TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Combinaison d'informations pour l'alignement monolingue

Houda Bouamor Aurélien Max Anne Vilnat
LIMSI-CNRS, Univ. Paris-Sud
Orsay, F-91403, France
{prénom.nom}@limsi.fr

Résumé. Dans cet article, nous décrivons une nouvelle méthode d'alignement automatique de paraphrases d'énoncés. Nous utilisons des méthodes développées précédemment afin de produire différentes approches hybrides (hybridations). Ces différentes méthodes permettent d'acquérir des équivalences textuelles à partir d'un corpus monolingue parallèle. L'hybridation combine des informations obtenues par diverses techniques : alignements statistiques, approche symbolique, fusion d'arbres syntaxiques et alignement basé sur des distances d'édition. Nous avons évalué l'ensemble de ces résultats et nous constatons une amélioration sur l'acquisition de paraphrases sous-phrastiques.

Abstract. In this paper, we detail a new method to automatic alignment of paraphrase of statements. We also use previously developed methods to produce different hybrid approaches. These methods allow the acquisition of textual equivalence from a parallel monolingual corpus. Hybridization combines information obtained by using advanced statistical alignments, symbolic approach, syntax tree based alignment and edit distances technique. We evaluated all these results and we see an improvement on the acquisition of sub-sentential paraphrases.

Mots-clés : Paraphrase sous-phrastique, corpus parallèle monolingue, hybridation.

Keywords: Phrasal paraphrase, monolingual parallel corpora, hybridization.

1 Introduction

Le traitement de corpus monolingues et multilingues constitue un champ d'investigation très animé dans le domaine du traitement automatique des langues. Ils sont souvent constitués d'unités de texte ayant des liens sémantiques forts, une information qui peut être exploitée pour acquérir des équivalences entre des mots ou des groupes de mots et construire des ressources linguistiques importantes pour diverses applications. Ces resssources peuvent être utilisées par la suite pour extraire des réponses à des questions (Duclaye *et al.*, 2003), par exemple, ou autoriser des formulations différentes en évaluation de la traduction automatique (Russo-Lassner .G & .P, 2005; Kauchak & Barzilay, 2006), ainsi qu'en génération, pour aider des auteurs à trouver des formulations plus adaptées (Max, 2008).

De nombreuses techniques ont été proposées pour l'acquisition de segments en relation de paraphrase. Ces techniques ont en commun d'être directement liées aux types de ressources sur lesquelles elles s'appliquent. Les plus nombreuses exploitent des corpus monolingues comparables disponibles en grandes quantités, et se fondent sur l'hypothèse que des unités linguistiques apparaissant de nombreuses fois dans des contextes similaires peuvent avoir la même signification. Restreindre 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.

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 engendrée par le coût de leur construction. Mais elles présentent des caractéristiques qui en font les candidates les plus naturelles pour l'étude de la paraphrase sous-phrastique : les phrases parallèles étant issues de la volonté d'exprimer la même idée, les équivalences apprises apparaissent comme beaucoup plus fiables que celles extraites indirectement via des textes comparables ou des équivalences de traduction. En outre, le contexte de ces équivalences peut être extrait de façon directe, ce qui est particulièrement important pour caractériser les

conditions de leur validité.

Ce travail porte sur l'acquisition de paraphrases sous-phrastiques depuis des corpus monolingues parallèles, et vise en particulier à extraire des paraphrases de qualité. Dans cet article, nous présentons DIST une nouvelle méthode symbolique optimisée pour l'alignement de bi-segments exploitant un corpus monolingue parallèle. Puis nous décrivons une approche hybride d'extraction de paraphrases sous-phrastiques par la combinaison d'informations issues de différentes techniques. Cet article est organisé comme suit : dans la section 2, nous passons en revue les travaux portant sur l'acquisition automatique de paraphrases puis nous détaillons, dans la section 3, le cadre expérimental de notre travail, l'approche suivie pour combiner des informations issues de différentes techniques et extraire des bi-segments à partir de corpus monolingues parallèles ainsi que les résultats obtenus. Nous terminerons par une description de nos prochains travaux (section 4).

2 Travaux précédents en acquisition de paraphrases

L'acquisition de paraphrases peut être réalisées à l'aide de diverses méthodologies. Langkilde & Knight (1998) se sont basés sur les connaissances sémantiques de WordNet (Miller, 1995) pour exploiter les relations de synonymie entre termes et les utiliser ensuite lors de la génération de paraphrases. Cependant, ces ressources ne sont pas nécessairement disponibles dans toutes les langues et ne comportent que des équivalences textuelles au niveau des mots. C'est la raison pour laquelle de nombreux autres travaux se sont basés sur des corpus monolingues et multilingues parallèles ou comparables.

La majorité des travaux menés sur des corpus monolingues parallèles se basent essentiellement sur l'hypothèse de distributionnalité (Harris, 1954), selon laquelle les mots apparaissant dans le même contexte tendent à avoir des sens similaires. Cette hypothèse a été appliquée, par exemple, à des chemins dans des arbres de dépendance pour la découverte de règles d'inférence à partir de textes (Lin & Pantel, 2001). Barzilay & McKeown (2001) utilisent des informations contextuelles basées sur des similarités lexicales pour extraire des paraphrases à partir d'un ensemble de corpus alignés. 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 rares et 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. Quirk *et al.* (2004) proposent une approche consistant à apprendre un système de traduction statistique sur un corpus monolingue de phrases alignées automatiquement à partir d'un corpus comparable qui opère par reformulations locales.

Outre les corpus monolingues, des corpus multilingues parallèles ont été exploités pour l'extraction des paraphrases en se basant sur l'hypothèse que des segments partageant des traductions dans une autre langue peuvent être des paraphrases dans certains contextes. Bannard & Callison-Burch (2005) ont décrit une approche par pivot exploitant plusieurs corpus parallèles. De la même manière, Max (2009) utilise des traductions de segments en pivot pour produire des reformulations et sélectionner parmi celles-ci celles qui sont préférées par différents types de modèles. La majorité de ces approches s'attaque au problème d'acquisition de paraphrases d'énoncés complets. Or, il est également intéressant de pouvoir extraire des reformulations pour des unités de texte plus petites à partir de plusieurs corpus quel que soit leur degré de parallélisme.

3 Combiner des informations pour l'alignement

Différentes approches peuvent être utilisées pour faire l'acquisition de paraphrases sous-phrastiques depuis des corpus monolingues parallèles (Bouamor *et al.*, 2010). Outre l'amélioration individuelle de ces techniques, il est possible de parvenir à une amélioration des performances obtenues en exploitant utilement les résultats de chacune. Dans cette section, nous commençons par décrire le cadre expérimental dans lequel s'ancre notre étude sur l'alignement monolingue dans des paires de paraphrases, puis nous présentons brièvement quatre techniques que nous utilisons pour cette tâche. Nous décrivons ensuite un cadre de combinaison des résultats qu'elles produisent et détaillons les résultats de nos expériences.

3.1 Cadre expérimental

Les paraphrases d'énoncés sont relativement rares à l'état naturel, car peu d'activités humaines en gardent la trace lorsqu'elles existent. En outre, certains types de réécritures, comme le résumé, altèrent de façon significative le contenu des textes. Des solutions pour l'acquisition de paraphrases ont cependant été proposées, par exemple à partir de corpus comparables (Dolan & Brockett, 2005) ou de traces d'éditions (Dutrey *et al.*, 2010), mais l'identification de ce qui constitue des paraphrases acceptables reste une difficulté majeure. Une solution plus directe consiste à faire produire de telles paraphrases par des humains dans le cadre naturel d'une traduction où une même phrase est traduite plusieurs fois indépendamment. Le corpus MultiTrad (Bouamor, 2010) a été construit selon ce principe en obtenant des traductions vers le français d'extraits du corpus des débats parlementaire européen.

Pour l'étude présentée ici, nous avons sélectionné un corpus de développement issu de MultiTrad constitué de 50 énoncés traduits 4 fois de l'anglais vers le français. Pour chaque groupe de quatre paraphrases, la paraphrase la plus similaire en moyenne aux autres paraphrases a été identifiée et associée aux trois autres. Cette similarité est calculée par une valeur moyenne d'édition mesurée par TER (*Translation Error Rate*) (Snover *et al.*, 2009). Les 150 paires de paraphrases obtenues ont alors été annotée au niveau des mots par 3 annotateurs à l'aide de YAWAT (Germann, 2008), un outil qui permet d'utiliser, au choix, une vue parallèle entre énoncés présentés sous forme de paragraphes ou de matrices d'alignement. Chaque paire a été annotée par un seul annotateur : Callison-Burch (2008) mentionne un accord inter-annotateur acceptable sur une telle tâche [1], mais l'ensemble des annotations a par la suite été vérifié par le même annotateur. À partir des matrices d'alignement produites, l'ensemble des bi-segments de référence est extrait en respectant la contrainte suivante : tous les mots du segment contenu dans la première paraphrase sont alignés avec au moins un mot du segment de la seconde paraphrase et ne sont alignés qu'avec des mots de ce segment, et réciproquement.

Pour évaluer la performance de nos techniques d'alignement monolingue, nous utilisons l'approche PARAMETRIC (Callison-Burch *et al.*, 2008), dans laquelle un ensemble de *bi-segments* (correspondant à des paires de paraphrases sous-phrastiques) de référence est comparé aux bi-segments produits par la méthode évaluée. La mesure PARAMETRIC se décompose en des valeurs usuelles de *précision* et de *rappel*, définies respectivement comme la proportion des candidats proposés appartenant à la référence et la proportion des éléments de la référence proposés, ainsi qu'en une *F-Mesure* combinant les deux à égalité. Notre évaluation portera sur un extrait du corpus de traductions multiples issus de la campagne CESTA [2] contenant 375 paires de paraphrases (comportant entre 15 et 25 mots) et obtenues par traduction de l'anglais vers le français. L'alignement de référence a été réalisé en suivant la même procédure que pour le corpus de développement avec 2 annotateurs. Notre étude a révélé un taux d'accord inter-annotateur global de 88,96% qu n'est plus, cependant, que de 67,35% lorsque les paraphrases "identité" ne sont pas prises en compte.

3.2 Techniques individuelles

Nous avons implémenté dans ce travail quatre techniques, développées pour des besoins différents. Nous les avons choisies parce qu'elles opèrent à différents niveaux ce qui devrait permettre de tirer parti de leur complémentarité potentielle. La première est fondée sur l'apprentissage statistique d'alignements entre mots (MOT), et requiert

1. Il faut cependant noter que les travaux de Callison-Burch (2008) portait sur des textes journalistiques en anglais et qu'un guide d'annotation avait été fourni aux annotateurs.
2. Corpus de la Campagne d'Evaluation de Systèmes de Traduction Automatique : `http://www.elda.org/article125.html`

donc des quantités de données d'apprentissage en nombre relativement important. La seconde exploite des règles de description de variantes de termes et des connaissances *a priori* sur la variation lexicale (TERME). La troisième utilise la structure syntaxique des énoncés pour mettre en correspondance des segments (SYNT), et requiert par conséquent un analyseur syntaxique. La quatrième, calcule une transformation au niveau des mots pour transformer une séquence de mots en une autre en mettant en jeu des opérations de transformation dont le coût est appris automatiquement (DIST). Une étude comparative des trois premières techniques a été faite dans (Bouamor *et al.*, 2010). Elle a, en particulier, mis en évidence des différences de performance notables sur deux types de corpus parallèles monolingues obtenus par traductions multiples à partir d'une même langue d'une part, et de plusieurs langues d'autre part. Dans cet article, une nouvelle technique est introduite et utilisée de façon originale, et une combinaison efficace sous forme d'adaptation de cette dernière technique est proposée.

3.2.1 Approche fondée sur l'apprentissage d'alignements entre mots (MOT)

La technique MOT consiste à apprendre des alignements entre mots en utilisant des modèles d'alignement statistique appliqués sur deux phrases parallèles, initialement conçus pour la tâche d'alignement bilingue entre mots en traduction automatique statistique. Une telle technique requiert typiquement des quantités de données importantes pour apprendre des alignements fiables [3]. Dans nos expériences, nous mettrons à disposition de MOT toutes les paires de paraphrases possibles (pour des groupes constitués de 4 paraphrases) afin d'améliorer ses capacités d'alignement, ce qui constitue pour elle un avantage car les autres techniques ne considèrent les paires de paraphrases qu'isolément (en d'autres termes, pour les autres techniques l'information acquise sur une paire de paraphrases n'est pas directement exploitée pour les alignements ultérieurs). Par ailleurs, ce type de technique fonctionne d'autant mieux que les phrases des corpus d'apprentissage utilisées sont *parallèles*, signifiant ici qu'un alignement mot à mot est facile à réaliser. Dans le cas bilingue, ce n'est évidemment pas le cas de langues très différentes, et dans le cas monolingue, nos expériences précédentes ont montré que MOT obtenait des résultats sensiblement meilleurs lorsque les paraphrases utilisées sont obtenues par traduction depuis une même langue.

Nous avons utilisé le programme GIZA++ (Och & Ney, 2003) pour réaliser l'alignement entre mots et les heuristiques du système de traduction statistique MOSES (Koehn *et al.*, 2007) pour extraire des bi-segments à partir des matrices d'alignement obtenues. Un exemple d'une matrice d'alignement produite par MOT est donné dans la figure 1. À partir de cette matrice, 12 bi-segments différents sont extraits en appliquant les critères décrits ci-dessus.

3.2.2 Approche fondée sur l'expression symbolique de la variation (TERME)

Pour chaque paire d'énoncés en relation de paraphrase, il est possible d'exprimer des règles régissant les variations syntagmatiques et paradigmatiques acceptables au niveau des segments. Les nombreux travaux qui ont porté sur les notions de *termes* et de *variantes de termes* offrent ainsi une solution assez directe à ce problème de mise en correspondance. L'approche symbolique TERME que nous utilisons exploite l'opération d'*indexation contrôlée* du système FASTR (Jacquemin, 1999) pour trouver les alignements sous-phrastiques possibles entre deux paraphrases d'une paire donnée. Cette opération définit les variations acceptables pour un terme par un système de métarègles décrivant ses réécritures morphosyntaxiques possibles. Les métarègles peuvent également mettre en jeu des relations lexicales définissant des variations morphologiques (mots d'une même famille morphologique) et sémantiques (synonymie). Ces ressources constituent donc des connaissances *a priori* utilisées par TERME qui ne sont pas accessibles aux autres techniques.

L'outil FASTR utilisé a été conçu pour rechercher efficacement des termes et leurs variantes dans de grands corpus de textes. Pour nos besoins, considérant une paire de paraphrases d'énoncés, nous recherchons 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 favorable à l'obtention d'une bonne précision. À l'inverse, les métarègles 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 (Dutrey *et al.*, 2010).

3. La technique développée par Lardilleux (2010) constitue une exception notable adaptée aux événements de basse fréquence, et sera naturellement considérée dans la suite de nos travaux.

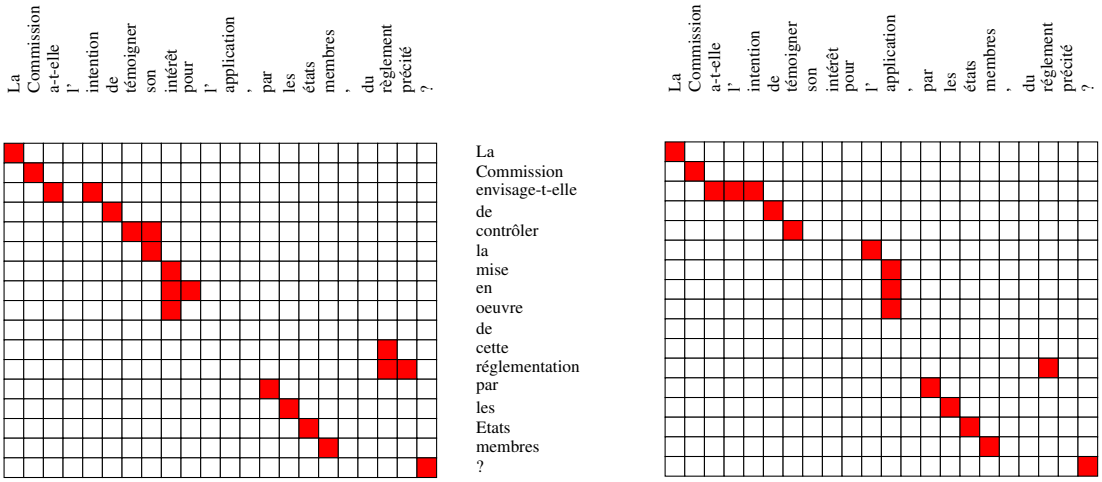

FIGURE 1 – Matrice d'alignement d'une paire de phrases dans MOT (à gauche), et sa matrice correspondante dans la base de référence.

3.2.3 Approche fondée sur l'alignement de structures syntaxiques (SYNT)

Lorsque deux énoncés en relation de paraphrase partagent une même structure syntaxique, il est possible de réaliser un alignement fin guidé par la syntaxe permettant de faire apparaître des correspondances sous-phrastiques fines. L'algorithme de Pang *et al.* (2003) décrit une *fusion syntaxique* consistant essentiellement à fusionner des arbres de constituants de deux énoncés là où les listes de catégories filles sont compatibles et qu'aucune évidence de non parallélisme syntaxique (via un mécanisme de *blocage lexical*) n'est détectée. La forêt d'arbres syntaxiques ainsi obtenue permet de construire un treillis de mots représentant des formulations alternatives qu'il est possible d'extraire par simple parcours du treillis.

Pour la méthode SYNT nous avons réimplémenté l'algorithme originel 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. Par ailleurs, 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 minimisant le nombre de nœuds dans le treillis obtenu avant réduction. Un exemple de treillis obtenu par application de SYNT est donné dans la figure 2.

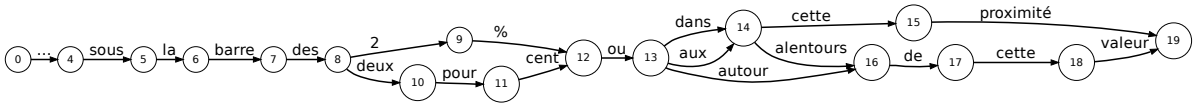

FIGURE 2 – Exemple d'un treillis obtenu par application de SYNT

Tout comme TERME, cette technique semble *a priori* plus adaptée à l'extraction précise de bi-segments monolingues, mais contrairement à TERME il est attendu qu'elle ne parvienne pas à extraire de correspondance lorsque les structures syntaxiques de haut niveau des paraphrases d'énoncés ne sont pas compatibles.

3.2.4 Approche fondée sur la distance d'éditions sur des séquences de mots (DIST)

Une relation entre deux paraphrases peut également s'exprimer sous forme de la séquence d'éditions la plus directe permettant de transformer l'une en l'autre. Une telle séquence d'éditions sur les mots est, par exemple,

implémentée dans la technique TERp (Translation Edit Rate plus) (Snover *et al.*, 2009), originellement développée pour le calcul d'une distance d'édition servant de mesure en traduction automatique pour évaluer une hypothèse de traduction relativement à une traduction de référence. Ce calcul met en jeu des opérations de transformation de chaîne incluant l'insertion, la suppression et la substitution de mots, ainsi que le déplacement et la substitution de segments. Chaque type d'opération est associé à une pondération optimisée sur un corpus de développement pour une mesure particulière, et l'algorithme effectue une recherche de la séquence d'opération la moins coûteuse. Les substitutions de mots ou segments sont optionnelles, mais peuvent exploiter des listes fournies à l'algorithme [4], et les substitutions de segments ont une probabilité associée.

Pour son calcul, TERp produit donc un alignement au niveau des mots entre deux énoncés. Pour nos besoins, nous avons implémenté une méthode DIST qui extrait l'ensemble des bi-segments (à concurrence d'une taille maximale) dérivables des alignements produits par TERp. Nous avons exploité la possibilité d'optimiser TERp pour nos besoins, en optimisant ses paramètres par la méthode du *hill climbing* [5]. Par la suite, nous dénoterons DIST_A l'optimisation originelle réalisée par Snover *et al.* (2009) pour l'évaluation de la traduction automatique (le « A » est pour « *adequacy* »). Les variantes DIST_P, DIST_R et DIST_{F_1} correspondent à des optimisations réalisées sur un corpus de développement maximisant respectivement la précision, le rappel et la F-mesure de PARAMETRIC exploitant des annotations de référence. L'ensemble de ces configurations n'utilisent pas de substitutions de segments, mais nous ferons appel à cette possibilité dans un cadre d'hybridation décrit plus loin. Un exemple de résultat d'alignement fourni par TERp est donné dans la figure 3.

Reference	**faisant**	**suite**	à	**la**	réponse	**donnée**	**le**	22	mai	1992	(| 1 |)	**.**	la
	S	**S**		**S**		**D**	**S**							**I**	
Hyp After Shifts	**en**	**complément**	à	**sa**	réponse		**du**	22	mai	1992	(| 1 |)	**,**	la

FIGURE 3 – Exemple d'un alignement résultat de DIST

3.2.5 Résultats expérimentaux et analyse

Nous avons évalué chacune des méthodes présentées ci-dessus sur le corpus de test décrit dans la section 3.1. Les techniques MOT, TERME et SYNT ont été utilisées telles que décrites. Pour DIST, nous avons exploité la possibilité d'optimiser la mesure selon nos propres objectifs. La variante DIST_A, évaluée pour référence, correspond à la version de TERp optimisée pour les besoins de l'évaluation de la traduction automatique. Les autres variantes DIST_P, DIST_R et DIST_{F_1} correspondent à TERp optimisée pour maximiser respectivement la précision, le rappel et la f-mesure de PARAMETRIC. La première partie de la table 1 donne les résultats obtenus sur les trois sous-mesures de PARAMETRIC. On constate tout d'abord que les résultats pour les 3 premières techniques sont cohérents avec ceux obtenus dans (Bouamor *et al.*, 2010). La seule différence notable est l'amélioration de la précision des deux techniques symboliques TERME et SYNT. La technique statistique d'alignement entre mots MOT obtient un rappel beaucoup plus important que les deux autres techniques qui se distinguent par une précision relativement forte (60,87 pour TERME et 66,96 pour SYNT). La précision de MOT reste toutefois dans une zone raisonnable à 47,02. Comme expliqué précedemment, MOT tire avantage des 3 paires de paraphrases sur lesquelles il peut réaliser son apprentissage, alors que les deux autres techniques, telles qu'implémentées, ne peuvent améliorer l'alignement à l'intérieur d'une phrase en exploitant des informations dérivées d'autres phrases.

L'ajout original pour notre tâche de DIST, technique fondée sur une distance d'édition sur des séquences de mots, révèle de nouveaux résultats intéressants. Tout d'abord, on constate qu'au niveau de la f-mesure, il n'existe qu'une faible différence entre DIST_A et la variante optimisée sur la f-mesure, DIST_{F_1}. Ceci signifie que nos objectifs sont très similaires à ceux de l'évaluation en traduction automatique tels que décrits par (Snover *et al.*, 2009). On constate cependant que des optimisations spécifiques en faveur de la précision ou du rappel mènent ici à des gains très importants de +8,69 en précision et de +7,42 en rappel. Ces résultats montrent que la technique DIST peine à améliorer simultanément la précision et le rappel, même si celle-ci obtient globalement des performances très proches de la meilleure technique envisagée jusque-là, MOT, avec une précision légèrement meilleure et un rappel

4. La version standard de TERp fournit des techniques de racinisation ainsi que des ressources de synonymie ainsi que de paraphrases locales pour l'anglais. TERp utilise jusqu'à 11 paramètres.

5. La première itération d'optimisation se fait avec des poids uniformes, puis nous réalisons 10 itérations avec des valeurs initiales tirées aléatoirement afin de diminuer le risque d'utiliser un minimum local.

légèrement inférieur. Il est possible que les modèles mis en jeu pour le calcul de la distance d'édition ne soient pas suffisamment expressifs pour nos besoins, et qu'en particulier, la non prise en compte de critères linguistiques pour opérer des transformations de séquences de mots soit à mettre en cause.

	Précision	**Rappel**$_{/13532}$	**F**$_1$
Mot	47,02	61,42	53,26
Terme	60,87	4,19	7,85
Synt	66,96	13,11	21,92
Dist$_A$	49,85	54,14	51,91
Dist$_P$	58,54	2,68	5,13
Dist$_R$	39,48	61,56	48,11
Dist$_{F_1}$	49,03	56,21	52,37
union(Mot,Terme,Synt,Dist$_{F_1}$**)**	38,99	**73,55**	50,97
intersection(Mot,Dist$_{F_1}$**)**	**70,38**	32,31	44,29

TABLE 1 – Résultats obtenus pour chaque technique

La dernière partie de la table 1 donne les résultats obtenus en opérant une combinaison élémentaire des résultats visant à maximiser d'une part la précision, et d'autre part le rappel. L'union sur le résultat de l'ensemble des techniques obtient un maximum de valeur de rappel de 73,55 (+12,13 relativement à MOT), avec une précision légèrement affectée (-2,29 relativement à MOT). Par ailleurs, réaliser l'intersection entre les différentes techniques peut raisonnablement mener à une précision améliorée. Cependant, le peu de résultats produits par TERME et SYNT nous ont fait préférer une mesure sur l'intersection de MOT et DIST$_{F_1}$: nous obtenons alors une valeur maximale de précision de 70,38 (+23,36 relativement à MOT et +21,35 relativement à DIST$_{F_1}$). Ces résultats montrent bien la complémentarité qui existe entre ces différentes techniques, et servent donc ici de motivation pour la recherche d'un mode de combinaison plus efficace des informations issues de chaque technique.

3.3 Approche hybride d'extraction de paraphrases locales

3.3.1 Observations et motivations

Les expériences présentées dans la section 3.2.5 ont révélé que les différentes techniques ont des performances variées, ce qui permet aussi de faire l'hypothèse qu'il est possible d'opérer une combinaison efficace de leurs résultats. Nous faisons ici une synthèse des points forts et des limitations de chacune de ces techniques orientée par la recherche d'un mode de combinaison plus efficace :

– MOT : très sensible à la fréquence de ses observations de mots et de cooccurrences entre mots, cette technique peut être informée par des connaissances d'association *a priori*, qui peuvent par exemple être transmises sous forme de données d'apprentissage additionnelles. En outre, il est possible, avec des données d'entraînement annotées, d'améliorer les performances des alignements statistiques par apprentissage discriminant (Tomeh *et al.*, 2010).
– TERME : cette technique est spécialisée dans l'extraction d'un type de bi-segments contraints par des règles de réécriture et de variation lexicale. Les métarègles, qui ont été développées manuellement, sont assez précises et ne peuvent couvrir tous les phénomènes de paraphrase. Leur apprentissage automatique peut améliorer la couverture, mais au détriment de la précision. L'enrichissement automatique des familles morphologiques et sémantiques devrait également permettre d'augmenter le rappel.
– SYNT : cette technique est très sensible au degré de parallélisme des énoncés qui décide de la fusion de constituants syntaxiques. Nous avons déjà pris en compte la qualité des analyses syntaxiques en autorisant la fusion à opérer sur les *k*-meilleures analyses syntaxiques. Le blocage lexical empêche une fusion lorsqu'un mot présent dans le constituant d'une phrase est attesté dans un constituant non aligné de l'autre phrase. Il pourrait être amélioré par la connaissance *a priori* de paraphrases locales, ce qui, néanmoins, ne pourrait bénéficier qu'à la précision.
– DIST : cette technique transforme une séquence de mots en une autre en un coût minimal, en utilisant des pondérations optimisées pour les différentes opérations utilisées. L'algorithme manipule des segments qui n'ont pas nécessairement de motivation linguistique, ce qui peut mener à des transformations aberrantes. En outre, des

opérations d'insertion et de suppression peuvent être utilisées à tort lorsque des correspondance au niveau des mots ou des segments ne sont pas connues. Ainsi, si de telles correspondances peuvent être fournies à TERp, il est possible d'espérer diminuer le nombre d'opérations de transformation aberrantes et ainsi d'augmenter la performance.

3.3.2 Présentation de l'hybridation des méthodes

Dans la section précédente nous avons montré qu'il existait plusieurs voies pour améliorer la performance de l'alignement monolingue auquel nous nous intéressons à partir des techniques décrites. Sans considérer davantage, à ce stade, l'amélioration individuelle de chacune des techniques, nous pouvons décrire les deux grandes familles d'approches possibles pour l'hybridation de la manière suivante : 1) les résultats produits indépendamment par chaque technique sont combinés *a posteriori*, et 2) une technique est *adaptée* par la connaissance des résultats produits par les autres techniques.

Nous avons déjà montré le résultat de l'évaluation d'une approche élémentaire par combinaison *a posteriori* dans la section 3.2.5, illustrée sur la partie gauche de la figure 4, qui révèle que la précision et le rappel peuvent ainsi être facilement améliorés. Nous considérons désormais la seconde approche. D'après nos observations, DIST est un candidat assez naturel pour l'adaptation. En effet, la connaissance d'alignements au niveau des mots ou des segments peut diminuer le nombre d'opérations effectuées à tort. Il s'agit précisément de la motivation majeure pour l'évolution de TER à TERp (Snover *et al.*, 2009), liée à la possibilité d'utiliser une base de paraphrases locales connues *a priori* et ainsi d'être plus robustes quant aux hypothèses de traduction acceptées par le système lorsqu'elle ne correspondent pas exactement à une traduction de référence.

Au contraire de ce qui est fait dans TERp, nous n'utiliserons pas une base de connaissances externe, même si nous ne rejetons pas cette hypothèse pour de futures expériences, mais nous adaptons dynamiquement la base de paraphrases utilisées en fonction des hypothèses extraites par les autres techniques, à savoir des hypothèses nombreuses et relativement précises pour MOT, et peu nombreuses mais précises pour TERME et SYNT. De plus, comme nous l'avons déjà montré, ces techniques peuvent être complémentaires quant aux types de paraphrases locales qu'elles permettent d'identifier, ce qui rejoint nos intuitions initiales liées à la nature de chacune d'elles.

Cette approche est illustrée sur la partie droite de la figure 4. Les bi-segments obtenus par MOT, TERME et SYNT sont combinés pour construire une table de paraphrases utilisée ensuite par DIST, que nous pouvons optimiser en fonction d'un besoin particulier (précision, rappel ou f-mesure). On remarque ici une analogie assez directe avec d'autres scénarios de combinaisons d'informations en TAL : en traduction automatique, l'approche de la partie gauche de la figure 4 correspond à la définition classique de la combinaison de systèmes (Matusov *et al.*, 2009), alors que l'approche de la partie droite correspond à l'adaptation d'un système par des sources externes telles que d'autres systèmes de traduction (Crego *et al.*, 2010).

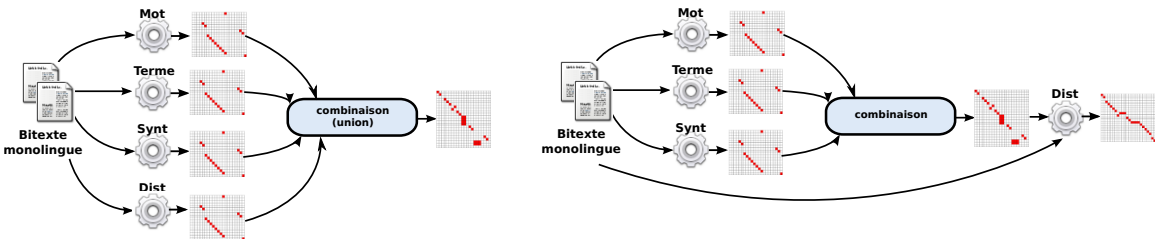

FIGURE 4 – Principales approches de combinaisons d'informations pour l'alignement multilingue. À gauche, plusieurs techniques produisent des résultats combinés pour produire une nouvelle sortie. À droite, un sous-ensemble des techniques fournissent leurs résultats à une dernière technique adaptée à l'exploitation de ces connaissances.

Un problème important à considérer concerne la manière dont la table de paraphrases utilisée par TERp est construite à partir des hypothèses produites par les différentes techniques. À ce stade de nos travaux, nous ne disposons pas de *mesures de confiance* données par chaque technique pour chacune de ses hypothèses, et nous sommes donc contraints de les considérer initialement comme équiprobables. De plus, pour assurer une comparaison plus directe avec la combinaison correspondant à la partie gauche de la figure 4, nous réalisons une combinaison simple

à base d'union : chaque hypothèse apparaissant au moins une fois parmi les hypothèses des différents systèmes est retenue et est associée à un poids constant uniforme. [6]

Un autre aspect important concerne là encore la pondération associée à chacune des paires de paraphrases *a priori* fournies à TERp. Considérons le cas où deux paraphrases sont fournies à TERp et où l'une est un sous-segment de l'autre : par exemple, (*ce degrévement* ↔ *cet allègement*) inclut (*dégrèvement* ↔ *allègement*). Si ces deux paraphrases sont fournies avec le même score à TERp, celui-ci préfèrera, dans de nombreux cas, utiliser la plus couvrante des deux, car cela minimisera souvent la quantité d'opérations de transformation restant à faire, et donc le coût global de transformation (voir partie gauche de la figure 5). Cela peut ne pas être un défaut en soi, car l'identification des plus longues sous-unités paraphrastiques peut être utile. Cependant, PARAMETRIC base ces mesures sur l'ensemble des bi-segments pouvant être extraits à partir d'alignement sur les mots. Ainsi, si dans l'exemple précédent l'alignement de référence inclut deux points d'alignement pour (*ce* ↔ *cet*) et (*dégrèvement* ↔ *allègement*), l'ensemble des bi-segments de référence sera constitué des deux bi-segments précédents et de leur combinaison ou « extension » (*ce degrévement* ↔ *cet allègement*). Si ce dernier est utilisé par TERp, il n'existe pas de moyen immédiat pour retrouver l'alignement sous-phrastique, et donc le rappel de la technique adaptée sera pénalisé.

Plusieurs solutions sont envisageables pour pallier ce problème. La pondération des paraphrases pourraient prendre en compte le nombre de mots/tokens couverts en favorisant les courts segments. Ne disposant néanmoins pas de solutions génériques applicables à toutes les techniques ni de moyen d'intégrer des scores de confiance motivés, nous préférons nous en remettre à une solution initiale plus simple, qui consiste à ne conserver que les sous-segments minimaux parmi l'union de ceux proposés par chacune des techniques. Ainsi, ne seront gardés pour construire la table de paraphrases utilisée par TERp que les bi-segments n'étant inclus dans aucun autre bi-segment, que nous appellons *bi-segments minimaux*.

FIGURE 5 – Exemple de deux alignements résultats de DIST_{F_1}, avec à gauche l'ensemble des bi-segments non filtrés, et à droite un ensemble de bi-segments minimaux

3.3.3 Résultats expérimentaux et analyse

Les résultats que nous obtenons en optimisant TERp sur les trois mesures de PARAMETRIC et en utilisant différentes sources de bi-segments sont présentés dans la table 2. Le résultat principal de ces expériences est la nouvelle f-mesure de 55,27 obtenue en optimisant sur cette mesure et en exploitant les bi-segments provenant des trois autres techniques. C'est la valeur la plus élevée sur l'ensemble de nos expériences, et elle correspond notamment à un gain de +4,3 par rapport à la combinaison par union des résultats de toutes les techniques, ou encore à un gain de +2,01 par rapport à MOT, la meilleure technique individuelle pour la f-mesure, et à un gain de +2,9 par rapport à DIST_{F_1}, la technique utilisée sans adaptation et optimisée selon le même critère. Ces résultats viennent confirmer notre hypothèse que TERp a pu ici tirer utilement profit des connaissances *a priori* qui lui ont été fournies.

Nous constatons de plus que des valeurs de précision et de rappel encourageantes peuvent être atteintes : une précision de 69,66 est obtenue en exploitant les prédictions de TERME et en optimisant sur la précision (+2,7 par rapport à la meilleure technique individuelle SYNT), et un rappel de 62,38 est obtenu en exploitant les prédictions de MOT en optimisant sur le rappel (+0.82 par rapport à la meilleure technique individuelle DIST_{R}).

Les cas de combinaisons où une seule technique est utilisée pour alimenter la base de paraphrases de TERp peuvent également être étudiés en comparant les valeurs des tables 1 et 2. Hormis les valeurs de rappel obtenues pour

6. Il serait bien sûr possible de pondérer *a priori* chaque hypothèse par le nombre de techniques l'ayant proposée, et/ou par la performance mesurée des techniques en question, dérivée par exemple de leur performance individuelle dans les différentes valeurs de PARAMETRIC. En outre, la contribution de chacune des techniques pourrait faire l'objet d'un paramètre optimisé simultanément aux paramètres de TERp. Toutes ces possibilités seront considérées dans notre travail futur.

Source de bi-segments	Critère d'optimisation								
	DIST$_P$			DIST$_R$			DIST$_{F_1}$		
	P	R$_{/13532}$	F$_1$	P	R$_{/13532}$	F$_1$	P	R$_{/13532}$	F$_1$
MOT	67,83	13,21	22,11	41,49	**62,38**	49,83	55,11	54,51	54,81
TERME	**69,66**	6,82	12,42	40,51	55,6	46,87	53,16	49,84	51,45
SYNT	68,08	8,11	14,48	29,99	56,84	39,26	51,25	50,14	50,69
comb(MOT, SYNT, TERME)	66,02	13,15	21,93	38,46	61,09	47,2	55,01	55,54	**55,27**

TABLE 2 – Résultats obtenus pour différentes optimisations et différentes sources de bi-segments. La fonction *comb* correspond à l'union avec pondération uniforme des bi-segments ne retenant que les bi-segments minimaux.

DIST$_R$ avec les paraphrases de TERME et SYNT, toutes les autres combinaisons de DIST avec les données d'une autre technique et optimisées selon un critère particulier améliorent la meilleure des deux valeurs précédentes. Par exemple, DIST$_P$ adapté avec les paraphrases de TERME obtient une précision de 69,66, qui est meilleure que celle de DIST$_P$ (58,54) et celle de TERME (60,87). Il est à noter qu'en combinaison de systèmes, comme c'est par exemple le cas en traduction automatique, des gains sont plus généralement obtenus lorsque un certain nombre de systèmes sont combinés. La complémentarité de nos sources d'information et l'impact assez immédiat d'une amélioration des informations *a priori* utilisées par TERp semblent donc ici avoir un rôle très bénéfique pour notre tâche.

Il est finalement instructif de considérer la performance des différentes configurations testées en fonction d'une certaine difficulté *a priori*. Celle-ci pourrait se mesurer par le degré d'accord inter-annotateurs pour chaque phrase, mais nous avons choisi d'utiliser un résultat en lien avec TERp : $(1 - TER(paraphrase_1, paraphrase_2))$, qui est donc d'autant plus grand que les phrases sont proches. Le résultat pour nos quatre techniques individuelles est présenté dans la figure 6. Pour la précision, on constate tout d'abord que MOT 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 DIST$_P$ 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, sont une moyenne qui ne rend pas compte du nombre d'éléments. Il est possible que SYNT extraie peu de bi-segments sur des paires de phrases difficiles, mais que lorsqu'elle parvient à trouver des structures syntaxiques compatibles, celles-ci permettent un alignement précis. Enfin, TERME est lui insensible à cette difficulté, ce qui était attendu puisqu'elle fonctionne sur de courts patrons morphosyntaxiques pouvant impliquer des mots différents. Nous déduisons donc 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 : MOT, DIST$_R$ et SYNT extraient d'autant moins de bi-segments de la référence que les phrases sont difficiles. À nouveau, TERME y semble insensible. On retiendra de cette analyse qu'il est préférable d'avoir des paraphrases d'énoncés les plus « parallèles » possibles pour obtenir une bonne performance en acquisition, mais que les techniques symboliques sont utiles pour extraire des paraphrases sous-phrastiques précises dans des paraphrases d'énoncés de formes très différentes.

4 Conclusion et travaux futurs

Dans cet article nous avons poursuivi deux objectifs. D'une part, nous avons présenté quatre méthodes d'acquisition de paraphrases sous-phrastiques à partir de corpus monolingues parallèles. Trois d'entre elles avaient déjà été évaluées, la dernière est nouvelle. Ces méthodes reposent sur des caractéristiques linguistiques différentes : MOT sur l'apprentissage statistique, TERME sur un approche symbolique de la variation de termes, SYNT sur des proximités syntaxiques et enfin DIST sur des distances d'édition. En évaluant ces méthodes, nous avons constaté qu'effectivement leurs résultats semblent complémentaires, ce qui nous a mené à notre second objectif, à savoir l'hybridation de ces méthodes. Plutôt que de combiner les résultats *a posteriori*, nous avons choisi d'utiliser les résultats de certaines méthodes comme données d'entrée d'une autre. Les résultats de cette approche ont confirmé notre hypothèse en montrant que la complémentarité de ces techniques donne un gain significatif.

De nombreuses pistes s'ouvrent à nous à la suite de ce travail. Nous souhaitons explorer toutes celles évoquées au cours de cet article. À court terme, nous comptons attribuer des scores de confiance à chacune des techniques afin

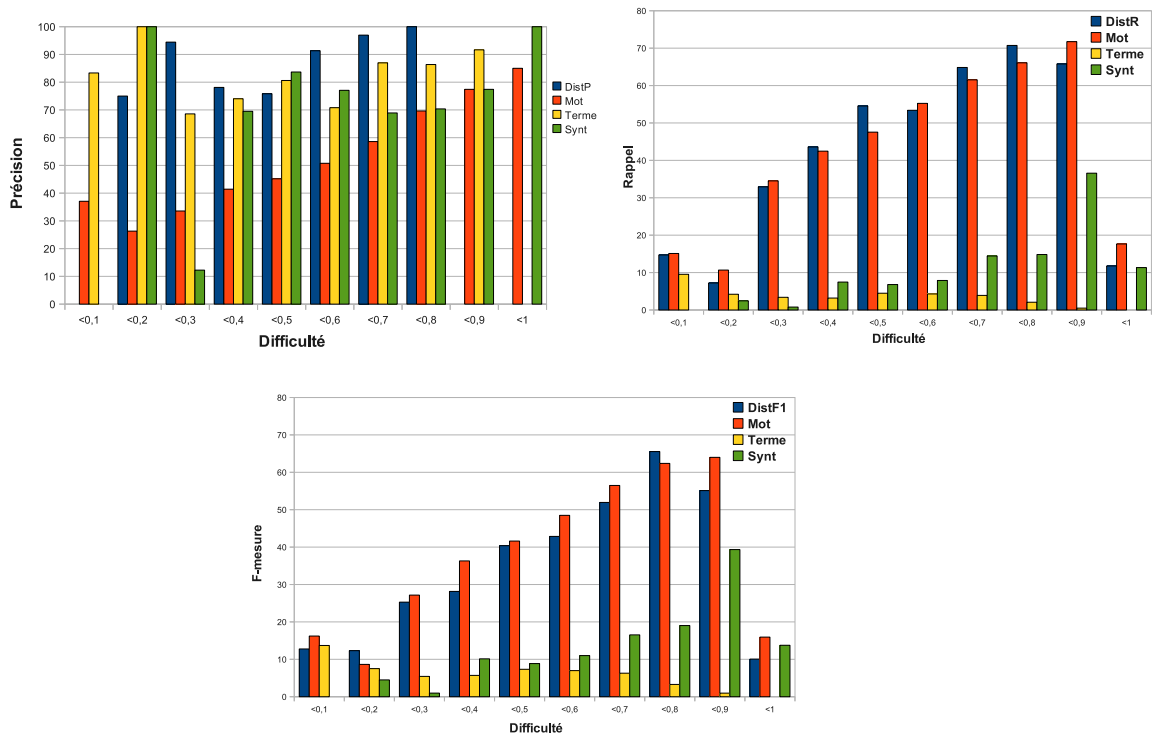

FIGURE 6 – Performance selon les différents critères de PARAMETRIC de différentes techniques. La valeur de chaque barre dans les intervalles discrétisés est une moyenne des éléments de cet intervalle, et ne rend pas compte du nombre de ces éléments. Pour la précision, une valeur de 0 peut indiquer soit l'absence de proposition pour les phrases de cet intervalle, soit de propositions toutes incorrectes.

de mieux tirer parti de leur complémentarité. Nous allons également utiliser des connaissances complémentaires. Il est important de noter que cette méthode peut s'adapter à la tâche requérant des paraphrases. Ainsi, on peut souhaiter en obtenir de nombreuses, au détriment de leur qualité pour de la recherche d'information, alors que la correction sera privilégiée pour le résumé automatique.

Références

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *Actes de ACL*, Ann Arbor, USA.

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

BOUAMOR H. (2010). Construction d'un corpus de paraphrases d'énoncés par traduction multilingue multi-source. In *Récital-TALN*, Montréal, Canada.

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

CALLISON-BURCH C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. In *Actes de EMNLP*, Hawai, USA.

CALLISON-BURCH C., COHN T. & LAPATA M. (2008). Parametric : An automatic evaluation metric for paraphrasing. In *Actes de COLING*, Manchester, UK.

CREGO J. M., MAX A. & YVON F. (2010). Local lexical adaptation in machine translation through triangulation : SMT helping SMT. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China.

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 W. B. & BROCKETT C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, Jeju Island, South Korea.

DUCLAYE F., COLLIN O. & YVON F. (2003). Apprentissage automatique de paraphrases pour l'amélioration d'un système de questions-réponses. In *Actes de TALN*, Batz-sur-mer, France.

DUTREY C., BOUAMOR H., BERNHARD D. & MAX A. (2010). Local modifications and paraphrases in wikipedia's revision history. In *Workshop on Corpus-Based Approaches to Paraphrasing and Nominalization,CBA 2010*, Barcelone, Espagne.

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

HARRIS Z. (1954). Distributional structure. *Word*.

IBRAHIM A., KATZ B. & LIN J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the second international workshop on Paraphrasing* : Association for Computational Linguistics.

JACQUEMIN C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Actes de ACL*, College Park, États-Unis.

KAUCHAK D. & BARZILAY R. (2006). Paraphrasing for automatic evaluation. In *Actes de NAACL-HLT*, New York, États-Unis.

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. (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of ACL, demo session*, Prague, Czech Republic.

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

LARDILLEUX A. (2010). *Contribution des basses fréquences à l'alignement sous-phrastique multilingue : une approche différentielle*. PhD thesis, Université de Caen, France.

LIN D. & PANTEL P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering*, **7**(4).

MATUSOV E., LEUSCH G. & NEY H. (2009). *Learning To Combine Machine Translation Systems*. MIT Press.

MAX A. (2008). Génération de reformulations locales par pivot pour l'aide à la révision. In *Actes de TALN*, Avignon, France.

MAX A. (2009). Sub-sentencial paraphrasing by contextual pivot translation. In *Proceedings of the ACL 2009 Workshop on Applied Textual Inference*, Singapore.

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

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*.

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.

QUIRK C., BROCKETT C. & DOLAN W. B. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of EMNLP*, volume 149, Barcelona, Spain.

RUSSO-LASSNER .G L. J. & .P R. (2005). *A Paraphrase-Based Approach to Machine Translation Evaluation*. Rapport interne TR-2005-57, UMIACS.

SNOVER M., MADNANI N., DORR B. & SCHWARTZ R. (2009). Fluency, adequacy, or HTER ? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece.

TOMEH N., ALLAUZEN A., WISNIEWSKI G. & YVON F. (2010). Refining word alignment with discriminative training. In *Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*.