Construction d'un corpus multilingue annoté en relations de traduction

Yuming Zhai

Construction d'un corpus multilingue annoté en relations de traduction

Yuming Zhai
LIMSI/CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay
yuming.zhai@limsi.fr

RÉSUMÉ

Les relations de traduction, qui distinguent la traduction littérale d'autres procédés, constituent un sujet d'étude important pour les traducteurs humains (Chuquet & Paillard, 1989). Or les traitements automatiques fondés sur des relations entre langues, tels que la traduction automatique ou la méthode de génération de paraphrases par équivalence de traduction, ne les ont pas exploitées explicitement jusqu'à présent. Dans ce travail, nous présentons une catégorisation des relations de traduction et nous les annotons dans un corpus parallèle multilingue (anglais, français, chinois) de présentations orales, les *TED Talks*. Notre objectif à plus long terme sera d'en faire la détection de manière automatique afin de pouvoir les intégrer comme caractéristiques importantes pour la recherche de segments monolingues en relation d'équivalence (paraphrases) ou d'implication. Le corpus annoté résultant de notre travail sera mis à disposition de la communauté.

ABSTRACT

Construction of a multilingual corpus annotated with translation relations

Translation relations, which distinguish literal translation from other translation techniques, constitute an important subject of study for human translators (Chuquet & Paillard, 1989). However, automatic processing techniques based on interlingual relations, such as machine translation or paraphrase generation exploiting translation equivalence, have not exploited these relations explicitly until now. In this work, we present a categorisation of translation relations and annotate them in a parallel multilingual (English, French, Chinese) corpus of oral presentations, the TED Talks. Our long term objective will be to automatically detect these relations in order to integrate them as important characteristics for the search of monolingual segments in relation of equivalence (paraphrases) or of entailment. The annotated corpus resulting from our work will be made available to the community.

MOTS-CLÉS : relations de traduction, annotation de corpus.

KEYWORDS: translation relations, corpus annotation.

1 Introduction

Dans le domaine des recherches en acquisition automatique de paraphrases, les premiers travaux guidés par les données ont exploité des corpus monolingues. Les méthodes proposées ont notamment reposé sur l'analyse des contextes environnants (Barzilay & McKeown, 2001), des calculs de similarité fondé sur des arbres de dépendances (Lin & Pantel, 2001; Ibrahim *et al.*, 2003), la fusion d'arbres de constituants (Pang *et al.*, 2003), ou le regroupement de documents par des critères de dates et de thèmes (Dolan *et al.*, 2004).

Une autre famille d'approches importante exploite des corpus multilingues 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 (Bannard & Callison-Burch, 2005), et sur l'hypothèse selon laquelle si deux segments dans la même langue partagent une ou plusieurs traductions communes (considérées comme des "pivots") dans une ou plusieurs langues étrangères, alors ils sont potentiellement des paraphrases (voir une illustration sur la figure 1). Cette méthode exploite les informations des tables de traduction statique générées par les systèmes de traduction automatique basés sur les segments (PBSMT). Le travail ultérieur de Callison-Burch (2008) a affiné cette approche en imposant que les segments partagent la même structure syntaxique *CCG (Combinatory Categorial Grammar)*, ce qui a permis d'améliorer la substituabilité grammaticale pour les paires produites. En se basant sur cette même approche, mais dans le but d'obtenir une meilleure généralisation, Zhao *et al.* (2008) ont utilisé des arbres de dépendances pour apprendre des patrons de paraphrases qui incluent des variables de partie du discours.

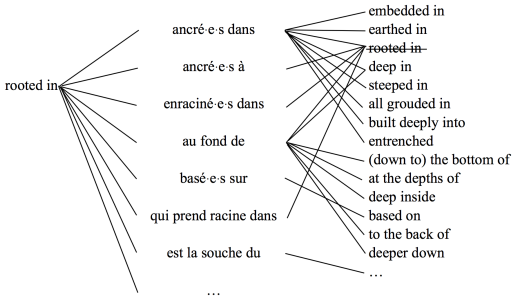

FIGURE 1: Exemple de génération de paraphrases par pivot en français pour le segment anglais «rooted in».

La méthode dite «par pivot» a été mise en œuvre pour la construction de la ressource PPDB (Paraphrase Database) [1], aujourd'hui la plus grande ressource de paraphrases disponible avec plus de 100 millions de paires de segments anglais (Ganitkevitch *et al.*, 2013). 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 (soit plus de 2 milliards de mots anglais) couvrant 22 langues pivot. La version multilingue de PPDB (Ganitkevitch & Callison-Burch, 2014) contient des paires de segments pour 23 langues, dont le français, obtenues en utilisant l'anglais comme langue pivot. Chaque paire de segments est associée à une trentaine de caractéristiques, notamment la probabilité de paraphrase (Bannard & Callison-Burch, 2005) et des scores de similarité distributionnelle monolingue (Ganitkevitch *et al.*, 2012). De plus, chaque paire partage une même catégorie grammaticale selon les contraintes imposées dans (Callison-Burch, 2008). Le score de classement dans la version initiale de PPDB est fondé sur un calcul combinant un sous-ensemble de caractéristiques avec des pondérations *ad hoc* fondées sur les intuitions des auteurs.

