RECITAL 2010, Montréal, 19–23 juillet 2010

Construction d'un corpus de paraphrases d'énoncés par traduction multiple multilingue

Houda Bouamor
LIMSI-CNRS, groupe ILES
Université Paris-Sud 11
Orsay, France
houda.bouamor@limsi.fr

Résumé. Les corpus de paraphrases à large échelle sont importants dans de nombreuses applications de TAL. Dans cet article nous présentons une méthode visant à obtenir un corpus parallèle de paraphrases d'énoncés en français. Elle vise à collecter des traductions multiples proposées par des contributeurs volontaires francophones à partir de plusieurs langues européennes. Nous formulons l'hypothèse que deux traductions soumises indépendamment par deux participants conservent généralement le sens de la phrase d'origine, quelle que soit la langue à partir de laquelle la traduction est effectuée. L'analyse des résultats nous permet de discuter cette hypothèse.

Abstract. Large scale paraphrase corpora are important for a variety of natural language processing applications. In this paper, we present an approach which collects multiple translations from several languages proposed by volunteers in order to obtain a parallel corpus of paraphrases in French. We hypothesize that two translations proposed independently by two volunteers usually retain the meaning of the original sentence, regardless of the language from which the translation is done. The analysis of results allows us to discuss this hypothesis.

Mots-clés : corpus monolingue parallèle, paraphrases, traductions multiples.

Keywords: monolingual parallel corpora, paraphrases, multiple translations.

1 Introduction

La richesse de la langue permet aux humains d'exprimer la même idée de façons très différentes. Cette variabilité d'expression est une source majeure de difficultés dans la plupart des applications de traitement automatique des langues. En effet, l'une des méthodes pour résoudre les problèmes engendrés par ce phénomène consiste à acquérir des paraphrases, à savoir un ensemble de phrases exprimant la même idée ou décrivant le même évènement. Le type de corpus le plus approprié pour les trouver naturellement est composé d'énoncés en relation de paraphrase, or ce genre de corpus est une ressource très rare, dont la constitution est une tâche compliquée et très coûteuse.

Cependant, étant donné leur utilité, plusieurs méthodes de construction de corpus de paraphrases d'énoncés dans différentes langues ont été mises en place. Barzilay & McKeown (2001) ont distingué trois manières différentes pour la collecte de paraphrases. La première est l'utilisation de ressources linguis-

tiques existantes. La seconde est l'extraction de mots ou d'expressions similaires en se basant sur un corpus. La troisième, enfin, est l'acquisition manuelle de paraphrases. Elle est sans doute la plus facile à implémenter, et ses résultats sont les plus fiables.

Dans cet article nous présentons une méthode qui consiste à contruire un corpus parallèle de paraphrases d'énoncés proposées par des contributeurs volontaires comme dans (Chklovski, 2005) sous la forme de traductions multiples à partir de plusieurs langues européennes vers le français. L'article est structuré de la façon suivante : dans la section 2 nous donnons un aperçu des travaux liés à notre approche, puis nous exposons notre méthode de création d'un corpus de paraphrases d'énoncés dans la section 3. La section 4 présente les résultats des analyses faites sur le corpus obtenu et la section 5 décrit quelques travaux futurs.

2 État de l'art

Les deux premières méthodes introduites par Barzilay & McKeown (2001) incluent les méthodes d'acquisition automatique de paraphrases. Par exemple, Langkilde & Knight (1998) se sont basés sur les connaissances sémantiques fournies par WordNet (Miller, 1995) pour exploiter les relations de synonymie entre termes et les utiliser ensuite lors de la génération de paraphrases. Ces ressources linguistiques ne sont pas nécessairement disponibles dans toutes les langues. C'est pour cela que de nombreux travaux d'acquisition de paraphrases se sont appuyés sur des corpus monolingues, comparables ou parallèles. Nous pouvons citer ainsi l'étude faite par Barzilay & McKeown (2001) dans laquelle les auteurs se servent d'un corpus parallèle monolingue et utilisent des informations contextuelles basées sur des similarités lexicales pour extraire des paraphrases. De manière similaire, Pang *et al.* (2003) exploitent la structure syntaxique d'un ensemble de phrases issues de corpus parallèles monolingues pour construire de nouvelles paraphrases d'énoncés par fusion syntaxique et regénération. Ibrahim *et al.* (2003) présentent eux une méthode non supervisée d'acquisition de paraphrases qui consiste à extraire des paraphrases structurelles, ou des fragments d'arbres syntaxiques sémantiquement équivalents, à partir de corpus monolingues parallèles.

