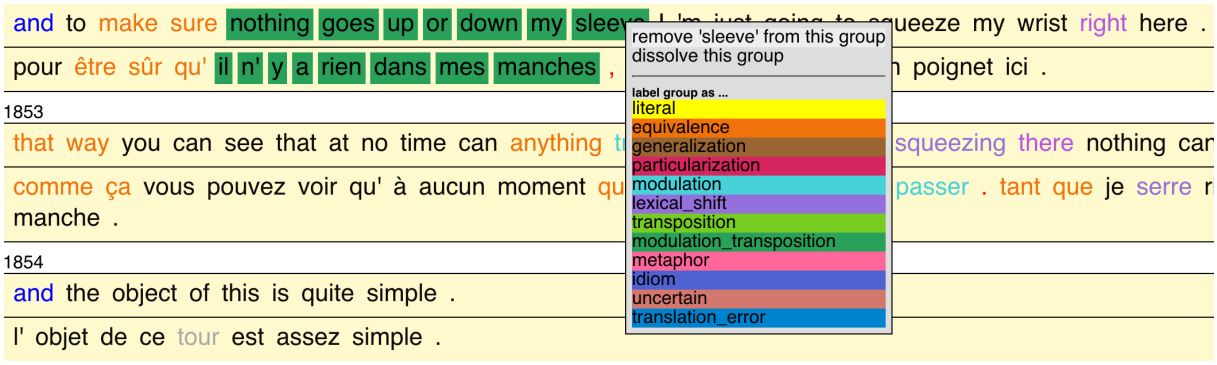

Figure 12: Choose category from the menu

8. Hover the mouse on words to check the boundary and the alignment of segments.
9. Make sure to save your annotations before logging out (click the **[save]** button in the top right corner, it appears only after changes).

Please contact us if you have any trouble in using this interface.

References

Michel Charolles. 2002. *La référence et les expressions référentielles en français*. Ophrys.

Hélène Chuquet and Michel Paillard. 1989. *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

Ulrich Germann. 2008. Yawat: Yet Another Word Alignment Tool. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Demo Papers*, pages 20–23. The Association for Computer Linguistics.

Klaudia Gibová. 2012. *Translation Procedures in the Non-literary and Literary Text Compared*. Books on Demand.

Lucía Molina and Amparo Hurtado Albir. 2002. Translation Techniques Revisited: A Dynamic and Functionalist Approach. *Meta*, 47(4):498–512.

Peter Newmark. 1981. *Approaches to Translation*. Language teaching methodology series // Pergamon Institute of English. State University of New York Press.

Peter Newmark. 1988. *A Textbook of Translation*. English language teaching. Prentice-Hall International.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: méthode de traduction*. Bibliothèque de stylistique comparée. Didier.

Annotation Guidelines of Translation Techniques for English-Chinese

Yuming Zhai
LIMSI, CNRS
Université Paris-Sud
Université Paris-Saclay
Orsay, France
zhai@limsi.fr

Lufei Liu
Université Paris Diderot
Paris, France
luffyliu1108@gmail.com

Xinyi Zhong
Université Sorbonne Nouvelle
Paris, France
zhongxinyisophie@gmail.com

Contents

1 Introduction

1.1 Task description

This document is a guide for annotating translation techniques in a parallel corpus. For example, in this work, our compiled English-Chinese parallel corpus contains 2 200 pairs of sentences, covering eleven different domains: *Art, Educational Materials, Literature, Law, Microblog, News, Official document, Spoken, Science, Subtitle* and *Scientific article*. Each sentence pair contains on average 24 English tokens and 38 Chinese characters. The translation direction is from English to Chinese, except for the domain of *Scientific article*.

The Chinese corpus has been tokenized by the tool *THULAC* (Li and Sun, 2009)[1] and English corpus by *Stanford Tokenizer*[2]. The parallel corpus has been further automatically aligned at word level by the tool *TsinghuaAligner* (Liu and Sun, 2015)[3].

1.2 Translation techniques

Consider this example (all examples in this guide are shown with tokenization):
I know what people think when they see this .
→ 我 知道 人们 想 什么 当 他们 看到 这个 的 时候 。

For this pair of sentences, we could conduct word alignments as follows, where all the source segments have been translated literally.
I → 我
know → 知道
what → 什么
people → 人们
think → 想
when → 当 ... 的 时候
they → 他们
see → 看到
this → 这个
. → 。

Translation techniques constitute an important subject of study for translators and linguists (Vinay and Darbelnet, 1958; Newmark, 1981; Newmark, 1988; Chuquet and Paillard, 1989; Molina and Hurtado Albir, 2002; Gibová, 2012), which distinguish literal translation from other translation techniques on word or phrase level. Based on the above cited work in translation techniques, by annotating and analyzing our English-Chinese parallel corpus, we have proposed a typology of translation techniques (see figure 1) in order to have a global view of these categories. The table 1 and 2 provide a recapitulation of definition and important rules for each translation technique.

2 General instructions

2.1 How to deal with minor misspelling?

In order to guarantee the corpus quality and to generate a clean data set for developing our automatic classifier, we correct minor spelling errors in the corpus, for example 作法 → 做法.

During the annotation, please note down the sentence ID and the misspelled pair.

2.2 How to deal with Chinese segmentation errors?

The word segmentation has first been done automatically with the tool *THULAC* (Li and Sun, 2009). The remaining errors should be corrected manually before or during the annotation. If you see segmentation errors, please **skip the sentence first**, and tell us to correct the segmentation. (Because if we align all the words and then correct the segmentation, the alignment indices will be influenced.) For example:

[1] http://thulac.thunlp.org/
[2] https://nlp.stanford.edu/software/tokenizer.html
[3] http://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html

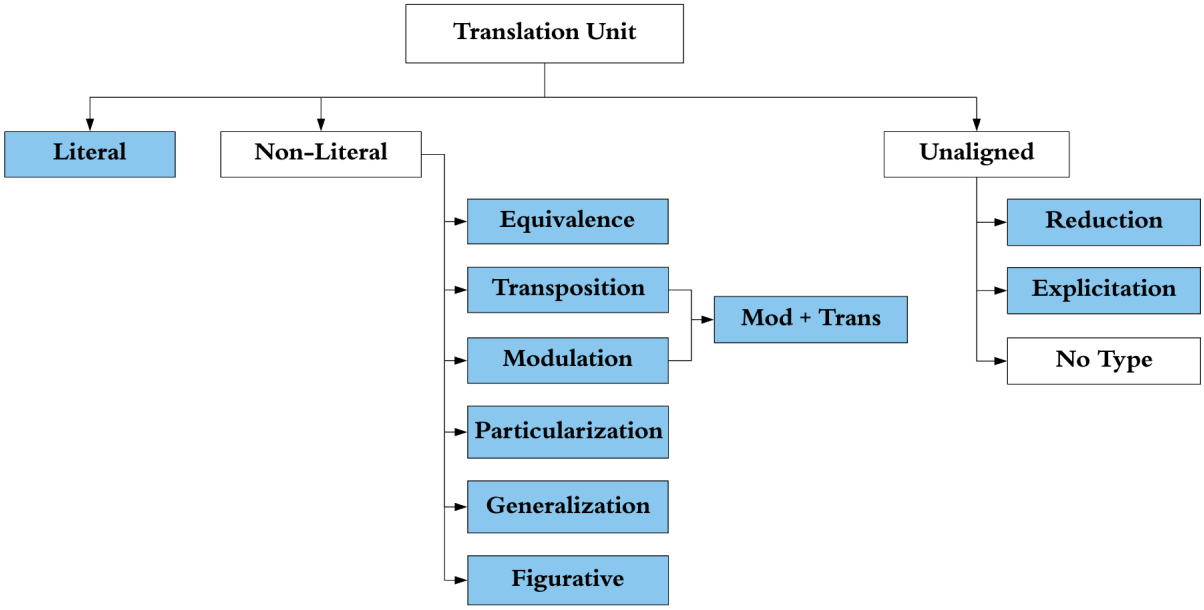

Figure 1: Typology of translation techniques

就是, 只是: should be tokenized into "就 是" (就: emphasise something)， "只 是"
们: it is a suffix and should be tokenized

2.3 Three conditions of annotation

- Raw text without any manual annotation: you should conduct segmentation of translation units, correct existing automatic word alignments, attribute categories of translation technique.
- Text which has already been annotated once: in order to guarantee the quality of our work, you need to verify the annotation. You can modify the alignment and category attribution if there is a disagreement. You can also modify the boundary of translation units.
- Text which has been annotated twice, and it was you who conducted the first pass: please try to reach consensus with the other annotator on the remaining differences between you two. We will provide you with the differences between the two versions to accelerate the reviewing.

