Une approche hybride pour la construction de lexiques bilingues d'expressions multi-mots à partir de corpus parallèles

Nasredine Semmar[1] Morgane Marchand[2]
(1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, 91191 Gif-sur-Yvette, France
(2) eXenSa, 41 rue Périer, 92120 Montrouge, France
nasredine.semmar@cea.fr, morgane.marchand@exensa.com

RESUME

Les expressions multi-mots jouent un rôle important dans différentes applications du Traitement Automatique de la Langue telles que la traduction automatique et la recherche d'information interlingue. Cet article, d'une part, décrit une approche hybride pour l'acquisition d'un lexique bilingue d'expressions multi-mots à partir d'un corpus parallèle anglais-français, et d'autre part, présente l'impact de l'utilisation d'un lexique bilingue spécialisé d'expressions multi-mots produit par cette approche sur les résultats du système de traduction statistique libre Moses. Nous avons exploré deux métriques basées sur la co-occurrence pour évaluer les liens d'alignement entre les expressions multi-mots des langues source et cible. Les résultats obtenus montrent que la métrique utilisant un dictionnaire bilingue amorce de mots simples améliore aussi bien la qualité de l'alignement d'expressions multi-mots que celle de la traduction.

ABSTRACT

A hybrid approach to build bilingual lexicons of multiword expressions from parallel corpora
Multiword expressions play an important role in numerous applications of Natural Language Processing such as machine translation and cross-language information retrieval. This paper presents, on the one hand, a hybrid approach to build bilingual lexicons of multiword expressions from an English-French parallel corpus, and on the other hand, the impact of using a domain-specific bilingual lexicon of multiword expressions on the results of the open source statistical machine translation Moses. We have explored two metrics based on co-occurrence to evaluate alignment links between multiword expressions of source and target texts. The obtained results show that scoring based on a bilingual dictionary improves the quality of both alignment and translation.

MOTS-CLES : Lexique bilingue, alignement d'expressions multi-mots, programmation linéaire.
KEYWORDS: Bilingual lexicon, multiword expressions alignment, linear programming.

1 Introduction

Une expression multi-mots est une combinaison de mots pour laquelle les propriétés syntaxiques ou sémantiques de l'expression entière ne peuvent pas être obtenues à partir de ses parties (Sag et al., 2002). Contrairement à l'alignement de mots simples qui est désormais une tâche bien maîtrisée plus particulièrement pour les langues à écriture latine, l'alignement d'expressions multi-mots continue à

susciter de nombreux travaux de recherche (DeNero, Klein, 2008; Lefever et al., 2009; Bouamor et al., 2012; Makcen et al., 2013). La plupart de ces travaux commencent tout d'abord par identifier les expressions multi-mots dans chaque partie du corpus parallèle (Ramisch, 2014), ensuite, utilisent différentes approches d'alignement pour les apparier. Les approches pour l'extraction monolingue d'expressions multi-mots peuvent être: (1) symboliques en reposant sur des patrons morpho-syntaxtiques (Okita et al., 2010; Dagan, Church, 1994), (2) statistiques en utilisant des mesures d'association pour classer les expressions multi-mots candidates (Vintar, Fisier, 2008), et (3) hybrides combinant (1) et (2) (Seretan, Wehrli, 2007; Semmar, Laib, 2010). Pour identifier les correspondances entre expressions multi-mots dans différentes langues, plusieurs travaux font appel à des outils d'alignement de mots simples pour guider l'alignement d'expressions multi-mots. L'outil de création de lexiques bilingues de termes techniques de Dagan et Church (1994) identifie ces termes sur la base d'un étiqueteur morpho-syntaxique. Ensuite, la liste de candidats trouvés est filtrée manuellement. L'application de filtres à base de catégories grammaticales permet une réduction importante du bruit dans les sorties par l'exclusion de candidats constitués de mots vides. Malgré leur simplicité, les approches symboliques restent difficiles à appliquer lorsque les données ne sont pas étiquetées morpho-syntaxiquement. Une autre limite de cette approche est que la définition de patrons d'extraction d'expressions multi-mots est dépendante de la langue. D'autres utilisent les algorithmes d'apprentissage statistique (Okita et al., 2010). Une hypothèse largement suivie pour acquérir des expressions multi-mots bilingues est qu'une expression multi-mots dans une langue source garde la même structure syntaxique que son équivalente dans une langue cible donnée (Seretan, Wehrli, 2007). Or, cette hypothèse n'est pas toujours vérifiée puisque certaines expressions multi-mots ne se traduisent pas forcément par des expressions ayant la même structure syntaxique (Bouamor et al., 2012). Nous présentons, dans cet article, une approche hybride pour l'acquisition d'un lexique bilingue d'expressions multi-mots à partir d'un corpus parallèle anglais-français. Contrairement aux travaux cités précédemment, notre approche permet de réaliser les tâches d'extraction des expressions multi-mots et d'alignement en une seule étape.