Pour la seconde version de cette ressource, PPDB 2.0 (Pavlick *et al.*, 2015b), un modèle de régression a été utilisé afin d'adapter le score de paraphrase à des jugements humains de la qualité des paraphrases, permettant une meilleure corrélation qu'avec le classement heuristique de PPDB 1.0. 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 par l'équivalence de traduction. Ce travail décrit une catégorisation automatique de diverses relations (*Équivalence, Implication (dans les deux sens), Exclusion, Autrement lié et Indépendant*) ayant ex-

1. http://paraphrase.org

ploité de nombreuses caractéristiques incluant : des informations de niveau lexical, des informations issues de WordNet (Miller, 1995), des patrons lexico-syntaxiques, des valeurs de similarité distributionnelle entre vecteurs de contexte de dépendances, des probabilités de paraphrase, le nombre total de traductions partagées pour chaque paire de phrases, etc. La meilleure combinaison, qui utilise à la fois des caractéristiques monolingues et des caractéristiques bilingues, permet d'atteindre une précision globale de 79%. Une estimation réalisée sur la plus grande taille de PPDB 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 ou pour obtenir de manière contrôlée d'autres types de variantes.

La traduction automatique, qui repose elle aussi sur des correspondances bilingues, a connu récemment des améliorations significatives avec l'avènement des techniques neuronales (NMT), permettant pour des langues proches d'atteindre des performances plus proches de traductions humaines que par les techniques statistiques (SMT) précédentes (Wu *et al.*, 2016). Le travail de Lapata *et al.* (2017) a consisté à réimplémenter la méthode de génération de paraphrases par pivot en utilisant des systèmes NMT. Les résultats expérimentaux obtenus dans le cadre de plusieurs tâches (prédiction de la similarité, identification des paraphrases et génération de paraphrases) montrent que leur approche améliore les approches précédentes.

Ces travaux en génération de paraphrases et en traduction automatique n'ont cependant, à notre connaissance, jamais pris en compte les relations de traduction entre paires de segments, alors que cela correspond à un sujet très 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 de mots, essentiel pour les méthodes statistiques, et elles peuvent faire dévier le sens originel du texte source. Le manque de modélisation de ces relations conduit à une perte de contrôle sur la sémantique produite, ce qui est attesté par les diverses relations sémantiques dans la ressource PPDB (Pavlick *et al.*, 2015a). De plus, les différences culturelles donneront éventuellement des distributions de relations de traduction différentes en fonction des langues. Pour la génération de variantes par équivalence de traduction, des langues pivots différentes peuvent éventuellement produire des résultats très différents, ce qui n'a pas été considéré jusque-là.

Dans cet article, nous catégorisons ces relations de traduction en modélisant le choix des traducteurs humains qui les ont produites, et nous les annotons dans un corpus multilingue de discours préparés, les *TED Talks* [2]. Nous décrivons les définitions des relations, le processus d'annotation ainsi que les principaux problèmes rencontrés. Notre objectif suivant portera sur la détection automatique de ces relations afin de les intégrer comme caractéristiques dans la recherche de paraphrases ou de paires en relation d'implication. Nous faisons l'hypothèse que cela permettra davantage de contrôle sémantique et de variété en génération. Nous présentons les travaux précédents en lien avec notre travail dans la Section 2, et décrivons notre corpus dans la Section 3. Les relations de traduction sont décrites dans la Section 4, suivies par le processus et les statistiques des annotations obtenues dans la Section 5. Nous donnons finalement nos conclusions et perspectives dans la Section 6.

2. https://www.ted.com/

2 Travaux précédents

Le travail de Deng & Xue (2017) a étudié les divergences présentes dans la traduction automatique anglais-chinois à l'aide d'un schéma d'alignement hiérarchique entre des arbres d'analyse pour ces deux langues. Sept types de divergences ont été identifiées, certaines posant des difficultés importantes pour l'alignement automatique de mots, notamment les différences lexicales résultant de traductions non littérales, et les différences de structures entre langues (avec ou sans changement de types de syntagmes). En vue de fournir un jeu de données particulier sur les expressions multi-mots pour la traduction automatique, Monti *et al.* (2015) ont annoté spécifiquement ces expressions dans le corpus *TED Talks* anglais-italien associées à leur traduction générée par un système automatique. Les phénomènes discutés dans les deux travaux que nous venons de mentionner sont inclus dans les relations de traduction que nous présentons dans ce travail.

Reprenant l'approche de génération de paraphrases par pivot, Kok & 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. Cette approche parcourt des chemins de longueur supérieure à 2 en utilisant l'information entre les nœuds représentant des segments dans une autre langue et en permettant d'intégrer des connaissances monolingues sous forme de nœuds spéciaux. Les résultats expérimentaux ont permis d'obtenir davantage de paraphrases correctes que l'approche de Callison-Burch (2008) [3] ainsi qu'une meilleure précision pour les paraphrases classées aux premiers rangs. Nous comptons par la suite poursuivre ce travail en nous intéressant spécifiquement aux relations de traduction non-littérales afin d'étudier si un changement de sens a lieu, dans l'espoir de mieux guider le parcours dans des corpus multilingues pour obtenir des paraphrases par équivalence de traduction.