2.4 External resources

Annotators are encouraged to use (online) authoritative resources, such as *Cambridge Dictionary, Longman Dictionary, TheFreeDictionary*, etc. to consult word senses, part of speech information, meaning of multiword expressions, terminology translation and so on. When annotators have doubt about word senses, they should refer to an English monolingual dictionary, instead of a bilingual dictionary.

For example, online dictionaries give all possible parts of speech of words with examples, which helps to choose the translation technique category (see figure 2).

3 Decision helper

In order to facilitate the annotation task, please see figure 3 to help you to make decisions on the most confusing categories. Please note that this table recapitulates the most distinguishing aspects for each category, and doesn't include all the definitions and rules presented below.

During the annotation, there exist different categories concerning idioms[4] and fixed expressions[5] in

[4]Idiom: a phrase or an expression that has a figurative, or sometimes literal, meaning. The figurative meaning is based on the whole rather than on the individual words in it. Idiomatic expressions are strongly cultural and have different meanings derived from the cultures they come from. For example: *a piece of cake, every cloud has silver lining*.

[5]Fixed expression: a standard form of expression that has taken on a more specific meaning than the expression itself. It is used as a part of a sentence, and is the standard way of expressing a concept or idea. Unlike idioms, they are generally transparent in meaning. For example: *as a matter of fact, all of a sudden, to whom it may concern*.

Translation Technique	Definition and important rules
	Aligned segments
Literal (6.1.1)	Word-for-word translation: *a **bronze** ring* → 一 个 **青铜** 戒指 Borrowing words using transliteration: *a cup of **coffee*** → 一 杯 **咖啡** Possible literal translation of idioms: ***ivory tower*** → **象牙 塔** Corresponding expression when absolute literal translation does not make sense: ***I give you my word.*** → **我 向 你 保证** 。 ("I promise you.")
Equivalence (6.1.2)	Non-literal translation of proverbs, idioms, or fixed expressions: ***A friend in need is a friend indeed.*** → **患难 见 真情** 。 ("Misfortune tests the sincerity of friends.") No change in meaning and point of view, a word-for-word translation makes sense but the translator has produced a different translation: *protect all locations **at all times*** → **日夜** ("day and night") 保护 所有 的 地点
Transposition (6.1.3)	Change grammatical categories without changing the meaning: *She **was careful** not to question him, **fearful** that he might leave them.* → 她 也 **小心 地** ("carefully") 从不 问起， **生怕** ("to fear") 他 走 了 。
Modulation (6.1.4)	Change the point of view, can be encountered both in lexis and syntactic structures: *I **like** the dreams of the future **better than** the history of the past .* → 我 **不** ("don't") 缅怀 ("recall") 过去 的 历史， **而** ("but") 致力于 ("devote myself to") 未来 的 梦想 。 Slight meaning change in lexical level according to the context: *he had **rudely** bellowed across the supper table to her* → 他 隔着 餐桌 对 她 **大声** ("loudly") 吼叫
Mod+Trans (6.1.5)	Combine the transformations in *Modulation* and *Transposition*: *One by one the other elders now timidly rise **with** innocuous requests.* → 其他 的 长老 一个 接 一个 怯生生 地 站起来， **提出 了** ("put forward") 一些 不关痛痒 的 要求
Particularization (6.1.6)	The source segment could be translated into several target segments with more specific meaning, and the translator has chosen one of them according to the context: *"Yes, **put** you to bed"* → *"是的， **服侍** ("serve") 你 上床 睡觉"* Specify the meaning of a segment in context: *On his **best days**, Gomes is a very nice, solid bench player.* → 当 他 **打 得 好 的 时候** ("play well")， 戈麦斯 是 很 优秀、很 得力 的 板凳 球员 。 Translate a pronoun by the thing(s) it references: *He then requested her to stay where **she** was* → 他 先 让 **苔丝** ("Tess") 在 外面 等 着
Generalization (6.1.7)	Several source words or expressions could be translated into a more general target word or expression, and the translator used the latter to translate: *a research that will be embraced by **millions of** bleary-eyed Britons* → 一 项 即将 被 **广 大** ("numerous") 睡眼惺忪 的 英国人 所 知道 的 研究 The translation of an idiom by a non-fixed expression: ***Every man has a fool in his sleeve.*** → **人人 都有 糊涂 的 时候** 。("Every man is a fool sometimes.") The removal of a metaphorical image: *But should clouds gather over the Atlantic, or **tempers rise** in the Middle East [...]* → 如果 大西洋 风云 再 起， 中东 **战火 重 燃** ("war resumes") 的 话 [...]
Figurative translation (6.1.8)	Introduce an idiom to translate a non-fixed expression, or a metaphorical expression to translate non-metaphor: *He gave the required information, **in words as suitable as he could find**.* → 他 **字斟句酌 地** ("weigh one's words") 作 了 回答 。 Use personification to translate: *For Joanne, new opportunities **are opening.*** → 对 乔安娜 而 言， 新 的 机遇 **现 已 向 她 招手 。** ("are waving to her")
Lexical shift (6.1.9)	Change of verbal tense, verbal modality or of determiner, changes between singular and plural forms, and other minor changes alike. *My cousin **is launching** a photography exhibit in the USA this week .* → 我 的 表哥 本周 在 美国 **推 出 一** 个 摄影 展 。
Translation error (6.1.10)	Obvious translation errors. *database **connection** method* → 数据库 **访问** 方式
Uncertain (6.1.11)	Not sure about which category to assign, need more discussion.

Table 1: Definition and important rules for aligned segments

the source or target language, the table 3 recapitulates them. Below is their definition.

4 Annotation tool

We use the web application Yawat (Germann, 2008) for our annotation, if you don't know how to use this tool, please read the section 7.

Translation Technique	Definition and important rules
Unaligned segments	
Explicitation (6.2.1)	Introduce clarifications that are implicit in the source text: *the building blocks of the universe* → 宇宙 **形成**("form") 的 **最** ("most") 基本 单位 Add Chinese-specific words: *the knife* → 这 **把** 刀 (Chinese measure word) *I will bring it to China.* → 我 可以 **把** 它 带到 中国 来。 (necessary addition due to syntactic order change in translation)
Reduction (6.2.2)	Deliberately remove certain words in translation: Removal of preposition: *A spokesman **from** the Ministry of National Defense* → 国防部 发言人 Removal of copula: *Peter **is** six years old.* → 彼得 六岁 。 Removal of anticipatory « it »: ***It** was a pleasant surprise to learn of her marriage.* → 得知 她 结婚 是 件 令 人 惊喜 的 事 。
No Type (6.2.3)	Function words necessary in English but not in Chinese: *The tragedy of the world is **that** those who are imaginative have but slight experience.* → 世界 的 悲剧 就 在于 有 想象力 的 人 缺乏 经验 。 Segments not translated but which do not impact the meaning: *The present state, application and development of coal mine hydraulic drill rig are described **in this paper**.* → 介绍 了 煤矿 用 液压 钻 车 现状，使用 情况 及 发展 。 Target segments added without reason, which do not correspond to any source segment.

