Une étude en 3D de la paraphrase : types de corpus, langues et techniques

Houda Bouamor Aurélien Max Anne Vilnat
LIMSI-CNRS
Univ. Paris-Sud
Orsay, France
prenom.nom@limsi.fr

RÉSUMÉ

Cet article présente une étude détaillée de l'impact du type du corpus sur la tâche d'acquisition de paraphrases sous-phrastiques. Nos expériences sont menées sur deux langues et quatre types de corpus, et incluent une combinaison efficace de quatre systèmes d'acquisition de paraphrases. Nous obtenons une amélioration relative de plus de 27% en F-mesure par rapport au meilleur système, en anglais et en français, ainsi qu'une amélioration relative à notre combinaison de systèmes de 22% pour l'anglais et de 5% pour le français quand tous les types de corpus sont utilisés pour l'acquisition depuis le type de corpus le plus couramment disponible.

ABSTRACT

A study of paraphrase along 3 dimensions : corpus types, languages and techniques

In this paper, we report a detailed study of the impact of corpus type on the task of sub-sentential paraphrase acquisition. Our experiments are for 2 languages and 4 corpus types, and involve an efficient machine learning-based combination of 4 paraphrase acquisition systems. We obtain relative improvements of more than 27% in F-measure over the best individual system on English and French, and obtain a relative improvement over the combination system of 22% for English and 5% for French when using all other corpus types as additional training data for our most readily available corpus type.

MOTS-CLÉS : acquisition de paraphrases, constitution de corpus.

KEYWORDS: paraphrase acquisition, corpus collection.

1 Introduction

La variation paraphrastique est probablement l'une des caractéristiques les plus fascinantes de la langue naturelle : différentes expressions peuvent être utilisés pour véhiculer des sens très proches. 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 de paraphrases acceptables pouvant être exploitées dans divers contextes.

Il s'agit cependant de l'une des principales sources de complexité pour les processus de traitement automatique des langues. Les thésaurus encodés manuellement sont par nature incomplets, et ne sont pas disponibles pour toutes les langues. De plus, ils ne comprennent souvent pas d'ex-

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 267–280,
Grenoble, 4 au 8 juin 2012.

pressions de plusieurs mots qui sont nécessaires pour produire ou reconnaître automatiquement des paraphrases plus complexes. Le besoin d'acquérir automatiquement des paraphrases à partir de corpus de textes a ainsi été à l'origine de nombreux travaux. L'acquisition de paraphrases sous-phrastiques, que nous appellerons simplement *paraphrases* dans la suite, repose la plupart du temps sur l'appariement préalable d'unités de plus grande taille (des paires de phrases ou des documents comparables). Ces unités peuvent être obtenues directement par un processus supervisé, tel que la traduction humaine multiple, ou l'appariement automatique fondé sur la similarité entre textes (Mihalcea *et al.*, 2006). On observe que les techniques pour l'acquisition de paraphrases sont généralement très dépendantes des types de corpus sur lesquels elles ont été développées (Madnani et Dorr, 2010). Dans l'ordre inverse de leur disponibilité, ces types de corpus peuvent être grossièrement définis comme :

1. **corpus monolingues parallèles** : des paires d'énoncés de sens équivalents alignées de façon supervisée (comme les traductions multiples de livres (Barzilay et McKeown, 2001) ou les groupes de questions ayant la même réponse (Bernhard et Gurevych, 2008)).
2. **corpus multilingues parallèles** : des paires d'énoncés disponibles dans deux langues ou plus (Bannard et Callison-Burch, 2005) (comme les transcriptions des débats parlementaires européens)
3. **corpus monolingues comparables** : des paires de textes associés en fonction de similarité textuelle (comme des extraits de documents du Web (Pasça et Dienes, 2005)) en suivant éventuellement certaines heuristiques (tels que les articles de journaux publiés dans le même intervalle de temps (Dolan *et al.*, 2004))

Les ressources dans lesquelles les paraphrases abondent, ce qui facilite généralement une extraction *précise*, sont peu fréquentes à l'état naturel, alors que les unités de textes *comparables* sont potentiellement très nombreuses à l'échelle du Web (Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008). Ces considérations nous conduisent à envisager d'améliorer les performances des techniques d'acquisition de paraphrases sur un type de corpus en utilisant du matériel d'apprentissage (i.e. des exemples annotés) à partir d'autres types de corpus.

Dans cet article, nous présentons une analyse détaillée de la tâche d'acquisition de paraphrases sur quatre types de corpus monolingues représentatifs, que nous avons nommés en fonction du *type de signal du contenu sémantique d'origine* :

– TEXTE : des paires de phrases résultant de traductions multiples d'un même texte.
– PAROLE : des paires d'énoncés résultant de traductions multiples de mêmes extraits de parole.
– SCÈNE : des paires d'énoncés résultant de descriptions multiples d'une même scène visuelle.
– ÉVÉNEMENT : des paires d'énoncés résultant de descriptions multiples d'un même événement ou de deux événements proches.

Notre étude sera menée sur des collections constituées d'un nombre identique de paires de phrases pour chacun des types de corpus, ceci pour deux langues, l'anglais et le français. Nous utiliserons quatre systèmes d'acquisition de paraphrases (Bouamor *et al.*, 2011) et décrirons une architecture efficace pour valider la combinaison de leurs hypothèses par apprentissage automatique. Nous détaillerons les quantités de paraphrases par type accessibles à partir de chacun des types de corpus étudiés, et nous donnerons les performances de chaque système individuel ainsi que notre système de combinaison sur chaque type de corpus pris indépendamment, et sur chaque type de corpus en ajoutant d'autres types de corpus comme données d'entraînement pour la validation.

L'un de nos principaux résultats est que, pour les deux langues étudiées, l'acquisition de paraphrases peut être significativement améliorée à l'aide des données d'entraînement de types de corpus différents. Ceci est notablement le cas pour le corpus ÉVÉNEMENT, source la plus facile à acquérir pour l'acquisition de paraphrases, ce qui ouvre d'intéressantes perspectives pour des études ultérieures sur l'acquisition de paraphrases à grande échelle et leur utilisation pour l'amélioration de la performance d'applications de TAL. Nous avons également élaboré une typologie, sur les deux langues, quantifiée des types de paraphrases sur chacun des types de corpus, à la fois pour les paraphrases de référence et pour celles que notre système de combinaison, le plus performant, parvient à acquérir sur chaque type de corpus, ce qui fournira des informations précieuses pour guider la suite de ces travaux.