La suite de l'article est organisée comme suit : dans la section 2, nous présentons notre approche d'alignement d'expressions multi-mots, nous décrivons en particulier les métriques que nous avons utilisées pour évaluer les liens d'alignement. Puis nous décrivons dans la section 3 les résultats obtenus. La section 4 conclut notre étude et présente nos travaux futurs.

2 Notre approche pour l'alignement d'expressions multi-mots

Comme nous l'avons mentionné précédemment, la plupart des approches d'alignement d'expressions multi-mots se déroulent en deux étapes. Tout d'abord les fragments de phrases représentant potentiellement des expressions sont extraits. Ensuite les potentielles expressions repérées sont associées deux à deux. Cela a l'inconvénient de propager d'éventuelles erreurs. Nous décrivons dans cette section une approche hybride permettant de réaliser les tâches d'extraction des expressions multi-mots et d'alignement en une seule étape sous la forme d'un problème d'optimisation (Marchand, Semmar, 2011). Ce problème d'optimisation peut être formulé comme suit : une bi-phrase consiste en deux phrases e et f en deux langues différentes qui sont traduction l'une de l'autre. Chacune de ces phrases peut être vue comme une séquence ordonnée de mots. On désigne par e_{ij} le fragment de phrase contenu entre les espaces i et j de la phrase e. f_{kl} identifie le fragment de phrase contenu entre les espaces k et l de la phrase f. Un lien est une paire de fragments (un bi-fragment) alignés que l'on note (e_{ij}, f_{kl}). Chaque e_{ij} peut être lié avec plusieurs f_{kl} et réciproquement. Un alignement a de la bi-phrase (e, f) est une segmentation des deux phrases en fragments contigus ainsi que l'ensemble des liens entre ces fragments. Nous utilisons une fonction réelle Φ pour donner des scores aux liens selon s'ils sont plus ou moins probables.

$$\phi : \{e_{ij}\} \times \{f_{kl}\} \rightarrow R$$

Le score d'un alignement *Φ(a)* est le produit des scores des liens qui le composent.

$$\phi(a) = \prod_{(e_{ij}, f_{kl}) \in a} \phi(e_{ij}, f_{kl})$$

Formellement, cette fonction n'a pas d'autres contraintes que celle d'être réelle. Dans la pratique on va choisir une fonction qui donne une idée sur la pertinence d'aligner ensemble tels ou tels fragments. Plus le score associé est élevé plus la pertinence de l'alignement est grand. On cherche donc l'alignement (segmentation + liens) qui va maximiser le score décrit ci-dessus. Cependant, trouver un alignement qui maximise ce score est un problème NP-complet (DeNero, Klein, 2008). C'est pourquoi on utilise une forme linéarisée du problème afin de pouvoir lui appliquer des méthodes de programmation linéaire en nombre entiers et de trouver une solution approchée rapidement. Pour écrire le problème sous forme de programme linéaire en nombres entiers nous devons introduire des variables binaires. Ces variables sont identifiées par des majuscules au contraire des fragments de phrases qui sont dénotés par des minuscules. Les variables E_{ij} et F_{kl} indiquent si les fragments e_{ij} et f_{kl} font partie de la segmentation considérée. Les variables binaires A_{ijkl} indiquent quant à elles s'il existe un lien entre les fragments e_{ij} et f_{kl}. On utilise $W_{ijkl} = log(\Phi(e_{ij}, f_{kl}))$ afin de pouvoir linéariser le score d'un alignement. Le problème se présente sous forme ci-dessous:

$$\begin{cases} \max \sum_{i,j,k,l} W_{i,j,k,l} A_{i,j,k,l} & \\ \forall\, x : 1 \leq x \leq |e| & \sum_{i,j : i < x \leq j} E_{i,j} = 1 \quad (1) \\ \forall\, y : 1 \leq y \leq |f| & \sum_{k,l : k < y \leq l} F_{k,l} = 1 \quad (2) \\ \forall\, i, j & \sum_{k,l} A_{i,j,k,l} \geq E_{i,j} \quad (3) \\ \forall\, k, l & \sum_{i,j} A_{i,j,k,l} \geq F_{k,l} \quad (4) \\ \forall\, i, j, k, l & 2 \cdot A_{i,j,k,l} \leq E_{i,j} + F_{k,l} \quad (5) \end{cases}$$

Avec les contraintes suivantes sur les variables:

$$\begin{cases} 0 \leq i < |e|, & 0 < j \leq |e|, & i < j \\ 0 \leq k < |f|, & 0 < l \leq |f|, & k < l \end{cases}$$

Les contraintes (1) et (2) signifient qu'un mot fait partie d'un et un seul fragment dans la segmentation choisie des phrases. La contrainte (3) assure que chaque fragment de la segmentation choisie de *e* est lié à au moins un fragment de la segmentation choisie de *f*. De même pour la contrainte (4) pour *f*. Un fragment peut tout à fait être lié à plusieurs fragments. Enfin, la contrainte (5) assure que si un lien existe entre e_{ij} et f_{kl} (c'est à dire que $A_{ijkl} = 1$), alors e_{ij} et f_{kl} sont dans les segmentations choisies de *e* et *f*. En effet, nous ne voulons lier entre eux que des fragments existants réellement. Bien que notre approche s'inspire des travaux de DeNero et Klein (2008), la formulation

des contraintes (3) à (5) nous différencie. En effet, DeNero et Klein (2008), ne considèrent que les alignements bijectifs alors que nous prenons en compte les alignements surjectifs. C'est un premier pas vers la non contiguïté des expressions. Comme mentionné plus haut, pour mettre ce problème sous forme linéaire nous avons été obligé de nous restreindre aux fragments continus. Hors certaines expressions sont en plusieurs parties. Cette surjectivité des alignements vise à contrebalancer en partie cette restriction en autorisant un fragment à être lié à plusieurs fragments pas forcément contigus. Nous avons utilisé la boite à outils GLPK[1] pour résoudre ce programme linéaire en nombres entiers. L'objectif étant de trouver comment attribuer un score pertinent aux bi-fragments. Ce programme fonctionne avec n'importe quelle fonction de score Φ à valeur réelle. Nous décrivons dans les sections suivantes les deux métriques que nous avons utilisées.

2.1 Co-occurrence

Pour estimer la distribution de probabilité d'alignement, nous avons utilisé une métrique fondée sur la co-occurrence. Cette métrique utilise l'information selon laquelle la traduction d'une partie d'une phrase se trouve forcément quelque part dans la phrase associée. Pour calculer la distance en terme de co-occurence, on utilise un corpus d'entrainement bilingue aligné phrase à phrase. Pour chacun des fragments, on vérifie s'il est ou non présent dans chaque phrase. Le score entre deux fragments e_{ij} et f_{kl} est calculé selon la formule suivante :

$$\phi_c(e_{ij}, f_{kl}) = \frac{\sum_{s' \in S} N_{s'}(e_{ij}) \times N_{s'}(f_{kl})}{\sum_{s \in S} N_s(e_{ij}) + N_s(f_{kl}) - N_s(e_{ij}) \times N_s(f_{kl})}$$

$N_s(e_{ij})$ vaut 1 si le fragment e_{ij} de la première langue est présent dans la phrase s du corpus S et 0 sinon. $N_s(f_{kl})$ se comporte de la même façon pour l'autre langue. Ce score calcule le nombre de présence commune des deux fragments divisé par le nombre total d'apparition de l'un ou de l'autre. Si e_{ij} ou f_{kl} n'apparaissent pas dans le corpus, le score vaut zéro. Ainsi si un fragment du corpus test est inconnu du corpus d'apprentissage son score sera nul et le score de n'importe quel alignement de la phrase le contenant sera nul également.

2.2 Dictionnaire bilingue

Nous voulons que notre alignement fragment à fragment respecte autant que possible les alignements mot à mot procurés par un dictionnaire bilingue amorce. Le dictionnaire anglais-français utilisé dans cette étude est composé de 243539 entrées. Ce dictionnaire n'est certainement pas un lexique amorce type mais nous avons utilisé uniquement les entrées qui sont sous la forme de mots simples (unitermes). Nous supposons que la présence d'une paire de mots dans ce dictionnaire renforcera la probabilité d'alignement des bi-fragments qui les contiennent. Ce dictionnaire peut aussi nous donner des informations d'alignement négatives lorsqu'un mot n'a pas de traduction. Nous prenons en compte ces informations pour pénaliser légèrement la création de liens qui n'étaient pas à la base dans le dictionnaire. Le score dictionnaire se calcule alors avec la formule suivante :

[1] GLPK est disponible sur http://www.gnu.org/s/glpk/.

$$\phi(e_{ij}, f_{kl}) = \frac{a \cdot R_1 + b \cdot R_0}{a \cdot R_1 + b \cdot R_0 + c \cdot N_1 + d \cdot N_0}$$

Avec, R_1: nombre de liens respectés; R_0: nombre de non-liens respectés; N_1: nombre de liens non-respectés; N_0: nombre de non-liens non respectés. Les coefficients *a*, *b*, *c* et *d* peuvent être adaptés dans le but d'ajuster l'importance relative des quatre termes. Nous avons analysé un petit corpus qui nous a permis de choisir empiriquement d'utiliser les valeurs suivantes : $a = b = c = 1$ et $d = ½$. Notons que le score dictionnaire est utilisé pour renforcer ou affaiblir le score co-occurrence.

3 Résultats expérimentaux et discussion

Nous avons évalué les résultats de notre approche d'alignement d'expressions multi-mots pour le couple de langues anglais-français selon deux stratégies différentes : (1) une évaluation manuelle comparant les résultats de notre approche d'alignement d'expressions multi-mots et ceux de l'outil Giza++ (Och, Ney, 2000) par rapport à un alignement de référence; (2) une évaluation automatique qui consiste à étudier l'impact de l'utilisation du lexique bilingue produit par notre approche d'alignement d'expressions multi-mots sur la qualité de traduction du système Moses[2].

3.1 Evaluation intrinsèque

L'évaluation manuelle a été réalisée en utilisant un lexique bilingue de référence construit à partir de 1992 phrases parallèles extraites du corpus Europarl (European Parliament Proceedings). Ce lexique composé de 7807 entrées a été construit manuellement à l'aide de l'outil Yawat (Germann, 2008) par deux annotateurs. Pour les métriques d'évaluation, nous avons utilisé celles du protocole défini lors de la conférence HLT/NAACL 2003 (Mihalcea et al., 2003). Le tableau 1 résume les performances des différentes approches.

Métrique utilisée	**Précision**	**Rappel**	**F-mesure**
Baseline (Giza++)	0,83	0,37	0,51
Co-occurrence	0,61	0,63	0,61
Co-occurrence + Dictionnaire bilingue	0,85	0,54	0,66

TABLE 1 : Performances des différentes approches d'alignement d'expressions multi-mots

Les résultats obtenus avec uniquement la métrique co-occurrence montrent que quand la structure des phrases est similaire dans les deux langues, l'alignement est efficace même si la traduction n'est pas mot à mot (Figure 1). La segmentation se fait mot à mot ou bien fragment par fragment en fonction de ce qui est plus fréquent dans le corpus.

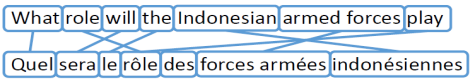

FIGURE 1: Exemple d'un alignement correct

[2] http://www.statmt.org/moses/

Ces résultats montrent aussi que la formulation surjective du problème permet de détecter les expressions en deux parties. Nous pouvons en effet voir que le mot français « rôle » est lié à la fois à « role » et « play », ce qui aurait été impossible avec les hypothèses de DeNero et Klein (2008), puisqu'elles ne permettent pas à un fragment d'une langue d'être aligné avec plusieurs fragments de l'autre langue. Dans ce cas précis, l'expression est partiellement reconnue car « role » et « play » sont deux mots pleins ayant une distribution de présence significative sur le corpus d'entrainement. Nous avons aussi constaté que l'ajout de l'information provenant du dictionnaire bilingue amorce améliore les résultats d'alignement. Dans l'exemple suivant (Figure 2), le dictionnaire donne les liens « be/être », « decided/décidé » et « there/y ». Ces liens permettent à notre approche d'alignement de reconstruire le bi-segment complet « is to be decided on there/doit y être décidé ». De plus, les liens « programme/programme » et « concrete/concret » sont renforcés. Notons que le lexique produit par notre approche d'alignement ne contient pas que les expressions multi-mots mais peut contenir des fragments de textes. Par exemple, le fragment « doit y être décidé » n'est pas une expression multi-mots mais un segment phrastique.