Table 2: Definition and important rules for unaligned segments

Figure 2: Annotators are encouraged to consult authoritative dictionaries

	word-for-word	non-literal translation of proverbs, idioms, fixed expressions	change point of view	change syntactic structure	change meaning	change grammatical category	more specific meaning	more general meaning
Literal	x							
Equivalence		x						
Transposition						x		
Modulation			perhaps	perhaps	perhaps			
Mod+Trans			perhaps	perhaps	perhaps	x		
Particularization					x		x	
Generalization					x			x
Lexical shift								

Figure 3: A table to help annotators to make decisions on the most confusing categories

5 Annotation conventions

5.1 How to decide the segment boundary?

In principle

Translation phenomenon	Category attributed
literal translation of an idiom	*Literal* (see subsection 6.1.1)
idiom → equivalent idiom	*Equivalence* (see subsection 6.1.2)
idiom → non-fixed expression	*Generalization* (see subsection 6.1.7)
non-idiom → idiom	*Figurative* (see subsection 6.1.8)
non-fixed expression → fixed expression	*Equivalence* (see subsection 6.1.2)

Table 3: Different categories concerning idioms and fixed expressions

The segment boundary is not provided to annotators and it should be fixed by respecting the given tokenization, while excluding the part not involved:

I 'll get my ***sleeve*** *back* → 然后 我 把 **袖子** 挽 起来

experience of reading → **阅读 体验** (include « of » in the boundary)

For simple literal lexical translation We annotate the smallest semantic unit as we can, for example given this pair:

this is not a Hollywood special effect → 这 不 是 好莱坞 特效

We should segment and align like this:

this → 这

not → 不

is → 是

Hollywood → 好莱坞

a special effect → 特效

Articles and prepositions

- In the following two examples, we annotate the article together with the noun, because the English article sometimes corresponds to an empty position in Chinese:
  1. *have* ***a screen*** *and* ***a wireless radio*** → 有 **显示屏** 和 **无线电 装置**
  2. *explore* ***the answer*** → 探寻 **答案**
- Here we align « to » with 《给》, both of which are followed by the indirect object of the action « show ».
  I 'll show it ***to*** *you* → 我 可以 展示 **给** 你们
- Align « want to » with《想》: regroup the preposition « to » with the verb « want ».
  I want to *start out by asking you* → 我 **想** 请 大家
- In this example, « based on » is aligned to《基于》. Adding the preposition « on » here changes the meaning of the verb, because « base » alone means « to situate at a specified place as the centre of operations ». Then « the content » is aligned to 《内容》, because « the » corresponds to an empty position in Chinese.
  based on the content *inside the image* → **基于** 图片 的 **内容**.
- Some English prepositions are translated by Chinese discontinuous segments:
  in → 在...中/里; 在...的时候
  for → 对...来说
  as → 在...时
  from → 从...中
  since → 从...开始
  if → 如果...的话
  dive in → 钻到...里
  related to → 跟...有关
  paring ... down to → 分解 到
  so...that → 如此...甚至
  stitched → (被)...拼合 (passive voice)

For non-literal translations Sometimes it is necessary to enlarge the boundary of the segments, in order to clarify the meaning, even though there are words which could be annotated as *Literal* inside this segment. For example :

1. ***you get to choose which one you want to use*** → **任 君 选择**
   (*Equivalence*, non-literal translation with no change in meaning, grammatical categories and point of view, subsection 6.1.2)
2. ***so many of*** → **不少**
   (*Modulation*, affirmative form → negative form, subsection 6.1.4)
3. *we* ***damaged*** *about 50 of the magnets* → 大概 有 *50* 个 磁铁 **受 损**
   (*Modulation*, active voice → passive voice)
4. *this question was* ***so compelling*** → 这样 的 疑问 **不断 萦绕** 在 我 的 脑海 中
   (*Transposition*, adjective → verb, subsection 6.1.3)
5. ***immense celebration*** → **欢呼雀跃**
   (*Transposition*, noun → verb)
6. ***become linked*** *together* → **交织** 在 一起
   (*Transposition*, adjective → verb)

Composition of categories In the above cases, the literal part is considered as neutral, when combined with another category, the latter becomes the translation technique for the entire segment.

Thus, in our work, the translation unit could be a word, a phrase or a short sentence (do not include the final punctuation):

Nice one → **好的** (*Equivalence*, subsection 6.1.2)

What happen is this **.** → **它 是 这样 发生 的** 。 (*Equivalence*)

5.2 How to align punctuation?

See table 4 for a recapitulation.

Punctuation	Alignment
if the punctuation doesn't change	
Final punctuation (period, exclamation mark, question mark), comma, colon, semicolon, quotation marks, angle quotes, brackets, ellipsis, etc.	Align them out of the segments.
Apostrophe, dash, hyphen	Respect the given tokenization. If you disagree with them, please contact us.
if the punctuation changes, or change between a conjunction word and punctuation	
For example, the translator replaces a double dash by a comma.	Annotate as *Lexical shift* (6.1.9).

Table 4: How to align punctuation

5.3 Mutually exclusive categories

There exist difficult borderline examples, but we do not allow multiple categorization in this task, which means that a pair of segments receive always one category listed in our typology (see figure 1).

For borderline examples, after discussion, annotators should agree on a category which better reflects the technique used by the translator.

5.4 How to annotate unaligned segments?

For the categories of *Unaligned - Explicitation* and *Unaligned - Reduction* (see definition in table 2), please annotate separately each span of reduced or added segment (separated by aligned segments), do not make a whole group.

For example in figure 4, please annotate these reduced instances separately: *« the terms of the », « by »*.

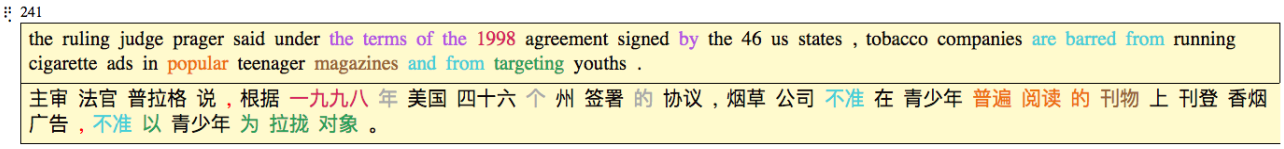

Figure 4: How to annotate unaligned reduced instances

5.5 How to deal with linguistic anaphora?

Linguistic anaphora is different from rhetorical anaphora which consists of repeating words at the beginning of phrases in order to make emphasis.

Linguistic anaphora can be the repetition of noun subject or words of other parts of speech. In case of using anaphora in translation, we annotate all the repeated words, which are not necessarily the same, with the English source word together. The annotation type is decided by the translation technique.

Here are some examples :

1. The noun subject is repeated and annotated as *Literal* (see figure 5). 《两国》("the two countries") refers to « Pakistan and India » which are not directly translated by their names.

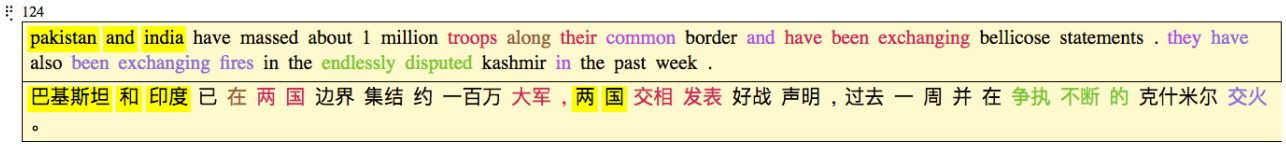

Figure 5: Repetition of subject, annotate as literal

2. In figure 6, the verb and adjective are repeated and annotated as *Particularization*. « many » is translated into《一 次 又 一 次》and 《无数 次》; « made » is translated into 《发》and 《下》.

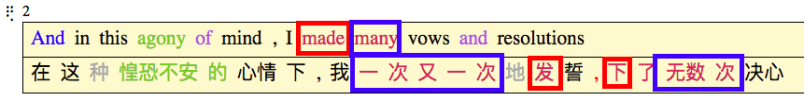

Figure 6: Repetition of verb and adjective, annotate as *Particularization*

3. The determiner is repeated by using the same word, annotate as *Literal*.
   I've loved another with all ***my*** *heart and soul .*
   → 我 用 **我 的** 心 和 **我 的** 灵魂 爱 过 另 一个 人 。

6 Annotation in practice

6.1 Annotation of aligned segments

Literal translations

6.1.1 Literal

Definition Word-for-word translation (including insertion or deletion of articles), or possible literal translation of some idioms:

the ***sunlight*** *is very warm* → **阳光** 很 温暖

I 've loved another ***with*** *all my heart and soul .* → 我 **用** 我 的 心 和 我 的 灵魂 爱过 另 一 个 人

(According to bilingual dictionaries, « with » can be translated by《用》according to the context, we consider this as a literal translation.)