Puisque les corpus monolingues parallèles sont des ressources difficiles à obtenir, d'autres techniques ont été implémentées en se basant sur des corpus monolingues comparables, corpus composés de textes dans la même langue partageant une partie du vocabulaire employé, ce qui implique généralement que les textes parlent d'un même sujet, durant la même période, afin d'obtenir des paraphrases. Notamment, certains travaux exploitent des corpus monolingues comparables, comme ceux de Deléger & Zweigenbaum (2009) dans le domaine médical visant la construction d'un corpus de paraphrases de segments opposant les langues de spécialité et de vulgarisation. Barzilay & Lee (2003) introduisent une technique d'alignement multi-séquence factorisant des phrases ayant la même structure syntaxique, extraites à partir d'un corpus comparable, sous forme de treillis contenant des équivalences locales. Dolan *et al.* (2004) présentent un corpus qui comprend plusieurs milliers de paires de paraphrases décrivant des événements similaires, collectées à partir de sites d'information. Les paires de paraphrases ont été annotées par des juges humains pour décider si elles sont sémantiquement équivalentes. Les paraphrases qu'ils ont obtenues sont souvent plus comparables que parallèles, dans le sens où il n'est pas toujours possible d'associer les éléments d'une phrase à ceux de l'autre phrase.

Ces méthodes ont le défaut d'être fondées sur des algorithmes nécessitant de grandes masses de données annotées de façon fiable. Pour pallier ces problèmes, Bouamor *et al.* (2009) ont formulé l'acquisition de paraphrases d'énoncés sous forme d'un jeu : une application pour acquérir des paraphrases demande à des participants de réécrire des phrases dans un premier temps, puis d'évaluer les paraphrases proposées par

les autres joueurs. Cette application permet à la fois de constituer un corpus de paraphrases, et d'acquérir des jugements humains multiples sur ces paraphrases. Les travaux de Chklovski (2005) s'inscrivent dans ce même cadre. Cet auteur a mis en place un jeu en ligne pour l'acquisition de paraphrases pour des expressions données[1]. Dans ce jeu, des joueurs doivent déterminer des paraphrases partiellement masquées pour une expression donnée. Chaque joueur a la possibilité de faire plusieurs essais pour trouver la paraphrase de « référence ».

Une autre manière de procéder consiste à demander à des contributeurs de paraphraser directement un énoncé dans la langue souhaitée. C'est ce qui a par exemple été fait pour la construction des fichiers de développement du corpus BTEC (Takezawa *et al.*, 2002), où des phrases en japonais sont tout d'abord traduites en anglais, puis paraphrasées par 15 locuteurs monolingues, créant ainsi un corpus monolingue parallèle où chaque énoncé dispose de 16 paraphrases, dont l'une peut être considérée comme « paraphrase de référence ». Nous émettons des réserves sur cette méthologie d'obtention de paraphrases, notamment à la vue des résultats présentés dans (Schroeder *et al.*, 2009). Ces auteurs rapportent des résultats en traduction multisource, où les différentes paraphrases d'un texte à traduire (ici, les 16 paraphrases de la partie anglaise du BTEC) sont utilisées simultanément. Les résultats montrent qu'utiliser l'ensemble de ces paraphrases *artificiellement*[2] obtenues sans l'énoncé de départ dans un scénario multisource mène à de moins bons résultats que traduire simplement l'énoncé de départ : cela peut donc dans une certaine mesure être interprété comme une déviation du sens de l'énoncé d'origine.