Une limite importante de l'approche par pivot est qu'elle ne distingue pas les différents sens possibles d'un segment lors de la génération de ses paraphrases potentielles. Le travail de Apidianaki *et al.* (2014) a analysé la sémantique des paraphrases lexicales obtenues avec l'approche de Callison-Burch (2008) et a mis en évidence la nécessité d'une étape de désambiguïsation. Le travail ultérieur de Cocos & Callison-Burch (2016) a introduit une méthode pour effectuer le regroupement des paraphrases par sens, laquelle a été appliquée à la ressource PPDB. Le travail que nous présentons dans cet article porte davantage sur les relations de traduction (bilingue) qui sont à l'origine des diversités sémantiques dans une ressource telle que PPDB, ainsi que sur leur exploitation pour la suite de notre travail. La prise en compte de la polysémie nous concernera dans un second temps.

3 Corpus

Afin de faire l'étude des relations de traduction pour plusieurs paires de langues, nous avons travaillé sur un corpus parallèle multilingue. Le corpus annoté est issu de l'inventaire Web *WIT*^{3} (Cettolo *et al.*, 2012) qui donne accès à une collection de conférences transcrites et traduites incluant le corpus *TED Talks* [4]. Ce corpus a été mis à disposition pour les campagnes d'évaluation IWSLT 2013 et 2014 [5]. La

3. La méthode d'évaluation est toutefois une version simplifiée de celle utilisée dans (Callison-Burch, 2008).

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

5. Nous avons utilisé le corpus d'entraînement de 2014 (160 656 lignes), de développement (880 lignes) et de test (1 556 lignes) de 2010.

langue d'origine, c'est-à-dire dans laquelle se sont originellement exprimés les orateurs, est l'anglais. Nous avons calculé l'intersection des corpus parallèles avec les traductions en français [6], chinois, arabe, espagnol et russe. La traduction des sous-titres de *TED Talks* est contrôlée par des bénévoles et des coordinateurs par langue [7], permettant une traduction d'un bon niveau de qualité en général. Le corpus annoté contient 2 436 lignes de phrases parallèles pour chaque paire de langues. À ce stade, nous décrivons le début de nos annotations pour les paires anglais-français et anglais-chinois. La table 3 décrit les statistiques principales des corpus correspondants.

Pour l'anglais et le français, la tokenisation est réalisée par l'outil Stanford Tokenizer [8]. Les lettres capitales au début de chaque ligne ont été transformées en minuscules, si et seulement si les mots en question ont par ailleurs leur première lettre en minuscule dans le corpus. Dans le cas contraire, les lettres capitales sont gardées telles quelles pour les mots qui apparaissent toujours avec des initiales en majuscule. Nous avons utilisé l'outil THULAC (Li & Sun, 2009) pour la segmentation du corpus chinois. L'alignement automatique de mots des corpus bilingues a été réalisé par l'outil FastAlign (Dyer *et al.*, 2013) avec ses paramètres par défaut et en l'entraînant sur l'intégralité de chaque corpus parallèle (soit 163 092 lignes et 3 303 660 tokens anglais).

4 Relations de traduction

Nous avons établi une hiérarchie décrivant les relations de traduction en nous fondant sur les théories explicitées dans l'ouvrage de Chuquet & Paillard (1989) et les phénomènes rencontrés pendant notre étude de corpus initiale (voir figure 2). Les nœuds colorés représentent nos catégories, les

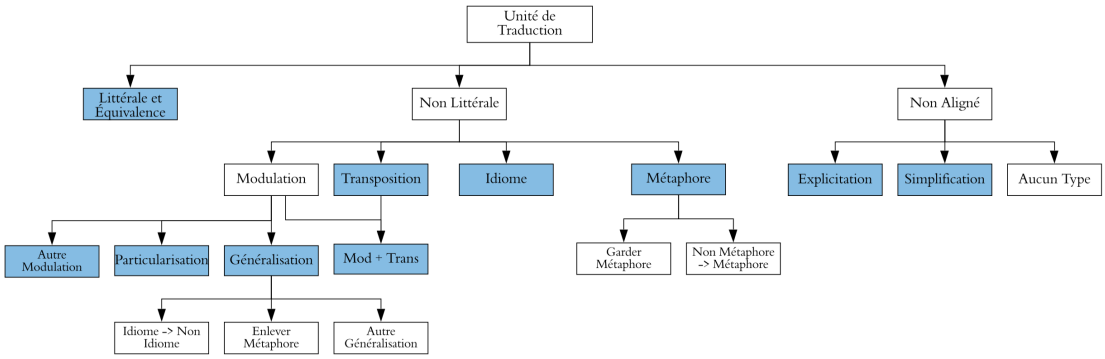

FIGURE 2: Hiérarchie de relations de traduction.