The *book is a collection of his reminiscences about the actress .*

→ **这** 本 书 辑录 了 他 对 那 位 女 演员 的 回忆 。

Rule 1 Borrowing English words using transliteration:
a cup of ***coffee*** → 一 杯 **咖啡**

Rule 2 When an absolute word-for-word translation does not make sense in the target language, the corresponding expression is deemed to be literal:
in other words *(?用 其它 话)* → **换 句 话 说**
I give you my word *. (?我 给 你 我的 话)* → **我 向 你 保证** 。

Rule 3 The change of numbers between the spelled out form and the Arabic form is annotated as *Literal*:
nearly ***4*** *million tons of food must be imported* → 必须 要 进口 接近 **四** 百万 吨 的 粮食

Rule 4 When the Chinese structural particle《的》marks the possession:

1. *Some people come into* ***our*** *life and quickly go .*
   → 一些 人 闯进 **我们 的** 生活 ， 又 匆匆 走 了 。
   (regroup《的》and the personal pronoun before it, annotate as *Literal*)
2. *in any deal , you need to know your opponent* ***'s*** *breaking point*
   → 在 任何 交易 中 你 都 需要 知道 对手 **的** 忍耐 度
   (when《的》is after a noun, align《的》with « 's »)
3. ***His*** *father emptied sacks of stale rye bread into the vat .*
   → **他** *(*的*)* 父亲 把 一袋袋 发霉 的 黑 面包 倒进 大 桶 里 。
   (Sometimes the《的》after a personal pronoun could be omitted in oral conversation or informal text, in this case, annotate as *Literal*.)

Rule 5 When the Chinese particle《的》is added after an attributive adjective, annotate《的》and the adjective together as *Literal*:
It 's a ***beautiful*** *3D graphics .* → 它 是 **漂亮 的** 3D 图画

Rule 6 The translation of proper noun is annotated as *Literal*:
Saltanov *will discuss Russian president* ***Putin*** *'s proposal .*
→ **沙福诺夫** 将 讨论 俄罗斯 总统 **普京** 的 提议 。

Rule 7 Fixed modulation is annotated as *Literal*:
a ***life jacket*** → 一 件 **救生 衣**
(The means《救生》("save life") substituted for the result « life », from a linguistic point of view, the category should be *Modulation*. However, since there is no other possible translation for « life jacket », and it's a recorded pair in bilingual dictionaries, we annotate this case by *Literal*.)

For each translation technique, we show some counterexamples and borderline examples:

Counterexamples:
the arms reduction treaty → 裁减 武器 条约
(The translation is word-for-word, but《裁减》("reduction") in Chinese is a verb. The category should be *Transposition*.)

Borderline examples:
in the past week → 上周
(This is a borderline example with *Equivalence*, but should be annotated as *Literal*.)

Non-literal translations

6.1.2 Equivalence

Definition There is no change of point of view like in *Modulation* (subsection 6.1.4). A word-for-word translation makes sense but the translator has expressed differently. However, if there exist changes of grammatical categories, the pair should be annotated with *Transposition* (see examples in subsection 6.1.3).

*She decides to string along with others **as** she has nothing else to do .*
→ 她 **反正** 也 无 事 可 做 ， 所以 就 跟着 大家 去 了 。

(« as » should be translated into《由于》(“since") but in this sentence, it’s translated into《反正》("anyway") to emphasize the reason.)

Rule 1 Non-literal translation of proverbs, idioms or fixed expressions:
A friend in need is a friend indeed . → **患难 见 真情** 。
Like father , like son → **有 其 父 必 有 其 子**

Rule 2 Cultural equivalence :
In the figure 7, the segment « great to » is translated into《荣幸》("it’s a great honour to ..."). A word-for-word translation makes sense here but it’s necessary to show the honoured side in the Chinese context.

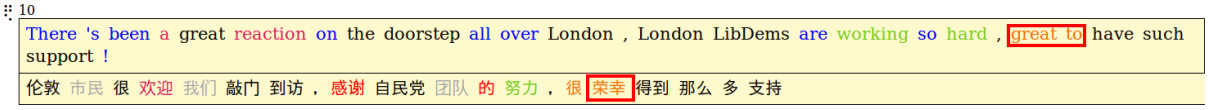

Figure 7: Cultural equivalence

swear to God → **对 天 发誓**

Rule 3 Change measure into the one used in the target culture:
*I bought **5 half kilo** of banana .* → 我 买了 **五 斤** 香蕉 。

Rule 4 Translate a non-fixed expression by a fixed expression:

Rule 5 Translate an abbreviation into a full version or vice versa:
SAFE → **国家外汇管理局**
(The full version of « SAFE » is « State Administration of Foreign Exchange ».)

Rule 6 Translate the title of a book:
THE COMPLETE FORTUNE-TELLER →《 算命大全 》

Counterexamples:
in other words → 换 句 话 说
(The word-for-word translation does not make sense in Chinese so the category is *Literal*.)

Borderline examples:
along the Kashmir demarcation of control separating India from Pakistani territories
→ 沿着 分隔 印度 与 巴基斯坦 辖区 的 克什米尔 控制 分界线
(Borderline with *Particularization*. According to Cambridge Dictionary, « territories » corresponds to《领土，领域，活动范围》, not 《辖区》. Annotate as *Equivalence* for now.)

6.1.3 Transposition

Definition Translating words or expressions by using other grammatical categories than the ones used in the source language, without altering the meaning of the utterance.

The change of grammatical categories could occur on a complete syntagm, for example:

- *" He who asserts must prove "* ***is rational*** → “ 谁 主张 谁 举证 ” **有 其 合理性**
  (« is rational » is translated into《有 合理性》 ("has rationality"))
- *Those who* ***are experienced*** *have feeble imagination .*
  → **有 经验** 的 人 缺乏 想象力 。
  (« are experienced » is translated into《有 经验》 ("have experience"))

verb -> noun
London LibDems are ***working*** *so* ***hard*** → 感谢 自民党 团队 的 **努力**
noun -> verb
It was a pleasant surprise to learn of her ***marriage*** *.*
→ 得知 她 **结婚** 是 件 令 人 惊喜 的 事 。
adj -> verb
I 'd like to talk about the oldest ***nutritional*** *method on Earth .*
→ 我 想 讲 一下 地球 上 最 原始 的 **获得 营养** 的 方式 。
adj -> noun
In figure 8, the English adjective « russian » is translated by a noun《俄罗斯》(the name of the country). Because in Chinese grammar, the noun is used as an adjective here (Fang, 2008).

Figure 8: In Chinese, a noun can be used as an adjective

prep -> verb
patients ***over*** *the age of 40* → **超过** 四十 岁 的 病人

When the Chinese particle《的》is used to replace a preposition or a relative pronoun in order to describe the subject, the annotation type should be *Transposition*.
prep -> particle
people ***of*** *Iran* → 伊朗 **的** 人们
The Irish representative team is currently training on Saipan ***in*** *the pacific ocean .*
→ 爱尔兰 代表队 目前 在 太平洋 **的** 塞班岛 集训 。
The air force ***from*** *the two nations will provide support .*
→ 两 国 **的** 空军 将 给以 支援 。
relative pronoun -> particle
Every citizen ***who*** *is able to work .* → 一切 有 劳动 能力 **的** 公民。
determiner -> pronoun
It was a pleasant surprise to learn of ***her*** *marriage .*
→ 得知 **她** 结婚 是 件 令 人 惊喜 的 事 。

Counterexamples:
Now boarding all passengers for flight 624 ***to*** *Newyork .*
→ 现在 所有 乘坐 *624* 航班 **飞往** 纽约 的 旅客 请 登机 。
(There is not only the change of grammatical category but also the meaning:《飞往》("fly to"). So it should be annotated as *Mod+Trans*.)

Borderline examples:
There is no need to see the little detailed words on it .
→ 不用 去 看 里面 具体 的 小字。

(Here, « on » is a preposition and《里面》("inside") is a noun. The meaning and the part of speech are both changed. Although it is borderline with *Lexical shift*, it's annotated as *Transposition* now.)

By the standards of history , his execution in 399 B.C. was singularly humane .
→ 按照 历史 的 标准 来看 ， 公元前 399 年 处死 他 的 方式 是 异常 人道 的 。
(Borderline with *Equivalence* (in this specific context). The point of view does not change.)

6.1.4 Modulation

Definition This translation technique consists of a change in the point of view that enables the translator to convey the same phenomenon in two languages in a different way.

- can be encountered both in lexis and in syntactic structures
- reveal a specific way of seeing things for the speakers of the target language
- circumvent translation difficulties
- the translator desires to achieve expression naturalness in the target language
- could bring meaning changes between source and target text

According to Vinay and Darbelnet (1958), there exist two types of modulation:

- metonymical modulations (the cause substituted for the effect, the container for the content, the part for the whole, etc. and vice versa)
- grammatical modulations (change between affirmative form and negative form; between injunction and interrogation; between passive voice and active voice; the subject becomes the object, etc.)

Examples:

Metonymical modulations:

1. *If you want , you can lay color in the mold , and get rid of* ***the paint shop*** *.*
   → 如果 你 愿意 ， 你 可以 在 模具 里 直接 加 颜色 ， 就 省得 **给 车 喷漆** 了 。
   (Use the concrete action《给车喷漆》("paint the car") to substitute for the abstract « the paint shop ».)
2. *I 'm thinking of becoming a suicide* ***bomber*** *.*
   → 我 正在 想 如何 成为 一个 自杀 **炸弹** 。
   (Use the object《炸弹》("bomb") to substitute for the « bomber ».)
3. Geographical modulation:
   In English « Downing street » is often used to refer to « the office of the UK prime minister ».
   By metonomy, English speaker often refer to 《中国政府》("Chinese government") by « Beijing ».

Change the point of view:

4. *Louis Hicks , what 's the matter ?* ***Work a little slow ?***
   → 路易斯希克斯 ， 怎么 了 ？ **没事 做 吗** ？
5. ***Just needs a little something more*** *.* → **还 少 了 一点 东西** 。

Change between passive voice and active voice:

6. *nearly 4 million tons of food must* ***be imported*** → 必须 要 **进口** 接近 四百万 吨 的 粮食

Affirmative form to negative form, the negation of the opposite:

7. *I* ***like*** *the dreams of the future* ***better than*** *the history of the past .*
   → 我 **不** 缅怀 过去 的 历史 ， **而** 致力于 未来 的 梦想 。
   (《不.....而》means « not...but ». Since the affirmative form « like...better than » is changed into negative form « not...but », the sentence's order is reversed : « the history of the past » is now at the beginning of the sentence.)
8. *It* ***'s difficult*** *.* → 这 **不 简单** 。

The subject becomes the object:

9. ***What do you hope to gain from*** *those colors ?*
   → 那 颜色 **能 给 你 什么 希望** ？

Obligatory syntactic change but no change in meaning:

10. ***There have been increased*** *contacts between the United States and India .*
    → 美 印 间 的 接触 **增加** 。
    (The English structure « There be + noun » can't be translated in a word-for-word way in Chinese, so the adjective « increased » is transformed to the intransitive verb 《增加》("increase"). Thus, the noun phrase « contacts between the United States and India » becomes the subject in Chinese.)

Circumvent translation difficulties, achieve expression naturalness:

11. ***Maybe*** *we should postpone till the weekend .* → **不如** 我们 推迟 到 周末 吧 。

Slight meaning change in lexical level according to the context:

12. *British airways* ***posts*** *first annual loss in 15 years .*
    → 英航 **出现** 十五 年 来 首次 年度 亏损 。
13. ***Stop by*** *tomorrow and drop off your ID badge and all that .*
    → 明天 **去** 店里 交上 你 的 工作证 还有 别的 。
    (《去》means « go to ». In this sentence, « stop by » is translated into « go to » which is more natural in the context.)
14. *I want to tell* ***you*** *three things today .* → 我 今天 想 告诉 **大家** 三 件 事 。
    (In this specific context, the speaker is talking to an audience, so « you » is translated into《大家》("everybody").)
15. *she remembered how frequently Gerald had* ***rudely*** *bellowed across the supper table to Suellen*
    → 她 想起 杰拉尔德 时常 隔着 餐桌 对 苏伦 **大声** 吼叫

Counterexamples:

Borderline examples:
My plan is to live here until I 'm 30 , and then enter a hermitage .
→ 我 的 计划 是 在 那里 一直 住到 我 三十 岁 ， 然后 过 起 隐居 生活 。
(Borderline with *Equivalence*, annotated as *Modulation* now.《隐居生活》("live in seclusion") substitutes « hermitage » which is the place to live in seclusion.)

6.1.5 Mod+Trans

This category combines the transformations in *Modulation* and *Transposition*, for example:
In the figure 9, the noun phrase « the attack on the national parliament in new delhi last december » is transformed into a verbal phrase《去年 十二月 新德里 国会 遭 攻击》. As a result, the preposition « on » is semantically aligned to the verb《遭》("encounter"). This is a case of transposition made necessary by the syntactic changes.

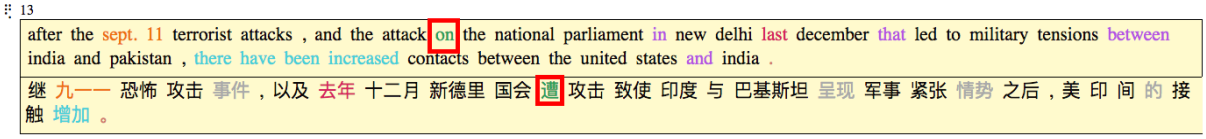

Figure 9: Necessary transposition caused by a syntactic change

1. *Tobacco companies are barred from running cigarette ads in popular teenager magazines and from* ***targeting*** *youth .*
   → 烟草 公司 不准 在 青少年 普遍 阅读 的 刊物 上 刊登 香烟 广告 ， 不准 **以** 青少年 **为 拉拢 对象** 。
   (The verb « target » in English becomes a noun 《拉拢对象》("target") in Chinese, and the verbal phrase « targeting youth » is transformed into a noun phrase 《以青少年为拉拢对象》 ("taking

youth as target"), so the whole part is annotated as *Mod+Trans*.)

2. When a preposition is translated into a verb to define the action, annotate the preposition with the verb as *Mod+Trans*.
   For example, in the figure 10, the preposition « with » is translated into 《提出了》("put forward").

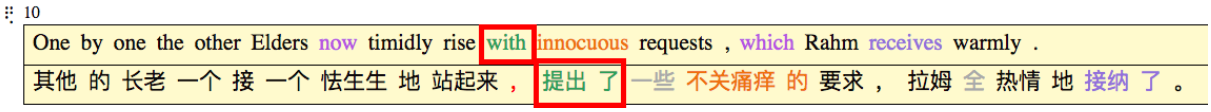

Figure 10: Translate a preposition into a verb

Counterexamples:
What I 'd love is for my highly paid publicist to take care of it .
→ 我 想 要 我 高 收费 的 公关 经理 搞定 这事 。
(There is only *Modulation* used in this translation. The point of view is changed from « highly paid » to《高 收费》("highly charged").)

Borderline examples:

6.1.6 Particularization

Definition The source word could be translated into several target words with more specific meaning, and the translator has chosen one of them according to the context.
*" Yes , **put** you to bed , " she added lightly* → " 是的 , **服侍** 你 上床 睡觉 , " 她 小声 补充 说

Longer boundary Annotators could enlarge the boundary to help disambiguate the meaning of the segment:

1. *On his **best days** , Gomes is a very nice , solid bench player .*
   → 当 他 **打 得 好 的 时候** , 戈麦斯 是 很 优秀 、很 得力 的 板凳 球员 。

Rule 1 Specify the meaning of a word in context is annotated as *Particularization*:
*drag yellow diamonds to adjust arrowhead or **tail*** → 拖动 黄色 菱形 可以 调整 箭头 或 **箭尾** 。

Rule 2 Translate a pronoun by the thing(s) it references is annotated as *Particularization*:
***Who is which** depending on whether the customer 's account is in credit or is overdrawn .*
→ 究竟 **谁 是 债务人 谁 是 债权人** , 要 看 储户 是 有 结余 还是 透支 。

Rule 3 Use different words to give colour to translation is annotated as *Particularization*:
*Pakistan and India have massed about 1 million **troops** along their common border .*
→ 巴基斯坦 和 印度 已 在 两 国 边界 集结 约 一百万 **大军** 。

Counterexamples:
*the plan includes a lay-off of 2,000 **employees*** → 这项 计划 括及 裁员 两千 人
(This is an instance of *Generalization*.)

