Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe

Houda Saadane[1] *Nasredine Semmar*[2]
(1) LIDILEM, Université de Grenoble, 38400 Grenoble Cedex 9
(2) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, 91191 Gif-sur-Yvette Cedex
houda.saadane@e.u-grenoble3.fr, nasredine.semmar@cea.fr

RESUME

Dans cet article, nous nous intéressons à l'utilisation de la translittération arabe pour l'amélioration des résultats d'une approche linguistique d'alignement de mots simples et composés à partir de corpus de textes parallèles français-arabe. Cette approche utilise, d'une part, un lexique bilingue et les caractéristiques linguistiques des entités nommées et des cognats pour l'alignement de mots simples, et d'autre part, les relations de dépendance syntaxique pour aligner les mots composés. Nous avons évalué l'aligneur de mots simples et composés intégrant la translittération arabe en utilisant deux procédés : 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 la traduction en faisant appel au système de traduction automatique statistique Moses. Les résultats obtenus montrent que la translittération améliore aussi bien la qualité de l'alignement que celle de la traduction.

ABSTRACT

Using Arabic transliteration to improve word alignment from French-Arabic parallel corpora

In this paper, we focus on the use of Arabic transliteration to improve the results of a linguistic word alignment approach from parallel text corpora. This approach uses, on the one hand, a bilingual lexicon, named entity and cognates linguistic properties to align single words, and on the other hand, syntactic dependency relations to align compound words. We have evaluated the word aligner integrating Arabic 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 Arabic transliteration improves the quality of both alignment and translation.

MOTS-CLES : Translittération, alignement de mots, construction de dictionnaires multilingues, traduction automatique, recherche d'information interlingue.
KEYWORDS : Transliteration, word alignment, multilingual lexicons construction, machine translation, cross-language information retrieval.

1 Introduction

La translittération consiste à substituer à chaque graphème un système d'écriture, un autre graphème ou un groupe de graphèmes d'un autre système d'écriture,

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 127–140,
Grenoble, 4 au 8 juin 2012.

indépendamment de la prononciation.

La translittération connait un essor important en raison du caractère de plus en plus multilingue de l'Internet et des besoins exponentiels dans le domaine de la recherche d'information interlingue. Cela est d'autant plus vrai pour la recherche d'entités nommées (noms de personnes, de lieux, de sociétés, d'organisations, etc.), mais ces dernières présentent une pluralité de formes écrites, d'orthographes et de transcriptions selon les langues et les pays. Le cas des noms propres en arabe illustre cette situation complexe et multiforme. Le meilleur exemple pour montrer cette pluralité est le nom معمر القذافي (Mouammar Kadhafi) qui est transcrit en latin par plus de 60 formes, parmi lesquelles : Muammar Qaddafi, Mo'ammar Gadhafi, Muammer Kaddafi, Moammar El Kadhafi, etc.

Cet article décrit un système de translittération automatique de noms arabes en écriture latine et montre une utilisation concrète de la translittération arabe en alignement de mots à partir de corpus de textes parallèles dans le but d'améliorer la qualité des lexiques bilingues ainsi construits.

Nous présentons dans la section 2 un résumé de l'état de l'art dans le domaine de la translittération. Dans la section 3, nous décrivons les approches que nous avons utilisées pour développer notre système de translittération automatique des noms arabes voyellés et non voyellés vers les différentes transcriptions possibles en écriture latine. Nous montrons dans la section 4 comment cette translittération est utilisée pour améliorer les résultats d'un outil d'alignement de mots. Nous présentons en section 5 les résultats que nous avons obtenus en précisant les taux d'amélioration de la qualité de l'alignement et de la traduction. La section 6 conclut notre étude et présente nos travaux futurs.

2 État de l'art

Le problème de la translittération a intéressé les spécialistes dans plusieurs langues, mais cet intérêt est relativement récent et lié au développement croissant de l'utilisation de l'Internet. 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 indispensables pour la recherche d'information interlingue et la traduction automatique. Citons notamment (Yaser et Knight, 2002) et (Sherif et Kondrak, 2007), qui ont travaillé sur l'alignement arabe-anglais, (Tao et al., 2006) qui ont travaillé sur l'arabe, le chinois et l'anglais ainsi que (Shao et 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.