autres nœuds décrivant la hiérarchie (*i.e. Non Littérale, Non Aligné, Modulation, Aucun Type*) ou des phénomènes plus précis mais pour lesquels une étiquette dédiée n'a pas été retenue (*e.g. Enlever Métaphore, Autre Généralisation*). Nous présentons ci-dessous leur définition ainsi que des exemples caractéristiques :

1. Traduction littérale : traduction mot à mot, ce qui concerne également les situations où des idiomes peuvent être traduits de façon littérale :

   facts are stubborn -> les faits sont têtus, *What time is it ? -> Quelle heure est-il ?*

6. Les frontières de phrases ont été corrigées dans le corpus de test français pour calculer l'intersection.
7. `https://www.ted.com/participate/translate/get-started`
8. `http://nlp.stanford.edu/software/tokenizer.shtml`

2. Équivalence :

   i) Traduction non-littérale de certains proverbes, idiomes ou expressions figées

   Birds of a feather flock together. -> Qui se ressemble s'assemble.

   on the brink of -> à deux doigts de

   ii) Équivalence sémantique au niveau supra-lexical, ou traduction des termes

   magic trick -> tour de magie, *hatpin -> épingle à chapeau*

3. Généralisation : cette catégorie inclut trois sous-types, mais nous les annotons par une seule étiquette :

   i) La traduction est plus générale ou neutre ; dans d'autres cas, ce procédé rend le sens plus accessible dans la langue cible :

   look carefully at -> regardez, *as we **sit here** in ... -> alors que nous **sommes** à ...*

   ii) Traduction d'un idiome par une expression non figée :

   trial and error -> procéder par tâtonnements

   iii) Suppression d'une image métaphorique :

   *ancient Tairona civilization which once **carpeted** the Caribbean coastal plain -> anciennes civilisations tyranniques qui **occupaient** jadis la plaine côtière des Caraïbes*

4. Particularisation : la traduction est plus précise ou présente un sens plus concret :

   *the director **said** -> le directeur **déclara***, *language loss -> l' extinction du langage*

5. Modulation : ce procédé consiste à changer le point de vue, soit pour contourner une difficulté de traduction, soit pour révéler une manière différente de voir les choses pour les locuteurs de la langue cible :

   this is a completely unsustainable pattern -> il est absolument impossible de continuer sur cette tendance, *I had an assignment -> on m'avait confié une mission*

6. Transposition : traduction 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 malades **ayant dépassé** l'âge de 40 ans*

7. Modulation plus Transposition : ce type peut contenir n'importe quel sous-type de modulation combiné avec la transposition :

   *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***

8. Idiome : cas de la traduction d'expressions non figées par un idiome (très fréquent dans la traduction de l'anglais en chinois) :

   at any given moment -> à un instant "t"

   died getting old -> 行将就木(getting closer and closer to the coffin)

9. Métaphore : cette catégorie inclut deux sous-types ramenés à une seule étiquette :

   i) 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*

   ii) Introduction d'une métaphore pour traduire des segments non métaphoriques :

   *if you **faint** easily -> si vous **tombez dans les pommes** facilement*

10. Non aligné - Explicitation : introduction dans la langue cible de clarifications pour des éléments implicites dans la langue source mais qui émergent du contexte ou de la situation :
    *feel their past in the wind -> ressentent leur passé **souffler** dans le vent*
11. Non aligné - Simplification : non traduction délibérée de certains mots pleins :
    *and you'll **suddenly** discover what it would be like -> et vous découvrirez ce que ce serait*
12. Non aligné et aucun type attribué : mots outils nécessaires dans une langue mais pas dans l'autre ; segments non traduits mais qui n'influencent pas le sens ; segments donnant des informations répétées en contexte :
    *minus 271 degrees, colder than -> moins 271 degrés, **ce qui est** plus froid*
    *the last example I have time to -> le dernier exemple **que** j'ai le temps de*

5 Annotation des relations

Outil et configuration Nous avons utilisé l'application Web Yawat [9] (Germann, 2008), qui nous permet d'aligner des mots ou des segments (continus ou discontinus), puis d'attribuer des étiquettes configurables adaptées pour notre tâche à des unités monolingues ou bilingues (voir figure 3) .

À des fins d'illustration, considérons l'exemple trilingue suivant :

well, we use that great euphemism, "trial and error", which is exposed to be meaningless.
eh bien, nous employons cet euphémisme, procédé par tâtonnements, qui est dénué de sens.
我们(nous) 普通人(les gens ordinaires) 会(particule du temps futur) 做(faire) 各种各样(divers) 的(particule pour attribut) 实验(expérience) 不断(sans arrêt) 地(particule pour adverbe) 犯错误(commettre une faute) 结果(par conséquent) 却(cependant) 一无所获 (ne rien gagner)

Les segments *well* et *we use that great euphemism* sont traduits littéralement en français mais sont omis en chinois. L'idiome *trial and error* est traduit par une généralisation dans les deux langues. Le segment *which is exposed to be* est traduit par une généralisation en français (*est*) et par une modulation en chinois (*结果(par conséquent) 却(cependant)*). L'adjectif *meaningless* est traduit par une transposition en français (*dénué de sens*) et par un idiome de quatre caractères en chinois (*一无所获(ne rien gagner)*).

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

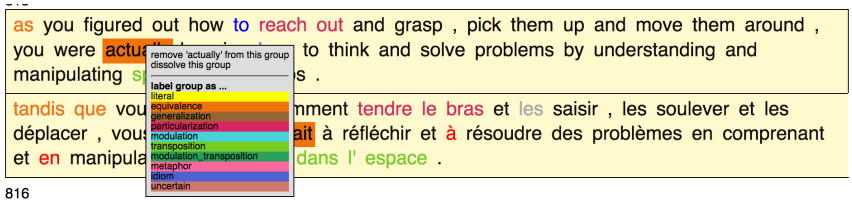

FIGURE 3: Interface de l'outil Yawat permettant d'effectuer l'annotation.

segmenté automatiquement contient des erreurs qui peuvent produire des alignements incorrects puis

9. *Yet Another Word Alignment Tool* ; cet outil est disponible pour la recherche sous la licence GNU Affero General Public License v3.0.

des attributions d'étiquette incorrectes. Certains mots chinois ont donc été resegmentés manuellement préalablement à 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 éventuelles fautes d'orthographe dans le corpus originel n'ont, par contre, pas été corrigées, car d'une part il n'en existe pas beaucoup et d'autre part cela n'empêche pas l'attribution de catégories en général. Toutefois, nous avons introduit une catégorie *Incertain* pour des paires pour lesquelles les annotateurs ne savent pas attribuer de catégorie ou pour des paires qui contiennent des erreurs de traduction manifestes.

Deux annotateurs [10] ont participé à ce travail préliminaire pour la paire anglais-français, et une seule annotatrice pour la paire anglais-chinois. La formation des futurs annotateurs s'appuiera sur un guide d'annotation qui définit l'ensemble des types illustrés par des exemples. La hiérarchie des catégories permet de donner une vue d'ensemble des relations entre les catégories, et des exemples discriminants permettent de mieux guider les annotateurs dans les étapes de décision. En vue de bien comprendre le contexte, les annotateurs peuvent regarder des vidéos de conférences [11] correspondantes avant d'annoter.

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). 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. Le nombre de tokens anglais annotés dans des segments de mêmes frontières est de 1 906 pour la catégorie *Littérale* 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 [12], la valeur de Kappa diminue à 0,617, ce qui correspond à un accord fort-modéré. Cependant, dans cette configuration, la couverture des tokens anglais augmente à 85,56%. Les tokens restant appartiennent eux à des segments aux frontières incompatibles.

Nous présentons la table de confusion pour ces segments (sans considérer les appariements flexibles) dans la figure 4. Sans surprise, la majorité des situations d'accord correspondent aux traductions littérales. Il existe tout de même certains désaccords pour ce type avec *Équivalence (e.g. in this way -> de cette façon), Modulation (e.g. this entire time -> tout ce temps), Particularisation (e.g. snuff -> tabac)* et *Transposition (e.g. their prayers **alone** -> **seulement** leurs prières)*. Néanmoins nos catégories *Équivalence* et *Littérale* sont très proches, et *Particularisation* est un sous-type de *Modulation*, confusions que nous aurions donc pu considérer comme admissibles dans une mesure plus flexible. *Modulation* présente le plus grand nombre de confusions avec *Littérale* et *Transposition (e.g. from the forest floor -> tombées par terre)*, ce qui indique qu'il est nécessaire de mieux expliciter leurs différences quand nous formons les annotateurs. *Mod+Trans* est un type combiné pour lequel certains annotateurs ne perçoivent parfois que l'un des deux types (*e.g. a great distance -> de loin*). Il existe très peu de confusions pour *Généralisation (e.g. because they're denatured -> étant dénaturés)* mais c'est moins le cas pour *Particularisation*. *Non aligné - Explicitation* et *Non aligné -*

10. Un annotateur français et une annotatrice chinoise.

11. Les corpus à annoter sont des transcriptions des *TED Talks* et leur traduction.

12. Par exemple, *I was asked by* et *I was asked by my professor at Havard* ont tous les deux été annotés par *Modulation*, et *my professor at Havard* a été annoté par *Littérale* par le premier annotateur. Nous considérons ici que les deux segmentations sont compatibles et qu'il y a accord entre les deux annotateurs sur un segment (le plus grand) du fait du type commun *Modulation*.

Simplification présentent très peu de confusions avec les autres types, mais sont parfois en compétition avec *Aucun Type*. *Métaphore* est à l'origine de quelques désaccords (*e.g. 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.

