

Détection des couples de termes translittérés à partir d'un corpus parallèle anglais-arabe

Wafa Neifar^{1,2} Thierry Hamon^{1,3} Pierre Zeigenbaum¹
Mariem Ellouze Khemakhem² Lamia Hadrich Belguith²

(1) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France

(2) MIRACL Laboratory, Sfax University, B.P-3018 Sfax, Tunisie

(3) Université Paris 13, Sorbonne Paris Cité, F-93430, Villetaneuse, France

neifar,hamon,pz@limsi.fr, mariem.ellouze@planet.tn, l.Belguith@fsegs.rnu.tn

RÉSUMÉ

Nous présentons une méthode pour extraire des couples de termes médicaux translittérés de l'anglais en caractères arabes. Nous avons proposé un processus de construction des translittérations de termes anglais en arabe. Celui-ci s'appuie sur une étude en corpus pour la création d'une table de correspondances des caractères anglais en arabe mais aussi sur des règles de conversion qui tiennent compte de certaines particularités de la langue arabe comme l'agglutination et la non-voyellation. Nous avons évalué l'apport de l'utilisation de la translittération pour identifier des couples de termes anglais-arabe sur un corpus parallèle de textes médicaux. Les résultats montrent que parmi 137 couples de mots anglais-arabe extraits, 120 sont jugés corrects (soit 87,59%), dont 107 représentent des couples de termes médicaux (soit 89,16% des translittérations correctes et 78,10% des résultats).

ABSTRACT

Detection of transliterated pairs of terms from an English-Arabic parallel corpus

We aim at extracting pairs of medical terms which are transliterated from English to Arabic characters. We propose a method for building transliteration of English terms in Arabic. This method relies on the study of a corpus in order to define a conversion table but also on rules which take into account specificities of Arabic language such as the agglutination and the absence of diacritisation. We have evaluate the use of this transliteration method for identify pairs of English-Arabic terms in a medical parallel corpus. The results show that among the 137 extracted English-Arabic word pairs, 120 are considered as correct transliterations (87.59%), among which 107 pairs represent medical terms (this represents 89.16% of the correct transliterations and 78.10% of the results).

MOTS-CLÉS : Extraction terminologique bilingue, alignement de mots, translittération, corpus parallèle..

KEYWORDS : Bilingual terminology extraction, word alignment, transliteration, parallel corpora..

1 Introduction

Bien que l'arabe standard moderne (MSA) soit la langue officielle de 26 pays, son usage est variable dans les domaines de spécialité. Ainsi, alors qu'en médecine, la langue utilisée lors de la pratique et de l'enseignement est la langue française ou anglaise (Samy *et al.*, 2012), l'arabe est également

une langue officielle de l'OMS (Organisation Mondiale de la Santé) ou d'autres organismes internationaux. De même, l'arabe est la langue de travail dans d'autres domaines comme l'agriculture, la géologie, la protection de l'environnement ou le droit (Massoud, 2003). De nombreux documents administratifs et techniques sont donc rédigés dans cette langue et il est important de disposer de systèmes de gestion terminologique et de procéder à un aménagement terminologique dans de nombreux autres domaines. L'étude des terminologies scientifiques et techniques en arabe montre que celles-ci sont imprégnées des terminologies établies en anglais ou en français, alors considérées comme des terminologies de référence (Lelubre, 2008). Ce constat n'est toutefois pas valable pour tous les domaines. Ainsi, certains termes médicaux anglais sont issus de l'arabe (Wulff, 2004).

Partant de ce constat mais aussi d'observations réalisées en corpus, nous avons choisi d'exploiter la présence de termes anglais dans les terminologies et les textes de spécialité rédigés en arabe. La méthode que nous proposons vise à exploiter la translittération des termes anglais en caractères arabes et le principe du transfert translingue (McDonald *et al.*, 2011). À notre connaissance, il n'existe pas de systèmes de translittération de l'anglais vers l'arabe. L'application de la méthode sur un corpus parallèle anglais-arabe aligné au niveau des mots permet d'extraire des termes arabes après reconnaissance ou extraction des termes anglais.

Afin d'évaluer l'apport de la translittération de termes anglais en caractères arabes sur la construction d'une terminologie bilingue à partir d'un corpus parallèle anglais-arabe de textes médicaux, nous présentons dans cet article un système de translittération qui permet d'extraire des couples de termes médicaux translittérés de l'anglais vers l'arabe. Après un état de l'art des travaux des systèmes de translittération (section 2), nous présentons le corpus que nous utilisons, sa constitution ainsi que son alignement au niveau des mots dans la section 3. Nous décrivons la méthodologie suivie lors de la construction de notre système de translittération (section 4). Nous présentons et discutons les résultats obtenus (section 5) avant de conclure et de fournir quelques perspectives à ce travail (section 6).