Dans la suite de cet article, nous commencerons par un rapide état de l'art sur l'acquisition de paraphrases (section 2), puis nous décrirons la méthodologie de construction de nos corpus et leurs caractéristiques (section 3). Nous détaillerons ensuite nos résultats de l'évaluation de l'acquisition de paraphrases (section 4). À la section 5.1, nous présenterons tout d'abord la performance d'un système de combinaison sur chacun des types de corpus, puis la performance de ce système lorsque sont utilisées des données d'entraînement additionnelles provenant des autres types de corpus (5.2). Enfin nous conclurons en évoquant différentes pistes de recherche ouvertes par nos travaux (section 6).

2 État de l'art

Au cours du temps, l'acquisition et la génération de paraphrases ont attiré un grand nombre de travaux de recherche, qui sont trop nombreux pour être correctement résumés ici : Madnani et Dorr (2010) présentent une revue relativement complète des principales approches. L'acquisition de paraphrases au niveau de phrases entières a été abordée à partir de ressources spécifiques augmentant la probabilité de trouver des phrases en relation de paraphrase (Dolan *et al.*, 2004; Bernhard et Gurevych, 2008; Wubben *et al.*, 2009), à partir de corpus monolingues comparables (Barzilay et Elhadad, 2003; Fung et Cheung, 2004; Nelken et Shieber, 2006) ainsi qu'à partir du Web (Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008).

Diverses techniques ont été proposées pour l'acquisition de paraphrases à partir de paires de phrases en relation (Barzilay et McKeown, 2001; Pang *et al.*, 2003) et à partir de corpus parallèles bilingues (Bannard et Callison-Burch, 2005; Kok et Brockett, 2010). Le lien entre la construction du corpus et le développement et l'évaluation des techniques d'acquisition est l'objet de (Cohn *et al.*, 2008; Callison-Burch *et al.*, 2008). À notre connaissance, il n'existe pas d'autres travaux portant sur l'acquisition de paraphrases qui soient menés sur plusieurs types de corpus et sur plusieurs langues de façon comparable. Pour sa part, le travail de Chan *et al.* (2011) explore la complémentarité de corpus bilingues et monolingues en acquisition de paraphrases. Faruqui et Padó (2011) s'intéressent à l'acquisition de *paires en relation d'implication* (prémisse et hypothèse), en menant des expériences dans trois langues sur des corpus journalistiques de différents domaines pour une langue. Bien que leur travail ne soit pas directement comparable au nôtre, ces auteurs montrent que la robustesse entre domaines est difficile à obtenir.

Enfin, l'évaluation de paraphrases générées automatiquement a été l'objet de quelques travaux récents (Liu *et al.*, 2010; Chen et Dolan, 2011; Metzler *et al.*, 2011), bien que ce problème reste difficile et globalement peu résolu. La génération de paraphrases motivée par une application

particulière offre un moyen indirect pour l'évaluation de la performance de la génération de paraphrases (Zhao *et al.*, 2009). Par exemple, le domaine de la Traduction Automatique Statistique est à l'origine de travaux montrant l'utilité à la fois de paraphrases produites par des humains (Schroeder *et al.*, 2009; Resnik *et al.*, 2010) et produites automatiquement (Madnani *et al.*, 2008; Marton *et al.*, 2009; Max, 2010) pour améliorer la performance en traduction.

3 Collecte de corpus de paires de phrases

Dans cet article, nous nous intéressons à l'acquisition de paraphrases à partir de paires de phrases en relation, caractéristiques de quatre types de corpus. Un corpus pour chaque type a été construit en deux langues, l'anglais et le français, et comporte 625 paires de phrases par langue. Nous détaillons maintenant la méthode de constitution de ces corpus.

TEXTE Pour l'anglais, nous avons utilisé le corpus MTC [1] (décrit dans (Cohn *et al.*, 2008)), qui regroupe des ensembles d'articles d'actualité traduits plusieurs fois depuis le chinois, et pour le français le corpus CESTA [2] regroupant des ensembles d'articles d'actualité traduits depuis l'anglais. Pour chaque groupe de phrases, nous retenons les paires de phrases ayant la plus petite distance d'édition au-dessus d'un seuil fixé empiriquement, en les extrayant d'abord de chacun des groupes et en reconsidérant par la suite des groupes déjà utilisés pour atteindre le nombre visé de paires de phrases.
Exemple : « *Dans l' autre cas , le gel des terres est destiné à maîtriser l' offre .* ⟷ *Le deuxième type de gel de terres doit servir à la gestion de l' offre .* »

PAROLE Pour l'anglais, nous avons utilisé des fichiers [3] librement disponibles de sous-titres de films tournés en français, *Le Fabuleux Destin d'Amélie Poulain* et *Les Choristes*, et pour le français nous avons pris les fichiers de la série télévisée tournée en anglais américain *Desperate Housewives*. Nous avons d'abord aligné chaque corpus parallèle en utilisant l'algorithme décrit dans (Tiedemann, 2007), basé sur des indices de durée et développé pour des sous-titres multilingues, puis nous avons extrait des paires de phrases en dessous d'un seuil minimal de distance d'édition, et filtré manuellement les erreurs apparentes de l'algorithme précédent.
Exemple : « *Vous pourriez passer ce soir et regarder ma tuyauterie ?* ⟷ *Pourriez-vous venir inspecter ma tuyauterie ce soir ?* »

SCÈNE Nous avons utilisé le *Multiple Video Description Corpus* (Chen et Dolan, 2011) obtenu à partir de descriptions multiples de courtes vidéos. De façon analogue à ce qui a été fait pour TEXTE, nous avons choisi des paires de phrases au sein de ces groupes en fonction d'une distance minimale d'édition au-dessus d'un certain seuil. Un point important est que pour l'anglais nous avons pu utiliser des descriptions qualifiées de "vérifiées". Cependant, les descriptions en français dans cette ressource sont disponibles dans des quantités bien moins importantes, et en outre aucune n'a le statut de "vérifiée". Nous avons tout de même décidé d'utiliser ce corpus, mais en gardant à l'esprit que cette source est de nettement moins bonne qualité. [4]
Exemple : « *une personne met du lait sur du riz.* ⟷ *un homme fait du riz au lait.* »

1. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T01`
2. `http://www.elda.org/article125.html`
3. `http://www.opensubtitles.org`
4. Ce type de corpus sera désigné entre parenthèse pour le français ("(SCÈNE)") dans tous les tableaux dans la suite pour rappeler son caractère particulier.

ÉVÉNEMENT Nous avons utilisé des titres de groupes d'articles d'actualité provenant du service d'agrégation Google News [5]. Nous avons ensuite affiné l'algorithme de regroupement en retenant les paires de titres dont les dates de publication des articles n'étaient pas séparées de plus d'un jour. Nous avons reproduit la même procédure de sélection que pour TEXTE et SCÈNE pour obtenir une couverture maximale sur l'ensemble des groupes.

Exemple : « *700 000 décès liés au Sida ont pu être évités en 2010* ⟷ *Forte baisse des décès et des infections liés au sida en 2010* »

	Statistiques du corpus 500 paires de phrases		**Accords inter-annotateur** 50 paires de phrases		**Stat. sur les formes dans les paraphrases** sans les paraphrases identiques			
					para. sûres		para. possibles	
	# formes	# formes par phrases	para. sûres	para. possibles	% formes	# formes	% formes	# formes
ANGLAIS								
TEXTE	21 473	21,0	66,1	20,4	18,6	4 004	12,3	2 651
PAROLE	11 049	10,5	79,1	10,9	17,5	1 942	31,6	3 500
SCÈNE	7 783	7,5	80,5	35,2	10,9	851	14,0	1 094
ÉVÉNEMENT	8 609	8,0	65,3	20,5	17,5	1 506	14,5	1 251
FRANÇAIS								
TEXTE	24 641	24,0	64,6	16,6	29,2	7 218	6,2	1 527
PAROLE	11 850	11,5	82,7	20,8	22,5	2 667	16,7	1 981
(SCÈNE)	7 012	6,5	42,8	9,3	3,9	275	9,4	664
ÉVÉNEMENT	9 121	9,1	67,8	3,8	19,6	1 793	9,6	876

TABLE 1 – Description de l'ensemble des corpus collectés et des annotations de référence pour les paraphrases en anglais et en français. Pour rappel, SCÈNE pour le français apparaît entre parenthèses car nous ne le considérons pas de la même qualité que les autres corpus.

Nous avons ensuite réalisé une annotation des paraphrases dans ces corpus, en suivant l'essentiel des consignes décrites dans (Cohn *et al.*, 2008) [6] à l'aide de l'outil YAWAT (Germann, 2008), mis à part le fait que nous ne sommes pas partis d'alignements initiaux obtenus automatiquement afin de ne pas biaiser le travail des annotateurs. Les principales consignes étaient que les paraphrases *sûres* et *possibles* devaient être distinguées, que les alignements les plus petits devaient être privilégiés sans décourager néanmoins les alignements groupe-à-groupe (i.e. *n-m*), et que les phrases devaient être alignées autant que possible. Nous ne considérerons dans la suite, pour toutes les statistiques et les expériences, que les paraphrases qui ne sont pas des paires identiques (telles que « *petit pont de bois* ⟷ *petit pont de bois* »), car on peut les considérer comme triviales au regard de la tâche d'acquisition.

La table 1 indique différentes statistiques pour les corpus collectés. La première observation est que TEXTE contient des phrases significativement plus longues que les autres types, plus de deux fois plus longues que celles de PAROLE par exemple. La table contient également les valeurs d'accords inter-annotateurs [7] calculées sur des sous-ensembles de 50 paires de phrases annotées indépendamment par deux annotateurs. Nous considérons comme acceptables les valeurs obtenues pour les paraphrases sûres, mais les valeurs obtenues pour les paraphrases possibles sont faibles. Ce dernier résultat était relativement prévisible, étant donné le nombre d'in-

5. `http://news.google.com`

6. Voir `http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_guidelines.pdf`

7. Pour chaque type de paraphrase, nous calculons la moyenne des valeurs de rappel obtenues par chaque annotateur comme référence et nous effectuons la moyenne.

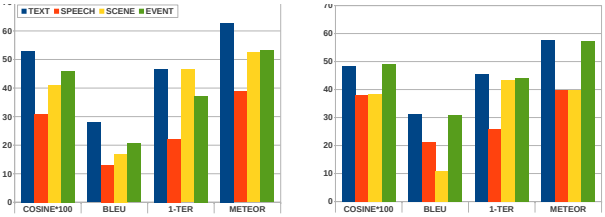

FIGURE 1 – Moyenne des similarités des paires de phrases pour tous les corpus pour l'anglais (à gauche) et le français (à droite) en utilisant le cosinus des vecteurs de formes, BLEU (Papineni *et al.*, 2002), TER (Snover *et al.*, 2006) et METEOR (Lavie et Agarwal, 2007) (à noter que les synonymes de WordNet ne pourraient être utilisés que pour l'anglais).

terprétations pour les paraphrases entrant possiblement dans cette catégorie. Cela ne constituera toutefois pas un problème dans la suite dans nos expériences : comme nous le verrons dans la section 4, notre métrique d'évaluation ne les considèrera pas comme des solutions attendues, et se limitera à ne pas les considérer comme fausses lorsqu'elles apparaitront parmi les hypothèses d'un système.

La table 1 montre enfin les pourcentages et les nombres de paraphrases pour chaque niveau de certitude pour chacun des corpus. Nous obtenons approximativement le même nombre total de paraphrases pour l'anglais (16 799) et le français (17 001). Les corpus anglais ont à peu près le même nombre de paraphrases sûres et possibles (8 303 par rapport à 8 496) alors qu'en français on trouve davantage de paraphrases sûres (11 953 contre 5 048). Ceci peut s'expliquer par le fait que les annotateurs ont travaillé indépendemment, avec des interprétations possiblement différentes de la tâche, et que les corpus, aussi *comparables* soient-ils entre langue, sont différents par nature. Les autres faits remarquables sont que TEXTE contient beaucoup plus de paraphrases que les autres corpus et que PAROLE comporte proportionnellement plus de paraphrases possibles que les autres corpus, et que SCÈNE contient nettement moins de paraphrases, en pourcentage et en nombre.

Dans la figure 1 différentes mesures typiquement utilisées, notamment en Traduction Automatique, de similarité entre paires de phrases sont données. TEXTE contient les paires de phrases les plus similaires selon toutes les métriques, suivi de près par ÉVÉNEMENT (dont les phrases sont beaucoup plus courtes). SCÈNE contient des paires de phrases qui sont plus similaires que celles de PAROLE pour l'anglais, ce qui n'est pas le cas pour le français.

4 Évaluation de l'acquisition de paraphrases

Nous adoptons la méthodologie PARAMETRIC de Callison-Burch *et al.* (2008) pour évaluer la performance des systèmes sur la tâche d'acquisition de paraphrases sur les corpus décrits dans la section précédente. Dans PARAMETRIC, un ensemble de paraphrases candidates extraites d'une paire de phrases en relation est comparé à un ensemble de paraphrases de référence, obtenues

par annotation manuelle, en calculant les mesures habituelles de *précision* (P) et *rappel* (R). La première valeur correspond à la proportion de paires d'hypothèses de paraphrases, ensemble noté H, produites par un système qui sont correctes par rapport à l'ensemble de référence contenant les paraphrases *sûres* et *possibles*, noté R_{tout}. Le rappel est obtenu en calculant la proportion de l'ensemble de référence de paraphrases *sûres*, noté $R_{\text{sûr}}$,qui sont trouvées par un système. Nous calculons également une valeur de F-mesure (F_1), qui considère le rappel et la précision comme également importants. Ces valeurs sont donc données par les formules suivantes :

$$P = \frac{|H \cap R_{\text{tout}}|}{|H|} \quad R = \frac{|H \cap R_{\text{sûr}}|}{|R_{\text{sûr}}|} \quad F_1 = \frac{2PR}{P+R}$$

Il est à noter que la façon dont les ensembles R_{tout} et $R_{\text{sûr}}$ de paires de paraphrases de référence sont définis garantit que les hypothèses de paraphrases incluant les paraphrases de référence annotées comme *possibles* ne pénaliseront pas la précision sans toutefois augmenter le rappel.

Toutes les valeurs de performance fournies dans les sections suivantes sont obtenues en effectuant une validation croisée 10 fois [8] au lieu d'utiliser le découpage des corpus en corpus de test/ corpus d'apprentissage. Nous moyennons, par la suite, les résultats sur chaque ensemble d'évaluation individuel pour obtenir des valeurs stables. Tous nos ensembles de données pour la validation croisée contiennent 500 paires de phrases. [9]

5 Expériences bilingues

5.1 Une architecture pour l'acquisition de paraphrases sous-phrastiques

Nous allons maintenant décrire les systèmes qui seront testés sur nos divers corpus décrits dans la section 3 utilisant la méthodologie décrite dans la section 4. Ces systèmes individuels sont décrits plus en détails dans (Bouamor *et al.*, 2011). Un système de combinaison est en outre utilisé pour valider automatiquement les hypothèses de paraphrases produites par les systèmes individuels en utilisant un ensemble de traits visant à reconnaître des paraphrases. Quatre systèmes individuels ont été utilisés et sont décrits ci-dessous : les raisons pour avoir retenu ces systèmes incluent leur libre disponibilité et/ou le coût raisonnable de leur développement, la possibilité d'utiliser des ressources comparables là où pertinent pour les deux langues étudiées, ainsi que les caractéristiques spécifiques à chaque technique.

Apprentissage statistique d'alignements de mots (GIZA) L'outil GIZA++ (Och et Ney, 2004) calcule des modèles statistiques d'alignement de mots de complexité croissante à partir de corpus parallèles. Il a été lancé sur chacun des corpus monolingues de paires de phrases dans les deux directions, les alignements ont été *symétrisés* puis les heuristiques classiques d'extraction de bi-segments *cohérents* ont été appliquées (Koehn *et al.*, 2003), sans toutefois agrandir les bi-segments par ajout de mots non alignés aux frontières.

Connaissances linguistiques sur la variation de termes (FASTR) L'outil FASTR (Jacquemin, 1999) permet de repérer des variantes de termes dans de grands corpus, les variations étant décrites à l'aide de métarègles spécifiant les dérivations morpho-syntaxiques possibles à partir

8. La validation croisée nous permet d'utiliser la totalité des données disponibles.

9. Il faut noter que, sur les 625 paires de phrases de départ pour chaque corpus, 125 paires de phrases sont extraites pour optimiser les paramètres d'un système basé sur la métrique TER_p (voir section 5.1).

d'un terme donné au moyen d'expressions régulières sur les catégories morpho-syntaxiques. La variation paradigmatique peut aussi s'exprimer au moyen de contraintes entre mots, imposant qu'ils appartiennent à la même famille morphologique ou sémantique en utilisant des ressources existantes disponibles pour nos deux langues. Les variantes pour tous les groupes de mots d'une des phrases d'une paire sont extraites dans l'autre phrase, et l'on conserve l'intersection des ensembles obtenus dans les deux directions.

Transformations optimales entre séquences de mots (TER$_p$) L'outil TER$_p$ (Snover *et al.*, 2010) peut être utilisé pour calculer un ensemble optimal (modulo quelques approximations) d'éditions au niveau des mots et des segments qui permettent de transformer une phrase en une autre. [10] Les types d'éditions sont paramétrés par un ou plusieurs poids qui sont optimisés par *hill climbing* pour maximiser la F-mesure, avec 100 redémarrages aléatoires, en utilisant les 125 paires de phrases réservées à cette fin dans chaque type de corpus.

Équivalence de traduction (PIVOT) Nous avons exploité la probabilité de paraphrase définie par Bannard et Callison-Burch (2005) pour des paraphrases extraites de corpus parallèles multilingues. Nous avons utilisé le corpus Europarl [11] de débats parlementaires en anglais et en français, comprenant environ 1,7 millions de phrases parallèles, en prenant chaque langue comme source et pivot à tour de rôle. GIZA++ a été utilisé pour aligner les mots et les probabilités de traduction de segments ont été estimées à partir de ces alignements par les méthodes standards du système de traduction statistique MOSES (Koehn *et al.*, 2007). Pour chaque segment d'une paire de phrases, nous avons construit son ensemble de paraphrases, et extrait sa paraphrase de l'autre phrase ayant la plus grande probabilité. Nous avons réitéré ce processus dans les deux directions, et finalement conservé pour chaque segment la paire de paraphrases issue d'une des deux directions avec la probabilité la plus forte.

Combinaison de systèmes par validation En calculant l'union de toutes les hypothèses de paraphrases issues de tous les systèmes précédents pour chaque paire de phrases, nous avons procédé à une classification en deux classes (soit, "paraphrase" ou "non paraphrase") en utilisant un classifieur à maximum d'entropie MAXENT [12]. Ceci permet d'inclure des traits qui n'étaient pas nécessairement pris en compte ou possibles à considérer dans les systèmes individuels. Plus généralement, ceci permet de tenter d'apprendre une caractérisation plus générique des paraphrases, qui pourrait s'adapter trivialement à un nombre quelconque de systèmes en entrée. Les exemples positifs pour le classifieur sont ceux provenant de l'union des hypothèses qui sont également présentes dans l'ensemble de référence $R_{\text{sûr}}$. Les exemples négatifs sont constitués du complément de cet ensemble dans l'union. Les traits que nous utilisons sont résumés dans la table 2.

Résultats expérimentaux Les résultats pour les systèmes individuels, leur union et nos systèmes de combinaison entraînés sur chaque type de corpus (colonne "appr.=C") sont donnés dans la Figure 3. Nous constatons tout d'abord que tous les systèmes obtiennent de meilleurs résultats sur TEXTE, pour lequel il y avait plus de données d'apprentissage disponibles et dans lequel les équivalences sémantiques entre les paires de phrases étaient plus probables. ÉVÉNEMENT apparaît comme le type de corpus le plus difficile, ce qui pourrait être considéré comme un

10. Il est à noter que contrairement à ce que TER$_p$ permet, nous n'utilisons pas les équivalents de mots ou de segments proposés par défaut car ceux-ci ne sont disponibles que pour l'anglais. Ce type de connaissance sera néanmoins apporté par les systèmes FASTR et PIVOT.

11. `http://statmt.org/europarl`

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

Traits dérivés des paires de segments – distance d'édition entre les paraphrases, racines identiques, mêmes formes, longueur de la phrase
Traits dérivés des paires de phrases – similarité entre paires de phrases (cosinus, BLEU, TER, METEOR), position relative des paraphrases, présence de formes communes aux frontières des paraphrases, présence d'une autre paire de paraphrases de chaque système aux frontières de la paraphrase, présence d'une paraphrase à un autre endroit dans l'autre phrase
Traits distributionnels – similarité (cosinus) des vecteurs de formes du contexte pour chaque segment d'une paraphrase (dérivée de fréquences obtenues dans le grand corpus parallèle anglais-français fourni pour la campagne d'évaluation WMT'11 (`http://www.statmt.org/wmt11/translation-task.html`, soit environ 30 millions de phrases parallèles)
Traits dérivés des systèmes – combinaison des systèmes individuels qui proposent la paire de paraphrase

TABLE 2 – Principaux traits utilisés par nos classifieurs. Des intervalles discrétisés basés sur les valeurs médianes sont utilisés pour les valeurs réelles, et des valeurs binarisées sont utilisées pour indiquer les configurations présentes pour les combinaisons.

Corpus type (C)	Systèmes individuels												Combinaison de systèmes								
	GIZA			FASTR			$TER_{\rightarrow F}$			PIVOT			union			appr.=C			appr.=tout		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
ANGLAIS																					
TEXTE	48,2	58,9	53,0	63,1	5,9	10,7	41,2	66,4	50,9	73,4	25,8	38,2	20,8	**80,8**	33,1	68,4	62,8	**65,5**	**77,5**	56,1	65,1
PAROLE	39,7	44,2	41,8	27,1	3,5	6,3	25,0	50,3	33,4	**79,2**	15,3	25,7	25,5	**71,4**	37,6	51,0	56,3	53,5	67,7	48,7	**56,6**
SCÈNE	44,8	57,7	50,5	47,4	5,2	9,5	40,1	67,9	50,4	**84,6**	14,6	25,0	36,2	**83,4**	50,5	44,9	66,8	**53,7**	33,2	59,7	42,7
ÉVÉN.	19,0	33,9	24,3	62,9	3,1	6,0	28,8	68,7	40,6	**97,4**	11,2	20,1	20,8	**75,5**	32,7	35,0	67,1	46,0	56,4	56,1	**56,2**
FRANÇAIS																					
TEXTE	52,5	58,9	55,5	56,9	4,9	9,1	46,4	61,4	52,8	64,5	30,3	41,2	41,5	**77,9**	54,1	**74,7**	61,0	**67,1**	**74,5**	60,2	66,6
PAROLE	44,0	54,9	48,9	30,7	4,3	7,6	34,8	60,2	44,1	**75,5**	19,0	30,4	31,4	**76,2**	44,5	60,2	59,7	**60,0**	55,1	61,0	57,9
(SCÈNE)	14,4	43,6	21,7	53,0	4,0	7,4	13,8	75,3	23,4	**94,6**	5,21	9,8	12,7	**86,4**	22,2	19,9	59,8	**29,8**	12,5	69,4	21,1
ÉVÉN.	28,7	44,2	34,8	34,4	2,3	4,3	29,9	58,9	39,7	**79,5**	15,0	25,2	25,2	**72,5**	37,4	40,0	56,3	46,8	62,4	40,7	**49,3**

TABLE 3 – Résultats de l'évaluation pour chaque système individuel (à gauche) et les systèmes combinés (à droite) sur tous les types de corpus, pour l'anglais (en haut) et le français (en bas). Les valeurs en gras indiquent les meilleurs résultats pour une métrique donnée pour chaque type de corpus et chaque langue.

résultat décevant dans la mesure où il s'agit du type pour lequel il existe le plus de données prêtes à être utilisées : nous reviendrons sur ce point dans la section 5.2.

En termes de performance en F-mesure par type de corpus, GIZA obtient de meilleurs résultats sur TEXTE et PAROLE, qui contiennent des phrases longues, avec d'éventuelles répétitions, alors que $TER_{\rightarrow F}$ a de meilleurs résultats sur SCÈNE et ÉVÉNEMENT, où les équivalences qui sont rares au niveau corpus sont plus fréquentes. Pour des raisons de place, nous ne détaillerons pas plus dans cet article les performances des systèmes individuels, pour nous concentrer sur nos combinaisons de systèmes.

Dans toutes les configurations, la combinaison de systèmes améliore de façon importante la F-mesure relativement au meilleur des systèmes individuels pour chaque type de corpus, ainsi que relativement à l'union des résultats de l'ensemble des systèmes. Les améliorations sont importantes sur TEXTE (respectivement +12,5 et +11,6 sur l'anglais et le français) et sur PAROLE (+11,7 et +11,1) et assez bonnes sur SCÈNE (+3,2 et +6,4) et sur ÉVÉNEMENT (+5,4 et +7,1).

	+Texte	+Parole	+Scène	+Événement	+Tous
ANGLAIS					
# ex+	7 342	2 296	1 784	1 171	12 593
Texte	65,5	**66,2**	65,1	**66,2**	65,1
Parole	56,0	53,5	52,8	54,8	**56,6**
Scène	49,7	**54,3**	53,7	53,8	42,7
Événement	51,1	45,3	42,5	46,0	**56,2**
FRANÇAIS					
# ex+	12 961	3 340	966	2 160	19 427
Texte	**67,1**	**67,2**	66,7	67,0	66,6
Parole	57,6	**60,0**	56,4	59,6	57,9
(Scène)	23,7	22,0	**29,8**	23,9	21,1
Événement	45,2	45,6	44,3	46,8	**49,3**

Table 4 – Résultats de l'évaluation (scores F_1) pour tous les types de corpus pour l'anglais (en haut) et le français (en bas) quand sont ajoutées les données d'entraînement des autres types de corpus (les valeurs sur fond grisé de la diagonale correspondent aux cas où aucune donnée n'est ajoutée). Les rangées "*#ex+*" indiquent le nombre d'exemples positifs de paraphrases apporté par chaque type de corpus supplémentaire sur le même nombre de paires de phrases.

Nous avons constaté (voir la table 1) que Texte et Parole sont les deux types de corpus ayant le plus grand nombre d'exemples de paraphrases sûres pour les deux langues : les résultats montrent que notre classifieur a été capable de les utiliser efficacement.

Les valeurs de rappel pour l'union sont assez grandes pour tous les types de corpus, allant de 71,4 (pour Parole en anglais) à 83,4 (pour Scène en anglais). Il y a, cependant, une nette baisse entre les valeurs de rappel pour les unions et pour les résultats de nos classifieurs, bien que ces dernières soient toutes autour de 6/10, ce qui peut être considéré comme une valeur acceptable pour une tâche de cette complexité. Une étude plus approfondie des faux négatifs pourrait nous aider à déterminer de nouveaux traits pour reconnaître des paraphrases plus difficiles à identifier. Enfin, nous pouvons noter que la précision est en général meilleure pour un des systèmes (Pivot), et atteint des valeurs intéressantes en particulier sur Texte, où nous disposons du plus grand nombre d'exemples (F-mesure de respectivement 68,4 et 74,6 pour l'anglais et le français).

5.2 Expériences sur l'apport des autres types de corpus

Nous considérons à présent la possibilité d'améliorer la performance de notre système de combinaison par l'utilisation de données d'apprentissage provenant d'autres types de corpus. Pour cela, nous construisons des systèmes en utilisant tout d'abord les données additionnelles provenant d'un autre type de corpus, puis de l'ensemble des types de corpus disponibles. Les résultats obtenus sont donnés dans la table 4 [13].

Nous observons qu'il existe deux cas de figure. Dans le premier, la performance en F-mesure est améliorée pour l'anglais sur Texte (+0,7), Parole (+3,1) et Scène (+0,6) en utilisant soit un seul type de corpus supplémentaire, soit l'ensemble des corpus disponibles, alors que pour le français

13. Nos résultats sont toujours donnés en procédant à une validation croisée qui réalise une moyenne des résultats obtenus sur 10 ensembles de test pour chaque type de corpus testé.

aucun ajout de données d'apprentissage n'améliore la performance pour ces types de corpus. Dans le second cas, ÉVÉNEMENT est amélioré à la fois pour l'anglais (+10,2) et pour le français (+2,5) en utilisant toutes les données d'apprentissage supplémentaires disponibles. Hormis la condition où les données provenant de TEXTE sont ajoutées pour l'anglais, tous les ajouts d'autres types de corpus diminuent la performance quand ils sont ajoutés individuellement : on observe donc ici nettement une contribution collective attribuable à l'ajout d'au moins deux sources. La nature des exemples pertinents ainsi ajoutés retiendra notre attention pour de futurs travaux : la sélection plus fine d'exemples pourrait effectivement repousser davantage la performance atteinte.

On peut encore noter que TEXTE n'est pratiquement pas touché par l'ajout de données supplémentaires, ce qui peut s'expliquer en partie par le fait que ce type de corpus contient à lui seul la moitié du nombre total d'exemples dans les deux langues. À l'opposé, SCÈNE, qui a le plus petit nombre d'exemples d'entraînement, voit sa performance baisser sensiblement, assez fortement par exemple avec l'ajout des données provenant de TEXTE (respectivement -4,0 et -6,1 pour l'anglais et le français) et par tous les corpus ensemble (respectivement -11,0 et -8,7). Ceci souligne à nouveau la nature spécifique de ce type de corpus : des descriptions indépendantes de la même scène vidéo peuvent être verbalisées de façons très diverses, à différents niveaux. Finalement, il y a nettement plus d'exemples positifs en français (19 427) qu'en anglais (12 593) : ceci peut s'expliquer par le fait que les phrases en français dans nos corpus contiennent plus de formes (voir Table 1) et que les paraphrases en français contiennent plus de variantes morphologiques telles que différentes formes conjuguées des verbes.

6 Discussion et perspectives

Dans cet article, nous nous sommes intéressés au problème de l'acquisition de paraphrases sur des types de corpus et entre ces types de corpus, en définissant les types de corpus à partir de l'origine du signal du contenu sémantique des paires de phrases utilisées : un texte dans différentes langues (TEXTE), de la parole transcrite dans une autre langue (PAROLE), une scène visualisée (SCÈNE), et une courte description (un titre d'article) d'un événement donné (ÉVÉNEMENT). Nous avons décrit un grand corpus annoté, contenant 2 500 paires de phrases pour l'anglais et pour le français, et nous avons réutilisé les principes généraux d'une méthodologie existante pour évaluer l'acquisition automatique de paraphrases (Callison-Burch *et al.*, 2008). Nous avons évalué un système efficace de combinaison exploitant les hypothèses de quatre systèmes, ainsi que l'impact produit par l'utilisation des données d'entraînement des autres types de corpus.

Notre résultat le plus prometteur est certainement l'amélioration obtenue sur le type de corpus ÉVÉNEMENT en utilisant les données d'entraînement de tous les corpus disponibles. Étant donné que les autres types de corpus sont beaucoup plus rares par nature, il semble que la disponibilité de tels corpus permet néanmoins d'apporter des connaissances utiles pour améliorer la reconnaissance des paraphrases sur ce qui s'est avéré être le type de corpus le plus difficile dans notre étude. Un résultat de cette nature incite à appliquer et améliorer nos techniques pour l'acquisition de paraphrases à l'échelle du Web (Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008), où les paires de phrases en relation peuvent être très nombreuses.

Une piste intéressante porte sur l'amélioration des traits de détermination du statut de paraphrases, en particulier si l'on travaille sur les résultats d'un moteur de recherche sur le Web,

incluant des mesures de similarités entre paires de texte plus informées, par exemple en exploitant les structures thématiques des documents (Barzilay et Elhadad, 2003), des mesures de similarité lexicale en contexte (Dinu et Lapata, 2010; Erk et Pado, 2010), ou des résultats de systèmes d'implication textuelle (Kouylekov et Negri, 2010) [14]

Une analyse fine des différents types de paraphrases serait nécessaire pour servir de guide pour des travaux futurs afin de repousser les limites des systèmes actuels : de premiers résultats d'une telle analyse quantitative, pour tous les types de corpus et les deux langues de notre étude, sont donnés dans la Table 5. La principale observation est que la synonymie (comme dans *dans l'affirmative ⟷ le cas échéant*) est le phénomène le plus courant, qui de plus représente le principal type d'hypothèses correctes proposées par nos systèmes. En revanche, il n'est pas surprenant de voir que nos systèmes ne sont pas compétents pour reconnaître des paraphrases dans la catégorie "pragmatique", ce qui requiert de nombreuses et coûteuses informations sur le monde et sur le contexte des paires de phrases. Enfin, il est intéressant de noter que le type de corpus ÉVÉNEMENT contient des paraphrases de référence de tous les types.

	synonymie		**typographie**		**inclusion**		**pragmatique**		**syntaxe**		**dérivation**		**flexion**	
	%réf	%sys	%réf	%sys	%réf	%sys	%réf	%sys	%réf	%sys	%réf	%sys	%réf	%sys
ANGLAIS														
TEXTE	51,2	43,5	7,6	7,0	12,1	16,4	0,6	0,0	4,4	4,7	12,1	10,5	11,5	17,6
PAROLE	39,8	34,0	25,6	38,2	12,3	6,3	1,7	0,0	3,5	0,0	3,5	2,1	13,2	19
SCÈNE	50,0	46,8	1,3	2,1	21,6	23,4	0,0	0,0	1,3	0,0	5,4	8,5	20,2	19,1
ÉVÉNEMENT	36,9	41,6	15,0	22,2	19,1	16,6	1,3	0,0	6,8	2,7	6,8	2,7	13,6	13,8
FRANÇAIS														
TEXTE	46,9	26,0	9,0	20,6	2,1	1,0	3,6	1,0	6,6	0,0	3,0	3,2	28,5	47,7
PAROLE	45,5	43,9	14,2	19,5	8,0	7,3	2,6	0,0	11,6	2,4	3,5	2,4	14,2	24,3
(SCÈNE)	46,4	51,3	5,3	2,7	8,9	5,4	0,0	0,0	5,3	0,0	0,0	0,0	33,8	40,5
ÉVÉNEMENT	28,3	16,6	19,7	27,7	16,0	11,1	7,4	0,0	8,6	5,5	7,4	0,0	12,2	38,8

TABLE 5 – Distribution des types de paraphrases mesurée dans 50 paires de phrases annotées (%réf) choisies aléatoirement et des hypothèses de paraphrases sur ces phrases pour notre meilleur système (%sys) pour l'anglais (en haut) et le français (en bas). À noter que %sys doit être examiné en relation avec le rappel du système donné dans la table 3. Les types sont illustrées par les exemples suivants : (*dans l'affirmative ⟷ le cas échéant*) (**synonymie**), (*Classement ⟷ Class.*) (**typographie**)„ (*BNP ⟷ BNP Paribas*) (**inclusion**), (*de plus en plus sales ⟷ ne se brossent plus les dents*) (**pragmatique**), (*il y a 6 mois ⟷ six mois avant*) (**syntaxe**), (*refroidie ⟷ froide*) (**dérivation**), (*crevette ⟷ crevettes*, *moque ⟷ moquait*) (**flexion**)

Références

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

BARZILAY, R. et ELHADAD, N. (2003). Sentence alignment for monolingual comparable corpora. *In Proceedings of EMNLP*, Sapporo, Japan.

14. Concernant les systèmes d'implication textuelle, nous nous trouvons face à un problème de dépendance circulaire, car ces systèmes se fondent typiquement sur des connaissances préalables, notamment des paraphrases. Nous pensons quand même que l'utilisation de telles connaissances doit être faite quand celles-ci sont disponibles, comme nous l'avons fait nous-mêmes par l'utilisation du système FASTR et de ses ressources lexico-sémantiques associées.

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

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

BHAGAT, R. et RAVICHANDRAN, D. (2008). Large scale acquisition of paraphrases for learning surface patterns. *In Actes de ACL-HLT*, Columbus, USA.

BOUAMOR, H., MAX, A. et VILNAT, A. (2011). Combinaison d'informations pour l'alignement monolingue. *In Actes de TALN*, Montpellier, France.

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

CHAN, T. P., CALLISON-BURCH, C. et VAN DURME, B. (2011). Reranking bilingually extracted paraphrases using monolingual distributional similarity. *In Proceedings of the EMNLP Workshop on Geometrical Models of Natural language Semantics*, Edinburgh, UK.

CHEN, D. et DOLAN, W. (2011). Collecting highly parallel data for paraphrase evaluation. *In Proceedings of ACL-HLT*, Portland, Oregon, USA.

COHN, T., CALLISON-BURCH, C. et LAPATA, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, 34(4).

DINU, G. et LAPATA, M. (2010). Measuring distributional similarity in context. *In Proceedings of EMNLP*, Cambridge, USA.

DOLAN, B., QUIRK, C. et BROCKETT, C. (2004). Unsupervised construction of large paraphrase corpora : Exploiting massively parallel news sources. *In Proceedings of Coling*, Switzerland.

ERK, K. et PADO, S. (2010). Exemplar-based models for word meaning in context. *In Proceedings of ACL*, Uppsala, Sweden.

FARUQUI, M. et PADÓ, S. (2011). Acquiring entailment pairs across languages and domains : A data analysis. *In Proceedings of IWCS*, Oxford, UK.

FUNG, P. et CHEUNG, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. *In Proceedings of COLING*, Geneva, Switzerland.

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

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

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of ACL*, Czech Republic.

KOEHN, P., OCH, F. J. et MARCU, D. (2003). Statistical Phrase-Based Translation. *In Proceedings of NAACL HLT*, Edmonton, Canada.

KOK, S. et BROCKETT, C. (2010). Hitting the Right Paraphrases in Good Time. *In Proceedings of NAACL*, Los Angeles, USA.

KOUYLEKOV, M. et NEGRI, M. (2010). An open-source package for recognizing textual entailment. *In Proceedings of the ACL*, Uppsala, Sweden.

LAVIE, A. et AGARWAL, A. (2007). METEOR : An automatic metric for MT evaluation with high levels of correlation with human judgments. *In Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, Czech Republic.

LIU, C., DAHLMEIER, D. et NG, H. T. (2010). PEM : A paraphrase evaluation metric exploiting parallel texts. *In Proceedings of EMNLP*, Cambridge, MA.

MADNANI, N. et DORR, B. J. (2010). Generating Phrasal and Sentential Paraphrases : A Survey of Data-Driven Methods . *Computational Linguistics*, 36(3).

MADNANI, N., RESNIK, P., DORR, B. et SCHWARTZ, R. (2008). Are multiple reference translations necessary ? investigating the value of paraphrased reference translations in parameter optimization. *In Proceedings of AMTA*, Waikiki, Hawaii.

MARTON, Y., CALLISON-BURCH, C. et RESNIK, P. (2009). Improved Statistical Machine Translation Using Monolingually-derived Paraphrases. *In Proceedings of EMNLP*, Singapore.

MAX, A. (2010). Example-based paraphrasing for improved phrase-based statistical machine translation. *In Proceedings of EMNLP*, Cambridge, MA.

METZLER, D., HOVY, E. et ZHANG, C. (2011). An empirical evaluation of data-driven paraphrase generation techniques. *In Proceedings of ACL-HLT*, Portland, USA.

MIHALCEA, R., CORLEY, C. et STRAPPARAVA, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *In Proceedings of AAAI*, Boston, USA.

NELKEN, R. et SHIEBER, S. M. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. *In Proceedings of EACL*, Trento, Italy.

OCH, F. J. et NEY, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4).

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

PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proceedings of ACL*, Philadelphia, USA.

PASÇA, M. et DIENES, P. (2005). Aligning Needles in a Haystack : Paraphrase Acquisition Across the Web. *In Proceedings of IJCNLP*, Jeju Island, South Korea.

RESNIK, P., BUZEK, O., HU, C., KRONROD, Y., QUINN, A. et BEDERSON, B. B. (2010). Improving translation via targeted paraphrasing. *In Proceedings of EMNLP*, Cambridge, MA.

SCHROEDER, J., COHN, T. et KOEHN, P. (2009). Word Lattices for Multi-Source Translation. *In Proceedings of EACL*, Athens, Greece.

SNOVER, M., DORR, B. J., SCHWARTZ, R., MICCIULLA, L. et MAKHOUL, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *In Proceedings of AMTA*, Boston, USA.

SNOVER, M., MADNANI, N., DORR, B. J. et SCHWARTZ, R. (2010). TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3).

TIEDEMANN, J. (2007). Building a multilingual parallel subtitle corpus. *In CLIN17*, Belgium.

WUBBEN, S., van den BOSCH, A., KRAHMER, E. et MARSI, E. (2009). Clustering and matching headlines for automatic paraphrase acquisition. *In EWNLG*, Athens, Greece.

ZHAO, S., LAN, X., LIU, T. et LI, S. (2009). Application-driven Statistical Paraphrase Generation. *In Proceedings of ACL-AFNLP*, Suntec, Singapore.