	Équivalence	Littérale	Modulation	Transposit.	Mod+Trans	Généralisat.	Particulari.	Explicitation	Simplificat.	Idiome	Métaphore	Incertain
Équivalence	21	4	0	5	0	0	1	0	1	0	0	0
Littérale	27	1857	26	6	0	0	10	0	0	0	0	7
Modulation	4	8	37	7	1	1	3	0	2	0	0	0
Transposition	6	7	10	30	0	1	0	0	0	0	0	0
Mod+Trans	0	1	6	2	2	2	0	0	0	0	0	0
Généralisation	0	1	0	0	0	17	0	0	0	0	0	1
Particularisation	4	13	6	2	0	1	29	0	0	0	0	2
Explicitation	0	0	0	0	0	0	0	10	0	0	0	0
Simplification	0	0	0	0	0	0	0	0	40	0	0	0
Idiome	0	0	0	0	0	0	0	0	0	0	0	0
Métaphore	1	0	0	0	0	1	1	0	0	0	1	0
Incertain	1	8	2	2	0	4	0	0	1	0	0	4

FIGURE 4: Table de confusion pour le corpus de contrôle (nombre d'instances).

Processus à plusieurs passes Le calcul d'une valeur d'accord inter-annotateurs permet certaines interprétations standard pour 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 types, nous avons adopté un processus d'annotation à plusieurs passes en vue d'obtenir une meilleure qualité d'annotation. Pour chaque sous-corpus [13], un premier annotateur réalise une première passe pour 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 3ème étant effectuée par le premier annotateur du corpus. Nous constatons que le nombre de modifications dans la 3ème passe décroît au fur et à mesure des documents annotés, rendant compte d'une adaptation progressive et rapide des annotateurs à la tâche. Ce mode d'annotation, plus coûteux, a toutefois été rendu nécessaire par la qualité visée ainsi que par les difficultés inhérentes à la segmentation observées sur le corpus de contrôle.

Certaines situations nécessitent l'établissement de conventions d'annotation pour garantir la cohérence des annotations (voir table 1). Par exemple, pour des articles français n'ayant pas de correspondance en anglais, nous attachons ceux-ci avec le nom modifié pour préciser leur appartenance, *e.g. play with **blocks** -> jouer avec **des cubes***. Tout changement en nombre, temps de verbe (sauf si rendu nécessaire par le contexte), pronom personnel, préposition (traduction non littérale) et ponctuation est considéré comme *Modulation*.

Comme nous l'avons vu, les frontières de segments peuvent différer selon les annotateurs. La procédure à plusieurs passes par alternance permet de faire disparaître progressivement ce type de désaccords. 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 le type *Modulation+Transposition*, ce qui a été finalement approuvé par le premier annotateur en 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 type *Généralisation* : *want to start by -> vais*.

13. Chaque sous-corpus représente une ou plusieurs interventions complètes aux *TED Talks*, afin de mieux comprendre le contexte.

Pour des types *Équivalence*, *Particularisation* et *Métaphore*, 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. Par ailleurs, une fonction pourrait être ajoutée à l'outil Yawat pour montrer où se situent les modifications à partir de la deuxième passe pour accélérer la révision à chaque nouvelle passe d'annotation.

Statistiques Nous présentons dans la figure 5 les statistiques portant sur les changements effectués pendant notre processus d'annotation en trois passes sur un sous-corpus [14]. Les chiffres dans la dernière colonne de la figure 5(b) signifient que, par exemple, le second annotateur était d'accord pour les 37 instances de type *Mod+Trans*, mais que le premier annotateur a corrigé ses premiers choix lors de la troisième passe. Ces deux tables montrent que les désaccords sur les frontières et sur les types diminuent progressivement grâce à cette procédure à plusieurs passes. Des exemples pour certains types de changements sont présentés dans la table 2.

À ce jour nous avons annoté un corpus de contrôle anglais-français, cinq sous-corpus anglais-français et deux sous-corpus anglais-chinois, dont les statistiques apparaissent dans la table 3. La table 4 donne elle les statistiques sur le nombre de tokens annotés par langue et par type pour la paire anglais-français [15]. Il apparaît que 73,5% des tokens anglais sont annotés avec les types *Littérale* et *Équivalence*, et 20,4% sont annotés avec *Modulation, Transposition, Mod+Trans, Généralisation* et *Particularisation*, qui sont les types de traduction les plus intéressants pour la génération de variantes non paraphrastiques par pivot.

Catégorie	Exemples
Littérale	I'll explain **it** -> je vais **l'**expliquer, refuses -> refuse de
Équivalence	here's -> voilà, no -> pas de
Particularisation	extend **it** -> allonge **la suite**
Modulation	we -> on, you -> on, about it -> en the reason [...] **is** -> la raison [...], **c' est** I encourage **all of you** -> je **vous** encourage **tous**
Transposition	humanity **'s** legacy -> héritage **de** l' humanité

TABLE 1: Exemples de conventions d'annotation.

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")
Littérale -> Équivalence	global warming pollution -> pollution à effet de serre
Rajouter *Simplification*	**most of** the last three years -> ces 3 dernières années
Littérale -> Modulation	shallow -> peu profond
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

TABLE 2: Exemples de changements de types survenus lors d'une deuxième passe.

	Nb lignes	Nb Tokens EN	Nb Tokens FR	Nb Caractères ZH
1	95	1,792	1,774	2,388
2	106	2,282	2,545	3,851
3	101	2,189	2,357	-
4	92	1,381	1,489	-
5	133	2,566	2,766	-
contrôle	100	3,055	3,238	-

TABLE 3: Statistiques sur les sous-corpus annotés.

	Anglais	Français	% EN tokens
Littérale	6,864	7,154	67,23%
Équivalence	645	822	6,32%
Modulation	1,173	1,221	11,49%
Transposition	189	263	1,85%
Mod+Trans	250	301	2,45%
Généralisation	172	121	1,68%
Particularisation	303	421	2,97%
Idiome	4	6	0,04%
Métaphore	16	19	0,16%
Simplification	122	0	1,20%
Explicitation	0	119	0,00%
Incertain	114	131	1,12%
Tous les types	9,852	10,578	96,49%
Aucun Type	358	353	3,50%
Nb tokens total	10,210	10,931	-

TABLE 4: Statistiques sur les annotations anglais-français (nombre de tokens).

14. Pour les 17 cas de désaccords de *Transposition* avec la même frontière, 11 cas concernent un changement de ponctuation, qui ont été par la suite annotés en *Modulation* selon une nouvelle convention rendue nécessaire.

15. Les chiffres ne sont pas encore définitifs, les deux derniers sous-corpus n'ayant pas encore subit une troisième passe d'annotation.

passe 1 à passe 2					
		même frontière		frontière différente	
	nb d'instances	même type	type différent	même type	type différent
Littérale	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
Simplification	4	4	0	0	0
Incertain	17	13	4	0	0
Total	1769	1651	93	7	18

(a) Passe 1 à passe 2

passe 2 à passe 3								
	même frontière			frontière différente				
	total	accord	désaccord	total	AFAT	AFDT	DFDT	correction du premier annotateur
Littérale	**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

(b) Passe 2 à passe 3

FIGURE 5: Statistiques sur les changements de types (nombre d'instances) pendant un processus d'annotation en trois passes sur un sous-corpus. AFAT : accord sur la frontière et le type ; AFDT : accord sur la frontière mais avec type différent ; DFDT : différente frontière et différent type.

Analyse contrastive entre langues cible L'annotation de ce corpus multilingue parallèle permet de révéler des contrastes entre les traductions vers des langues différentes (voir figure 6). Nous présentons des statistiques préliminaires basées sur l'annotation de deux sous-corpus anglais-français et anglais-chinois (voir table 3), et nous comparons la traduction des segments anglais strictement identiques (i.e. avec la même frontière). Nous trouvons ainsi qu'il existe moins de traductions littérales vers le chinois, et que la différence principale avec la traduction vers le français consiste en l'utilisation de différents types de *Modulation*. Les traducteurs chinois ont beaucoup plus recours à des phénomènes d'*Explicitation*. Nous poursuivons l'annotation de la paire anglais-chinois pour établir une analyse plus fine des contrastes obtenus.

	Anglais	Français	Anglais	Chinois
Littérale	2796	2896	2141	3414
Équivalence	247	310	326	478
Modulation	343	357	386	549
Transposition	106	150	112	171
Mod+Trans	128	126	29	47
Généralisation	58	37	158	154
Particularisation	132	180	210	505
Idiome	0	0	0	0
Métaphore	10	15	6	10
Simplification	74	-	178	-
Explicitation	-	46	-	459
Incertain	35	31	128	238
Tous les types	3929	4148	3674	6025
Aucun Type	145	171	400	214
Nb tokens total	4074	4319	4074	6239

FIGURE 6: Table de contraste entre les traductions vers le français et vers le chinois (nombre de tokens).

Nous constatons que parfois la qualité de la traduction chinoise n'est pas aussi bonne que la traduction française dans notre corpus, pour des raisons multiples : manque de connaissances du domaine spécifique d'une intervention ; manque d'édition finale pour corriger des erreurs évidentes, résultant notamment en des mots anglais laissés non traduits, des traductions erronées (voir figure 6) ou des ponctuations absentes. En outre, des traductions extrêmement libres posent même de réelles difficultés pour l'alignement manuel de mots.

Un dernier point qui n'est pas illustré dans la figure 6 concerne l'introduction d'idiomes en chinois pour traduire des expressions non figées. Par exemple :

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)

L'utilisation d'idiomes chinois en traduction est considérée comme une bonne pratique qui permet d'obtenir des textes concis adaptés à la culture chinoise. 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.

Le corpus complet (2 436 lignes dans chaque langue) sera distribué dans un format XML avec les métadonnées d'accord inter-annotateurs et de modifications entre passes d'annotation.

6 Conclusion et perspective

Dans ce travail nous nous sommes intéressée aux relations de traduction car celles-ci n'avaient pas jusque-là été prises en compte à notre connaissance dans l'approche de génération de paraphrases par équivalence de traduction, ou pivot, ainsi qu'en traduction automatique. Nous avons catégorisé ces relations et les avons annotées dans un corpus parallèle multilingue de *TED Talks*. Nous avons choisi ce genre spécifique (discours transcrit et traduit) afin d'obtenir plus de diversité qu'avec des corpus techniques. Le travail d'annotation est en phase préliminaire pour finaliser le guide d'annotation. L'accord inter-annotateurs mesuré est fort pour les segments de mêmes frontières, mais nous avons adopté un processus d'annotation à plusieurs passes, plus coûteux en temps, afin de garantir une meilleure qualité d'annotation. La faisabilité de la tâche étant confirmée, nous étendrons les annotations sur le corpus entier afin de les mettre à disposition de la communauté.