2 État de l'art

Plusieurs travaux de recherche ont été menés ces dernières années sur la translittération concernant l'arabe. La plupart des travaux se sont focalisés sur le passage de l'arabe vers l'écriture latine (Al-Onaizan & Knight, 2002 ; Sherif & Kondrak, 2007 ; Saadane & Semmar, 2012).

Ainsi, Semmar & Saadane (2014) proposent un système de translittération des noms propres de l'écriture arabe vers l'écriture latine afin d'améliorer leur processus d'alignement des mots simples et composés. Les équivalences graphémiques utilisées sont établies à partir d'une étude de corpus de textes parallèles français-arabe. La première étape consiste à supprimer les voyelles avant de translittérer le nom. Des règles contextuelles définies pour rendre compte le plus précisément possible des formes observées sont appliquées selon le nombre de consonnes du nom considéré dans un ordre de priorité déterminé. La liste des noms en écriture latine est ensuite normalisée et une pondération est définie comme le nombre d'occurrences retourné par le moteur de recherche Google. Ce système produit en sortie une liste triée de noms arabes écrits en caractères latins. L'évaluation est effectuée sur 283 phrases du corpus du Monde Diplomatique français-arabe. Les taux de précision, rappel et F-mesure du processus d'alignement augmentent respectivement de 85% , 80% et 82% à 88%, 85% et 86%.

Mubarak & Abdelali (2016) présentent une nouvelle approche pour la construction d'un système de translittération. Leur corpus est composé de 881 310 paires de noms de personnes de l'arabe vers l'anglais collectées à partir de Twitter. Les données sont nettoyées pour extraire le nom complet écrit

en arabe ayant un chevauchement au dessus d'un certain seuil, le nom d'utilisateur écrit en caractères latins (sa translittération anglaise), ainsi que le pays de l'utilisateur. Les noms écrits en arabe et en latin et la localisation des utilisateurs sont normalisés selon la méthode décrite dans (Darwish *et al.*, 2012). Ensuite, les marques typographiques et les titres comme Mr, Mme,.. sont supprimés. Les lettres arabes qui n'ont pas de correspondance phonétique exacte dans les langues latines seront remplacées par des chiffres. Les variantes multiples de translittération des caractères arabes en leurs équivalents latins sont extraites. Les noms arabes sont translittérés différemment selon les variations dialectales régionales. Pour mesurer et quantifier la similitude entre les noms en arabe et ceux en latin, le score de similarité est calculé en utilisant la distance d'édition de Levenshtein. Les scores obtenus sur 1000 paires de noms sélectionnées au hasard sont importants : 0,96 pour la précision, 0,97 pour le rappel et de nomss'agit 0,965 pour la F1-mesure.

Les travaux présentés ci-dessus s'intéressent à la translittération des noms propres arabe en anglais, la langue arabe étant considérée comme la langue source. Si notre objectif comporte des similarités avec ces travaux, il diffère sur deux points : nous nous intéressons ici à la translittération inverse (anglais vers arabe), ainsi qu'aux mots, plus particulièrement aux termes de manière générale.

3 Corpus

3.1 Constitution du corpus médical

Comme indiqué dans (Al-Sulaiti & Atwell, 2006), très peu de corpus de spécialité sont disponibles pour l'arabe. C'est particulièrement le cas lorsqu'il s'agit de corpus parallèles. Pour cela, nous avons constitué notre corpus à partir de 133 documents produits par la *National Library of Medicine* (NLM) et disponibles en ligne sur MedlinePlus¹. Ces documents sont des brochures de quelques pages à destination des patients. Ces brochures apportent des informations sur des problèmes médicaux, décrivent les conditions de réalisation d'examen, fournissent des conseils de comportement face à une maladie, ou pour l'amélioration du bien-être. L'anglais est la langue source, et ces documents sont traduits dans de nombreuses langues cibles comme le français ou l'arabe. Pour la définition de notre approche, nous avons utilisé 71 brochures. Les corpus anglais et arabe contiennent respectivement 35 521 et 33 402 mots. Pour évaluer notre système de translittération, nous avons ajouté 62 documents, soit 25 379 mots pour le corpus anglais et 21 950 mots pour le corpus arabe. Contrairement aux documents en anglais et en français, la conversion au format texte des documents en arabe, en vue de réaliser des traitements automatiques, pose des nombreux problèmes (Habash, 2010) que nous avons tenté de résoudre mais qui rendent la tâche de nettoyage très coûteuse en temps : erreur de forme des caractères, utilisation d'un caractère persan ressemblant graphiquement à un caractère arabe, etc. L'ensemble des documents du corpus est ensuite segmenté par MADAMIRA (Pasha *et al.*, 2014) et aligné au niveau des paragraphes et au niveau des phrases. La qualité de l'alignement a été vérifiée manuellement.

