Étude bilingue de l'acquisition et de la validation automatiques de paraphrases sous-phrastiques

Houda Bouamor — Aurélien Max — Anne Vilnat

LIMSI-CNRS et Univ. Paris Sud
BP 133 91403 Orsay cedex
prenom.nom@limsi.fr

RÉSUMÉ. *Dans ce travail nous présentons une étude détaillée de la tâche d'acquisition de paraphrases sous-phrastiques à partir de corpus monolingues parallèles. Nous démontrons empiriquement que ces corpus, bien qu'extrêmement rares, constituent le type de ressources le mieux adapté pour cette étude. Nos expériences mettent en jeu cinq techniques d'acquisition, représentatives de différentes approches et connaissances, en anglais et en français. Afin d'améliorer la performance en acquisition, nous réalisons la combinaison des paraphrases produites par ces techniques par une validation reposant sur un classifieur automatique biclasse. Un résultat important de notre étude est l'identification de paraphrases qui défient actuellement les techniques étudiées, lesquelles sont classées et quantifiées en anglais et français.*

ABSTRACT. *This work uses parallel monolingual corpora for a detailed study of the task of subsentential paraphrase acquisition. We argue that the scarcity of this type of resource is compensated by the fact that it is the most suited type for studies on paraphrasing. We propose a large exploration of this task with experiments on two languages with five different acquisition techniques, selected for their complementarity, and their combinations. We report a significant gain over all techniques by validating candidate paraphrases using a maximum entropy classifier. An important result of our study is the identification of difficult-to-acquire paraphrase pairs, which are classified and quantified in a bilingual typology.*

MOTS-CLÉS : *acquisition de paraphrases, classification automatique de paraphrase, typologie de paraphrases.*

KEYWORDS: *paraphrase acquisition, paraphrase automatic classification, paraphrase typology.*

1. Introduction

La variabilité en langue est une source majeure de difficultés dans la plupart des applications du traitement automatique des langues. Elle se manifeste dans le fait qu'une même idée ou un même événement peut être exprimé avec des mots ou des groupes de mots différents ayant la même signification dans leur contexte respectif. Capturer automatiquement les équivalences sémantiques entre ces unités de texte est une tâche complexe, mais qui s'avère indispensable dans de nombreux contextes. On observe un intérêt croissant porté à l'acquisition et à l'utilisation de paraphrases (Madnani et Dorr, 2010), même si la définition de cette notion reste difficile à cerner : de nombreuses définitions existent en effet dans la littérature (Fuchs, 1994). L'acquisition *a priori* de listes d'équivalences met à disposition des ressources utiles pour, par exemple, améliorer le repérage d'une réponse à une question (Ravichandran et Hovy, 2002), autoriser des formulations différentes en évaluation de la traduction automatique (Kauchak et Barzilay, 2006), ou encore aider des auteurs à trouver des formulations plus adaptées (Bouamor *et al.*, 2011a). Dans ces travaux, la définition de ce que peut être une paraphrase est fortement liée à l'application elle-même. La définition que nous utiliserons dans cet article pour des *paraphrases sous-phrastiques* est que deux segments sont en relation de paraphrase si un annotateur humain les a reconnus comme ayant le même sens dans leur contexte respectif. Par exemple, les segments soulignés dans les phrases *elle semblait heureuse*$_1$ *de retrouver sa famille*$_2$ et *elle avait l'air contente*$_1$ *d'être à nouveau parmi les siens*$_2$ constituent des paires acceptables de paraphrases pouvant être exploitées dans divers contextes [1].

De nombreuses techniques ont été proposées pour l'acquisition de segments en relation de paraphrase (Madnani et Dorr, 2010). Ces techniques ont en commun d'être directement liées aux types de ressources auxquels elles s'appliquent. La plupart exploitent des corpus monolingues disponibles en grande quantité, et se fondent sur l'hypothèse que des unités linguistiques apparaissant dans des contextes similaires peuvent avoir la même signification. Peu de travaux ont, en comparaison, porté sur l'exploitation de corpus monolingues parallèles, constitués de phrases alignées en relation de paraphrase. Cela peut certainement s'expliquer par la faible disponibilité de telles ressources : celles-ci sont coûteuses à construire et ne sont pas produites dans le cadre d'activités naturelles.

L'acquisition et l'utilisation de paraphrases requiert une bonne compréhension de ces objets linguistiques complexes. Cela repose sur la disponibilité de corpus appropriés, d'annotations humaines fiables et de mesures d'évaluation convenables. Dans notre étude, nous utiliserons des *corpus monolingues parallèles*, constitués de paires de phrases sémantiquement équivalentes. Si ces ressources sont évidemment rares,

1. Nous ne nous intéressons pas, dans cet article, au problème de la *pertinence* des paraphrases par rapport à un contexte donné, problème très dépendant des applications visées (Zhao *et al.*, 2009). De plus, nous ne limitons pas notre étude au problème de *substituabilité grammaticale* (Bouamor *et al.*, 2011a) des paraphrases exigeant que la substitution d'une paraphrase par une autre en contexte conserve la grammaticalité des énoncés.

nous défendons le fait qu'elles sont les candidates les plus naturelles pour l'observation de la paraphrase sous-phrastique. En outre, les phrases parallèles étant issues de la volonté d'exprimer les mêmes idées, les équivalences apprises seront par nature beaucoup plus fortes qu'en utilisant des ressources davantage « comparables ». De plus, le contexte de ces équivalences peut être extrait de façon directe, ce qui est particulièrement important pour pouvoir caractériser par la suite les conditions de leur substituabilité. Nous montrerons que ces corpus contiennent une grande variété de phénomènes paraphrastiques, y compris un certain nombre qui défient les techniques automatiques actuelles. Nous avons suivi les principes généraux de l'approche décrite par Cohn *et al.* (2008), dans laquelle des paires d'énoncés en relation de paraphrase sont alignées manuellement au niveau des mots, et des techniques d'acquisition sont comparées sur leur capacité à trouver les paires de paraphrases sous-phrastiques de la référence (*rappel*) et sur la qualité des paires qu'elles prédisent (*précision*).

Le but de cette étude est de répondre à des questions relatives à l'acquisition de paraphrases sous-phrastiques : De quel type de connaissance a-t-on besoin ? Quelles techniques doit-on implémenter ? À quel point ces techniques devraient-elles être complémentaires ? Quel est l'impact de la langue et de la comparabilité des paires de phrases utilisées sur le degré paraphrastique des bisegments qu'on extrait ? Quelles sont les caractéristiques des paraphrases difficiles à extraire automatiquement ?

Cet article est organisé de la façon suivante : nous allons d'abord passer en revue les différents types de corpus de paraphrases utilisés en recherche sur la paraphrasage (section 2). Nous allons ensuite décrire la tâche d'acquisition de paraphrases sous-phrastiques sur laquelle nous nous concentrons ici (section 3). Nous présenterons alors notre cadre expérimental (section 4.1), puis les différentes techniques étudiées (section 4.2), qui se fondent respectivement sur : *a*) des modèles statistiques d'alignement de mots ; *b*) des métarègles de variation de termes ; *c*) une similarité de structures syntaxiques ; *d*) une distance d'édition sur des séquences de mots ; *e*) des équivalences de traduction. Dans le but de généraliser nos résultats, nous avons menés toutes nos expériences en français et en anglais, ce qui correspond en outre à deux niveaux différents de *comparabilité* de corpus. Nous détaillerons et analyserons les résultats obtenus (section 4.3), puis nous introduirons une méthode de combinaison des résultats de ces techniques par validation (section 4.4). Des analyses portant sur l'impact du degré de comparabilité des énoncés sur les performances des techniques d'acquisition seront décrites (section 5.1), puis nous présenterons une typologie des paraphrases « difficiles à acquérir », dans laquelle chaque classe sera illustrée par des exemples représentatifs et quantifiée dans les deux langues étudiées (section 5.2). Nous terminerons par une conclusion et une description de nos travaux futurs (section 6).

2. Construction de corpus de paraphrases d'énoncés

Contrairement à la traduction, pour laquelle il existe de grandes quantités d'exemples, les activités humaines ne produisent pas explicitement des quantités importantes de paraphrases qui pourraient servir de données d'apprentissage. Par conséquent, les

techniques automatiques se sont principalement appuyées sur l'observation *indirecte* d'unités de texte en relation d'équivalence partielle, ce qui soulève un certain nombre de questions, incluant le fait que les paires de variantes textuelles extraites peuvent ne pas être toujours contextuellement liées.

De nombreuses approches d'acquisition de paraphrases ont donc limité leurs corpus à des paires de phrases « en relation », typiques des corpus comparables, où des textes sont souvent appariés en se fondant sur un sujet commun ou une période d'apparition commune, et/ou en utilisant des techniques de recherche d'information comme première étape de sélection. Barzilay et Elhadad (2003) construisent des règles exploitant des structures thématiques et des alignements locaux pour extraire des paires de phrases à partir de dépêches de presse de différentes sources. Exploitant ce même type de données, Dolan *et al.* (2004) apparient des phrases en utilisant des distances d'édition entre énoncés.

L'acquisition de paires d'énoncés en relation de paraphrase peut également être le résultat d'une étape de traduction. Des corpus construits pour l'évaluation des systèmes de traduction automatique associant plusieurs traductions possibles à une même phrase source mettent donc à disposition des paraphrases d'énoncés. Par exemple, le corpus MTC (Multiple-Translation Chinese corpus) [2] associe 11 traductions en anglais produites indépendamment pour chaque phrase source en chinois. Ce corpus constitue une ressource très utile pour l'acquisition de paraphrases ; il a par exemple été utilisé dans l'étude de Pang *et al.* (2003) où l'acquisition est guidée par la structure syntaxique commune d'énoncés. Un tel type de corpus est, par exemple, utilisé dans l'étude de Barzilay et McKeown (2001), où sont utilisées des traductions indépendantes en anglais d'œuvres littéraires.

Dans un travail précédent (Bouamor *et al.*, 2010), nous avons mesuré l'impact de la langue source de traductions multiples sur le degré de comparabilité des paraphrases obtenues. L'expérience a consisté à soumettre un ensemble de phrases du corpus Europarl [3] disponible dans dix langues pour que des locuteurs de ces langues les traduisent vers le français. Nous avons mesuré un degré de similarité lexicale entre les paraphrases obtenues à partir des différentes langues pour des paires de langues contenant au moins vingt paires de traductions. Le tableau 1 regroupe les moyennes des similarités obtenues entre différentes paires de langues d'origine. Par exemple, nous observons que les 172 paraphrases obtenues à partir de l'anglais comportent 90 % de formes communes en moyenne. En revanche, les paraphrases provenant de deux langues différentes comportent entre 36 % et 42 % [4] de formes différentes en moyenne. Ces valeurs montrent que nous obtenons, comme attendu, davantage de variations lexicales en traduisant à partir de différentes langues.

2. Linguistic Data Consortium (LDC) Catalog Number LDC2002T01

3. `http://www.statmt.org/europarl`

4. Ces valeurs représentent les complémentaires de celles figurant dans la partie gauche du tableau 1.

	Toutes formes					Lemmes des mots pleins				
	en	**es**	**de**	**it**	**pt**	**en**	**es**	**de**	**it**	**pt**
en	0,90 $_{172}$	0,64 $_{69}$	0,59 $_{89}$	0,63 $_{84}$	0,62 $_{58}$	0,90 $_{172}$	0,65 $_{69}$	0,61 $_{89}$	0,66 $_{84}$	0,64 $_{58}$
es	*	-	0,62 $_{57}$	0,63 $_{57}$	0,64 $_{51}$	*	-	0,57 $_{57}$	0,68 $_{57}$	0,68 $_{51}$
de	*	*	-	0,58 $_{67}$	0,61 $_{53}$	*	*	-	0,59 $_{67}$	0,62 $_{53}$
it	*	*	*	-	0,65 $_{50}$	*	*	*	-	0,66 $_{50}$
pt	*	*	*	*	-	*	*	*	*	-

Tableau 1. *Valeurs de similarité lexicale entre groupes d'au moins vingt paires de paraphrases, calculées comme la moyenne de la valeur suivante entre toutes les paires de paraphrases* (p_1, p_2) *:* $\frac{|P_1 \cap P_2|}{min(|P_1|,|P_2|)}$*, où* P_1 *et* P_2 *représentent le vocabulaire de chaque phrase. Les résultats sont donnés pour tous types de formes (partie gauche) et uniquement pour les lemmes de mots pleins (partie droite).*

Il existe d'autres activités humaines qui produisent des paires d'unités textuelles qui sont parfois exploitées pour constituer des corpus contenant des paires d'énoncés plus ou moins parallèles. Par exemple, Bernhard et Gurevych (2008) ont construit un corpus de paraphrases de questions en exploitant des ensembles de questions associées à des paraphrases sur des sites de recherche d'information communautaires. L'historique des révisions des ressources collaboratives telles que Wikipédia rend possible l'extraction de certains types de modifications locales reflétant l'évolution, la maturation et la correction de la forme linguistique des articles, et constitue donc une importante source de connaissances de plus en plus exploitée à ce jour (Nelken et Yamangil, 2008 ; Max et Wisniewski, 2010). Une étude détaillée des types de modifications faites dans les révisions de Wikipédia est présenté dans (Dutrey *et al.*, 2011) : nous montrons qu'une quantité importante de ces modifications peuvent être considérées comme des paraphrases, bien que plusieurs exemples défient les techniques d'identification automatique de paraphrases actuelles. Il est aussi possible d'acquérir directement des paraphrases ciblées *via* des sites de *crowdsourcing*, pour, par exemple, fournir aux systèmes de traduction automatique des formulations alternatives (Resnik *et al.*, 2010).

Enfin, diverses techniques ont été développées pour générer automatiquement des paraphrases (Madnani et Dorr, 2010). Celles-ci ont notamment été utilisées pour enrichir des corpus d'apprentissage de systèmes de traduction automatique (Nakov, 2008), pour optimiser leurs paramètres (Madnani *et al.*, 2008) et d'autres types d'applications (Quirk *et al.*, 2004 ; Zhao *et al.*, 2008a ; Max, 2009 ; Zhao *et al.*, 2010). Cependant, outre le fait que les performances actuelles de ces techniques sont limitées, il est à noter qu'elles posent un problème de *dépendance circulaire*, car elles sont, pour la plupart, tributaires de la disponibilité des ressources d'apprentissage.

Les principaux types de corpus de paraphrases d'énoncés décrits ci-dessus sont illustrés dans la figure 1.

Corpus comparables (articles de journaux)
(Barzilay et Lee, 2003)

Prague is a centuries-old city with a wealth of historic landmarks.
The physical attractions and landmarks of Prague are many.

Corpus comparables (articles de journaux)
(Dolan *et al.*, 2004)

The Hartford Courant reported %%day%% that Tony Bryant said two friends were the killers.
A lawyer for Skakel says there is a claim that the murder was carried out by two friends of one of Skakel's school classmates, Tony Bryan.

Corpus monolingues parallèles (traductions multiples de romans)
(Barzilay et McKeown, 2001)

Emma burst into tears and he tried to comfort her, saying things to make her smile.
Emma cried, and he tried to console her, adorning his words with puns.

Corpus parallèles alignés manuellement (questions extraites de sites communautaires)
(Bernhard et Gurevych, 2008)

How many ounces are there in a pound ?
What's the number of ounces per pound ?

Modifications extraites de l'historique de révision (Wikipédia)
(Dutrey *et al.*, 2011)

Ce vers de Nuit rhénane d'Apollinaire qui paraît presque sans structure rythmique
Ce vers de Nuit rhénane d'Apollinaire dont la césure est comme masquée

Paraphrases générées automatiquement (corpus parallèles bilingues)
(Madnani *et al.*, 2008)

(hong kong, macau and taiwan) macau passed legalization to avoid double tax.
macao adopted bills to avoidance of double taxation (hong kong, macao and taiwan)

Paraphrases générées automatiquement (systèmes de traduction automatique multiple)
(Zhao *et al.*, 2010)

he said there will be major cuts in the salaries of high-level civil servants .
he said there are significant cuts in the salaries of high-level officials .

Figure 1. *Exemples extraits de sources et techniques représentatives pour la collecte de paraphrases d'énoncés*

3. Acquisition de paraphrases sous-phrastiques : état de l'art

Les approches présentées dans la section précédente portent principalement sur l'acquisition de paraphrases d'énoncés. Or, il est également utile d'extraire des reformulations pour des unités de texte plus fines : de telles paraphrases sous-phrastiques peuvent être exploitées d'une manière plus générale par les techniques d'identification et de génération de paraphrases. Dans cette section, nous décrivons les principales approches essayées pour l'acquisition de paraphrases sous-phrastiques ; un panorama plus large de ces techniques se trouve dans (Madnani et Dorr, 2010).

Deux mots ou, par extension, deux segments, apparaissant dans des contextes similaires, peuvent être interchangeables : c'est l'*hypothèse de distributionnalité*, introduite par Harris, qui a été adoptée dans plusieurs travaux d'acquisition de paraphrases. Lin et Pantel (2001) l'ont, par exemple, appliquée à des chemins dans des arbres de dépendance pour la découverte de règles d'inférence. Les résultats obtenus prennent la forme de patrons d'équivalences avec deux arguments tels que : {*X asks for Y*, *X requests Y*, *X's request for Y*, *X wants Y*, *Y is requested by X*...}.

Des corpus comparables monolingues, dans lesquels un même contenu est probablement décrit sous plusieurs formes, permettent de guider la mise en correspondance d'équivalences locales. Ainsi, par exemple, Barzilay et Lee (2003) introduisent une technique d'alignement multiséquence factorisant des phrases structurellement similaires de ces corpus sous forme de graphes, qui contiennent par nature des équivalences locales. Les travaux de Bhagat et Ravichandran (2008) s'inscrivent dans ce même cadre mais à une plus large échelle. Limiter les corpus utilisés à des textes comparables, sélectionnés sur la base d'un genre ou de thèmes communs, permet d'augmenter la probabilité que les correspondances obtenues seront effectivement valides grâce aux contextes plus restreints.

Outre les corpus monolingues, des corpus multilingues parallèles ont été exploités pour l'extraction de paraphrases en se fondant sur l'hypothèse que des segments partageant des traductions dans une autre langue peuvent être des paraphrases dans certains contextes. Bannard et Callison-Burch (2005) ont décrit une approche par pivot exploitant plusieurs corpus parallèles. Zhao *et al.* (2008b) ont proposé une extension de cette approche permettant d'extraire des patrons à partir de tels corpus multilingues. Callison-Burch (2008) et Max (2009) utilise des traductions de segments en pivot prenant en compte une modélisation du contexte de la phrase d'origine.

Une autre approche possible consiste à décrire des règles d'identification de paraphrases locales exploitant des variations terminologiques. Ces dernières sont fréquemment utilisées en recherche d'information pour l'expansion de requêtes par des formes entretenant une même relation syntaxique (Sparck-Jones et Tait, 1984) ou encore pour l'identification de variantes de termes, dans un domaine de spécialité (Jacquemin, 1999). L'approche de Jacquemin (1999), repose sur des métarègles de réécritures morphosyntaxiques écrites manuellement ainsi que sur des ressources énumérant des variantes morphologiques (mots ayant le même lemme ou liés par dérivation morphologique) et sémantiques (synonymes). Ces travaux peuvent être étendus pour l'extraction de variantes à partir de corpus monolingues.

L'acquisition de paraphrases sous-phrastiques peut également être exprimée sous la forme d'une tâche d'alignement de mots entre deux énoncés liés, lorsque de tels corpus monolingues parallèles sont disponibles (Cohn *et al.*, 2008). Les travaux précédents ont principalement exploité des traductions multiples. Barzilay et McKeown (2001) utilisent des informations contextuelles fondées sur des similarités lexicales pour extraire des paraphrases à partir de paires de traductions produite indépendamment. Pang *et al.* (2003) exploitent la structure syntaxique d'un ensemble de traductions pour acquérir des équivalences locales. Ibrahim *et al.* (2003) proposent une

méthode non supervisée consistant à extraire des fragments d'arbres syntaxiques sémantiquement équivalents.

Il est à noter que la plupart des approches proposées pour la tâche d'acquisition de paraphrases sous-phrastiques ont en commun d'être fortement liées aux types de ressources auxquels elles s'appliquent. En fait, dans la plupart des travaux mentionnés, le type de corpus utilisé a un impact direct sur les performances, mais aucune modélisation plus générale de ce que sont les paraphrases sous-phrastiques ne semble avoir clairement émergé, les liens entre ces différents travaux étant souvent difficiles à établir.

4. Expériences en acquisition de paraphrases sous-phrastiques

L'étude présentée dans cet article a pour objectif initial d'étudier les caractéristiques des paraphrases sous-phrastiques difficiles à obtenir par des techniques d'acquisition représentatives des approches proposées, afin de mettre en évidence les types de connaissances requises. Pour cela, il est nécessaire d'établir une distinction claire entre les paraphrases et les autres phénomènes de reformulation. Nous commençons par décrire les données et la méthodologie d'évaluation choisies pour effectuer nos expériences dans la section 4.1. Nous détaillons ensuite les approches d'acquisition implémentées dans la section 4.2, dont une évaluation des performances est donnée dans la section 4.3. Finalement, nous décrirons une méthode de validation des paraphrases proposées par ces différentes techniques fondée sur un apprentissage automatique exploitant divers modèles dans la section 4.4. Cette dernière approche permettra des gains significatifs relativement aux meilleures techniques individuelles.

4.1. *Cadre expérimental*

Nous suivons l'approche introduite par Cohn *et al.* (2008), dans laquelle un ensemble de paraphrases sous-phrastiques de référence est comparé aux paraphrases produites par une méthode évaluée. Une annotation de référence décrivant les alignements entre mots est supposée disponible pour un ensemble d'énoncés en relation de paraphrase. La performance d'une technique se décompose en des valeurs usuelles de *précision* et de *rappel*, définies respectivement comme la proportion des candidates proposées appartenant à la référence et la proportion des éléments de la référence proposés parmi les candidates, ainsi qu'en une *f-mesure* combinant les deux avec une importance égale. Les paraphrases *atomiques* de référence sont notées $\mathcal{R}_{\text{atom}}$, et des paraphrases *composites*, notées $\mathcal{R}$, ajoutent à l'ensemble précédent les paraphrases obtenues par concaténation de paraphrases atomiques adjacentes. En notant respectivement $\mathcal{H}_{\text{atom}}$ et $\mathcal{H}$ les hypothèses atomiques et composites proposées par une technique, les mesures d'évaluation sont définies de la manière suivante :

$$p = \frac{|\mathcal{H}_{\text{atom}} \cap \mathcal{R}|}{|\mathcal{H}_{\text{atom}}|} \quad r = \frac{|\mathcal{H} \cap \mathcal{R}_{\text{atom}}|}{|\mathcal{R}_{\text{atom}}|} \quad f_1 = \frac{2pr}{p+r} \qquad [1]$$

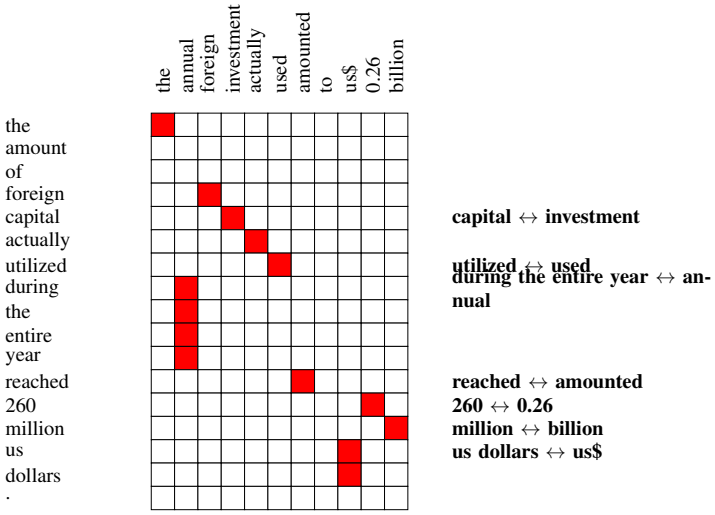

Figure 2. *Alignements de référence pour une paire de paraphrases en anglais extraite à partir du corpus de référence de Cohn* et al. *(2008) et liste des paraphrases atomiques extraites à partir de ces alignements*

L'exemple donné dans la figure 2 illustre différents cas qui expliquent la complexité de la tâche d'alignement au niveau des mots entre deux énoncés en relation de paraphrase. Les alignements présentés ont été obtenus par annotation manuelle réalisée par des locuteurs natifs auxquels un guide détaillé d'annotation avait été fourni. Bien que Cohn *et al.* (2008) annoncent un accord entre annotateurs acceptable pour cette tâche, de nombreux défauts apparaissent. Dans la figure 2, la paire de paraphrases *reached*↔*amounted* pourrait être corrigée par *reached*↔*amounted to*. De même, *260* ↔ *0.26* et *million* ↔ *billion* ne peuvent pas être considérés comme paraphrases : seule la paire *260 million* ↔ *0.26 billion* est acceptable.

Nos expériences ont été menées en deux langues, l'anglais et le français. Pour chaque langue, nous avons constitué un corpus contenant 150 paires de paraphrase, utilisées pour le développement et le réglage des paramètres. Les techniques sont évaluées sur un corpus de test constitué de 375 paires d'énoncés. Pour l'anglais, nous avons utilisé le corpus MTC, décrit dans (Cohn *et al.*, 2008), qui contient des traductions multiples en anglais depuis le chinois. Pour le français, nous avons utilisé une partie du corpus de référence de traduction en français à partir de l'anglais construites dans le cadre de la campagne CESTA [5]. L'alignement de référence a été réalisé en suivant la même procédure que pour l'anglais. Un sous-corpus a été annoté in-

5. Corpus de la Campagne d'Evaluation de Systèmes de Traduction Automatique : `http://www.elda.org/article125.html`

dépendamment par deux annotateurs natifs, révélant un taux d'accord interannotateur global de 88,96 % pour toutes les paraphrases, ou de 67,35 % si l'on ne considère pas les paraphrases « identité ». Les annotations ont été réalisées à l'aide de l'outil YAWAT (Germann, 2008).

4.2. *Acquisition de paraphrases sous-phrastiques : approches suivies*

Nous avons implémenté dans ce travail cinq techniques proposées dans la littérature ayant été initialement développées pour des besoins différents. Ces techniques ont été choisies parce qu'elles opèrent à des niveaux différents, reposent sur des principes différents et exploitent des ressources distinctes, ce qui notamment devrait permettre de tirer parti de leur complémentarité potentielle. La première, utilisée en traduction automatique statistique, est fondée sur l'apprentissage statistique d'alignement entre mots (GIZA), et requiert donc des données d'apprentissage en quantité relativement importante. La seconde, développée initialement dans le domaine de l'acquisition terminologique, exploite des règles de description de variantes de termes et des connaissances *a priori* sur la variation lexicale (FASTR). La troisième utilise la structure syntaxique des énoncés pour mettre en correspondance des segments (SYNT), et requiert par conséquent un analyseur syntaxique. La quatrième, introduite à l'origine pour l'évaluation des systèmes de traduction automatique, calcule une transformation au niveau des mots [6] en mettant en jeu des opérations d'édition dont le coût est appris automatiquement (TER_p). La cinquième, enfin, exploite des équivalences de traduction obtenues *via* une langue pivot (PIVOT).

4.2.1. *Approche fondée sur l'apprentissage d'alignement entre mots (*GIZA*)*

L'aligneur au niveau des mots GIZA++ (Och et Ney, 2003) estime des modèles de complexité croissante à partir de corpus parallèles. Bien que cet outil soit conçu à l'origine pour la tâche d'alignement bilingue entre mots pour la traduction automatique statistique, rien n'empêche sont utilisation dans un cadre monolingue. Une telle technique requiert typiquement des quantités de données importantes pour apprendre des alignements fiables. Afin d'améliorer ses capacités d'alignement, nous avons mis à disposition de GIZA toutes les paires de paraphrases possibles (pour des groupes constitués de quatre paraphrases) obtenues par des traductions multiples. Ceci constitue donc un avantage pour cette technique, car les autres techniques n'exploiteront pas l'information provenant d'autres paires d'énoncés pour construire leurs alignements.

À partir des matrices d'alignements obtenues au niveau des mots, nous appliquons les heuristiques d'extraction de bisegments du système de traduction automatique MOSES (Koehn *et al.*, 2007). La figure 3 présente un exemple de matrice d'alignement produite par GIZA : dans cet exemple, douze paraphrases différentes sont ainsi extraites.

6. Transformer une séquence de mots en une autre.

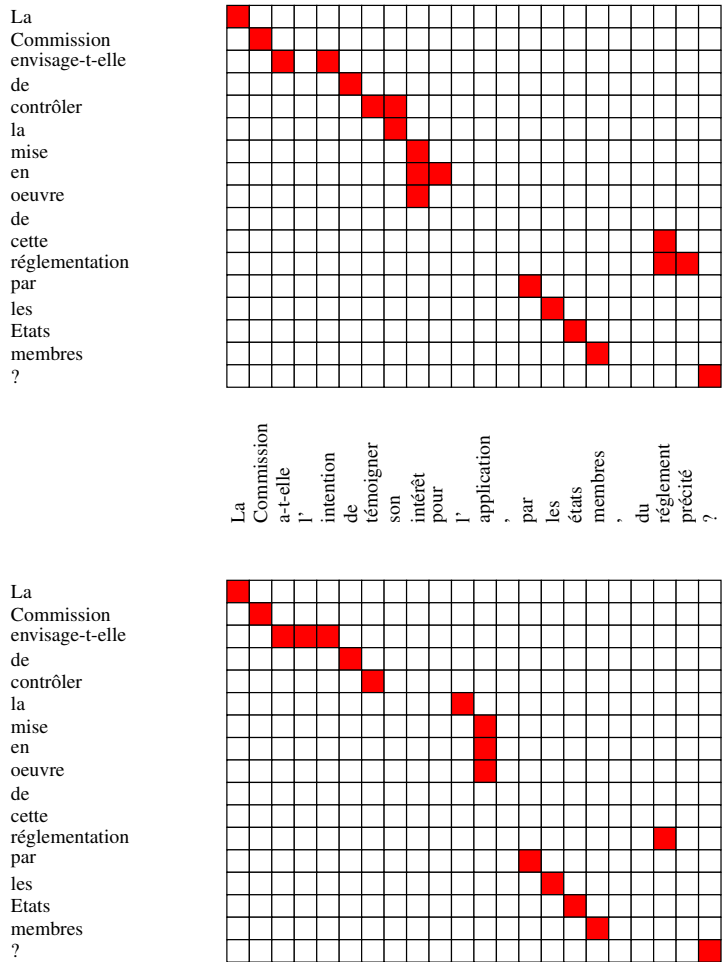

Figure 3. *Matrice d'alignement pour une paire d'énoncés en relation de paraphrase produite par la technique* GIZA *(partie supérieure), et matrice correspondante dans la base de référence (partie inférieure)*

4.2.2. *Approche fondée sur l'expression symbolique de la variation (*FASTR*)*

Dans le domaine de l'acquisition terminologique, les termes d'un domaine peuvent connaître des variations linguistiques importantes, dont le repérage automatique a été l'objet de nombreux travaux. Le système FASTR (Jacquemin, 1994) effectue une opération d'indexation contrôlée, qui permet de repérer des variantes en se fondant sur des patrons de réécriture morphosyntaxiques ainsi que des variations lexicales.

Nous utilisons cet outil de la manière suivante : considérant une paire de paraphrases d'énoncés, nous recherchons avec FASTR dans la première phrase (notre « corpus ») des variantes pour chacun des segments possibles de l'autre phrase (à concurrence d'une certaine taille), puis nous inversons la recherche et retenons l'intersection des résultats. L'usage que nous faisons du moteur de détection de variantes de termes semble *a priori* favorable à l'obtention d'une bonne précision. À l'inverse, les règles définies pour le repérage de variantes de termes ne sont pas nécessairement les mieux adaptées pour assurer une bonne couverture des phénomènes paraphrastiques entre segments de nature quelconque, comme démontré dans une étude précédente (Dutrey *et al.*, 2011).

La figure 4 illustre un exemple de métarègle (écrite manuellement) de FASTR. Cette métarègle, nommée « NAtoVASyn », porte sur un segment originel formé par un nom suivi d'un adjectif, qui peut être réécrit en un segment formé au minimum d'un verbe suivi d'un nom et d'un adjectif. La réécriture n'est autorisée que si le nom et le verbe appartiennent à la même famille morphologique (attribut `root`) et que les deux adjectifs sont connus comme synonymes (attribut `syn`). Une telle métarègle permet par exemple de reconnaître le segment *protéger de façon permanente* comme une variante du terme *protection constante*.

```
Metarule NAtoVAsyn
  ( X1 -> N1 A1) = X1 -> V1 <ART? | PRON? | PREP?> N A2:
    <N1 root> = <V1 root>
    <A1 syn> = <A2 syn>
    <X1 metaLabel> = 'XX'.
```

Figure 4. *Exemple de métarègle de* FASTR

4.2.3. *Approche fondée sur l'alignement de structures syntaxiques (*SYNT*)*

La relation entre deux paraphrases d'énoncés peut être exprimée en terme de correspondances syntaxiques plus ou moins fines permettant de délimiter des paraphrases sous-phrastiques là où celles-ci partagent une même structure syntaxique. L'algorithme introduit par Pang *et al.* (2003) décrit une opération de *fusion syntaxique* qui prend en entrée une paire de paraphrases d'énoncés et fusionne leurs arbres de constituants quand leurs listes de catégories filles sont compatibles. L'objectif de leur algorithme est de construire des automates représentant plusieurs paraphrases. L'algorithme met également en jeu un mécanisme de *blocage lexical* qui vise à empêcher toute fusion si un mot plein, présent dans la descendance d'une catégorie fille du premier arbre fusionné, se retrouve dans la descendance d'une autre catégorie fille du second arbre.

Pour la technique SYNT nous avons réimplémenté l'algorithme originel de Pang *et al.* (2003) et avons amélioré sa robustesse et sa correction en ajoutant un mode de fusion flexible dans lequel les parties de la phrase non concernées par un blocage lexical sont tout de même fusionnées. De plus, étant donné que l'algorithme est très

dépendant de la qualité des analyses syntaxiques produites, nous avons également ajouté un mode exploitant les *k* meilleures analyses produites par un analyseur probabiliste. La combinaison retenue entre une analyse du premier énoncé et une analyse du second parmi les k^2 combinaisons possibles est celle qui minimise le nombre de nœuds dans le treillis obtenu avant réduction. Nous avons utilisé l'analyseur syntaxique probabiliste de Berkeley (Petrov *et al.*, 2006) appris sur le français pour produire les cinq meilleures analyses pour chaque énoncé, et nous avons effectué une recherche exhaustive de la meilleure fusion pour chaque paire d'énoncés.

Un exemple de treillis obtenu par application de SYNT est donné dans la figure 5 : ont été fusionnées les trois phrases commençant par *La BCE veut conserver l'inflation sous la barre des...*, *La BCE veut garder l'inflation sous la barre des...* et *La Banque Centrale Européenne veut maintenir l'inflation sous la barres des...* Un parcours des chemins possibles dans le treillis obtenu permet d'extraire les paraphrases suivantes pour l'extrait donné : *BCE* ↔ *Banque Centrale Européenne*, *maintenir* ↔ *garder* ↔ *conserver*. Tout comme FASTR, cette technique semble *a priori* plus adaptée à l'extraction précise de paraphrases, mais contrairement à FASTR il est attendu qu'elle ne parvienne pas à extraire de correspondances lorsque les structures syntaxiques de haut niveau des paraphrases d'énoncés ne sont pas compatibles.

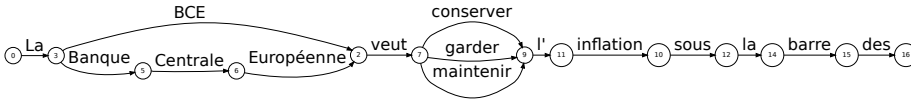

Figure 5. *Exemple d'une partie de treillis obtenue par application de* SYNT *sur trois énoncés en relation de paraphrase*

4.2.4. *Approche fondée sur une distance d'édition sur des séquences de mots (*TER_p*)*

La mesure TER_p (*Translation Edit Rate plus*) (Snover *et al.*, 2009) est une mesure permettant de calculer une distance d'édition entre une hypothèse de traduction et une traduction de référence. Elle a été originellement développée comme mesure en traduction automatique. Ses opérations de transformation de séquences de mots sont l'insertion, la suppression et la substitution de mots, ainsi que le déplacement et la substitution de segments. Le coût de chaque opération est pondéré par au minimum un poids optimisé par un algorithme de *hill climbing* [7] Les substitutions de mots ou de segments, qui sont optionnelles, peuvent exploiter des listes fournies à l'algorithme [8], et les substitutions de segments ont une probabilité associée.

7. Nous avons réimplémenté l'algorithme itératif utilisé dans les optimisations décrites par Snover *et al.* (2010). Nous avons utilisé dans nos expériences une première itération avec des poids uniformes puis cent itérations avec des poids aléatoires.

8. La version standard de TER_p implémente des techniques de racinisation ainsi que des ressources de synonymie et de paraphrases, mais pour l'anglais uniquement : nous ne les avons donc pas utilisées.

Pour nos expériences, nous avons exploité la tâche d'alignement sous-phrastique de deux énoncés sous forme de la séquence d'éditions la moins coûteuse permettant de transformer l'une en l'autre. Par la suite, nous dénoterons $\text{TER}_{p\to P}$, $\text{TER}_{p\to R}$ et $\text{TER}_{p\to F_1}$, les variantes TER_p correspondant à des optimisations réalisées sur un corpus de développement maximisant respectivement la précision, le rappel et la F-mesure (voir la section 4.1) pour des annotations de référence. Un exemple de résultat d'alignement avec TER_p est donné dans la figure 6. Sont obtenues une substitution de segments (P), *ce dégrèvement* ↔ *cet allongement*, une substitution lexicale (S), *équivaut* ↔ *revient* et des paraphrases composites, *ce dégrèvement fiscal* ↔ *cet allongement fiscal*.

Reference	ce	dégrèvement	fiscal	équivaut
	P	P		S
Hyp After Shifts	cet	allègement	fiscal	revient

Figure 6. *Exemple d'un alignement obtenu par* TER_p *entre deux paraphrases d'énoncés*

4.2.5. *Approche fondée sur des équivalences de traduction par langue pivot (*PIVOT*)*

Les équivalences de traduction peuvent être exploitées pour déterminer si deux unités textuelles constituent des paraphrases sous-phrastiques. Bannard et Callison-Burch (2005) ont proposé une *probabilité de paraphrasage* entre deux segments seg_1 et seg_2 qui exploite l'existence d'un alignement commun avec un segment $pivot$ dans une autre langue :

$$P_{para}(seg_1, seg_2) = \sum_{Pivot} P_{trad}(pivot|seg_1)P_{trad}(seg_2|pivot) \quad [2]$$

Des tables de traduction de segments sont créées en utilisant ces probabilités entre segments. Un segment source seg_1 est tout d'abord traduit dans une langue pivot avant d'être traduit à nouveau dans la langue d'origine pour obtenir la paraphrase candidate seg_2.

Pour implémenter cette approche, nous avons utilisé le corpus parallèle multilingue des débats parlementaires européens EUROPARL en anglais et français, constitué d'environ 1,7 million de phrases parallèles : ceci nous permet d'utiliser la même ressource pour construire les paraphrases dans les deux langues, en utilisant chaque langue comme pivot pour l'autre langue. Nous utilisons le système de traduction automatique statistique MOSES (Koehn *et al.*, 2007) qui a recours à l'outil GIZA++ (Och et Ney, 2003) pour aligner les phrases au niveau des mots. Pour chaque paire de paraphrases d'énoncés, nous appliquons l'algorithme suivant : pour chaque segment, nous

construisons l'ensemble de ses paraphrases potentielles selon la technique décrite. Nous retenons éventuellement la paraphrase présente dans l'autre énoncé maximisant l'équation 2 si celle-ci obtient une valeur supérieure à un seuil fixé à 10^{-4}.

4.3. *Évaluation des techniques individuelles*

Nous avons évalué chacune des méthodes présentées ci-dessus sur nos données d'évaluation décrites dans la section 4. Nous détaillons dans le tableau 2 les résultats obtenus pour les cinq techniques décrites précédemment.

	GIZA	PIVOT	FASTR	SYNT	TER_p $\rightarrow p$	$\rightarrow r$	$\rightarrow f_1$
	Français						
p	28,99	29,53	52,48	**62,50**	31,35	30,26	31,43
r	**45,98**	26,66	8,59	8,65	44,22	44,60	44,10
f_1	35,56	28,02	14,77	15,20	**36,69**	36,05	**36,70**
	Anglais						
p	31,01	31,78	37,38	**52,17**	50,00	29,15	33,37
r	38,30	18,50	6,71	2,53	5,83	45,19	**45,37**
f_1	34,27	23,39	11,38	4,83	10,44	35,44	**38,46**

Tableau 2. *Résultats obtenus pour chaque technique d'acquisition de paraphrases individuelle sur le français (partie supérieure) et l'anglais (partie inférieure)*

Tout d'abord, nous constatons que toutes les techniques, à l'exception de TER_p, obtiennent de meilleures performances sur le corpus français. Cela peut s'expliquer par le fait que ce corpus, est obtenu par traductions multiples à partir d'une langue plus proche du français (l'anglais) que ne l'est le chinois de l'anglais. Il n'est donc pas surprenant qu'il soit plus facile d'aligner des énoncés plus similaires, ce qui est clairement illustré dans les résultats de l'aligneur statistique GIZA, qui obtient un avantage de 7,68 en rappel pour le français relativement à l'anglais.

La technique statistique d'alignement entre mots, GIZA, obtient ainsi un rappel beaucoup plus important que les deux techniques exploitant des informations linguistiques, FASTR et SYNT. Ces dernières se distinguent, en outre, par une précision relativement forte (62,50 pour SYNT et 52,48 pour FASTR sur le français). Il faut noter que ces techniques en particulier peuvent reconnaître de longs segments qui ne correspondront pas nécessairement à des paraphrases atomiques de la référence. Les résultats de la technique TER_p sont comparables à ceux de GIZA pour le français, avec des valeurs de précision et de rappel moyennes. TER_p obtient les meilleures valeurs de f-mesure, avec un avantage sur GIZA de + 1,14 pour le français et + 4,19 et l'anglais. Le fait que TER_p ait de meilleures performances pour l'anglais, avec un avantage de 1,76 en f-mesure, n'est pas contradictoire : la distance d'édition implémentée est en mesure d'aligner des mots et des segments assez distants, indépendamment de la syn-

taxe, et d'identifier des correspondances entre les mots proches restants. Étant donné que l'anglais est une langue relativement peu fléchie, les indices d'alignement entre deux énoncés peuvent être plus nombreux que pour le français.

PIVOT est au même niveau que GIZA en précision, mais cette méthode ne trouve que peu de paraphrases, avec une différence de – 19,80 et – 19,32 en rappel sur l'anglais et le français, respectivement. Ceci peut être dû en partie au seuil des scores de paraphrases utilisés ainsi qu'au corpus de paraphrases d'énoncés choisi. En effet, PIVOT extrait ses paraphrases candidates d'un corpus de débats parlementaires, alors que notre ensemble de test provient du domaine des dépêches d'actualité. On peut donc notamment observer le résultat de différences relatives aux domaines et donc la présence de nombreux bisegments « hors vocabulaire », en particulier pour les entités nommées qui sont largement utilisées dans le domaine journalistique.

La meilleure valeur de rappel obtenue est de 45,98 sur le français (GIZA) et de 45,37 pour l'anglais (TER_p). Ce résultat vient confirmer la complexité de notre tâche d'identification de paraphrases sous-phrastiques : nous présenterons dans la section 5.2 une typologie des paraphrases difficiles à acquérir automatiquement, qui s'avèrent représenter ici plus de la moitié de celles contenues dans la référence. Rappelons toutefois que dans nos mesures les paraphrases identité ne sont pas considérées, et que donc les mesures ne portent que sur les paraphrases lexicalement différentes. De plus, nous avons déjà souligné le fait que la constitution du corpus de référence est un processus difficile (Cohn *et al.*, 2008). Ces remarques fournissent à nos yeux une justification supplémentaire pour l'utilisation des corpus monolingues parallèles, en dépit de leur rareté, pour mener des études fines sur les paraphrases sous-phrastiques.

Étude de la complémentarité des techniques

Les techniques présentées précédemment, opèrent à des niveaux d'appréhension du texte différents et exploitent des ressources distinctes. Avant d'examiner une éventuelle combinaison de leurs sorties, il est intéressant d'étudier leur éventuelle complémentarité. Pour cela, nous examinerons l'apport d'une technique dans une combinaison d'autres techniques en terme de nombre de paraphrases correctes. Nous avons alors estimé complémentarité entre deux techniques à l'aide de la différence entre le rappel de leur union et le meilleur de leurs rappels individuels. Pour cela, nous avons donc utilisé la formule suivante entre un ensemble de candidats t_i extraits en utilisant une technique i, et l'ensemble t_j des paraphrases proposées par une autre technique j :

$$C(t_i, t_j) = \text{rappel}(t_i \cup t_j) - \max(\text{rappel}(t_i), \text{rappel}(t_j)) \quad [3]$$

Dans le tableau 3, nous détaillons les résultats obtenus sur l'ensemble de test pour les deux langues. Il apparaît que plusieurs paires de techniques sont assez fortement complémentaires. La plus forte valeur de complémentarité est obtenue, pour les deux langues, en combinant GIZA et TER_p, avec des gains supérieurs à 10 points sur les deux langues. SYNT et FASTR présentent elles aussi une forte complémentarité en

français malgré leur faible nombre de paraphrases communes. L'apport de chaque technique relativement à l'ensemble des autres techniques montre que GIZA apporte le plus grand nombre de paraphrases inédites pour le français et TER_p pour l'anglais.

	GIZA	PIVOT	FASTR	SYNT	$\text{TER}_{p\rightarrow R}$	*toutes les autres*
	Français					
GIZA	–	9,79	3,64	2,20	**10,73**	9,91
PIVOT	**9,79**	–	2,26	5,22	7,84	3,39
FASTR	3,64	2,26	–	**7,28**	3,01	0,19
SYNT	2,20	5,22	**7,28**	–	1,76	0,44
$\text{TER}_{p\rightarrow R}$	**10,73**	7,84	3,01	1,76	–	5,65
	Anglais					
GIZA	–	4,65	2,83	0,59	**10,31**	9,31
PIVOT	**4,65**	–	2,30	1,88	3,12	3,72
FASTR	**2,83**	2,30	–	2,42	1,71	0,53
SYNT	0,59	1,88	**2,42**	–	0,59	0,00
$\text{TER}_{p\rightarrow R}$	**10,31**	3,12	1,71	0,59	–	12,20

Tableau 3. *Valeurs de* complémentarité *pour un ensemble de test dans les deux langues mesurées par l'équation 3. Les valeurs de complémentarité sont calculées entre chaque paire de techniques individuelles, et pour chaque technique individuelle relativement à l'ensemble de toutes les autres techniques. Les valeurs données en gras indiquent les valeurs les plus élevées pour chaque technique.*

4.4. *Validation de paraphrases par classification automatique*

Dans un travail précédent (Bouamor *et al.*, 2011b), nous avons décrit deux types de combinaisons possibles des techniques d'acquisition individuelles. Dans la première, les résultats produits indépendamment par chaque technique sont combinés *a posteriori* par une union naïve. Dans la seconde, une technique particulière, TER_p, est adaptée en y intégrant les résultats des autres techniques. Nous avons évalué ces deux méthodes et avons obtenu des résultats encourageants, principalement en rappel, la prise en compte d'un grand nombre de paraphrases candidates augmentant mécaniquement la quantité de bruit en entrée. Nous avons mis en place un processus de validation des paraphrases, formulé comme une classification binaire (« paraphrase » Vs « pas paraphrase ») des paraphrases candidates produites par l'ensemble des techniques individuelles considérées.

Nous avons abordé ce problème avec une classification discriminante à maximum d'entropie MAXENT (Berger *et al.*, 1996) [9]. Un tel classifieur cherche à maximiser la

9. Nous avons utilisé l'implémentation disponible sur : `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`

probabilité conditionnelle $P(y|x)$ en faisant l'hypothèse qu'elle suit une loi exponentielle.

Cette tâche de classification permet d'inclure des traits qui n'étaient pas nécessairement pris en compte ou possibles à considérer. Plus généralement, ceci permet de tenter d'apprendre une caractérisation plus générique des paraphrases, qui pourrait s'adapter trivialement à un nombre quelconque par les techniques individuelles en entrée. Les traits que nous utilisons, calculés pour toutes les paires de paraphrases candidates indépendamment de la ou les techniques l'ayant proposée, sont les suivants :

1) **sources** : trait indiquant quelle technique ou combinaison de techniques a proposé une paire de paraphrases candidate donnée ;

2) **séquence de POS** : trait décrivant la séquence de catégories morphosyntaxiques (POS) pour une paire de paraphrases donnée. Nous avons utilisé les symboles préterminaux des arbres syntaxiques fournis par l'analyseur syntaxique utilisé dans SYNT ;

3) **similarité contextuelle** : trait indiquant le degré de similarité entre les contextes dans lesquels les paraphrases d'une paire donnée apparaissent. Pour cela, nous avons utilisé l'ensemble complet du corpus bilingue français-anglais disponible pour la dernière version de l'atelier de traduction automatique WMT [10]. Ce corpus comporte environ 30 millions de phrases parallèles : cela permet à nouveau de garantir que les mêmes ressources ont été utilisées pour les expériences menées sur les deux langues. Nous avons collecté toutes les occurrences des segments apparaissant dans une paire de paraphrases, puis nous avons construit des vecteurs de mots pleins en ne gardant que les mots du voisinage distant de moins de 10 formes du segment considéré. Nous calculons enfin le cosinus entre les vecteurs représentant les deux paraphrases, modélisant ainsi l'hypothèse classique de distributionnalité ;

4) **distance entre caractères** : distance d'édition de Levenshtein entre les deux paraphrases d'une paire donnée ;

5) **similarité entre racines** : trait indiquant si les racines des paraphrases d'une paire donnée sont identiques [11] ;

6) **identité d'ensemble de formes** : trait indiquant si les deux ensembles de formes représentant une paire de paraphrases donnée comportent les mêmes formes, éventuellement dans un ordre différent ;

7) **autres** : différents traits indiquant la position relative des deux paraphrases dans leurs phrases d'origine, si un segment se trouve dans la phrase contenant l'autre segment, ratio de longueur des paraphrases, etc.

Pour le développement des différents traits, nous avons utilisé le corpus comportant 125 paires d'énoncés utilisé dans l'optimisation des paramètres de TER_p. Pour

10. http://www.statmt.org/wmt11/translation-task.html

11. Nous avons utilisé une implémentation de l'outil de racinisation Snowball pour les deux langues, disponible sur : http://snowball.tartarus.org

l'apprentissage de notre classifieur, nous avons constitué un nouveau corpus comportant 150 paires de paraphrases d'énoncés, selon les mêmes principes que nos corpus de développement et test décrits précédemment. Nos exemples positifs sont l'ensemble des paraphrases proposées par au moins l'une des cinq techniques d'acquisition étudiées et qui appartiennent à la référence. Nous avons extrait le même nombre d'exemples négatifs à partir des paires de paraphrases proposées ne figurant pas dans la référence.

	GIZA	PIVOT	FASTR	SYNT	TER_p $\rightarrow p$	$\rightarrow r$	$\rightarrow f_1$	**Union**	**Validation**
	Français								
p	28,99	29,53	52,48	**62,50**	31,35	30,26	31,43	17,58	40,77
r	**45,98**	26,66	8,59	8,65	44,22	44,60	44,10	**63,36**	45,85
f_1	35,56	28,02	14,77	15,20	**36,69**	36,05	**36,70**	27,53	**43,16**
	Anglais								
p	31,01	31,78	37,38	**52,17**	50,00	29,15	33,37	21,44	50,51
r	38,30	18,50	6,71	2,53	5,83	45,19	**45,37**	**60,87**	41,19
f_1	34,27	23,39	11,38	4,83	10,44	35,44	**38,46**	31,71	**45,37**

Tableau 4. *Résultats obtenus pour les techniques individuelles ainsi que pour leur union et notre validation sur le français (partie supérieure) et l'anglais (partie inférieure)*

Nous détaillons dans le tableau 4 les résultats obtenus pour l'union naïve et notre validation [12]. Nous obtenons le meilleur résultat de cette étude en terme de f-mesure à l'aide de notre processus de validation. Pour le français, cette classification apporte des améliorations de + 6,46 relativement au résultat de la meilleure technique individuelle (TER_p) et de + 15,63 relativement à l'union de toutes les techniques individuelles. Pour l'anglais, un gain de + 6,91 en f-mesure est obtenu par rapport au résultat de TER_p et de + 13,66 par rapport à l'union. Sans surprise, les valeurs maximales pour le rappel sont obtenues par l'union : on constate ainsi que nos techniques individuelles étudiées permettent de trouver dans les deux langues plus de six paraphrases sur dix de la référence.

Bien que ces résultats soient satisfaisants, étant donné la complexité de notre tâche, une analyse plus poussée des faux positifs et des faux négatifs pourra éventuellement nous permettre d'améliorer les performances obtenues. Nous revenons sur cette possibilité dans la section 5.2 ainsi que dans nos conclusions. La section suivante propose deux premiers axes d'analyse : l'étude de la performance des différentes techniques en fonction du degré de comparabilité des énoncés (section 5.1), et la description des paraphrases n'ayant été trouvées par aucune des techniques considérées (section 5.2).

12. Nous avons également décidé de recopier les résultats des techniques individuelles pour faciliter l'analyse.

5. Analyse des résultats

5.1. *Performance en fonction du degré de comparabilité des énoncés*

Les paires de paraphrases d'énoncés que nous étudions dans ce travail posent des problèmes de difficultés variables. En effet, ces paraphrases peuvent être très proches et ne différer que de quelques mots. Certaines paraphrases peuvent, en revanche, avoir des structures syntaxiques très différentes et/ou peuvent être lexicalement très distantes. Il est ainsi instructif de considérer les performances des différentes méthodes testées en fonction de ces difficultés. La difficulté d'alignement pourrait se mesurer par un accord entre annotateurs au niveau de chaque phrase, mais nous avons choisi d'utiliser une mesure fondée sur une distance d'édition sur les mots, $(1 - TER(paraphrase_1, paraphrase_2))$, qui sera d'autant plus grande que les phrases seront proches. Les résultats obtenus pour l'ensemble des techniques étudiées sont présentés dans la figure 7.

Pour la précision, on constate tout d'abord que GIZA est très sensible à la difficulté telle que nous la définissons, et que les alignements que cette technique produit sont d'autant moins bons que les phrases sont différentes. De façon un peu plus surprenante, SYNT et $\text{TER}_{p \to P}$ ne semblent pas trop affectés par cette difficulté. Cependant, ceci est peut-être dû au fait que les valeurs des barres, pour chaque intervalle discrétisé, représentent une moyenne qui ne rend pas compte du nombre d'éléments. Il est possible que SYNT extrait peu de paraphrases sur des paires de phrases difficiles, mais que, lorsqu'elle parvient à trouver des structures syntaxiques compatibles, celles-ci permettent un alignement précis. Enfin, FASTR est insensible à cette difficulté, ce qui était attendu, puisque cette technique fonctionne avec des patrons morphosyntaxiques pouvant impliquer des mots différents. Nous déduisons de ces remarques que ces différentes techniques peuvent être utilisées à bon escient pour différents niveaux de parallélisme des corpus d'acquisition. Le rappel fait apparaître une tendance beaucoup plus marquée : GIZA, $\text{TER}_{p \to R}$ et SYNT extraient d'autant moins de paraphrases de la référence que les phrases sont difficiles. À nouveau, FASTR y semble insensible. PIVOT présente un comportement différent du reste des techniques : sa précision ne semble pas être très affectée par le degré de parallélisme des phrases, alors qu'elle présente un meilleur rappel sur les paraphrases d'énoncés les plus parallèles.

Cette analyse confirme l'hypothèse qu'il est préférable d'avoir des paraphrases d'énoncés les plus « parallèles » possibles pour obtenir une bonne performance en acquisition car plus les textes à aligner sont différents et plus il sera difficile d'identifier des paraphrases correctes. Un autre enseignement intéressant concerne la technique symbolique FASTR : celle-ci est particulièrement utile pour extraire des paraphrases sous-phrastiques précises dans des paraphrases d'énoncés de formes très différentes, ce qui reflète assez bien son usage originel.

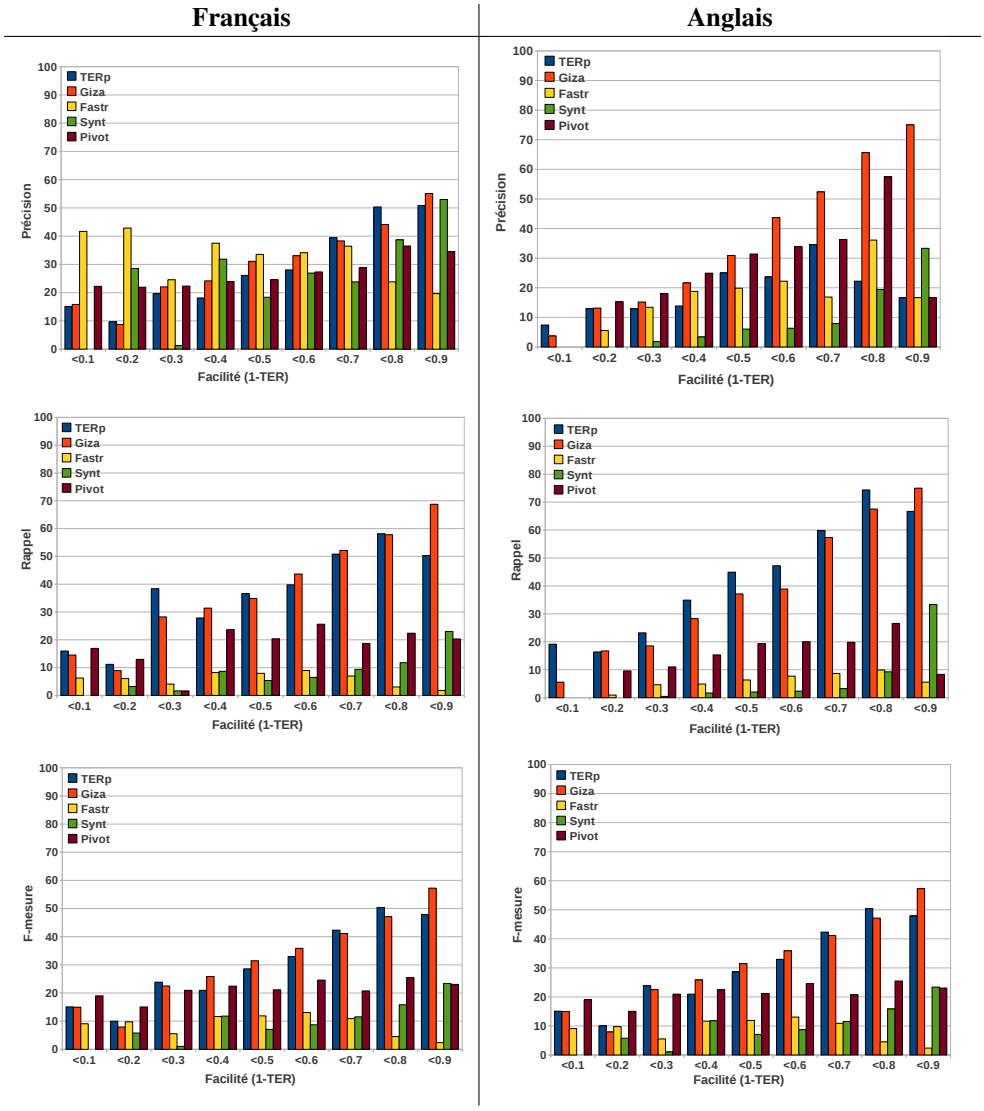

Figure 7. *Performance en précision, rappel et f-mesure des techniques étudiées pour les deux langues en fonction de la complexité des paires de paraphrases mesurée par la valeur de (1-TER). La valeur de chaque barre dans les intervalles discrétisés est une moyenne des éléments de cet intervalle, et ne rend pas compte du nombre de ces éléments. Pour la précision, une valeur de 0 peut indiquer soit l'absence de proposition pour les phrases de cet intervalle, soit des propositions toutes incorrectes.*

5.2. *Typologie des paraphrases difficiles à acquérir*

Un objectif important de cette étude consiste à caractériser les paraphrases qui sont difficiles à acquérir automatiquement par les techniques étudiées. Les rappels de l'ordre de 60 % obtenus par l'union naïve sur les deux langues nous semblent à la fois souligner la complexité de la tâche visée et indiquer la bonne performance relative obtenue par les techniques choisies. Ce dernier point confirme donc l'intérêt d'étudier les paraphrases de la référence qu'aucune des techniques n'a pu identifier.

Nous avons décidé de nous concentrer sur le cas où l'on considère les résultats de l'union de chacune des techniques, puisque, naturellement, c'est ainsi qu'est obtenu le meilleur rappel (voir le tableau 4). Nous avons extrait le sous-ensemble du corpus de référence contenant des paraphrases ne faisant pas partie des paraphrases obtenues par l'union des techniques. Ces paraphrases constituent le sous-ensemble de paraphrases de référence qu'aucune technique individuelle n'a pu capturer. Nous avons défini les grandes classes de paraphrases au moyen d'une annotation itérative effectuée par deux annotateurs. Nous décrivons la classification obtenue dans la figure 8 : chaque classe est illustrée par des exemples représentatifs dans les deux langues, et ont été ordonnées par ordre décroissant de leurs fréquences dans le corpus anglais.

Tout d'abord, nous remarquons qu'il existe une corrélation claire entre les deux langues, à quelques exceptions près. Par exemple, il existe plus d'équivalences lexicales et sous-phrastiques dans notre corpus français, mais moins de variations typographiques et morphologiques. L'équivalence lexicale et sous-phrastique est de loin la catégorie la plus représentée avec un peu moins d'un tiers de toutes les paraphrases difficiles à repérer dans le corpus anglais et $4/10$ de ces paraphrases dans le français.

Quelques autres paires impliquant des segments assez longs tel que *en train d'être réalisé à grands pas* ↔ *en cours* sont difficiles à identifier pour toutes les techniques à l'exception de SYNT dans le cas où elles apparaissent dans des structures syntaxiques compatibles. L'approche d'alignement statistique (GIZA) peut, en particulier, avoir des difficultés à capturer des équivalences entre des éléments rares dans une paire de paraphrases comme *ignore* ↔ *be blind to*. Quelques autres exemples, tel que *bonne* ↔ *appropriée*, auraient pu être capturés si l'on dispose de suffisamment de connaissances lexicales *a priori*. Ces connaissances pourraient provenir de dictionnaires ou de ressources lexico-sémantiques (tel que Wordnet pour l'anglais) et être intégrées dans la liste des synonymes exploitée par FASTR.

La classe des variations pragmatiques, représentant 8,67 % en anglais et 6,97 % en français, correspond aux segments qui ne sont des paraphrases que dans certains contextes très spécifiques et qui sont, par conséquent, difficiles à capturer sans passer par une analyse profonde (*to south korea* ↔ *home*). Il n'est donc pas surprenant que ces phénomènes, résultant des différents choix faits par les traducteurs humains impliqués, soient difficiles à acquérir automatiquement, bien que des techniques comme TER_p ou SYNT soient capables de les capturer dans certaines conditions. Néanmoins, il peut s'avérer difficile de réutiliser ces paraphrases dans des applications concrètes, ce qui limite par conséquent leur intérêt à l'étude des phénomènes de paraphrase. Il

Classes	Exemples	# ang.	% ang.	# fr.	% fr.
Équivalences lexicales et sous-phrastiques	businesses ↔ entreprises at a rapid rate ↔ fast maintenant ↔ à présent ignore ↔ be blind to	192	27,75	270	40,3
Inclusions	the hopewell group ↔ hopewell pfizer now is ↔ pfizer is centrale ↔ centrale thermique interne ↔ d'ordre interne	93	13,44	74	11,05
Variations typographiques	hong kong ↔ hongkong 11 ↔ eleven programme-cadre ↔ programme cadre UPU ↔ Union postale universelle	72	10,41	52	7,76
Variations morphosyntaxiques	british ↔ by Great Britain research of aids ↔ aids research postaux ↔ de postes environnementales ↔ relatifs à l'environnement	67	9,68	61	9,10
Variations pragmatiques	to south korea ↔ home plans ↔ was considering endémiques locaux ↔ sédentaires mettant en place ↔ soumettant	60	8,67	46	6,97
Variations morphologiques	to resign ↔ resigning iraqi ↔ iraq Réglementation ↔ Règlement le terme ↔ la terminologie	60	8,67	25	3,73
Variations syntaxiques	temperature on the surface ↔ surface temperature it is an urgent task ↔ has become urgent pour quel montant ↔ quel était le montant qui lui sont soumis ↔ lui ayant été soumis	47	6,79	35	5,22
Anaphores	Pinochet ↔ he Somalie ↔ pays	15	2,17	16	2,39
Autres catégories	in the ↔ of at ↔ in du ↔ d'un	86	12,43	91	13,58
Total		692	100	670	100

Figure 8. *Classes et exemples de paraphrases sous-phrastiques difficiles à obtenir par les techniques automatiques étudiées*

est également à noter que certains cas correspondent à des choix difficiles, et parfois incorrects, faits par les annotateurs humains lors de la construction du corpus de référence, ou leur incapacité à suivre correctement les directives de la tâche complexe d'annotation.

6. Conclusion

Dans ce travail, nous nous sommes intéressés à la tâche d'acquisition de paraphrases sous-phrastiques à partir de corpus monolingues parallèles. Bien que ce type de corpus soit extrêmement rare, nous avons montré empiriquement qu'il représente un type de corpus approprié pour l'étude des phénomènes paraphrastiques et des al-

gorithmes d'acquisition de paraphrases. Nous avons défendu, en outre, le fait qu'une meilleure connaissance des paraphrases défiant les techniques automatiques est nécessaire pour l'amélioration des techniques automatiques et la réalisation de progrès significatifs dans la recherche sur les phénomènes de paraphrase. Nous avons pour cela proposé une exploration de différentes techniques pour la tâche d'acquisition de paraphrases sous-phrastiques : cinq techniques ont été choisies pour leurs ressources et algorithmes mis en œuvre et on été appliquées sur des corpus monolingues parallèles en deux langues. Nous avons présenté les résultats détaillés des performances de chaque technique individuelle. Nous avons ensuite proposé une méthode de validation de paraphrases en formulant ce problème sous la forme d'une classification automatique exploitant différents traits linguistiques et statistiques. Ceci nous a permis d'obtenir des gains significatifs relativement aux techniques individuelles : nous obtenons une amélioration relative de + 18 % environ en f-mesure sur les deux langues. Un résultat important de notre étude est l'identification des paraphrases qui défient les techniques automatiques employées.

Bien que nous ayons identifié un manque de ressources appropriées et suffisantes comme un frein pour les recherches abordant les phénomènes paraphrastiques, nous croyons que la disponibilité d'un grand nombre de corpus monolingues parallèles nous permettra déjà de poursuivre nos travaux dans plusieurs directions dans le but d'améliorer les résultats obtenus. En particulier, nous souhaitons extraire des paraphrases à partir de corpus de différents degrés de comparabilité et caractériser les contextes dans lesquels deux unités en relation de paraphrase peuvent être substituées.

7. Bibliographie

Bannard C., Callison-Burch C., « Paraphrasing with Bilingual Parallel Corpora », *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, États-Unis, p. 597-604, 2005.

Barzilay R., Elhadad N., « Sentence Alignment for Monolingual Comparable Corpora », *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japon, p. 25-32, 2003.

Barzilay R., Lee L., « Learning to Paraphrase : an Unsupervised Approach Using Multiple-Sequence Alignment », *Proceedings of NAACL-HLT*, Edmonton, Canada, p. 16-23, 2003.

Barzilay R., McKeown K. R., « Extracting Paraphrases from a Parallel Corpus », *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, p. 50-57, 2001.

Berger A. L., Pietra V. J. D., Pietra S. A. D., « A maximum entropy approach to natural language processing », *Computational Linguistics*, vol. 22, p. 39-71, 1996.

Bernhard D., Gurevych I., « Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites », *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, Columbus, États-Unis, p. 44-52, 2008.

Bhagat R., Ravichandran D., « Large Scale Acquisition of Paraphrases for Learning Surface Patterns », *Proceedings of ACL-08 : HLT*, Columbus, États-Unis, p. 674-682, 2008.

Bouamor H., Max A., Illouz G., Vilnat A., « Web-based Validation for Contextual Targeted Paraphrasing », *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, Portland, États-Unis, p. 10-19, 2011a.

Bouamor H., Max A., Vilnat A., « Comparison of Paraphrase Acquisition Techniques on Sentential Paraphrases », *Proceedings of IceTAL*, Rejkavik, Iceland, 2010.

Bouamor H., Max A., Vilnat A., « Monolingual Alignment by Edit Rate Computation on Sentential Paraphrase Pairs », *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT11)*, Portland, États-Unis, p. 395-400, 2011b.

Callison-Burch C., « Syntactic Constraints on Paraphrases Extracted from Parallel Corpora », *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Hawai, États-Unis, p. 196-205, 2008.

Cohn T., Callison-Burch C., Lapata M., « Constructing Corpora for Development and Evaluation of Paraphrase Systems », *Computational Lingustics*, vol. 34, n° 4, p. 597-614, 2008.

Dolan B., Quirk C., Brockett C., « Unsupervised Construction of Large Paraphrase Corpora : Exploiting Massively Parallel News Sources », *Proceedings of Coling 2004*, Geneva, Suisse, p. 350-356, 2004.

Dutrey C., Bouamor H., Bernhard D., Max A., « Local modifications and paraphrases in Wikipedia's revision history », *SEPLN Journal*, 2011.

Fuchs C., *Paraphrase et énonciation*, Ophrys, Paris, 1994.

Germann U., « Yawat : Yet Another Word Alignment Tool », *Proceedings of the ACL-08 : HLT Demo Session*, Columbus, États-Unis, p. 20-23, 2008.

Ibrahim A., Katz B., Lin J., « Extracting Structural Paraphrases from Aligned Monolingual Corpora », *Proceedings of the Second International Workshop on Paraphrasing*, Sapporo, Japon, p. 57-64, 2003.

Jacquemin C., « Recycling terms into a partial parser », *Proceedings of the fourth conference on Applied natural language processing*, Stuttgart, Germany, 1994.

Jacquemin C., « Syntagmatic and Paradigmatic Representations of Term Variation », *Proceedings of ACL*, College Park, États-Unis, 1999.

Kauchak D., Barzilay R., « Paraphrasing for Automatic Evaluation », *Proceedings of the Human Language Technology Conference of the NAACL*, New York, États-Unis, p. 455-462, 2006.

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., Herbst E., « Moses : Open Source Toolkit for Statistical Machine Translation », *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, République tchèque, p. 177-180, 2007.

Lin D., Pantel P., « Discovery of Inference Rules for Question Answering », *Natural Language Engineering*, vol. 7, n° 4, p. 343-360, 2001.

Madnani N., Dorr B. J., « Generating Phrasal and Sentential Paraphrases : A Survey of Data-Driven Methods », *Computational Linguistics*, vol. 36, n° 3, p. 341-387, 2010.

Madnani N., Resnik P., Dorr B., Schwartz R., « Are multiple reference translations necessary ? Investigating the value of paraphrased reference translations in parameter optimization », *Proceedings of AMTA*, Hawai'i, États-Unis, 2008.

Max A., « Sub-sentencial Paraphrasing by Contextual Pivot Translation », *Proceedings of the ACL Workshop on Applied Textual Inference*, Suntec, Singapour, p. 18-26, 2009.

Max A., Wisniewski G., « Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History », *Proceedings of LREC*, Valetta, Malte, 2010.

Nakov P., « Improved Statistical Machine Translation Using Monolingual Paraphrases », *Proceeding of the 18th European Conference on Artificial Intelligence (ECAI08)*, Patras, Grèce, p. 338-342, 2008.

Nelken R., Yamangil E., « Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms », *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence : An Evolving Synergy*, Chicago, États-Unis, 2008.

Och F. J., Ney H., « A Systematic Comparison of Various Statistical Alignment Models », *Computational Linguistics*, 2003.

Pang B., Knight K., Marcu D., « Syntax-based Alignement of Multiple Translations : Extracting Paraphrases and Generating New Sentences », *Proceedings of NAACL-HLT*, Edmonton, Canada, p. 102-109, 2003.

Petrov S., Barrett L., Thibaux R., Klein D., « Learning Accurate, Compact, and Interpretable Tree Annotation », *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australie, 2006.

Quirk C., Brockett C., Dolan W., « Monolingual Machine Translation for Paraphrase Generation », *Proceedings of EMNLP 2004*, Barcelone, Espagne, p. 142-149, 2004.

Ravichandran D., Hovy E., « Learning surface text patterns for a Question Answering System », *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, États-Unis, p. 41-47, 2002.

Resnik P., Buzek O., Hu C., Kronrod Y., Quinn A., Bederson B. B., « Improving Translation via Targeted Paraphrasing », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, Cambridge, États-Unis, p. 127-137, 2010.

Snover M., Madnani N., Dorr B. J., Schwartz R., « TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate », *Machine Translation*, 2010.

Snover M., Madnani N., Dorr B., Schwartz R., « Fluency, Adequacy, or HTER ? Exploring Different Human Judgments with a Tunable MT Metric », *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athènes, Grèce, 2009.

Sparck-Jones K., Tait J., « Automatic search term variant generation », *Journal of Documentation*, vol. 40, n° 1, p. 50-66, 1984.

Zhao S., Lan X., Liu T., Li S., « Application-driven Statistical Paraphrase Generation », *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*, Suntec, Singapour, 2009.

Zhao S., Niu C., Zhou M., Liu T., Li S., « Combining Multiple Resources to Improve SMT-based Paraphrasing Model », *Proceedings of ACL-08 : HLT*, Columbus, États-Unis, p. 1021-1029, 2008a.

Zhao S., Wang H., Lan X., Liu T., « Leveraging Multiple MT Engines for Paraphrase Generation », *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Pékin, Chine, p. 1326-1334, 2010.

Zhao S., Wang H., Liu T., Li S., « Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora », *Proceedings of ACL-08 : HLT*, Association for Computational Linguistics, Columbus, États-Unis, p. 780-788, June, 2008b.