On trouve ainsi 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 (Katakana) vers le système d'écriture latin (Knight et Graehl, 1997).

Cette approche a été adaptée par (Stalls et 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.

Cela a conduit certains chercheurs à pallier cette carence par un recours à la technique statistique. C'est le cas du système de translittération des noms anglais vers l'arabe proposé par (Abduljaleel et Larkey, 2003). Mais celui-ci a montré également ses limites parce qu'il est basé sur le calcul de la forme la plus probable, censée être la forme correcte, ce qui n'est pas vrai pour tous les pays arabes ni pour tous les dialectes.

Pour contourner la difficulté de la prononciation et le problème des variantes dialectales, (Alghamdi, 2005) a proposé un système de translittération en écriture anglaise des noms arabes voyellés. 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. Mais cette approche cumule les inconvénients des deux précédentes : non seulement elle ne prend pas en compte les prononciations non répertoriées dans le dictionnaire, mais en plus elle est normative par le fait qu'elle ne propose qu'une seule translittération pour un nom donné. L'objectif de l'auteur semblerait être de favoriser l'adoption d'un standard de translittération, mais cela ne peut être le résultat d'une initiative individuelle et isolée.

En réalité, l'état actuel de la recherche dans ce domaine ne rend pas compte de la complexité du problème de la transcription et de la translittération, lequel touche autant à l'oralité qu'à la scripturalité dans deux ou plusieurs systèmes linguistiques en même temps. En effet, transcrire un nom ou un prénom d'un système linguistique source vers un système d'écriture cible, est une tâche délicate qui nécessite un certain nombre d'opérations exigeant de prendre en considération un ensemble de propriétés morphologiques, phonologiques et sémantiques. Ces opérations sont nécessaires pour assurer un processus de translittération robuste, notamment pour des applications de sécurité, de vérification d'identité, ou encore de recherche d'informations sur Internet.

Or, très peu d'études prennent en considération le lien :

- entre phonologie comparée et transcription interlingue;
- entre graphématique comparée et translittération multilingue;
- 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. C'est le cas notamment des travaux de (Guidère, 2004) et de (Barkat-Defradas et al., 2004).

3 Translittération en caractères latins des noms écrits en arabe standard

Le système d'écriture de la langue arabe standard est constitué d'un alphabet de 28 lettres, dont 25 consonnes et 3 voyelles, celles-ci pouvant être courtes ou longues en fonction du mot.

Il existe également des phénomènes morphologiques et phonologiques particuliers dont il faut tenir compte dans la translittération tels que le dédoublement des consonnes,

parfois matérialisé dans l'écriture arabe par la «shadda», et le redoublement des voyelles, parfois matérialisé dans l'écriture arabe par le «tanwin». Mais l'écriture arabe moderne présente la particularité de ne pas marquer dans les textes –de manière générale– ni le dédoublement ni les voyelles courtes, ce qui constitue l'une des principales sources d'ambiguïté pour les systèmes de translittération.

3.1 Méthodologie de construction du translittérateur

Nous avons choisi une méthodologie ascendante pour la construction de notre translittérateur. En d'autres termes, nous avons commencé par faire un recensement des translittérations existantes pour chaque lettre de l'alphabet arabe standard à partir des normes et des usages observés sur Internet. 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 actuellement en usage dans les écrits utilisant l'alphabet latin.

Nous faisons figurer dans le tableau suivant quelques équivalences graphématiques établies à partir de cette étude sur corpus :

Lettre arabe	Équivalent en écriture latine	Lettre arabe	Équivalent en écriture latine
ء	a	غ	Gh, gh, Ǧ, ǧ, ḡ
ا	A, a, ä, â, á, ā, e, ê	ف	F, f, ph
ب	B, b	ق	Q, q, C, c, K, k
ت	T, t	ك	K, k, C, c
ث	Th, th, t, t̲	ل	L, l

TABLE 1 – Exemples d'équivalences graphématiques entre les alphabets arabe et latin