3.2 Alignement au niveau des mots

Un alignement au niveau des mots a été obtenu en utilisant l'outil GIZA++ (Och & Ney, 2003). Nous avons défini trois alignements entre le corpus anglais et le corpus arabe qui prennent en compte diffé-

¹http://www.nlm.nih.gov/medlineplus/languages/all_healthtopics.html

rentes informations morphologiques fournies par MADAMIRA sur le corpus arabe : (Alignement1) l'alignement est réalisé sans traitement morphologique particulier ; il s'agit de notre base de comparaison ; (Alignement2) avant l'alignement, les enclitiques et proclitiques sont désagglutinés du mot arabe auquel ils se rapportent ; (Alignement3) avant alignement, les enclitiques, les proclitiques et les articles sont désagglutinés du mot arabe auquel ils se rapportent. Après l'alignement 3, le corpus arabe comporte 51 951 mots, et la partie ajoutée pour l'évaluation, 27 952 mots.

4 Extraction de paires de termes par translittération

L'un des facteurs les plus importants qui ont contribué à la modernisation rapide de la langue arabe a été l'assimilation du vocabulaire d'origine étrangère. Ainsi, nous observons dans notre corpus que certains termes médicaux arabes sont le produit de la transcription des termes anglais. Pour extraire ces couples de termes anglais-arabe, nous nous sommes basés sur la translittération complète fixée par la norme ISO 233-1 (translittération des caractères arabes en caractères latins) pour créer la correspondance inverse (translittération des caractères latins en caractères arabes). Nous avons ainsi créé notre propre table de correspondance des caractères anglais vers l'arabe en nous appuyant également sur les observations faites sur notre corpus. Celle-ci est ainsi adaptée aux caractéristiques phonologiques utilisées lors du passage de l'anglais à la langue arabe. Nous nous sommes inspirés aussi de la translittération complète fixée par la norme ISO 233-1 (translittération des caractères arabes en caractères latins), sauf que nous faisons le passage à l'inverse. A notre connaissance, il n'existe pas de table de correspondance des caractères latin en arabe.

La translittération vers l'arabe peut être ambiguë. Pour chaque caractère anglais, nous essayons d'identifier les différents caractères arabes qui peuvent lui correspondre. Par exemple, le caractère *g* peut être translittéré par le caractère ج comme dans le couple *glucophage* (جلوكوفاج) ou par le caractère غ comme dans le couple *gluconate* (غلوكونات). La table 1 montre un extrait de notre table de correspondance des caractères. Certains couples de mots ont été directement récupérés à partir de notre table de correspondance, alors que d'autres ont nécessité l'enrichissement de notre processus d'extraction par certains traitements complémentaires.

Caractère anglais	Caractères arabes
g	ج غ
r	ر
c	ق ك س

TABLE 1 – Un extrait de la table de correspondance des caractères

Notre méthode d'extraction de couples de termes anglais-arabe par translittération tient compte de certaines particularités de la langue arabe. Comme nous l'avons déjà signalé dans (section 3.2), l'étape de l'alignement repose sur le corpus arabe désagglutiné au niveau des enclitiques, des proclitiques et des articles. Cependant, lorsqu'il s'agit d'un terme emprunté d'une langue étrangère, l'étiqueteur morpho-syntaxique MADAMIRA ne parvient pas à identifier ou à séparer les articles et les proclitiques du mot en question. Ainsi, la présence de ces particules provoque des erreurs lors de l'extraction des couples de termes anglais-arabe transcrits. La prise en compte de leur présence éventuelle a nécessité l'ajout de certains traitements supplémentaires au sein de notre processus d'extraction. Il s'agit d'identifier les articles et les propositions situés au début des termes arabes pour avoir une

meilleure correspondance de couple de termes anglais-arabe transcrits. La table 2 présente quelques exemples de couples extraits grâce à cette étape.

TABLE 2 – Amélioration de l'extraction suite au traitement de l'agglutination

Terme anglais	Terme arabe agglutiné	Terme arabe traité
titanium	التيتانيوم (<i>et le titanium</i>)	تيتانيوم (<i>titanium</i>)
progesterone	والبروجسترون (<i>et le progestérone</i>)	بروجسترون (<i>progestérone</i>)
steroid	بالاستيرويد (<i>par le stéroïde</i>)	ستيرويد (<i>stéroïde</i>)

Outre la prise en compte de l'agglutination telle que décrite précédemment, nous tenons compte de la non-voyellation. En effet, comme la plupart des textes en MSA, notre corpus arabe n'est pas voyellé². