Etude de l'impact de la translittération de noms propres sur la qualité de l'alignement de mots à partir de corpus parallèles français-arabe

Nasredine Semmar[1] Houda Saadane[2]

(1) Institut CEA LIST, DIASI, Laboratoire Vision et Ingénierie des Contenus, CEA Saclay – Nano-INNOV, 91191 Gif-sur-Yvette Cedex
(2) LIDILEM, Université Stendhal-Grenoble III, Domaine Universitaire, 1180, avenue centrale, 38400 Saint Martin d'Hères
nasredine.semmar@cea.fr, houda.saadane@e.u-grenoble3.fr

Résumé. Les lexiques bilingues jouent un rôle important en recherche d'information interlingue et en traduction automatique. La construction manuelle de ces lexiques est lente et coûteuse. Les techniques d'alignement de mots sont généralement utilisées pour automatiser le processus de construction de ces lexiques à partir de corpus de textes parallèles. L'alignement de formes simples et de syntagmes nominaux à partir de corpus parallèles est une tâche relativement bien maîtrisée pour les langues à écriture latine, mais demeure une opération complexe pour l'appariement de textes n'utilisant pas la même écriture. Dans la perspective d'utiliser la translittération de noms propres de l'arabe vers l'écriture latine en alignement de mots et d'étudier son impact sur la qualité d'un lexique bilingue français-arabe construit automatiquement, cet article présente, d'une part, un système de translittération de noms propres de l'arabe vers l'écriture latine, et d'autre part, un outil d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe. Le lexique bilingue produit par l'outil d'alignement de mots intégrant la translittération a été évalué en utilisant deux approches : une évaluation de la qualité d'alignement à l'aide d'un alignement de référence construit manuellement et une évaluation de l'impact de ce lexique bilingue sur la qualité de traduction du système de traduction automatique statistique Moses. Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement de mots que celle de la traduction.

Abstract. Bilingual lexicons play a vital role in cross-language information retrieval and machine translation. The manual construction of these lexicons is often costly and time consuming. Word alignment techniques are generally used to construct bilingual lexicons from parallel texts. Aligning single words and nominal syntagms from parallel texts is relatively a well controlled task for languages using Latin script but it is complex when the source and target languages do not share the same written script. A solution to this issue consists in writing the proper names present in the parallel corpus in the same written script. This paper presents, on the one hand, a system for automatic transliteration of proper names from Arabic to Latin script, and on the other hand, a tool to align single and compound words from French-Arabic parallel text corpora. We have evaluated the word alignment tool integrating transliteration using two methods: A manual evaluation of the alignment quality and an evaluation of the impact of this alignment on the translation quality by using the statistical machine translation system Moses. The obtained results show that transliteration of proper names from Arabic to Latin improves the quality of both alignment and translation.

Mots-clés : Lexique bilingue, translittération, alignement de mots, traduction automatique statistique, évaluation.
Keywords: Bilingual lexicon, transliteration, word alignment, statistical machine translation, evaluation.

1 Introduction

Les lexiques bilingues jouent un rôle important dans les applications de Traitement Automatique des Langues (TAL) telles que la Recherche d'Information Interlingue (RII) et la Traduction Automatique (TA). La construction manuelle de ces lexiques est lente et coûteuse. C'est la raison pour laquelle depuis quelques années de nombreux travaux ont fait appel aux techniques d'alignement pour automatiser le processus de construction de lexiques bilingues. Ces travaux ont montré que l'alignement de formes simples et de syntagmes nominaux à partir de corpus parallèles est une tâche relativement bien maîtrisée pour les langues à écriture latine. En revanche, l'appariement de textes parallèles n'utilisant pas la même écriture demeure une opération complexe. Ce qui a conduit plusieurs chercheurs à exploiter la transcription

ou la translittération de certains mots des textes parallèles comme « points d'ancrage » pour améliorer la mise en correspondance bilingue. La transcription consiste à substituer à chaque son ou à chaque phonème d'un système phonologique, un graphème ou un groupe de graphèmes d'un système d'écriture, tandis que la translittération consiste à substituer à chaque graphème d'un système d'écriture un autre graphème ou un groupe de graphèmes d'un autre système d'écriture, indépendamment de la prononciation.

Dans la perspective d'évaluer l'impact de l'utilisation de la translittération de noms propres sur la qualité d'un lexique bilingue français-arabe construit automatiquement, nous présentons dans cet article, d'une part, un système de translittération de noms propres de l'arabe vers l'écriture latine et un outil d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe, et d'autre part, les résultats d'évaluation de ce lexique bilingue selon deux approches (intrinsèque et extrinsèque) et utilisant deux corpus différents (ARCADE II et OPUS).

La suite de l'article est organisée comme suit : dans la section 2, nous présentons l'approche de la translittération des noms propres écrits en arabe vers l'écriture latine. Puis nous décrivons dans la section 3, l'outil d'alignement de mots à partir d'un corpus de textes parallèles français-arabe en nous focalisant plus particulièrement sur l'étape d'appariement de cognats qui exploite la translittération. La section 4 sera consacrée aux expérimentations effectuées ainsi que la présentation des résultats obtenus et la section 5 conclut notre étude et présente nos travaux futurs.

2 Translittération

Les évolutions rapides des nouvelles technologies d'information et de communication sont accompagnées d'un essor important de la quantité et la diversité d'information générée et manipulée notamment celle disponible sur le web. Cette dernière, étant destinée à un public large et varié, est transcrite dans différentes langues ce qui a fait émerger la nécessité d'internationaliser les contenus afin de permettre un partage de données le plus large possible, entre des utilisateurs manipulant des langues différentes. Ainsi, les techniques de translittération trouvent tout leur intérêt afin de rendre cette perspective de partage possible.

2.1 Etat de l'art

Plusieurs travaux de recherche sur la transcription et la translittération ont été menés ces dernières années. Nous citons à titre d'exemple les travaux de (Jiang et al., 2007) pour la translittération des entités nommées (ENs) du chinois vers l'anglais, qui utilisent un modèle d'entropie maximale pour déterminer la translittération candidate, en se basant sur la similarité phonétique avec l'EN dans la langue source. Ces méthodes fonctionnent bien avec les entités nommées qui sont traduites phonétiquement, mais ce n'est pas toujours le cas. Pour ce type d'ENs, il est plus recommandé d'explorer les similitudes sémantiques entre les ENs dans les différentes langues. Ce constat a été approuvé dans les travaux de (Huang et al., 2004) qui combine les similitudes sémantiques et phonétiques. Les expérimentations effectuées montrent que cette approche réalise une précision de 67%. Par ailleurs, (Huang et al., 2003) ont travaillé sur l'extraction des paires d'ENs hindi-anglais grâce à l'alignement d'un corpus parallèle. Des paires chinois-anglais sont d'abord extraites à l'aide d'une programmation dynamique. Ce modèle chinois-anglais est alors adapté à l'hindi-anglais de manière itérative, en utilisant les paires hindi-anglais d'entités nommées déjà extraites pour l'amorçage du modèle. On trouve aussi des propositions de systèmes visant à attribuer une seule translittération à un nom donné : c'est le cas du modèle génératif proposé pour les noms d'origine anglaise écrits en japonais vers le système d'écriture latin (Knight, Graehl, 1997). Cette approche a été adaptée par (Stalls, Knight, 1998) à la façon dont un nom anglais écrit en arabe est transcrit en anglais. Le système de génération de translittérations s'appuie sur un dictionnaire d'apprentissage et ne prend pas en compte les prononciations non répertoriées ou inconnues du dictionnaire. Pour pallier cette limitation, certains travaux utilisent un modèle non supervisé. C'est le cas du système de translittération des noms anglais vers l'arabe proposé par (Abduljaleel, Larkey, 2003). Ce système est fondé sur le calcul de la forme la plus probable, censée être la forme correcte. Or cette hypothèse n'est pas vérifiée pour tous les pays arabes ni pour tous les dialectes. (Alghamdi, 2005) a proposé un système de translittération en écriture anglaise des noms arabes voyellés pour contourner la difficulté de la prononciation et le problème des variantes dialectales. Ce système est basé sur un dictionnaire de noms arabes dans lequel la prononciation est réglée au moyen de voyelles ajoutées aux noms répertoriés, avec indication en vis à vis de leur équivalent en écriture anglaise. Cependant, cette approche, non seulement ne prend pas en compte les prononciations non répertoriées dans le dictionnaire, mais, de plus, elle est normative par le fait qu'elle ne propose qu'une seule translittération pour un nom donné.

En conclusion, la plupart des travaux actuels ne prennent pas en compte la complexité du problème de la transcription et de la translittération qui concerne aussi bien l'oralité que le modèle scriptural des systèmes linguistiques impliqués. En effet, très peu de travaux prennent en considération le lien entre phonologie comparée et transcription interlingue, entre

graphématique comparée et translittération multilingue et entre dialectologie arabe et systèmes de translittération latins. Les rares études qui proposent une solution prenant en compte partiellement l'une de ces problématiques sont dédiées à l'identification automatique de l'origine du locuteur à partir de son dialecte (Guidère, 2004) (Barkat-Defradas et al., 2004). Dans le cadre de cette étude, notre objectif est de proposer un système automatique de translittération qui tient compte du lien entre phonologie, graphématique et dialectologie, dans la transcription des noms et des prénoms arabes vers l'écriture latine et plus particulièrement pour le français et l'anglais (Pouliquen, Steinberger, 2007).

2.2 Approche proposée pour la translittération de noms propres de l'arabe vers l'écriture latine

Afin de renvoyer la totalité des cas possibles de la translittération d'un nom arabe en écriture latine, nous nous sommes intéressés aux questions et aux problèmes liés à la translittération basée sur le système phonétique de l'arabe littéraire ainsi que sur la majorité des familles de dialectes, en prenant en compte des nombreuses variantes régionales et locales. Nous avons commencé par recenser les translittérations existantes pour chaque lettre de l'alphabet arabe standard depuis les normes et usages observés sur le Web et sur les dictionnaires de lieux géographiques de GeoNames. Nous avons constaté qu'au sein du même dictionnaire géographique un nom propre peut avoir plusieurs translittérations différentes. Cette investigation empirique est basée sur un corpus de textes qui a été recueilli dans les différentes langues cibles visées par le translittérateur. Elle a permis de constituer une librairie des équivalents graphématiques utilisés dans les écrits utilisant l'alphabet latin. Ci-dessous quelques équivalences graphématiques établies à partir de cette étude sur différents corpus :

- La lettre ش est transcrite en S dans DIN-31635, Sh selon UN, EI & ALA-LC, š suivant ISO/R 233 et (ch) dans le corpus d'apprentissage.
- La lettre ظ est transcrite en ẓ dans les différentes normes de translittération et en z, dh et d dans le corpus d'apprentissage.

Nous avons défini un certain nombre de règles syntaxiques et contextuelles afin de recenser les différentes translittérations. Parmi les règles syntaxiques que nous avons considérées dans notre translittération, le fait que le nom arabe ne prend pas en compte la dernière voyelle courte ou tanwin (marqueur du cas) à la fin du mot. Par exemple : زار بلال محمداً, le prénom محمداً est transcrit par Mohammed et non pas Mohammedan. Le module de translittération de l'écriture arabe vers l'écriture latine tient compte du lien entre la phonologie, la graphématique et la dialectologie en utilisant un certain nombre de règles issues d'une étude expérimentale. Il est fondé sur les automates à états finis pondérés de type transducteurs. Nous avons utilisé l'outil HTFST qui est constitué d'une interface basée sur la librairie open-source OpenFst (Reley et al., 2009). Cet outil sert à créer les automates de règles morphologiques, syntaxiques, et autres, et les appliquer ensuite à des textes. HTFST possède aussi une syntaxe propre aux « règles de replacements parallèles et contextuelles » offrant les mêmes possibilités que celles de XFST (Xerox Finite State Tool) (Beesley, Karttunen, 2003) implémentées en utilisant la librairie FOMA (Hulden, 2009). Le fonctionnement de notre approche de translittération est déterminé par la nature du mot fourni en entrée : l'automate passe d'état en état suivant les transitions, à la lecture de chaque lettre arabe de l'entrée. La Figure 1 décrit l'organigramme de notre module de translittération:

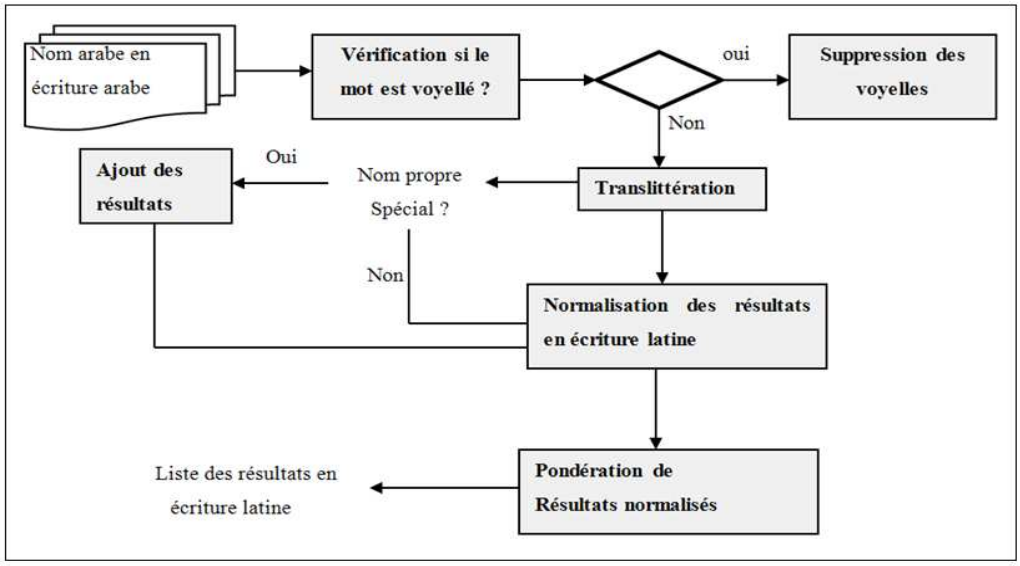

FIGURE 1: Organigramme du fonctionnement du translittérateur de l'arabe vers le latin

A l'issue de la lecture, un premier automate traite l'entrée de la manière suivante: si l'entrée est voyellée, il supprime les voyelles avant de translittérer le nom; si l'entrée est non-voyellée, il procède directement à la translittération du nom. Nous supprimons les voyelles afin de générer toutes les translittérations françaises et anglaises possibles. Ceci est dû à l'influence des dialectes sur les voyelles où les translittérations des mots issus du dialecte du Macherek sont orientées vers la translittération anglaise et ceux du dialecte du Maghreb sont plus orientées vers la translittération française. Enfin, le module produit en sortie une liste triée de noms arabes écrits en caractères latins.

Le cœur du système de translittération est constitué de règles contextuelles qui permettent le remplacement des lettres arabes en lettres latines ainsi que l'ajout des voyelles latines, en prenant en compte les lettres situées devant et/ou derrière la lettre à ajouter ou remplacer. Ces règles visent aussi à rendre compte de la manière la plus précise possible des formes observées en entrée : S'agit-il d'une «kunya» ? D'un nom précédé d'un article ? Ou bien d'un prénom seul ? Selon la forme d'entrée, nous appliquons d'abord des règles adéquates pour transcrire la partie qui ne constitue pas le nom à proprement parler (particules). Ainsi, des noms propres (spéciaux) comme (Ibn) إبن, (Abd) عبد, (Taha) طه, etc. seront transcrits directement. Prenons, par exemple, le prénom عبد qui peut être translittéré de plusieurs façons différentes. Nous attribuons un poids pour chaque translittération, sachant que le poids le plus bas indique la solution la plus probable. La règle ci-dessous indique que, lorsqu'un mot débute par ع ب د, il est transcrit le plus souvent par « Abd », ou bien moins souvent par « Abed ». Plus rarement, il sera transcrit « 3abd » ou « 3Abd », et dans quelques cas il sera transcrit par « Abd ».

R = ((ع ب د) .x. (((\` A b d <3000>) | A b d | (A b e d <1000>) | (3 (a|A) b d <2000>))).*;

Après avoir traité la partie (spéciale) du nom propre, nous appliquons les règles pour la translittération des noms eux-mêmes. Les règles pour la translittération des noms s'appliquent à leur tour selon le nombre de consonnes du nom considéré, et dans un ordre de priorité déterminé.

Par exemple, pour translittérer en écriture latine le nom propre arabe عبد الرشيد qui est composé par Abd (عبد) + Al (ال) + Nom (رشيد), le système procède de la manière suivante :

- Translittération de la particule عبد «Abd»;
- Translittération de l'article ال «Al»;
- Concaténation de la particule «Abd» et de l'article «Al» en les reliant au nom par un trait d'union ou en insérant un blanc entre les deux : Abd Al Rachid (عبد الرشيد);
- Génération de toutes les formes de translittération possibles pour ces trois éléments (Table 1):

Nom propre arabe	Translittérations
عبد الرشيد	Abd Al-Rachid
	Abdul Rashid
	abd al-Rashid
	3abd El Rachid
	abd Al Rashid
	Abdar-Rashid
	Abdel Rachid

TABLE 1 : Quelques translittérations pour le nom propre composé «عبد الرشيد»

- Normalisation de la liste des noms en écriture latine en supprimant les caractères spéciaux (diacritiques et chiffres) et en ajoutant la majuscule au début de chaque nom propre;
- Pondération de la liste des noms en écriture latine en attribuant un poids aux règles qui ont servi à la génération de la liste. Cette pondération est réalisée en utilisant divers moteurs de recherche en notant à chaque fois le nombre d'occurrences pour chaque forme générée du nom propre.

3 Alignement de mots

L'alignement de mots ou l'extraction de lexiques bilingues à partir de corpus de textes parallèles peut se décomposer conceptuellement en deux aspects: il s'agit de repérer les mots du texte source et du texte cible, puis de les mettre en correspondance.

3.1 Etat de l'art

Il existe principalement trois approches pour l'alignement de mots à partir de corpus de textes parallèles alignés phrase à phrase:

- Les approches à dominante statistique qui s'appuient sur les modèles IBM (Brown et al., 1993). L'outil d'alignement GIZA++ (Och, Ney, 2000) implémente notamment ce type d'approche. Cet outil implémente divers modèles de traduction (IBM 1, 2, 3, 4, 5 et HMM). GIZA++ est un outil efficace pour aligner les mots simples, mais il est moins performant, d'une part, lorsque les langues source et cible ont des morphologies et des structures syntaxiques différentes, et d'autre part, pour aligner les expressions multimots (Allauzen, Wisniewski, 2009) (Abdulhay, 2012).
- Les approches linguistiques qui utilisent généralement des dictionnaires bilingues déjà disponibles mais aussi les résultats de l'analyse morpho-syntaxique des phrases source et cible (Debili, Zribi, 1996). Les méthodes proposées par (Debili, Zribi, 1996) utilisent des ressources linguistiques externes (lexiques, règles, etc.) pour apparier les mots des textes parallèles alignés au niveau de la phrase. Ces méthodes font l'hypothèse que pour que des phrases soient en correspondance de traduction, il faut que les mots qui les composent soient également en correspondance. Elles n'utilisent qu'une information interne, c'est-à-dire que toute l'information nécessaire (et en particulier les correspondances lexicales) est dérivée des textes à aligner eux-mêmes (ancrage lexical).
- Une combinaison des méthodes statistiques avec différentes sources d'information linguistique (Daille et al., 1994) (Gaussier, Langé, 1995) (Ozdowska, Claveau, 2006) (Semmar et al., 2010). La méthode proposée par Gaussier (1995) est fondée sur des modèles statistiques pour établir les associations entre mots anglais et mots français, et ce en exploitant la propriété de dépendance entre les mots et leurs traductions respectives. La prise en compte des positions des mots dans les phrases permet de constituer un modèle de distorsion qui aide à la construction des associations. Ensuite, les structures morpho-syntaxiques représentant les séquences admissibles d'étiquettes grammaticales et de mots ont été recensées. Les correspondances et non-correspondances entre les structures anglaises et françaises sont utilisées pour élaborer les modèles statistiques permettant de retrouver les équivalences entre termes anglais et termes français. Quant à l'approche développée par Ozdowska et Claveau (2006), elle consiste d'abord à apparier les mots à un niveau global grâce au calcul des fréquences de cooccurrence dans des phrases alignées. Ensuite, ces mots constituent les couples amorces qui servent de point de départ à la propagation des liens d'appariement à l'aide des différentes relations de dépendance identifiées par un analyseur syntaxique dans chacune des deux langues.

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 multimots continue à susciter de nombreux travaux de recherche (Ozdowska, Claveau, 2006) (Lefever et al., 2009) (Bouamor et al., 2012). La plupart de ces travaux commencent tout d'abord par identifier les expressions multimots dans chaque partie du corpus parallèle, ensuite, utilisent différentes approches d'alignement pour les apparier. Les approches pour l'extraction monolingue d'expressions multimots peuvent être: (1) symboliques en reposant sur des patrons morpho-syntaxtiques (Okita et al., 2010), (2) statistiques en utilisant des mesures d'association pour classer les expressions multimots candidates (Vintar, Fisier, 2008), et (3) hybrides combinant (1) et (2) (Seretan, Wehrli, 2007). Pour identifier les correspondances entre expressions multimots dans différentes langues, plusieurs travaux font appel à des outils d'alignement de mots simples pour guider l'alignement d'expressions multimots. D'autres se basent sur des algorithmes d'apprentissage statistique. Une hypothèse largement suivie pour acquérir des expressions multimots bilingues est qu'une expression multimots dans une langue source garde la même structure syntaxique que son équivalente dans une langue cible donnée (Seretan, Wehrli, 2007) (Tufis, Ion, 2007). Or, cette hypothèse n'est pas toujours vérifiée puisque certaines expressions multimots ne se traduisent pas forcément par des expressions ayant la même structure syntaxique. De même, certaines expressions ne se traduisent pas systématiquement par une expression de même longueur.

Pour les langues n'utilisant pas l'écriture latine, de nombreux travaux ont été réalisés pour aligner automatiquement les translittérations à partir de corpus de textes multilingues en vue de l'enrichissement de lexiques bilingues. Citons notamment les travaux de (Yaser, Knight, 2002) et (Sherif, Kondrak, 2007) sur l'alignement arabe-anglais, (Tao et al., 2006) sur l'utilisation de la translittération pour l'extraction d'entités nommées à partir de corpus comparables ainsi que (Shao, Ng, 2004) qui utilisent l'information apportée par les translittérations sur la base de leur prononciation. Ils combinent l'information apportée par le contexte des traductions avec l'information apportée par les translittérations entre l'anglais et le chinois. L'intérêt de ce travail réside dans le fait qu'il permet l'alignement de mots très spécifiques mais rares.

Nous décrivons, dans la section suivante, notre démarche pour extraire un lexique bilingue de mots simples et de mots composés à partir d'un corpus parallèle français-arabe aligné au niveau de la phrase.

3.2 Approche proposée pour l'alignement de mots à partir de corpus de textes parallèles français-arabe

La démarche que nous proposons pour la construction de lexiques bilingues à partir de corpus de textes parallèles, est composée des trois étapes suivantes:

- alignement de mots simples,
- alignement de mots composés se traduisant mot à mot,
- alignement d'expressions multimots.

Notre approche pour l'alignement de mots est basée, d'une part, sur un modèle linguistique utilisant un dictionnaire bilingue, les caractéristiques des cognats, les catégories grammaticales, les relations de dépendance syntaxique et les règles de reformulation pour l'alignement de mots simples et composés, et d'autre part, sur un modèle hybride combinant patrons morpho-syntaxiques et méthodes statistiques pour l'alignement d'expressions multimots. Les entrées de l'outil d'alignement, implémentant notre approche, sont les sorties normalisées d'une analyse morpho-syntaxique effectuée à l'aide de la plate-forme d'analyse linguistique LIMA (Besançon et al., 2010) sur le corpus de textes parallèles. Cette plate-forme fournit pour chaque couple de phrases source et cible :

- la liste des lemmes et des formes fléchies des mots ainsi que leur position dans la phrase,
- les catégories grammaticales des mots,
- les relations de dépendance syntaxique entre les mots et les mots composés.

Le processus de normalisation consiste à supprimer les mots vides de la liste des lemmes des mots retournés par la plate-forme LIMA. Les mots vides sont identifiés à partir de leur catégorie grammaticale (prépositions, articles, ponctuations et certains adverbes). Nous considérons les mots restants comme des mots significatifs (pleins).

Nous décrivons ci-dessous uniquement les principaux modules composant l'aligneur de mots simples et nous nous focalisations sur l'étape qui concerne l'alignement de mots utilisant la détection de cognats et d'entités nommées dans les phrases source et cible. C'est cette étape qui utilise la translittération des noms propres de l'arabe vers l'écriture latine. Les modules d'alignement de mots composés et d'expressions multimots sont décrits respectivement dans (Semmar et al., 2010) et (Bouamor et al., 2012). L'alignement de mots simples se déroule selon les trois étapes suivantes:

- alignement utilisant le dictionnaire bilingue préexistant,
- alignement utilisant la détection de cognats et d'entités nommées dans les phrases source et cible,
- alignement utilisant les catégories grammaticales des mots des phrases source et cible.

L'alignement en utilisant le dictionnaire bilingue préexistant consiste, d'une part, à extraire les traductions des lemmes significatifs des phrases de la langue source en interrogeant le dictionnaire bilingue, et d'autre part, à rechercher la traduction dans la phrase cible et en comparant sa position avec celle du lemme à aligner. Si les positions des deux lemmes source et cible sont dans une même fenêtre de taille *n* respectivement dans les phrases source et cible, alors ils seront considérés traduction l'un de l'autre. Nous avons fixé expérimentalement la valeur de *n* à 6. Ainsi, le mot de la phrase source Mot_{source} est considéré comme traduction du mot de la phrase cible Mot_{cible} si les conditions [1] et [2] sont vérifiées :

$$Position\ (Mot_{source}) - 3 <= Position\ (Mot_{cible}) \quad [1]$$

$$Position\ (Mot_{cible}) <= Position\ (Mot_{source}) + 3 \quad [2]$$

Nous avons constaté aussi que beaucoup de noms arabes ne sont pas reconnus comme entités nommées par la plate-forme LIMA. Cela vient du fait que cette plateforme utilise des listes ainsi que des règles de déclencheurs pour reconnaître des entités telles que les noms de personnes, d'organisations, de lieux… mais ces listes sont limitées et plus particulièrement pour les langues peu dotées comme l'arabe. C'est pour cette raison que nous avons ajouté une étape supplémentaire à notre outil d'alignement de mots simples. Cette étape est utilisée pour permettre l'appariement des cognats présents dans les phrases source et cible. En linguistique, les cognats sont des paires de mots de langues différentes qui partagent des propriétés phonologiques, orthographiques et sémantiques. Nous pouvons étendre cette définition aux noms propres et aux expressions numériques puisqu'ils varient en général légèrement d'une langue à une autre. Plusieurs travaux ont montré que la détection et la mise en correspondance des cognats dans les textes source et cible permettent d'améliorer les résultats d'alignement au niveau des phrases (Simard et al., 1993) mais aussi des mots (Al-Onaizan, Knight, 2002) (Kondrak, 2005) (Kraif, 2001). Récemment, Frunza et Inkpen (2009) ont évalué une

méthode qui utilise 13 mesures de similarité orthographique pour identifier les cognats et les « faux amis ». Nous considérons dans une première étape comme cognats les mots dont les quatre premiers caractères sont identiques. Cette étape est simple à implémenter lorsque les phrases source et cible sont écrites avec le même script ou dans deux scripts proches. Dans notre étude, l'alignement de mots est réalisé à partir de corpus de textes parallèles français-arabe. Or ces deux langues sont écrites avec deux scripts différents. Pour détecter les cognats présents dans ces textes, nous avons utilisé le système de translittération décrit précédemment pour transformer les noms propres écrits en arabe vers l'écriture latine. Cette première étape a permis de détecter que les noms propres « Garner » et « Irak » et leur translittération respective en écriture latine «garnir» (du nom propre « غارنر ») et « irak » (du nom propre «العراق») sont des cognats. En revanche, cette étape ne permet pas d'aligner des couples de mots comme « Algérie » et « aljezeyr » (translittération du nom propre « الجزائر »). Pour ce faire, nous avons utilisé la distance Jaro–Winkler (Winkler, 1990), une mesure de similarité basée sur le nombre de lettres en commun entre le mot de la langue source *ms* et le mot de la langue cible *mc*.

$$DJ(ms,mc) = \begin{cases} 0 & si\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|ms|} + \frac{m}{|mc|} + \frac{m-t}{m}\right) & sinon \end{cases}$$

Où:

- *m* est le nombre de caractères correspondants. Deux caractères identiques des mots *ms* et *mc* sont considérés comme correspondants si leur éloignement (la différence entre leurs positions dans leurs chaînes respectives) ne dépasse pas :

$$\left(\frac{max(|ms|,|mc|)}{2}\right) - 1$$

- *t* est le nombre de transpositions. Ce nombre est obtenu en comparant le i^{ème} caractère correspondant du mot *ms* avec le i^{ème} caractère correspondant du mot *mc*. Le nombre de fois où ces caractères sont différents, divisé par deux, donne le nombre de transpositions.
- |*ms*|, |*mc*| correspondent aux longueurs en nombre de caractères des mots *ms* et *mc*.

La mesure de similarité Jaro–Winkler est une variante de la distance Jaro *DJ* (Jaro, 1989).

$$DJW(ms,mc) = DJ(ms,mc) + (lp(1 - DJ(ms,mc)))$$

Où *l* est la longueur du préfixe commun et *p* est un coefficient qui permet de favoriser les chaînes avec un préfixe commun.

Pour fixer les valeurs de *l* et *p* ainsi que le seuil pour lequel deux mots sont considérés comme cognats, nous avons utilisé un échantillon de 100 noms propres arabes translittérés en écriture latine. Dans cet échantillon, un nombre propre écrit en arabe peut avoir en moyenne 37 translittérations en écriture latine mais il existe des noms propres qui peuvent dépasser les 1 000 translittérations comme c'est le cas du mot « الجزائر » (Algérie) qui en a 1 120. Nous avons constaté que les valeurs de *l* et *p* qui permettent d'accepter le plus grand nombre de translittérations pour un nom propre sont respectivement 2 et 0,1 pour un seuil de cognats égal à 0,9. Ces paramètres fixés empiriquement permettent certes d'identifier comme cognats le mot «Algérie» et la translittération «aljezeyr» mais génèrent aussi des erreurs puisque cet aligneur considère par exemple que les mots «mohamed» et la translittération «mahmoud» du nom propre arabe «محمود» sont des cognats. Pour réduire ce type d'erreurs, nous vérifions les conditions [1] et [2] relatives aux positions des mots respectivement dans les phrases source et cible.

Certes, la détection de cognats améliore significativement les résultats de l'alignement mais ça concerne uniquement les corpus de textes ayant une forte présence de noms propres. Pour détecter de nouvelles correspondances, nous prenons en compte les paires de mots des langues source et cible qui ont les mêmes catégories grammaticales et dont les positions vérifient les conditions [1] et [2] décrites précédemment. Cette étape est particulièrement performante pour identifier les traductions des mots entourés par des mots déjà traduits.

Le tableau ci-dessous (Table 2) présente le résultat de l'alignement de mots simples et de mots composés se traduisant mot à mot de la phrase source « Le général Garner a laissé entendre que l'occupation de l'Irak ne serait pas éternelle. » et de sa traduction en langue cible «اشار الجنرال غارنر الى ان احتلال العراق لن يدوم الى الابد.».

Lemmes des mots de la phrase en langue source	Lemmes des mots de la phrase en langue cible	Etape d'alignement utilisée
général	جنرَال	Appariement de catégories grammaticales
Garner	غارنر	Appariement de cognats
laisser	أشَارَ	Appariement de catégories grammaticales
occupation	إحْتِلاَل	Dictionnaire bilingue
Irak	العِرَاق	Appariement de cognats
général_garner	جنرَال_غارنر	Mise en correspondance de mots composés
occupation_Irak	إحْتِلاَل_العِرَاق	Mise en correspondance de mots composés

TABLE 2 : Résultat de l'alignement de mots simples et composés

Ce tableau montre, d'une part, que les lemmes « entendre », « être » et « éternel » de la phrase source n'ont pas été alignés, et d'autre part, que l'alignement du lemme « laisser » n'est pas correct. En vérifiant dans le dictionnaire bilingue, nous avons trouvé plusieurs traductions pour ces lemmes, mais ils n'ont pas été alignés car ces traductions ne sont pas présentes dans la phrase cible. Cet exemple montre bien l'intérêt des alignements n:m (dans notre exemple il s'agit d'un alignement 2:1 pour le lemme « laisser entendre » qui aurait du être aligné avec le lemme «أشَارَ») même s'ils ne sont pas aussi fréquents que les alignements 1:1. Notons que le lexique bilingue construit à l'issue du processus d'alignement de mots contient les alignements corrects et incorrects, mais, les lemmes qui n'ont pas été alignés ne seront pas pris en compte. Les symboles « _ » séparant les lemmes des mots composés seront remplacés par des espaces.

4 Résultats expérimentaux et discussion

Pour illustrer l'apport de la translittération sur la qualité du lexique bilingue produit par l'alignement de mots simples et composés, nous avons évalué les résultats de l'alignement selon deux approches différentes :

- une évaluation manuelle comparant les résultats de notre aligneur de mots par rapport à un alignement de référence,
- une évaluation automatique en intégrant les résultats de notre aligneur de mots dans le corpus d'apprentissage du modèle de traduction du système de traduction statistique libre Moses (Koehn et al., 2007).

L'évaluation manuelle de l'aligneur de mots a été réalisée sur une partie composée de 1 000 phrases du corpus MD (Monde Diplomatique) français-arabe de la campagne ARCADE II (Véronis et al., 2008). Cet alignement de référence au niveau des mots simples et composés a été construit manuellement à l'aide de l'outil Yawat (Germann, 2008). Pour les métriques d'évaluation, nous avons utilisé celles du protocole défini lors de la conférence HLT/NAACL 2003 (Mihalcea, Pedersen, 2003). La table 3 résume nos résultats en termes de précision et de rappel selon que l'aligneur de mots utilise ou non l'appariement de cognats avec la translittération de noms propres. Ces résultats montrent que l'utilisation de la translittération arabe permet d'augmenter aussi bien la précision que le rappel et confirment les résultats que nous avons obtenus précédemment sur un petit corpus de 283 phrases (Saadane, Semmar, 2012) ainsi que ceux de (Kondrak et al., 2003) qui ont pu réduire de 10% le taux d'erreurs de l'alignement de mots en utilisant l'appariement de cognats. Le lexique bilingue extrait à partir des 1 000 paires de phrases en utilisant notre outil d'alignement de mots contient 16 291 entrées dont 2 023 noms propres. L'analyse de ce lexique montre qu'il contient un nombre important de doublons plus particulièrement pour les noms propres mais aussi quelques traductions de mots polysémiques. En outre, environ 53% des mots alignés se trouvaient dans le dictionnaire bilingue et 12% ont été alignés à l'aide du module d'appariement de cognats qui utilise la translittération.

Alignement de mots	Précision	Rappel	F-Mesure
sans l'appariement de cognats (sans translittération)	0,82	0,86	0,83
avec l'appariement de cognats (avec translittération)	0,87	0,88	0,87

TABLE 3 : Résultats de l'évaluation de l'alignement de mots

L'évaluation automatique de notre aligneur de mots a été réalisée en utilisant le corpus OPUS (Tiedemann, 2009) pour la paire de langues français-arabe. Ce corpus regroupe 74 067 paires de phrases parallèles extraites des résolutions des

Nations Unies. Ces résolutions citent certains noms de dirigeants, et beaucoup de noms de pays et d'organisations. Nous avons divisé ce corpus en trois parties : 70 067 paires de phrases pour l'apprentissage du modèle de traduction, 3 500 paires de phrases pour la construction du lexique bilingue en utilisant notre aligneur de mots et 500 paires de phrases pour l'évaluation du système de traduction Moses. Pour estimer le modèle de traduction du système de référence, nous avons construit un corpus d'apprentissage contenant 70 067 paires de phrases auquel nous avons ajouté les 3 500 paires de phrases utilisées pour l'alignement de mots. Pour étudier l'impact du lexique bilingue produit par l'outil d'alignement de mots intégrant la translittération sur le modèle de traduction du système Moses, nous avons ajouté ce lexique bilingue construit à partir des 3 500 paires de phrases au corpus d'apprentissage. Le modèle de traduction utilisé est appris sur les lemmes des mots composant le corpus parallèle d'apprentissage et les lemmes des mots produits par notre aligneur. Nous avons aussi entraîné un modèle de langue (tri-grammes) sur la totalité du corpus OPUS en langue arabe (74 067 phrases) en utilisant la boite à outils IRSTLM (Federico et al., 2008). Deux types de corpus de test ont été utilisés pour mener nos expérimentations : *Tout-Corpus-Test* et *Noms-propres-Corpus-Test*. Le premier corpus de test *Tout-Corpus-Test* est constitué de 500 paires de phrases parallèles extraites aléatoirement du corpus OPUS. Pour mesurer l'apport réel du lexique bilingue des noms propres translittérés, nous avons constitué un corpus de test noté *Noms-propres-Corpus-Test* où nous ne conservons que les phrases du corpus *Tout-Corpus-Test* contenant au moins un nom propre. Ce corpus contient 173 paires de phrases parallèles. La qualité de traduction du système de référence (celui qui n'intègre pas les translittérations) ainsi que celui intégrant les translittérations est évaluée sur les deux corpus de test sur la base de la métrique BLEU (Papineni et al., 2002). Nous avons préféré utiliser la métrique BLEU car elle est la plus appropriée pour évaluer les systèmes de traduction statistique à base de séquences (n-grammes) tels que Moses. Nous avons considéré qu'à chaque phrase source correspond une seule phrase de référence en langue cible. Les résultats de traduction obtenus pour les deux configurations sont regroupés dans la table 4.

Corpus d'apprentissage	Tout-Corpus-Test	Noms-propres-Corpus-Test
sans les résultats de l'appariement de cognats (sans translittération)	15,79	17,67
avec les résultats de l'appariement de cognats (avec translittération)	16,49	19.52

TABLE 4 : Résultats de traduction selon le score BLEU

Tout d'abord, nous constatons que le score BLEU obtenu est satisfaisant compte tenu de la taille du corpus d'apprentissage et du modèle de traduction utilisé et qui a été estimé sur des lemmes plutôt que sur des formes de surface (Sadat, Habash, 2006). Ce score varie en fonction du type du jeu de test. Le corpus de test *Noms-propres-Corpus-Test* qui ne considère que les phrases contenant des noms propres du lexique bilingue rapporte des scores BLEU plus élevés que le corpus de test *Tout-Corpus-Test* dans les deux configurations (corpus d'apprentissage sans l'ajout de translittération ou avec translittération). Les résultats obtenus montrent que l'intégration dans le corpus d'apprentissage du modèle de traduction des alignements obtenus par le module d'appariement de cognats utilisant la translittération a permis d'obtenir un gain de +0,70 points BLEU pour le corpus de test *Tout-Corpus-Test* et un gain de +1,85 pour le corpus de test *Noms-propres-Corpus-Test*. Ces résultats confirment ceux de (Huang et al., 2003) qui ont obtenu une F-Mesure de 81% pour l'alignement d'entités nommées à partir d'un corpus parallèle chinois-anglais et un gain de +0,06 en score NIST pour la traduction.

Pour évaluer la significativité statistique des résultats obtenus, nous utilisons la méthode par ré-échantillonnage par amorce décrite par (Koehn, 2004). Cette méthode estime la probabilité (p-valeur) qu'une différence mesurée entre les scores BLEU surgit par hasard et ce par la création à plusieurs reprises (10 fois) d'échantillons uniformes avec remise à partir des corpus de tests. Nous exploitons cette méthode pour comparer les deux configurations (corpus d'apprentissage sans l'ajout de translittération ou avec translittération) selon le corpus de test utilisé. Sur un intervalle de confiance (IC) de 95%, les résultats varient de non significatifs (quant $p > 0.05$) à hautement significatifs. Les p-valeurs obtenues sur les corpus de test *Tout-Corpus-Test* et *Noms-propres-Corpus-Test* sont respectivement de 0,02 et 0,01. Par conséquent, les améliorations apportées par l'utilisation de la translittération sont significatives dans les deux configurations de test.

5 Conclusion et travaux futurs

Nous avons décrit dans cet article, d'une part, un système de translittération des noms propres de l'écriture arabe vers l'écriture latine, et d'autre part, un outil d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe. Nous nous sommes particulièrement intéressés à l'étude de l'impact de l'utilisation de la translittération sur la qualité du lexique bilingue produit par l'outil d'alignement de mots. Pour réaliser cette étude, nous avons évalué l'outil d'alignement de mots intégrant la translittération en utilisant deux approches : une évaluation de la qualité

d'alignement à l'aide d'un alignement de référence construit manuellement et une évaluation de l'impact de cet alignement sur la qualité de traduction du système de traduction automatique statistique Moses. Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement de mots que celle de la traduction. Dans nos expérimentations sur l'outil d'alignement de mots, le modèle de traduction a été estimé sur des lemmes plutôt que sur des formes de surface qui généralement diminue la qualité de traduction plus particulièrement pour une langue morphologiquement riche comme l'arabe. De même, les traductions du lexique bilingue produit par l'outil d'alignement de mots ne sont pas pondérées, ce qui nous prive d'intégrer ce lexique directement dans la table de traduction. Nos travaux futurs sur l'alignement de mots 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 une amélioration des résultats de notre outil d'alignement en lui intégrant l'appariement d'expressions multimots et en pondérant les traductions du lexique bilingue qu'il produit. Par ailleurs, nos expérimentations sur le système de translittération ont montré que les corpus étudiés contenaient aussi des noms propres latins et que la précision de l'alignement de mots est très élevée lorsque des noms propres arabes sont présents dans les phrases source et cible. Nos travaux futurs en translittération s'orientent vers une prise en compte plus large des noms propres latins.

Références

ABDULHAY A. (2012). Constitution d'une ressource sémantique arabe à partir d'un corpus multilingue aligné. *Thèse de Doctorat de l'Université Stendhal – Grenoble III.*

ABDULJALEEL N., LARKEY L. (2003). Statistical transliteration for English-Arabic Cross Language Information Retrieval. *Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management*, New Orleans, Louisiana, 139-146.

ALGHAMDI M. (2005). Alghorithms for Romanizing Arabic names. *Journal of King Saud University - Computer and Information Sciences*,Volume 17, Riyadh, 105-128.

ALLAUZEN A., WISNIEWSKI G. (2009). Modèles discriminants pour l'alignement mot à mot. *TAL Volume 50 – n° 3/2009,* 173 – 203.

AL-ONAIZAN Y., KNIGHT K. (2002). Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th ACL Conference*, USA.

BARKAT-DEFRADAS M., HAMDI R., PELLEGRINO F. (2004). De la caractérisation linguistique à l'identification automatique des dialectes arabes. *Proceedings of MIDL 2004,* 51-56.

BEESLEY K. R., KARTTUNEN L. (2003). Finite State Morphology. *Stanford, CA: CSLI Publications*.

BESANÇON R., DE CHALENDAR G., FERRET O., GARA F., LAIB M., MESNARD O., SEMMAR N. (2010). LIMA: A Multilingual Framework of Linguistic Analysis and Linguistic Resources Development and Evaluation. *Proceedings of LREC 2010,* 3697-3704.

BOUAMOR D., SEMMAR N., ZWEIGENBAUM P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine. *Proceedings of the Eighth international conference on Language Resources and Evaluation (LREC), Turkey.*

BROWN P. F., PIETRA S. A. D., PIETRA V. J. D., MERCER R. L. (1993). The mathematics of statistical machine translation : parameter estimation. *Computational Linguistics*, *Volume 19, Number 2,* 263-311.

DAILLE B., GAUSSIER E., LANGE J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, 515-521.

DEBILI F., ZRIBI A. (1996). Les dépendances syntaxiques au service de l'appariement des mots. *Actes du 10ème Congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*.

FEDERICO M., BERTOLDI N., CETTOLO M. (2008). IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models. *Proceedings of Interspeech, Australia, 2008.*

FRUNZA O., INKPEN D. (2009). Identification and Disambiguation of Cognates, False Friends, and Partial Cognates Using Machine Learning Techniques. *International Journal of Linguistics, Vol. 1*.

GAUSSIER E., LANGE J. M. (1995). Modèles statistiques pour l'extraction de lexiques bilingues. *Traitement Automatique des Langues, Volume 36. ATALA,* 133-155.

GERMANN U. (2008). Yawat: Yet Another Word Alignment Tool. *Proceedings of ACL 2008, Columbus,* 20-23

GUIDERE M. (2004). Le traitement de la parole et la détection des dialectes arabes. *Langues stratégiques et défense nationale, Publications du CREC, Saint-Cyr,* 53–75.

HUANG F., VOGEL S., WAIBEL A. (2004). Improving named entity translation combining phonetic and semantic similarities. *Proceedings of HLT-NAACL 2004,* 281-288.

HUANG F., VOGEL S., WAIBEL A. (2003). Automatic Extraction of Named Entity Translingual Equivalence Based on Multi-feature Cost Minimization. *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics (ACL'03), Workshop on Multilingual and Mixed-language Named Entity Recognition, Sapporo, Japan.*

HULDEN M. (2009). Foma: a Finite-State Compiler and Library. *Proceedings of: EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics,* Athens, Greece, 29-32.

JARO M. A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Association 84*, 414–420.

JIANG L., ZHOU M., CHIEN L. F., NIU C. (2007). Named entity translation with web mining and transliteration. *Proceedings of the 20th International Joint Conference on Artif cial Intelligence*, 1629-1634.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORGAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A., HERBST E. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of ACL* 2007, 177-180.

KOEHN P. (2004). Statistical significance tests for machine translation evaluation. *Proceedings of EMNLP 2004.*

KNIGHT K., GRAEHL J. (1997). Machine transliteration. *Journal version Computational linguistics,* 24(4), 599-612.

KONDRAK G. (2005). Cognates and Word Alignment in Bitexts. *Proceedings of the Tenth Machine Translation Summit (MT Summit X), Thailand*.

KONDRAK G., MARCU D., KNIGHT K. (2003). Cognates Can Improve Statistical Translation Models. *Proceedings of HLT-NAACL 2003*, 46-48.

KRAIF O. (2001). Exploitation des cognats dans les systèmes d'alignement bi-textuel: architecture et évaluation. *TAL, 42(3),* 833-867.

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

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

OZDOWSKA S., CLAVEAU V. (2006), Inférence de règles de propagation syntaxique pour l'alignement de mots. *TAL, Volume 47, n°1 ATALA,* 167-186.

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

OKITA T., GUERRA M., ALFREDO GRAHAM Y., WAY A. (2010). Multi-word expression sensitive word alignment. *Proceedings of the 4th International Workshop on Cross Lingual Information Access at COLING 2010, 26–34*.

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

RILEY M., ALLAUZEN C., MARTIN J. (2009). OpenFst: An Open-Source, Weighted Finite-State Transducer Library and its Applications to Speech and Language. *Proceedings of NAACL HLT 2009: Tutorials,* 9–10.

POULIQUEN B., STEINBERGER R. (2007). Acquisition and Use of Multilingual Name Dictionaries. *Proceedings of the Workshop Acquisition and Management of Multilingual Lexicons (AMML'2007) - RANLP'2007, Bulgaria.*

SAÂDANE H., SEMMAR N. (2012). Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe. *Actes TALN 2012*, 127-140.

SADAT F., HABASH N. (2006). Combination of Arabic Preprocessing Schemes for Statistical Machine Translation. *Proceedings of ACL 2006,* 1-8.

SEMMAR N., SERVAN C., DE CHALENDAR G., LE NY B. (2010). A Hybrid Word Alignment Approach to Improve Translation Lexicons with Compound Words and Idiomatic Expressions. *Proceedings of the 32nd Translating and the Computer conference, England*.

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

SHAO L., NG H. T. (2004). Mining new word translations from comparable corpora. *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04),* 618-624.

SHERIF T., KONDRAK G. (2007). Bootstrapping a stochastic transducer for Arabic-English transliteration extraction. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007),* 864-871.

SIMARD M., FOSTER G. F., ISABELLE P. (1993). Using cognates to align sentences in bilingual corpora. *Proceedings of the Conference of the Centre for Advanced Studies on Collaborative Research,* 1071-1082.

STALLS B., KNIGHT K. (1998). Translating names and technical terms in Arabic text. *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages,* Montreal, Québec, 34-41.

TAO T., YOON S. Y., FISTER A., SPROAT R., ZHAI C. (2006). Unsupervised named entity transliteration using temporal and phonetic correlation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06),* 250-257.

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,* 237-248.

TUFIS I., ION R. (2007). Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. *Proceedings of the 4th International Conference on Speech and Dialogue Systems,* 183–195.

VERONIS J., HAMON O., AYACHE C., BELMOUHOUB R., KRAIF O., LAURENT D., NGUYEN T. M. H., SEMMAR N., STUCK F., ZAGHOUANI W. (2008). Arcade II Action de recherche concertée sur l'alignement de documents et son évaluation. *Chapitre 2, Editions Hermès.*

VINTAR S., FISIER D. (2008). Harvesting multi-word expressions from parallel corpora. *Proceedings of LREC, Morocco.*

WINKLER W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Section on Survey Research Methods, American Statistical Association,* 354–359.

YASER A. O., KNIGHT K. (2002). Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL'02),* 400-408.