À court terme, nous allons réaliser des annotations plus fines sur les blocs de type *Modulation, Transposition* et *Modulation+Transposition*. Pendant les premières annotations, nous avons en effet privilégié l'objectif de garder une information complète à celui d'obtenir un alignement des unités les plus petites possibles, comme illustré par les paires de segments suivantes : *is believed in enough -> peut être tellement crédible*, *make it seem -> lui donner l'apparence*, *they're able to be moved around -> on peut les déplacer*, *have this kind of reception -> être reçu de cette manière*, *give you good close look at this -> vous montrer de près*. Cependant, ces segments particuliers sont souvent peu réutilisables, et il est nécessaire de réaliser un alignement plus fin pour détailler leur structure, notamment dans le but d'apprendre des patrons puis de vérifier si ceux-ci ont des attestations dans des corpus parallèles de grande taille.

Une fois ce corpus annoté finement avec l'ensemble des relations de traduction, nous développerons un classifieur pour les détecter automatiquement. De telles informations n'ont pas été prises en compte à notre connaissance dans les travaux précédents portant sur la génération de paraphrases ou la traduction automatique.

Références

APIDIANAKI M., VERZENI E. & MCCARTHY D. (2014). Semantic clustering of pivot paraphrases. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Ninth International Conference*

on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014., p. 4270–4275 : European Language Resources Association (ELRA).

BANNARD C. J. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, p. 597–604.

BARZILAY R. & 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.*, p. 50–57.

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*, p. 196–205.

CETTOLO M., GIRARDI C. & 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)*, p. 261–268, Trento, Italy.

CHUQUET H. & PAILLARD M. (1989). *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

COCOS A. & CALLISON-BURCH C. (2016). Clustering paraphrases by word sense. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL 2016)*, San Diego, California : Association for Computational Linguistics.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

DENG D. & XUE N. (2017). Translation divergences in chinese-english machine translation : An empirical investigation. *Computational Linguistics*, **43**(3), 521–565.

DOLAN B., QUIRK C. & 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.

DYER C., CHAHUNEAU V. & SMITH N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In L. VANDERWENDE, H. D. III & K. KIRCHHOFF, Eds., *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*, p. 644–648 : The Association for Computational Linguistics.

GANITKEVITCH J. & 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.*, p. 4276–4283.

GANITKEVITCH J., DURME B. V. & CALLISON-BURCH C. (2012). Monolingual distributional similarity for text-to-text generation. In E. AGIRRE, J. BOS & M. T. DIAB, Eds., *Proceedings of the First Joint Conference on Lexical and Computational Semantics, *SEM 2012, June 7-8, 2012, Montréal, Canada.*, p. 256–264 : Association for Computational Linguistics.

GANITKEVITCH J., DURME B. V. & CALLISON-BURCH C. (2013). PPDB : the paraphrase database. In *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*, p. 758–764.

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*, p. 20–23 : The Association for Computer Linguistics.

IBRAHIM A., KATZ B. & LIN J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second International Workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, p. 57–64, Stroudsburg, PA, USA : Association for Computational Linguistics.

KOK S. & 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*, p. 145–153.

LAPATA M., SENNRICH R. & MALLINSON J. (2017). Paraphrasing revisited with neural machine translation. In M. LAPATA, P. BLUNSOM & A. KOLLER, Eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1 : Long Papers*, p. 881–893 : Association for Computational Linguistics.

LI Z. & SUN M. (2009). Punctuation as implicit annotations for Chinese word segmentation. *Comput. Linguist.*, **35**(4), 505–512.

LIN D. & 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)*, p. 323–328, New York, NY, USA : ACM.

MILLER G. A. (1995). Wordnet : A lexical database for english. *Commun. ACM*, **38**(11), 39–41.

MONTI J., SANGATI F. & ARCAN M. (2015). Ted-MWE : a bilingual parallel corpus with MWE annotation towards a methodology for annotating mwes in parallel multilingual corpora.

PANG B., KNIGHT K. & MARCU D. (2003). Syntax-based alignment of multiple translations : Extracting paraphrases and generating new sentences. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, p. 102–109, Stroudsburg, PA, USA : Association for Computational Linguistics.

PAVLICK E., BOS J., NISSIM M., BELLER C., DURME B. V. & 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1 : Long Papers*, p. 1512–1522.

PAVLICK E., RASTOGI P., GANITKEVITCH J., DURME B. V. & 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*, p. 425–430.

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. & DEAN J. (2016). Google's neural machine translation system : Bridging the gap between human and machine translation. *CoRR*, **abs/1609.08144**.

ZHAO S., WANG H., LIU T. & LI S. (2008). Pivot approach for extracting paraphrase patterns from bilingual corpora. In K. MCKEOWN, J. D. MOORE, S. TEUFEL, J. ALLAN & S. FURUI, Eds., *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, p. 780–788 : The Association for Computer Linguistics.