Une autre solution est d'exprimer ce problème d'acquisition comme une tâche de traduction multiple d'un même corpus de phrases. Cette méthode a été suivie pour constituer le MTC (Multiple-Translation Chinese corpus)[3]. Ce corpus a été développé pour la traduction automatique, afin de permettre l'utilisation de plusieurs traductions de référence en traduction vers l'anglais. Il contient 105 articles (993 phrases) extraits de 3 journaux écrits en mandarin. Les phrases sources ont été traduites indépendamment en anglais par 11 traducteurs. Chaque groupe de phrases traduites comporte donc 11 traductions sémantiquement équivalentes, qui peuvent être considérées comme étant des paraphrases d'énoncés. Ce corpus constitue une source riche pour l'apprentissage de paraphrases lexicales et structurelles, et a en particulier été utilisé dans l'étude de Pang *et al.* (2003). Barzilay & McKeown (2001), quant à elles, ont exploité un corpus de paraphrases en anglais contruit via 11 traductions vers l'anglais de 5 romans pour la génération automatique de paraphrases d'énoncés. Cohn *et al.* (2008) utilisent également des traductions multiples afin de construire un corpus monolingue parallèle pour faire de l'annotation manuelle de paraphrases.

Dans le cadre de l'amélioration des systèmes de Questions-Réponses, plusieurs travaux ont porté sur la construction de corpus de paraphrases de questions. Bernhard & Gurevych (2008) ont construit un corpus de paraphrases de questions en exploitant des réseaux sociaux sur le Web. Dans cette expérience, elles ont recueilli un corpus de questions et leurs paraphrases à partir du site WikiAnswers[4]. Lorsqu'un utilisateur entre une question qui ne fait pas déjà partie des questions référencées sur WikiAnwers, le site web affiche une liste de questions déjà existantes sur le site et semblables à celle que vient de poser l'utilisateur. L'utilisateur choisit la question qui paraphrase sa propre question. Ces reformulations de questions sont stockées et peuvent être récupérées de manière automatique pour constituer un corpus de paraphrases de questions. Dans le même domaine, Max & Wisniewski (2010) décrivent la construction d'un corpus issu des révisions de l'encyclopédie collaborative Wikipedia, qui compte de nombreuses paraphrases parmi les

[1] 1001 Paraphrases (`http://ai-games.org/paraphrase.html`)

[2] Nous les considérons comme artificielles parce qu'elles sont le résultat d'un processus de paraphrasage qui tel que formulé n'est pas une activité naturelle pour les humains

[3] Linguistic Data Consortium (LDC) Catalog Number LDC2002T01, ISBN 1-58563-217-1

[4] `http://wiki.answers.com/`

reformulations obtenues.

Dans les sections suivantes, nous allons détailler notre méthode d'acquisition de paraphrases d'énoncés qui se situe, comme le corpus MTC, dans la troisième catégorie de construction de corpus de paraphrases, selon la catégorisation de Barzilay & McKeown (2001).

3 Acquisition de paraphrases d'énoncés

3.1 Approche générale

Une façon d'obtenir des paraphrases est de traduire plusieurs fois un même texte. En effet, deux traductions d'un même texte doivent conserver le sens d'origine, ceci indépendamment de la langue source. Néanmoins, chaque traducteur pourra faire des choix de traduction différents, qui mèneront donc dans la majorité des cas à des traductions différentes. C'est cette hypothèse que nous mettons en œuvre dans ce travail. Notre objectif est de constituer un corpus de paraphrases d'énoncés en français, en collectant des traductions multiples d'un même ensemble de phrases proposées par des participants volontaires, et ce à partir de plusieurs langues.

3.2 Campagne de collecte de paraphrases

3.2.1 Collecte de paraphrases

L'expérience a consisté à soumettre un ensemble de 500 phrases, exprimées chacune dans 10 langues européennes (anglais, allemand, italien, portugais, espagnol, néerlandais, danois, suédois, finnois et grec) à un groupe de participants francophones en leur demandant de traduire chacune des phrases tout en conservant du mieux possible le sens d'origine.