Borderline examples:

6.1.7 Generalization

Definition This translation technique is the opposite of *Particularization*. Several source words or expressions could be translated into a more general target word or expression, and the translator used the

latter to translate. Or some semantic properties have been lost via the generalized translation.

1. *a research that will be embraced by **millions of** bleary-eyed Britons*
   → 一 项 即将 被 **广 大** 睡眼惺忪 的 英国人 所 知道 的 研究
   (《广 大》("vast") is much more general than « millions of ».)
2. *the plan includes a lay-off of 2,000 **employees*** → 这项 计划 括及 裁员 两千 **人**
   (« employees » is translated into《人》("people") which is a more general word .)

Rule 1 The translation of an idiom by a non-fixed expression is annotated as *Generalization*:
Every man has a fool in his sleeve .
→ **人人 都有 糊涂 的 时候** 。
(Literally, the translation means « everyone has a confused moment » which is not an idiom like in English.)
be a far cry from → 和.....完全不同

Rule 2 The removal of a metaphorical image is annotated as *Generalization*:

- *Don 't make me **go through all of this and not make it** .*
  → 别 让 我 的 **辛苦 白费 了** 。
  (« Go through all of this and not make it » is translated into《辛 苦》("hard work")《白 费》("waste"), so there is no more metaphorical image like in the source sentence.)
- *But should clouds gather over the Atlantic, or **tempers rise** in the Middle East, the price could jump again.*
  → 如果 大西洋 风云 再 起 ， 中东 **战火 重 燃** 的 话， 无疑， 油价 会 再度 上涨 。
  (« tempers rise » is translated into 《战火》("flames of war") 《重燃》("burn again"), so there is no more metaphorical image like in the source sentence.)

Rule 3 Use pronoun to translate the thing(s) that it references:
the couple *broke up shortly before this interview* → **他们** 在 接受 这次 采访 的 不久 前 分手 了

Counterexamples:
*He is trying to fight off the **advances** of his rival .*
→ 他 正 试图 击溃 对手 的 **进攻** 。
(This is an instance of *Particularization*.)

Borderline examples:
*All relevant departments will co-operate to publicise this **MPF system** which affects the general public .*
→ 有关 部门 会 齐心合力 ， 推广 这项 影响 广 大 市民 的 **计划** 。
(It's a borderline with *Modulation*, annotated as *Generalization* for now. (MPF : Mandatory Provident Fund System.))

6.1.8 Figurative translation

Rule 1 Introduce an idiom to translate a non-fixed expression, or a metaphorical expression to translate non-metaphor:
pressured the people a little bit about it → 刨根问底 *("inquire into the root of the matter")*
*He gave the required information, **in words as suitable as he could find**.*
→ 他 **字斟句酌 地** 作 了 回答 。

Rule 2 Use personification to translate:
*For Joanne, new opportunities **are opening** .* → 对 乔安娜 而 言 ， 新 的 机遇 **现 已 向 她 招手** 。
(《现已向她招手》means « is waving to her ».)

Counterexamples:

Borderline examples:

6.1.9 Lexical shift

Definition The translation is not literal, but there is no change in meaning. They are minor lexical level changes, which do not involve any translation technique.

change verbal tense or verbal modality (without changing the verb):
*He also **indicated** that the United States will hold negotiations with Cuba*
→ 他 还 **表示** ， 美国 将 与 古巴 举行 谈判

differences between plural and singular form:
*include the following additional **responsibilities*** → 包括 下列 新增 的 **职责**

remove the passive voice:
*The game **was** temporarily **suspended** for 20 minutes.*
→ 比赛 **暂停** 了 二十 分钟 。
(« was suspended » should be translated into 《被暂停》, but in Chinese, it's not always necessary to translate 《被》("be"). In this case, the pair of verb is annotated as *Lexical shift*.)

Counterexamples:
*The book is a **collection** of his reminiscences about the actress .*
→ 这 本 书 **辑录** 了 他 对 那位 女演员 的 回忆 。
The meaning of the word doesn't change but the part of speech is altered, so it is a *Transposition*.)

6.1.10 Translation error

Definition This category concerns obvious translation errors.
1. *a funny video **about** a newly married couple .* → 一个 有趣 的 视频 **对** 新婚 夫妇 。
2. *database **connection** method* → 数据库 **访问** 方式

6.1.11 Uncertain

Please attribute this category when you do not know which label to assign: because it is a difficult borderline example, or because there is something not clear in this annotation guide, and you want more discussion about this pair of segments.

6.2 Annotation of unaligned segments

Definition Certain segments are left unaligned, there exist these three cases:

6.2.1 Unaligned - Explicitation

Definition Translations that could not be aligned to any source segment.
Rule 1 Resumptive anaphora (Charolles, 2002): add a phrase or sentence summarizing the preceding information (which could be present in the previous sentence), to help understanding the present sentence.
and it does n't matter how much information we 're looking at , how big these collections are or how big the images are

→ 不管 所 见到 的 数据 有 多少 、 图像 集 有 多大 以及 图像 本身 有 多 大 ， ***Seadragon*** **都 拥有 这样 的 处理 能力** 。

(The last phrase is an example of resumptive anaphora, which refers to information present in the previous sentence (in the corpus of TED Talks).)

Rule 2 Introduce in the target language clarifications that remain implicit in the source language but emerge from the situation; Personal interpretation or explanation of the translator:

- *the building blocks of the universe* → 宇宙 **形成** 的 **最** 基本 单位
- *most of them are* → 大部分 **都** 是
  (to emphasise)
- *hold negations with Cuba on resuming direct mail service*
  → 与 古巴 就 **重新** 恢复 直接 通邮 举行 谈判
- *London LibDems are working so hard , great to have such support !*
  → **感谢** ("Thanks to") 自民党 团队 的 努力 ， 很 荣幸 得到 那么 多 支持 ！

Rule 3 Add Chinese-specific words:

- Chinese measure words:
  the knife → 这 **把** 刀
  a pharmaceutical → **一 种** 药物 作用
- Verbs added in special text genre, for example in newspaper article:
  AFP , Washington → 法新社 华盛顿 **电**
- Add obligatory word for the date:
  October 13 → 十月 十三 **日**

Rule 4 Add logical connectives:

Get 'boring' done as early as possible today and you will be prepared for anything .
→ 今天 **只要** 尽快 搞定 那些 无聊 的 东西 ， 你 **就** 能 应付 各种 事情 了 。

Rule 5 Necessary addition due to syntactic order change in translation. In Chinese, we often use preposition like 《把》, 《将》,《使》etc. to advance the object to the position before the verb. :

- *I will bring it to China .* → 我 可以 **把** 它 带到 中国 来 。
  (By adding 《把》， « bring it » is translated into 《把》 《它》("it") 《带到》("bring to").)
- *Strong political commitment has significantly reduced the burden of tuberculosis borne by India and China .*
  → 坚定 的 政治 承诺 **使得** 印度 和 中国 的 肺结核 疾病 负担 显著 缩减 。
  (By adding 《使得》， the transitive verb « has reduced » becomes an intransitive verb 《缩减》("decrease") and is moved to the end.)

Rule 6 When the relative pronoun is absent in the source side while《的》exists in the target side, annotate《的》as *Explicitation*.

The US federal government has grossly violated the promise it made in the trade negotiation ***(which was)*** *held last November in Doha .*
→ 美国 联邦 政府 违背 了 去年 十一月 在 卡塔尔 首都 多哈 举行 **的** 贸易 谈判 上 所 作 的 承诺 。

6.2.2 Unaligned - Reduction

Definition Remove certain words that could have been translated. See rules below for each specific case.