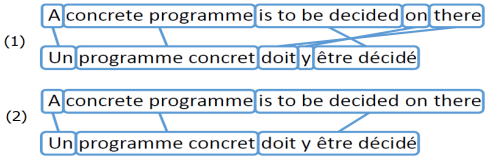

FIGURE 2: Exemples d'améliorations : (1) Métrique co-occurence, (2) Métrique co-occurence + dictionnaire bilingue

3.2 Evaluation extrinsèque

La non disponibilité d'un protocole commun permettant d'évaluer les résultats d'alignement d'expressions multi-mots nous a conduit à réaliser une évaluation extrinsèque dans laquelle nous étudions l'impact de l'utilisation du lexique bilingue produit par notre approche d'alignement dans le système Moses. Pour ce faire, nous avons réalisé des expérimentations sur deux corpus parallèles anglais-français (Table 2): Europarl et Emea (European Medicines Agency Documents). Ces deux corpus ont été extraits de la base libre de corpus parallèles OPUS (Tiedemann, 2012).

N° du run	Apprentissage (nombre de phrases)	Développement (nombre de phrases)
1	150K+10K (Europarl+Emea)	2K+0,5K (Europarl+Emea)
2	150K+20K (Europarl+Emea)	2K+0,5K (Europarl+Emea)
3	150K+30K (Europarl+Emea)	2K+0,5K (Europarl+Emea)

TABLE 2 : La taille (en nombre de phrases) de l'ensemble des corpus bilingues utilisés pour l'apprentissage et le développement du système de traduction Moses

Nous avons effectué trois runs avec, pour chacun d'entre eux, un test avec des données du domaine et un autre hors-domaine. Pour cela, nous avons extrait de manière aléatoire un ensemble de 500 paires de phrases à partir du corpus Europarl comme un corpus correspondant au domaine et 500 autres paires de phrases à partir du corpus Emea comme un corpus hors-domaine. Ces tests ont pour but de montrer l'impact que peut avoir le lexique du domaine produit par notre approche d'alignement d'expressions multi-mots à partir du corpus parallèle spécialisé (Emea) sur le modèle

de traduction du système Moses. Ce lexique est constitué en utilisant les deux métriques : co-occurrence et co-occurrence + dictionnaire bilingue. Nous comparons les résultats de traduction du système Moses après l'injection de ces deux lexiques spécialisés avec les résultats de ce système après l'injection du corpus ayant servi à produire ce lexique spécialisé (Emea) dans les données d'apprentissage (Europarl). Comme les entrées du dictionnaire bilingue sont sous forme de lemmes, nous avons d'une part, lemmatisé le corpus d'apprentissage à l'aide de la plate-forme LIMA (Besançon et al., 2010), et d'autre part, estimé le modèle de traduction sur des lemmes plutôt que sur des formes de surface. Les performances de Moses ont été évaluées en utilisant le score BLEU (Papineni et al., 2002) sur les deux ensembles de test pour les trois runs sachant que nous considérons une référence par phrase, les résultats obtenus sont présentés dans le tableau 3 (Table 3).

N° du run	**Domaine (Europarl)**			**Hors-Domaine (Emea)**		
	Baseline (Giza++)	**Co-occurrence**	**Co-occurrence + Dictionnaire bilingue**	**Baseline (Giza++)**	**Co-occurrence**	**Co-occurrence + Dictionnaire bilingue**
1	32,62	32,69	32,71	22,96	23,03	23,06
2	33,81	33,88	33,89	23,30	23,37	23,39
3	34,25	34,30	34,32	24,55	24,59	24,62

TABLE 3 : Scores BLEU de Moses selon l'approche d'alignement d'expressions multi-mots utilisée

Tout d'abord, nous constatons que le score BLEU obtenu est satisfaisant compte tenu de la taille du corpus d'apprentissage utilisé. Ce score varie en fonction du type du jeu de test. Le corpus de test du Domaine (Europarl) rapporte des scores BLEU plus élevés que le corpus de test Hors-Domaine (Emea) dans les trois configurations (alignement avec Giza++, alignement avec notre approche en utilisant la métrique fondée sur la co-occurrence, alignement avec notre approche en utilisant une combinaison des deux métriques : co-occurrence et dictionnaire bilingue). Les résultats obtenus montrent que le score fondé sur la combinaison des deux métriques produit les meilleures performances pour tous les runs.

4 Conclusion

Dans cet article, nous avons présenté une approche hybride pour l'acquisition d'un lexique bilingue d'expressions multi-mots à partir d'un corpus parallèle permettant de réaliser les tâches d'extraction des expressions multi-mots et d'alignement en une seule étape. Nous avons, en particulier, montré que l'ajout d'un lexique bilingue spécialisé (construit à l'aide de cette approche) au corpus d'apprentissage du système de traduction statistique Moses améliore la qualité de traduction aussi bien des textes du domaine de spécialité que les textes du domaine général. Dans nos expérimentations, le modèle de traduction a été estimé sur des lemmes plutôt que sur des formes de surface. De même, nos diverses métriques calculent des probabilités jointes. Nos travaux futurs s'orientent, d'une part, vers l'utilisation d'un modèle de génération pour produire les formes de surface adéquates à partir des résultats de traduction présentés en lemmes dans cette étude, et d'autre part, vers l'extraction des probabilités conditionnelles associées qui sont les primitives de la probabilité jointe afin de pouvoir intégrer les scores fournies par nos métriques dans la table de traduction de Moses.

Références

BESANÇON R., DE CHALENDAR G., FERRET O., GARA F., LAIB M., MESNARD O., SEMMAR N. (2010). LIMA: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. Actes de *LREC 2010*.

BOUAMOR D., SEMMAR N., ZWEIGENBAUM P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine. Actes de *LREC 2012*.

DAGAN I., CHURCH K. (1994). Termight : Identifying and translating technical terminology. Actes de *the 4th Conference on ANLP*, pages 34–40, Stuttgart, Germany.

DENERO J., KLEIN D. (2008). The complexity of phrase alignment problems. Actes de *the 46th annual meeting of the association for computational linguistics on human language technologies*.

GERMANN U. (2008). Yawat: Yet Another Word Alignment Tool. Actes de *ACL 2008*.

LEFEVER E., MACKEN L., HOSTE V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. Actes de *the 12th Conference of the European Chapter of the Association for Computational Linguistics, Greece*.

MACKEN L., LEFEVER E., HOSTE V. (2013). TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology, Volume 19, Issue 1*, pages: 1-30.

MARCHAND M., SEMMAR N. (2011). A Hybrid Multi-Word Terms Alignment Approach Using Word Co-occurrence with a Bilingual Lexicon. Actes *de 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'11).*

MIHALCEA R., PEDERSEN T. (2003). An evaluation exercise for word alignment. Actes de *the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, pages: 1-10.

OCH F. J., NEY H. (2000). Improved Statistical Alignment Models. Actes de *the 38th Annual Meeting of the Association for Computational Linguistics*, pages: 440-447.

OKITA T., GUERRA M., ALFREDO GRAHAM Y., WAY A. (2010). Multi-word expression sensitive word alignment. Actes de *4th International Workshop on Cross Lingual Information Access*.

PAPINENI K., ROUKOS S., WARD T., ZHU W. J. (2002). Bleu: a method for automatic evaluation of machine translation. Actes de *the 40th Annual meeting of the Association for Computational Linguistics*, pages: 311-318.

RAMISCH C. (2014). *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer.

SAG I., BALDWIN T., FRANCIS BOND F., COPESTAKE A., FLICKINGER D. (2002). Multiword expressions: a pain in the neck for NLP. Actes de *CICLing 2002*.

SEMMAR N., LAIB M. (2010). Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons. Actes de *Workshop on Language Resources and Human Language Technologies for Semitic Languages co-located with the seventh international conference on Language Resources and Evaluation (LREC 2010)*.

SERETAN V., WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. *Actes de TALN 2007*.

TIEDEMANN J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing, Volume V*, pages: 237-248.

VINTAR S., FISIER D. (2008). Harvesting multi-word expressions from parallel corpora. Actes de *LREC 2008.*