L'étude sur corpus a également permis de constater que certaines lettres arabes, sans équivalent graphématique dans l'écriture latine, étaient transcrites par le biais de chiffres arabes dans les textes écrits en caractères latins. Ce type de translittération est particulièrement utilisé dans les messages téléphoniques (SMS) et dans les sites web sociaux en Europe et au Moyen Orient. Le tableau suivant récapitule ces équivalences alphanumériques pour les lettres concernées de l'alphabet arabe :

Lettre	ء	ح	خ	ص	ض	ط	ظ	ع	غ	ق
Équivalence alphanumérique	2	7	7'	9	9'	6	6'	3	3'	8

TABLE 2 – Équivalences alphanumériques dans les textes écrits en alphabet latin

Ainsi, en combinant ces deux types de représentation symbolique, on peut rencontrer dans les textes des translittérations qui illustrent ces différentes équivalences pour des noms et des prénoms courants dans le monde arabe :

Nom en arabe	منى	عدنان	حنان	طارق
Exemple d'équivalents en écriture latine	Mouna ou Mona…	Adnane ou 3adnan…	Hanane ou 7anan…	Tarek ou 6ariq…

TABLE 3 – Exemples de noms et prénoms arabes

Cette variation dans les usages translittérationnels, source d'ambiguïté lors du traitement automatique et de la recherche d'information, s'explique par trois types de raisons :

Tout d'abord, des raisons historiques puisque certains pays arabes ont été colonisés ou placés sous mandat français ou britannique pendant une période plus ou moins longue selon les pays et ont, par conséquent, gardé de cette période des traces dans leur vocabulaire, dans leur prononciation et dans la manière dont ils ont tendance à translittérer les noms et les prénoms. Ainsi, l'influence du système linguistique et graphématique du français est perceptible dans les usages translittérationnels des pays du Maghreb, de manière plus ou moins forte selon les pays. Il en est de même des pays du Proche et du Moyen-Orient par rapport à l'influence britannique ou américaine.

Ensuite, pour des raisons politiques puisqu'il n'existe pas de norme commune ni de stratégie unifiée dans le domaine de la translittération pour ce qui est de la langue arabe. Cela a conduit chaque écrivain ou scripteur à s'appuyer sur la prononciation dialectale qui lui était la plus familière pour transcrire les noms arabes. L'exemple le plus célèbre est celui de Laurence d'Arabie qui, pour transcrire le nom de la ville de Djeddah (جدة) en Arabie Saoudite, utilise : 25 fois l'orthographe « Jeddah », 6 fois l'orthographe « Jidda », et 1 fois l'orthographe « Jedda », et cela dans le même ouvrage (1926). Laurence d'Arabie justifie cette variation dans la translittération de la manière suivante : « On ne peut pas transcrire correctement et de la même façon un nom arabe à cause des consonnes qui diffèrent des consonnes latines et des voyelles dont la prononciation diffère d'une région à une autre. » (Alsalman et al., 2007). Cela est d'autant plus vrai que les différentes orthographes données par Laurence d'Arabie diffèrent de l'usage actuel en Arabie Saoudite pour la transcription du nom de cette même ville : « Jaddah ».

Enfin, pour des raisons dialectologiques puisqu'il existe une telle variété de parlers régionaux et locaux dans le monde arabe qu'il est impossible de retrouver la même prononciation d'un pays à l'autre et d'une région à l'autre. Ainsi par exemple, l'un des prénoms arabes les plus répandus, celui du Prophète Muhammad (محمد) – transcrit en français Mahomet depuis l'époque moderne – possède une dizaine de prononciations – et donc de transcriptions – différentes. Citons notamment : Mohamed, Mouhammad, Muhamed, Mhamed, M'Hamed, Muhammad, etc. Même lorsque ce prénom est voyellé (مُحَمَّدٌ), il présente plusieurs translittérations dans les textes : Muhamad, Mouhamad, Mohamad, Mehammad, Mehammade.

Cette variation dans les translittérations possibles selon les dialectes est parfois accompagnée par l'utilisation de caractères spéciaux dans certaines régions ou pays

arabes. Citons comme exemples les noms suivants qui présentent des formes non conventionnelles en écriture latine : Mu`ammar, Mabrūk, Muṣṭafá, Ismā`īl, Hâdî.

Tous ces phénomènes nécessitent une observation fine en amont du traitement pour identifier les cas problématiques et construire des règles efficaces permettant l'automatisation du processus de translittération des noms arabes en temps réel.

3.2 Fonctionnement du translittérateur de l'arabe vers le latin

Le module de translittération de l'écriture arabe vers l'écriture latine est fondé sur les automates d'états finis. Cela signifie qu'il est constitué d'états et de transitions. Son fonctionnement 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.

A l'issue de la lecture, l'automate produit une réponse « oui » ou « non », c'est-à-dire qu'il accepte (oui) ou rejette (non) l'entrée en question : voyellée ou non-voyellée. Ensuite, il traite l'entrée de la manière suivante : si voyellée, il supprime les voyelles avant de translittérer le nom; si non-voyellée, il procède directement à la translittération du nom. 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. Ces règles visent 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 ?

On sait à cet égard que le nom d'une personne contient plusieurs éléments en arabe. Il est constitué en principe de quatre composants principaux :

1. La « Kunya » (particule d'usage) : généralement composée de « Abou » (père de…), suivi du nom d'un enfant ou bien de « Oum » (mère de + nom d'un enfant de la famille). Exemple : « Abou Omar » (Père d'Omar), «Oum Mohamed» (Mère de Mohamed), etc.
2. Le « Ism » (Prénom) : par exemple, Omar, Ali, Mohamed, Khaled, Abdallah, etc. Il indique parfois l'origine ethnique ou confessionnelle de celui qui le porte : par exemple, « Omar » est un prénom typiquement sunnite ; « Rustam » est un prénom typiquement iranien ; « Arslan » est typiquement turc, etc.
3. Le « Nasab » (particule généalogique) : chaque nom est précédé par « Ibn » ou «Bin/Ben» («Bint/Bent» pour les femmes). Il indique la filiation généalogique exacte de l'individu concerné. Les Arabes remontent parfois très loin dans l'indication des ancêtres pour éviter les confusions entre personnes : ex. Muhammad Bin Abdallah Bin Salih Bin Said, etc.
4. La « Nisba » (suffixe d'origine) : ce suffixe renvoie en principe à la tribu ou au clan dans la généalogie ancienne mais aujourd'hui, il désigne surtout le lieu de naissance des individus : Maghribi (né au Maroc), Libi (né en Libye), Masri (né en Égypte), etc. La « Nisba » est toujours précédée de l'article [Al-] et se termine par le suffixe [i]. Elle indique la résidence territoriale initiale des personnes, ou encore leur nationalité.

Selon la forme d'entrée, on applique d'abord des règles adéquates pour transcrire la

partie qui ne constitue pas le nom à proprement parler (particules), puis on applique 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, Si le mot est composé par Abd (عبد) + Al (ال) + Nom (رحيم), 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-Rahim (عبد الرحيم) ;
- Génération de toutes les formes de translittération possibles pour ces trois éléments :

Nom propre arabe	Translittérations
عبد الرحيم	Abd Al-Rahim
	Abd Al Rahim
	Abd al-Rahim
	Abd al Rahim
	Abd El-Rahim
	Abd El Rahim
	Abd el-Rahim
	Abd el Rahim
	Abd Ar-Rahim
	Abd Ar Rahim
	Abd Ar-Rahîm
	Abd ar-Rahim

TABLE 4 – Quelques formes de translittération pour le nom propre عبد الرحيم

Une étape intermédiaire s'ajoute afin de procéder à d'autres traitements, pour ne pas occulter l'un des problèmes très difficile de la transcription, comme la transcription de certains noms propres qui changent totalement phonétiquement pour des raisons religieuses ou autres : c'est le cas de Moussa qui est traduit par Moïse, Yussuf par Josef,

Yaakoub par Jackoub, Hawa par Eve, etc. Cette étape consiste à fournir ces transcriptions dans une liste.

Une fois générée la liste triée des noms translittérés, on procède à deux types de traitements :

- Normalisation de la liste des noms en écriture latine : cette phase consiste à effectuer certains traitements sur la sortie du nom en écriture latine tels que la suppression des caractères spéciaux (diacritiques et chiffres) et l'ajout de la majuscule au début de nom propre, étant donné que les majuscules n'existent pas dans l'écriture arabe des noms. Cette notion de majuscule est conservée seulement dans le cas d'une utilisation dans des bases de données, mais elle n'est pas ajoutée pour les moteurs de recherche usuels, qui ne considèrent pas la casse comme pertinente;
- Pondération de la liste des noms en écriture latine : cette étape consiste à attribuer un poids aux règles qui ont servi à la génération de la liste, afin de pouvoir afficher les résultats en sortie du plus probable vers le moins probable, ou inversement. Pour réaliser cette pondération, nous utilisons le moteur de recherche Google en notant à chaque fois le nombre d'occurrences pour chaque forme générée du nom propre : par exemple pour le prénom arabe جمال (jamal), le système génère trois translittérations distinctes et attestées dans les textes (Djamel, Jamel, Gamel) et le calcul de fréquences fournit les résultats suivants :

Forme translittérée du nom en écriture latine	Nombre moyen d'occurrences du nom sur le moteur de recherche Google
Djamel	4000000
Jamel	5500000
Gamel	500000

TABLE 5 – Résultats pour les formes translittérées du prénom جمال

Du point de vue de la pondération, cet exemple permet de constater que la lettre arabe (ج) est transcrite, en termes de fréquence, majoritairement par la lettre (J), puis par la graphie (Dj), puis par la lettre (G).

Cette procédure a été appliquée à toutes les formes de translittération des caractères arabes. Elle a permis d'établir une liste d'équivalences pondérée au niveau des graphèmes, qui sert à afficher les résultats en sortie du plus probable vers le moins probables.

4 Utilisation de la translittération en alignement de mots

L'outil d'alignement de mots simples et composés utilisé dans cette étude est décrit dans (Semmar et Laib, 2010). Cet outil utilise les ressources et les modules suivants :

- un lexique bilingue français-arabe composé de 124581 entrées. Les entrées de ce

lexique sont utilisées comme des points d'ancrage pour réduire l'espace de recherche des mots à aligner dans les phrases source et cible ;
- un module pour l'appariement des entités nommées présentes dans les phrases source et cible ;
- un module permettant d'apparier les catégories grammaticales des mots composant les phrases source et cible. Ce module utilise les positions des mots à apparier par rapport aux entrées du lexique bilingue et des entités nommées déjà alignées ;
- un module pour l'appariement des mots composés identifiés à partir des mots simples déjà alignés et les relations de dépendance syntaxique entre ces mots.

Les entrées de cet outil d'alignement sont les sorties (résultats) 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 :

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

Nous avons constaté lors de l'alignement de mots à partir de corpus de textes parallèles anglais-arabe ou français-arabe que beaucoup de noms arabes ne sont pas reconnus comme entités nommées par la plate-forme LIMA. Cela vient du fait que cette plate-forme 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 un module supplémentaire à notre outil d'alignement de mots. Ce module est utilisé pour permettre l'appariement des cognats présents dans les phrases source et cible. Nous considérons comme cognats les mots dont les quatre premiers caractères sont identiques.

Cette étape utilise la translittération des noms propres et permet de détecter, par exemple, que le nom propre « Jackson » et la translittération du mot arabe « جاكسون » («jackson») sont des cognats. En revanche, cet algorithme ne permet pas de détecter des couples de mots comme « blair » et « bleer » (translittération du mot arabe «بْلِير»). Pour ce faire, nous avons défini une similarité basée sur le nombre de lettres en commun. Ceci permettra de détecter les couples de mots cités précédemment ainsi que les noms propres et les expressions numériques. L'algorithme de détection de cognats a été adapté pour ne sélectionner que les mots de taille proche et avec un nombre important de caractères en commun sans tenir compte de l'ordre de ces caractères. L'adaptation de cet algorithme a été réalisée en ajoutant les deux paramètres « Ratio_mots » et « Ratio_cognats » définis comme suit :

$$Ratio\_mots = (Nombre\ de\ caractères\ du\ mot\ court)/(Nombre\ de\ caractères\ du\ mot\ long)$$

$$Ratio\_cognats = (Nombre\ de\ caractères\ en\ commun)/(Nombre\ de\ caractères\ du\ mot\ court)$$

Deux mots sont cognats si Ratio_mots est supérieur à 0,8 et Ratio_cognats est supérieur à

0,5. Les valeurs de ces deux paramètres ont été fixées empiriquement.

Cette adaptation permet certes d'identifier comme cognats le mot « blair » et la translittération « bleer » mais il génère aussi des erreurs comme c'est le cas du couple de mots « mohamed » et la translittération « mahmoud ». Pour réduire le taux d'erreurs de ce module, nous avons ajouté un critère supplémentaire relatif aux positions des deux mots dans les phrase source et cible.

Le tableau 6 présente le résultat de l'alignement des mots simples et composés de la phrase source « M. Blair a imposé des frais d'inscription élevés à l'université qui ont introduit une sélection par l'argent. » et sa traduction en langue cible « فرض بلير رسوم تسجيل مرتفعة في الجامعة مما ادى الى اختيار الطلاب على قاعدة المال.».

Lemmes des mots simples et composés de la phrase source	Lemmes des mots simples et composés de la phrase cible
Blair	بْلير
imposer	فْرَضَ
frais	رَسْم
inscription	تَّسْجِيل
élevé	مُرْتَفِع
université	جَامِعَة
introduire	أدَّى
sélection	إخْتِيَار
argent	مَال
frais_inscription	رَسْم_تَّسْجِيل

TABLE 6 – Résultats de l'alignement de mots simples et composés

Le mot « Blair » a été aligné à l'aide de l'appariement de cognats après translittération, les mots « frais », « élevé » et « introduire » ont été alignés à l'aide de l'appariement de catégories grammaticales et les autres mots existent dans le lexique bilingue. Le mot composé « frais_inscription » a été aligné en utilisant les alignements de ces composants « frais » et « inscription ». Notons que le mot arabe « قاعدة » (base) n'a pas de mot qui lui correspond dans la phrase en langue source (français).

5 Résultats expérimentaux

Pour illustrer l'apport de la translittération sur la qualité de 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 Moses (Koehn et al., 2007).

L'évaluation manuelle de l'aligneur de mots a été réalisée sur une partie composée de 283 phrases du corpus MD (Monde Diplomatique) français-arabe de la campagne ARCADE II (Veronis et al., 2008). Le choix d'une telle taille de corpus s'explique par le fait que la constitution de l'alignement de référence est une tâche coûteuse puisque l'identification des alignements des mots simples et composés est réalisée manuellement sur les 283 phrases. Pour les métriques d'évaluation, nous avons utilisé celles du protocole défini lors de la conférence HLT/NAACL 2003 (Mihalcea et Pedersen, 2003).

Le tableau 7 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 arabes. Ces résultats montrent que l'utilisation de la translittération arabe permet d'augmenter aussi bien la précision que le rappel.

Alignement de mots	Précision	Rappel	F-mesure
sans l'appariement de cognats	0,85	0,80	0,82
avec l'appariement de cognats	0,88	0,85	0,86

TABLE 7 – Résultats de l'évaluation de l'alignement de mots

Certes, la taille insuffisante du corpus utilisé pour l'évaluation de notre aligneur de mots ne permet pas de mesurer quantitativement l'apport de la translittération mais les résultats obtenus indiquent clairement qu'il y a une amélioration de la qualité de l'alignement.

La non disponibilité d'un alignement de référence d'une taille significative pour les mots simples et composés ne nous permet pas de comparer notre approche avec les différents travaux de l'état de l'art. C'est la raison pour laquelle, nous avons décidé d'étudier l'apport de l'utilisation de la translittération en alignement de mots en intégrant les résultats de notre aligneur de mots dans le corpus d'apprentissage du modèle de traduction du système Moses. 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 (Koehn et Hoang, 2007).

Le corpus initial d'apprentissage est composé de 10000 paires de phrases français-arabe issues du corpus ARCADE II auquel nous avons ajouté environ 10000 paires de mots simples et composés correspondant aux résultats de l'aligneur de mots intégrant l'appariement de cognats à l'aide de la translittération sur 500 paires de phrases français-arabe. Nous avons aussi spécifié un modèle de langue pour la langue cible en utilisant la

totalité des phrases arabes du corpus ARCADE II.

La performance du système de traduction statistique Moses est évaluée à l'aide du score BLEU sur un corpus de test composé de 250 paires de phrases. Pour chaque phrase source, une seule phrase de référence en langue cible est considérée. Les résultats de traduction obtenus sont regroupés dans le tableau 8.

Corpus d'apprentissage	BLEU
sans les résultats de l'appariement de cognats (sans translittération)	12,50
avec les résultats de l'appariement de cognats (avec translittération)	12,82

TABLE 8 – Résultats de traduction selon le score BLEU

Ces résultats 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 $+0{,}32$ points BLEU.

Il est difficile de dire à ce stade si ce gain en score BLEU induit une amélioration significative de la qualité de la traduction au vu de la faible valeur de ce score liée à la taille des corpus d'apprentissage utilisés (uniquement 10000 paires de phrases pour l'apprentissage du modèle de traduction et environ 11000 phrases pour l'apprentissage du modèle de la langue cible). Nous pourrons conclure tout de même que la translittération améliore la performance de l'aligneur de mots, quelle que soit la manière d'évaluer les résultats, manuellement ou automatiquement.

6 Conclusion

Dans cet article, nous avons décrit un système de translittération des noms propres de l'écriture arabe vers l'écriture latine. Ce système a été utilisé dans un processus d'alignement de mots à partir de corpus de textes français-arabe. Ce processus se déroule en deux phases : d'abord, les mots simples sont alignés en utilisant un lexique bilingue et certaines propriétés de ces mots (positions, catégories grammaticales, entités nommés et cognats), ensuite les mots composés sont alignés en les identifiant à l'aide des relations de dépendance syntaxique reliant leurs composants. Ce processus donne des résultats très satisfaisants lorsque la translittération arabe est utilisée pour apparier les noms propres présents dans les phrases source et cible. Nos travaux futurs s'orientent, d'une part, vers une évaluation à une large échelle de notre outil d'alignement en vue de consolider les résultats déjà obtenus, et d'autre part, vers une translittération géo-localisée pour identifier comment les différentes translittérations peuvent fournir des indications sur l'origine et/ou sur le profil de celui qui les utilise (francophone ou anglophone, du Maghreb ou du Macherek, du nord ou du sud...).

Références

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

ALGHAMDI, M. (2005). Alghorithms for Romanizing Arabic names. *In Journal of King Saud University: Computer Sciences and Information*, n° 17, 2005, Riyadh, pages 1–27.

ALSALMAN, A., ALGHAMDI, M., ALHUQAYL, K. et ALSUBAI, S. (2007). Romanization System for Arabic Names. *In Proceedings of The First International Symposium on Computer and Arabic Language (ISCAL – 07)*, Riyadh, pages 214–227.

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

BESANÇON, R., DE CHALENDAR, G., FERRET, O., GARA, F., LAIB, M., MESNARD, O. et SEMMAR, N. (2010). Lima: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. *In Proceedings of LREC 2010*, Malta.

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

KNIGHT, K. et GRAEHL, J. (1997). Machine transliteration. *In Journal version Computational Linguistics, 24(4), 1997*, pages 599–612.

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

KOEHN, P. et HOANG, H. (2007). Factored Translation Models. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (ACL 2007),* Prague, pages 868–876.

MIHALCEA, R. et PEDERSEN, T. (2003). An evaluation exercise for word alignment. *In Proceedings of the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, pages 10–10.

SEMMAR, N. et LAIB, M. (2010). Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons. *In Proceedings of LREC 2010: Workshop on Language Resources and Human Technologies for Semitic Languages,* Malta.

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

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

STALLS, B. et KNIGHT, K . (1998). Translating names and technical terms in arabic text. *In Proceedings of the COLING/ACL Workshop on Computational Approches to Semitic Languages,* Montreal.

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

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

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