1. *we started this* ***over*** *15 years ago* → 我们 于 *15* 年 前 开始 这 项 工作
2. *one by one the other elders* ***now*** *timidly rise*
   → 其他 的 长老 一个 接 一个 怯生生 地 站起来
3. ***let*** *me show you* → 我 演示 给 你们 看

4. *Irish soccer team captain Keane announces he **will** retire after the World Cup .*
   → 爱尔兰 足球队 队长 基恩 宣布 世界杯 之后 退休 。

Rule 1 Removal of preposition:
*A spokesman **from** the Ministry of National Defense* → 国防部 发言人
(In the translation, there is no preposition between《发言人》("spokesman") and《国防部》("the Ministry of National Defense").)

Rule 2 Removal of determiner:
*Bush proposed in **his** speech to step up the humanitarian assistance to Cuba*
→ 布什 在 演说 中 提议， 增加 对 古巴 的 人道 援助
(There is no determiner before《演说》("speech") .)

Rule 3 Removal of noun:
*Declaration on the Protection of Right of **Persons** belonging to National , Ethnic , Religious and Linguistics Minorities .*
→ 保护 民族， 种族， 语言， 宗教 上 属于 少数 人 的 权利 宣言 。
(There noun « Persons » is not repeated in the translation.)

Rule 4 Removal of pronoun:
*When she gets there , **she** finds out that there are no doctors .*
→ 但是 当 她 到 那里 时 却 发现 并 没有 医生 。
(The subject « she » is not repeated in the translation.)

Rule 5 Removal of copula:
*Peter **is** six years old .* → 彼得 六岁 。
(The copula « is » is not translated.)
*Its glimmer **was** yet dim in the plain below .*
→ 暗淡 的 微光， 在 平原 依稀 可 见 。

Rule 6 Removal of anticipatory « it » : In English, « it » is used to introduce or 'anticipate' the subject or object of a sentence, especially when the subject or object of the sentence is a clause. It's not translated in Chinese.
***It** was a pleasant surprise to learn of her marriage .*
→ 得知 她 结婚 是 件 令 人 惊喜 的 事 。

6.2.3 Unaligned - No Type

Rule 1 Function words necessary in English but not in Chinese:
*The tragedy of the world is **that** those who are imaginative have but slight experience .*
→ 世界 的 悲剧 就 在于 有 想象力 的 人 缺乏 经验 。
the two gentlemen leaned forward and looked at each other
→ 这 两 位 绅士 身体 前倾， 互相 你 看看 我 、 我 看看 你
(« and » is often replaced by a comma)

Rule 2 Segments not translated but which do not impact the meaning:
*The present state , application and development of coal mine hydraulic drill rig are described **in this paper** .*
→ 介绍 了 煤矿 用 液压 钻 车 现状， 使用 情况 及 发展 。
***you know** , we do n't know necessarily what it 'll look like*
→ 我们 并不 一定 知道 它 看起来 是 怎样

Rule 3 Target segments added without reason, which do not correspond to any source segment:
then , after about 30 seconds , it reshuffles , and you have a new set of letters and new possibilities to try .

→ 大约 30 秒 后 Siftables 会 自动 根据 正确 的 单词 拼写 顺序 改变 原来 显示 的 字母 ， 组 成 正确 的 单词 ， 你 还 可以 尝试 用 其他 的 字母 拼 出 新 的 单词 。

(The content is added by the translator but not expressed by the speaker.)

Rule 4 《鼓掌》("applause"),《笑声》("laughter"): specific in the subtitles of TED Talks.

7 Tutorial for using Yawat

In order to conduct all the operations available in Yawat, please use Firefox (version 67 or before)[6] as navigator instead of Chrome or Safari.

Url : `https://pakin.limsi.fr/cgi-bin/cgi/yawat.cgi`

Every annotator has an account created by the administrator. After logging in, according to the current task, please find the corresponding sub-corpus in different directories (e.g. pass1, pass2, pass3). Then click the corpus name to begin the annotation. (We also prepare a small test corpus for you to be familiar with the functions in Yawat.)

Here are some instructions about using the tool:

1. You can change the layout to show pairs of sentences in two lines or in two columns, by clicking on the button in the top right corner.
2. The black color is by default which means the category *Literal*. (But this does not mean that the boundary and alignment are all correct. Annotators should verify the alignment of all words.)
   At source side: blue means *unaligned and no type*.
   At target side: red means *unaligned and no type*.
3. Automatic word alignments have been imported, annotators should correct them when necessary. For example, in figure 11, there was just *open* and *ouvre* aligned automatically, and we want to add *up* to *open*.
   First, left click on either *open* or *ouvre*, the pair will be in a state of "being chosen", then left click *up*, we can see a new discontinuous boundary is created. Finally right click on any of these three words, the new alignment will be confirmed.

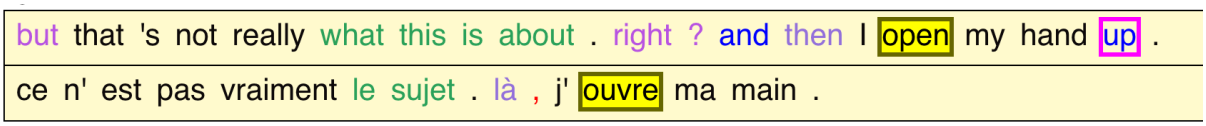

Figure 11: How to correct alignment in Yawat

4. If we want to remove a word from a segment, right click on it, and click *remove ... from this group*. If it is a lexical pair like the example in figure 12, when removing one word from one side, the link between them is broken in consequence. The corresponding word will change into the *unaligned and no type* state a moment later, you can hover your mouse on it to confirm it.

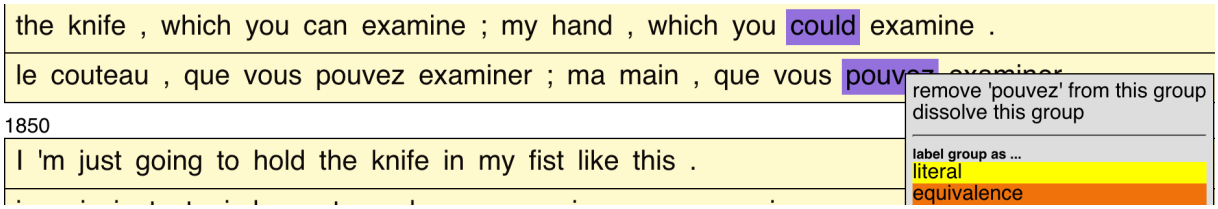

Figure 12: How to remove a word from a segment

5. To facilitate the alignment of a pair of long segments, for example in figure 13, first left click *nothing* and *il*, then right click either of them to create this pair.

[6] If you need to use higher version of Firefox for other applications, you can continue to use Yawat by downloading the Firefox ESR version here: `https://www.mozilla.org/en-US/firefox/organizations/`.

Next, left click this pair (by clicking either word inside) and add the rest of the segment by left clicking the other words, and finally right click on any word of them to confirm this long segment: *nothing goes up or down my sleeve → il n' y a rien dans mes manches*.

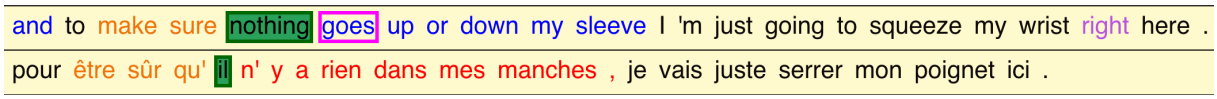

Figure 13: How to align long segments in Yawat

6. In order to annotate unaligned segments, *i.e. Explicitation* and *Reduction*, left click words to fix the boundary then right click the segment to choose the category from the menu (figure 14).

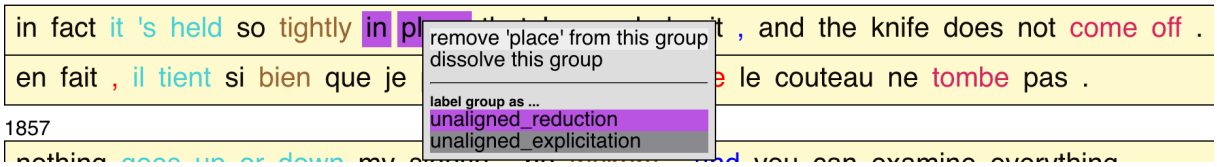

Figure 14: How to annotate unaligned segments

7. Once the boundary and the alignment are fixed, we can attribute a category. Right click any word in the pair, and choose a category from the menu (figure 15).

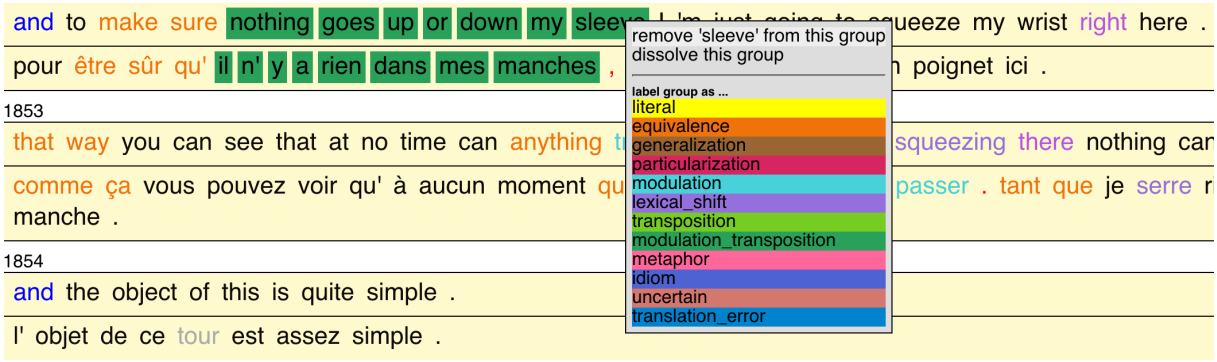

Figure 15: Choose category from the menu

8. Hover the mouse on words to check the boundary and the alignment of segments.
9. Make sure to save your annotations before logging out (click the **[save]** button in the top right corner, it appears only after changes).

Please contact us if you have any trouble in using this interface.

References

Michel Charolles. 2002. *La référence et les expressions référentielles en français*. Ophrys.

Hélène Chuquet and Michel Paillard. 1989. *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

Yuqing Fang. 2008. 实用汉语语法(第二次修订本) *(Practical Chinese Grammar)*. 北京语言大学出版社.

Ulrich Germann. 2008. Yawat: Yet Another Word Alignment Tool. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Demo Papers*, pages 20–23. The Association for Computer Linguistics.

Klaudia Gibová. 2012. *Translation Procedures in the Non-literary and Literary Text Compared*. Books on Demand.

Zhongguo Li and Maosong Sun. 2009. Punctuation As Implicit Annotations for Chinese Word Segmentation. *Computational Linguistics*, 35(4):505–512, December.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2295–2301. AAAI Press.

Lucía Molina and Amparo Hurtado Albir. 2002. Translation Techniques Revisited: A Dynamic and Functionalist Approach. *Meta*, 47(4):498–512.

Peter Newmark. 1981. *Approaches to Translation (Language Teaching Methodology Senes)*. Oxford: Pergamon Press.

Peter Newmark. 1988. *A textbook of translation*, volume 66. Prentice Hall New York.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: méthode de traduction*. Bibliothèque de stylistique comparée. Didier.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Reconnaissance des procédés de traduction sous-phrastiques : des ressources aux validations

Mots clés : création de corpus, reconnaissance automatique, application en traitement automatique des langues

Résumé : Les procédés de traduction constituent un sujet important pour les traductologues et les linguistes. Face à un certain mot ou segment difficile à traduire, les traducteurs humains doivent appliquer les solutions particulières au lieu de la traduction littérale, telles que l'équivalence idiomatique, la généralisation, la particularisation, la modulation syntaxique ou sémantique, etc.
En revanche, ce sujet a reçu peu d'attention dans le domaine du traitement automatique des langues (TAL). Notre problématique de recherche se décline en deux questions : est-il possible de reconnaître automatiquement les procédés de traduction ? Certaines tâches en TAL peuvent-elles bénéficier de la reconnaissance des procédés de traduction ?
Notre hypothèse de travail est qu'il est possible de reconnaître automatiquement les différents procédés de traduction (par exemple littéral versus non littéral). Pour vérifier notre hypothèse, nous avons annoté un corpus parallèle anglais-français en procédés de traduction, tout en établissant un guide d'annotation. Notre typologie de procédés est proposée en nous appuyant sur des typologies précédentes, et est adaptée à notre corpus. L'accord inter-annotateur (0,67) est significatif mais dépasse peu le seuil d'un accord fort (0,61), ce qui reflète la difficulté de la tâche d'annotation. En nous fondant sur des exemples annotés, nous avons ensuite travaillé sur la classification automatique des procédés de traduction. Même si le jeu de données est limité, les résultats expérimentaux valident notre hypothèse de travail concernant la possibilité de reconnaître les différents procédés de traduction. Nous avons aussi montré que l'ajout des traits sensibles au contexte est pertinent pour améliorer la classification automatique.
En vue de tester la généricité de notre typologie de procédés de traduction et du guide d'annotation, nos études sur l'annotation manuelle ont été étendues au couple de langues anglais-chinois. Ce couple de langues partagent beaucoup moins de points communs par rapport au couple anglais-français au niveau linguistique et culturel. Le guide d'annotation a été adapté et enrichi. La typologie de procédés de traduction reste identique à celle utilisée pour le couple anglais-français, ce qui justifie d'étudier le transfert des expériences menées pour le couple anglais-français au couple anglais-chinois.
Dans le but de valider l'intérêt de ces études, nous avons conçu un outil d'aide à la compréhension écrite pour les apprenants de français langue étrangère. Une expérience sur la compréhension écrite avec des étudiants chinois confirme notre hypothèse de travail et permet de modéliser l'outil. D'autres perspectives de recherche incluent l'aide à la construction de ressource de paraphrases, l'évaluation de l'alignement automatique de mots et l'évaluation de la qualité de la traduction automatique.

Title : Recognition of sub-sentential translation techniques: from resources to validation

Keywords : corpus creation, automatic recognition, application in natural language processing

Abstract : Translation techniques constitute an important subject in translation studies and in linguistics. When confronted with a certain word or segment that is difficult to translate, human translators must apply particular solutions instead of literal translation, such as idiomatic equivalence, generalization, particularization, syntactic or semantic modulation, etc.
However, this subject has received little attention in the field of Natural Language Processing (NLP). Our research problem is twofold: is it possible to automatically recognize translation techniques? Can some NLP tasks benefit from the recognition of translation techniques?
Our working hypothesis is that it is possible to automatically recognize the different translation techniques (e.g. literal versus non-literal). To verify our hypothesis, we annotated a parallel English-French corpus with translation techniques, while establishing an annotation guide. Our typology of techniques is proposed based on previous typologies, and is adapted to our corpus. The inter-annotator agreement (0.67) is significant but slightly exceeds the threshold of a strong agreement (0.61), reflecting the difficulty of the annotation task. Based on annotated examples, we then worked on the automatic classification of translation techniques. Even if the dataset is limited, the experimental results validate our working hypothesis regarding the possibility of recognizing the different translation techniques. We have also shown that adding context-sensitive features is relevant to improve the automatic classification.
In order to test the genericity of our typology of translation techniques and the annotation guide, our studies of manual annotation have been extended to the English-Chinese language pair. This pair shares far fewer linguistic and cultural similarities than the English-French pair. The annotation guide has been adapted and enriched. The typology of translation techniques remains the same as that used for the English-French pair, which justifies studying the transfer of the experiments conducted for the English-French pair to the English-Chinese pair.
With the aim to validate the benefits of these studies, we have designed a tool to help learners of French as a foreign language in reading comprehension. An experiment on reading comprehension with Chinese students confirms our working hypothesis and allows us to model the tool. Other research perspectives include helping to build paraphrase resources, evaluating automatic word alignment and evaluating the quality of machine translation.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France