Les contributeurs, au nombre de 132, ne sont majoritairement pas des traducteurs professionnels mais ont très majoritairement le français comme langue maternelle (quelques contributeurs ont une des autres langues comme langue maternelle, mais ont tous un bon niveau de français écrit). Nous leur avons demandé de traduire en français des phrases qui leur sont proposées dans une des autre langues pour lesquelles ils se sont déclarés compétents pour traduire lors de leur inscription sur le système.

Les 500 phrases proposées sont extraites aléatoirement du corpus parallèle multilingue Europarl (Koehn, 2002) composé de transcriptions des débats parlementaires européens. Nous avons choisi Europarl car il comporte un ensemble de phrases disponibles en 11 langues[5] (les 10 langues d'origine des phrases à traduire et le français), ce qui est une caractéristique rare pour un corpus parallèle.

[5]La version d'Europarl que nous avons utilisée ne nous permet pas de connaître la langue d'origine pour chaque phrase.

3.2.2 Description de l'outil de collecte

Afin de collecter ces données, nous avons mis en place une interface Web[6], illustrée sur la Figure 1. La traduction se fait en deux étapes : dans une première étape, le contributeur saisit une proposition de traduction initiale. Une fois cette proposition soumise, une étape d'amélioration est proposée dans laquelle l'ensemble des traductions déjà effectuées pour cette phrase, dont celle d'un traducteur professionnel (version originale d'Europarl), sont présentées[7]. Les utilisateurs sont sensibilisés au fait que l'objectif n'est pas de transformer leur proposition en la traduction de référence, mais de simplement corriger les éléments qu'ils auraient pu mal traduire. Notre objectif principal est en effet d'obtenir le plus d'équivalence sémantique possible entre les paraphrases candidates collectées, tout en ayant des variations lexicales ou syntaxiques.

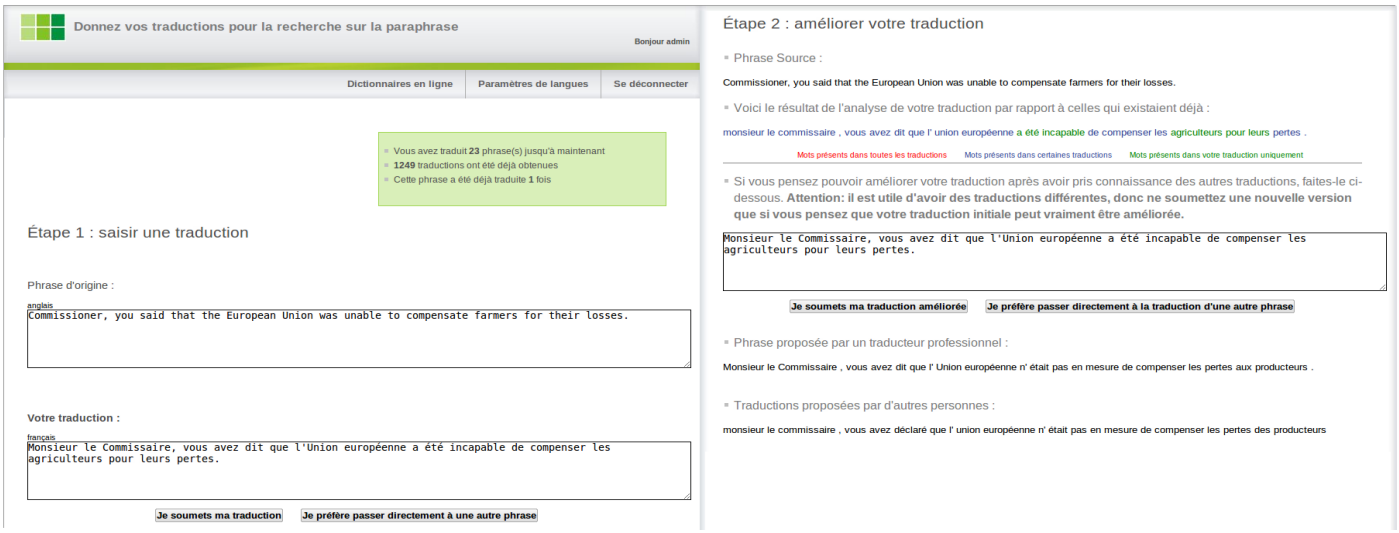

FIG. 1 – Interface de collecte de paraphrases : les 2 étapes

Pour ne pas perdre d'information sur le processus de traduction, notre outil conserve la trace des différentes versions proposées, ces données pouvant se révéler utiles pour des études sur les mécanismes de traduction. Nous avons ainsi pu noter que 75% (99 sur 132) de nos contributeurs utilisent la fonctionnalité proposée pour améliorer leurs traductions, afin de corriger des erreurs d'orthographe (*citoyeneté* → *citoyenneté*), de réécrire un segment (*au sens littéral* → *au sens propre du terme*) ou des structures plus complexes (*Ce sont tous deux des emplois pénibles et prenants.* → *Ce sont des emplois à la fois pénibles et exigeants.*), ou encore de corriger des erreurs de traduction (*Ce procédé n'est pas particulièrement difficile.* → *Ce sont des emplois à la fois pénibles et exigeants.*).

Afin de maintenir l'équilibre du corpus en terme de langues et de nombre de traductions par phrase, nous avons mis au point un algorithme de choix de la prochaine phrase soumise à la traduction pour un contributeur. Celle-ci est sélectionnée en fonction des langues à partir desquelles celui-ci sait traduire, des phrases qu'il a déjà traduites (un utilisateur ne traduit qu'une seule fois une même phrase, dans n'importe

[6]Accessible sur `http://perso.limsi.fr/hbouamor/Para`

[7]Des entretiens informels nous ont permis de constater que cette étape de localisation rapide des éventuelles difficultés de traduction était réalisée de façon assez efficace par la plupart des contributeurs. De fait, cela permet d'avoir des traductions dont la structure correspond souvent à celle initialement proposée par les contributeurs, indépendamment donc des autres traductions disponibles, et qui sont parfois corrigées au niveau lexical.

quelle langue), et de la langue pour laquelle on dispose du plus petit nombre de traductions. Une fois la langue de la phrase source identifiée, la phrase traduite le plus petit nombre de fois et n'ayant pas déjà été traduite par ce contributeur est retenue[8].

4 Analyse des résultats obtenus

À la date du 15 mars 2010, nous avons répertorié 1196 traductions dont la répartition sur les différentes langues d'origine est précisée dans le Tableau 1[9].

Langue	**en**	**es**	**de**	**it**	**pt**	**fi**	**el**	**sv**	**nl**	**da**
Nombre de traductions	730	112	100	94	58	40	15	13	10	5
Nombre de contributeurs	76	19	14	11	9	5	2	5	7	4

TAB. 1 – Nombre de traductions et de contributeurs par langue

Langue	**Phrase source et traduction proposée**
en	There should be absolutely no ambiguity about our message.
fr	**Notre message devrait être parfaitement clair.**
es	No debe haber la menor ambigüedad en nuestro mensaje.
fr	**Il ne doit pas y avoir la moindre ambiguïté dans notre message.**
de	Unsere Botschaft muß eindeutig sein.
fr	**Notre message doit être clair.**
it	Non può esserci ambiguità alcuna nel nostro messaggio.
fr	**Il ne peut y avoir aucune ambiguïté dans notre message.**
fi	Viestissämme ei saa olla minkäänlaista monitulkintaisuutta.
fr	**En aucun cas notre message ne devra contenir plusieurs interprétations possibles.**

TAB. 2 – Exemples de traductions obtenues à partir de plusieurs langues sources pour une phrase courte

Les contributeurs qui savent traduire à partir de l'allemand, de l'espagnol ou d'autres langues, savent en général traduire aussi à partir de l'anglais, ce qui justifie le fait que la somme des contributeurs est supérieure aux 132 participants. Dans le corpus obtenu, 490 phrases ont été traduites au minimum 2 fois. Nous avons donc des paraphrases pour 94% des phrases contenues de notre corpus. Des exemples de traductions d'une même phrase source à partir de différentes langues sont donnés dans le Tableau 2.

Le corpus obtenu contient, en plus des paraphrases correctes, des paraphrases comportant des erreurs causées par des fautes d'orthographe (*nécéssaire, peche, apprôché*), des erreurs de traductions (*M. Le Président* au lieu de *Mme La Présidente*), ou encore la présence de phrases incomplètes.
Pour mesurer le degré de similarité lexicale entre les paraphrases obtenues à partir des différentes langues,

[8]En outre, cette phrase ne doit pas être en cours de traduction par un autre contributeur.
[9]Codes de la représentation des noms de langues selon la norme ISO 639-1

nous proposons d'utiliser le coefficient de chevauchement (*overlap coefficient*) *CO* qui représente le pourcentage de chevauchement lexical entre les vocabulaires P_1 et P_2 de deux phrases, défini comme :

$$CO = \frac{|P_1 \cap P_2|}{min(|P_1|, |P_2|)} \quad (1)$$

Le nombre minimal de paires de paraphrases pour un groupe est un paramètre, qui a été fixé à 20 dans notre expérience. La Table 3 regroupe les moyennes du coefficient de chevauchement lexical pour l'ensemble des *tokens* sur les groupes de paraphrases retenus, ceci pour différentes langues d'origine. Les 172 [10] paraphrases obtenues à partir de l'anglais proposées par différents traducteurs comportent ainsi 90% de *tokens* communs en moyenne. En revanche, nous constatons que celles provenant de deux langues différentes comportent entre 36% et 42% de *tokens* différents en moyenne. Ces valeurs montrent que nous obtenons davantage de variation lexicale en utilisant des langues sources différentes pour obtenir des phrases sémantiquement équivalentes, ce qui correspond à ce que nous attendions.

	en	**es**	**de**	**it**	**pt**	**fi**
en	0,90 $_{172}$	0,64 $_{69}$	0,59 $_{89}$	0,63 $_{84}$	0,62 $_{58}$	0,64 $_{32}$
es	* [11]	- [12]	0,62 $_{57}$	0,63 $_{57}$	0,64 $_{51}$	-
de	*	*	-	0,58 $_{67}$	0,61 $_{53}$	-
it	*	*	*	-	0,65 $_{50}$	-
pt	*	*	*	*	-	-
fi	*	-	-	-	-	-

TAB. 3 – Calcul de similarité entre les paraphrases obtenues à partir de différentes langues (tous *tokens*)

Nous avons refait ces mesures, mais en ne considérant cette fois que les mots pleins pour la mesure de chevauchement, ce qui élimine de fait les nombreux *tokens* correspondant à des signes de ponctuation et à des mots grammaticaux. Les résultats sont donnés dans le Tableau 4. Nous observons dans ce cas qu'en moyenne le degré de similarité entre les paires de phrases est plus faible.

	en	**es**	**de**	**it**	**pt**	**fi**
en	0,88 $_{172}$	0,57 $_{69}$	0,52 $_{89}$	0,57 $_{84}$	0,54 $_{58}$	0,62 $_{32}$
es	*	-	0,53 $_{57}$	0,54 $_{57}$	0,56 $_{51}$	-
de	*	*	-	0,52 $_{67}$	0,52 $_{53}$	-
it	*	*	*	-	0,58 $_{50}$	-
pt	*	*	*	*	-	-
fi	*	-	-	-	-	-

TAB. 4 – Similarité entre les paraphrases obtenues à partir de différentes langues (mots pleins)

Nous avons fait une dernière mesure, qui cette fois-ci ne porte que sur les mots pleins, responsables du contenu, et sur leur forme lemmatisée, neutralisant ainsi des variations de surface. Le degré de similarité,

[10] C'est le nombre indiqué en indice dans la table 3 représentant le nombre de traductions communes obtenues à partir de deux langues

[11] Cette valeur est indiquée ailleurs dans le tableau.

[12] Nous n'avons pas assez de paraphrases provenant des 2 langues.

légèrement plus fort que dans l'expérience précédente, varie tout de même entre 57% et 68%, ce qui montre une variation raisonnablement importante sur ce type d'unité et reflète donc des choix de traduction lexicaux variés.

	en	**es**	**de**	**it**	**pt**	**fi**
en	0,90 $_{172}$	0,65 $_{69}$	0,61 $_{89}$	0,66 $_{84}$	0,64 $_{58}$	0,68 $_{32}$
es	*	-	0,57 $_{57}$	0,68 $_{57}$	0,68 $_{51}$	-
de	*	*	-	0,59 $_{67}$	0,62 $_{53}$	-
it	*	*	*	-	0,66 $_{50}$	-
pt	*	*	*	*	-	-
fi	*	-	-	-	-	-

TAB. 5 – Similarité entre les paraphrases obtenues à partir de différentes langues (lemmes de mots pleins)

Nous avons, ensuite, extrait deux sous-corpus comportant des groupes de 4 paraphrases pour les mêmes 50 phrases sources. Le premier sous-corpus *corpusEN* est constitué de paraphrases obtenues par des traductions multiples en français d'une même phrase en anglais, et le deuxième *corpusMulti* est constitué de traductions à partir de l'allemand, de l'espagnol, de l'italien et du portugais.

Afin de constituer un corpus de paraphrases de référence servant à la comparaison des méthodes d'acquisition de paraphrases sous-phrastiques présentée dans (Bouamor *et al.*, 2010), nous considérons une paraphrase de chaque groupe comme une phrase de référence, qui devra être alignée avec chacune des 3 autres paraphrases de son groupe. Nous avons, alors, constitué 300 paires de paraphrases (2 corpus * 50 groupes * 3 alignements), puis nous avons demandé à 3 annotateurs d'aligner au niveau des mots chacune de ces paires en utilisant l'outil d'alignement YAWAT (Germann, 2008). Nous avons comparé

	Alignement	**Chevauchement**
CorpusEN	0,95	0,90
CorpusMulti	0,77	0,61

TAB. 6 – Pourcentage des mots alignés et des mots communs entre 2 paires de paraphrases dans chacun des sous-corpus

les deux sous-corpus en terme du pourcentage de mots alignés par nos annotateurs et de chevauchement lexical. Les résultats obtenus sont présentés dans le Tableau 6. Nous remarquons que pour *corpusEN*, les annotateurs ont réussi à aligner 95% des mots dont 90% sont des mots communs dans les paires de paraphrases obtenues à partir du français. En revanche, nous obtenons plus de variations dans le vocabulaire utilisé dans *corpusMulti* avec 61% de mots communs et 77% des mots ayant été alignés.

Le Tableau 7 indique le degré d'intérêt des mots alignés dans la collecte des paraphrases. En effet, plus nous obtenons de correspondances entre des mots différents, plus nous avons de variations lexicales dans le corpus de paraphrases construits. Ce critère a été utilisé dans (Bouamor *et al.*, 2010) afin de comparer des méthodes d'acquisition de paraphrases sous-phrastiques.

	mêmes mots	**mots différents**
Alignement humain OK	pas intéressant internationales ↔ internationales	très intéressant alimentation pour nourrissons ↔ aliments pour bébés
Alignement humain non OK	? ? (sans objet) internationales ↔ internationales	étude nécessaire Madame ↔ Monsieur

TAB. 7 – Intérêt des mots alignés manuellement dans l'étude des paraphrases d'énoncés obtenues

5 Conclusion et travaux futurs

Dans cet article, nous avons présenté une méthode permettant de créer un corpus d'énoncés en relation de paraphrases en français au travers de la collection de traductions proposées par plusieurs contributeurs à partir de différentes langues. Nous avons formulé l'hypothèse que deux traductions soumises indépendamment par deux participants conservent généralement le sens de la phrase d'origine, quelle que soit la langue à partir de laquelle la traduction est effectuée.

Nous avons décrit quelques mesures dans le but d'évaluer le degré de similarité entre les paraphrases du corpus obtenu. Notre principal résultat est que les paraphrases obtenues à partir de différentes langues comportent plus de variations lexicales que celles obtenues à partir d'une seule langue[13]. Nous prévoyons d'effectuer d'autres mesures complémentaires sur la variation, en particulier au niveau syntaxique. Nos expériences en acquisition de paraphrases sous-phrastiques sur un sous-ensemble de ce corpus (Bouamor *et al.*, 2010) répondent déjà en partie à cette question : en particulier, une approche opérant par fusion syntaxique de paires de paraphrases est plus efficace pour l'extraction de paraphrases sous-phrastiques sur un corpus obtenu à partir d'une seule langue cible, laissant entendre que les structures syntaxiques sont beaucoup plus comparables que lorsque les traductions proviennent de plusieurs langues.

Une première perspective de ce travail est de nettoyer le corpus obtenu, en commençant par corriger les fautes d'orthographe qu'il contient, ainsi que les erreurs de traduction et les phrases incomplètes. Nous pourrons alors mettre ce corpus nettoyé à la disposition de la communauté. Nous envisageons également de mettre notre outil à la disposition d'élèves dans des écoles de traduction, afin d'obtenir des traductions de meilleure qualité et d'enrichir ainsi notre corpus. Un deuxième type de perspective concerne l'automatisation d'une partie de l'acquisition des traductions en intégrant des systèmes de traduction automatique et en offrant la possibilité aux contributeurs de transformer ces hypothèses en la traduction la plus proche, à la manière de la mesure HTER (Snover *et al.*, 2009).

Remerciements Nous tenons à remercier les nombreux contributeurs volontaires participant à la construction de ce corpus de traductions multiples.

Références

BARZILAY R. & LEE L. (2003). Learning to paraphrase : an unsupervised approach using multiple-

[13]Ce résultat vaut pour le type de contributeurs qui utilisent notre système, à savoir des non-spécialistes de la traduction, qui en outre ne sont pas rémunérés. Il serait intéressant de pouvoir répondre aux mêmes questions si les contributions provenaient de traducteurs professionnels.

sequence alignment. In *Actes de NAACL-HLT*, Edmonton, Canada.

BARZILAY R. & MCKEOWN K. (2001). Extracting paraphrases from a parallel corpus. In *Actes de ACL*, Toulouse, France.

BERNHARD D. & GUREVYCH I. (2008). Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, USA.

BOUAMOR H., MAX A. & VILNAT A. (2009). Amener des utilisateurs à créer et évaluer des paraphrases par le jeu. In *Actes de TALN, session de démonstrations*, Senlis, France.

BOUAMOR H., MAX A. & VILNAT A. (2010). Acquisition de paraphrases sous-phrastiques depuis des paraphrases d'énoncés. In *Actes de TALN 2010*, Montréal, Canada.

CHKLOVSKI T. (2005). Collecting paraphrase corpora from volunteer contributors. In *Proceedings of the 3rd international conference on Knowledge capture*.

COHN T., CALLISON-BURCH C. & LAPATA M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, **34**.

DELÉGER L. & 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*.

DOLAN B., QUIRCK 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*.

GERMANN U. (2008). Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL-08 : HLT Demo Session*, Columbus, Ohio.

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* : Association for Computational Linguistics.

KOEHN P. (2002). Europarl : A multilingual corpus for evaluation of machine translation. *Ms., University of Southern California*.

LANGKILDE I. & KNIGHT K. (1998). Generations that Exploits Corpus-based Statistical Knowledge. In *Proceedings of the 36th International Conference on Computational Linguistics*.

MAX A. & WISNIEWSKI G. (2010). Mining naturally-occurring corrections and paraphrases from wikipedia's revision history. In *Proceedings of LREC*.

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

PANG B., KNIGHT K. & MARCU D. (2003). Syntax-based alignement of multiple translations : Extracting paraphrases and generating new sentences. In *Actes de NAACL-HLT*, Edmonton, Canada.

SCHROEDER J., COHN T. & KOEHN P. (2009). Word lattices for multi-source translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece.

SNOVER M., MADNANI N. & DORR, B.J. ET SCHWARTZ R. (2009). Fluency, adequacy, or HTER ? : exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th Workshop on SMT*.

TAKEZAWA T., SUMITA E., SUGAYA F., YAMAMOTO H. & YAMAMOTO S. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of LREC 2002*.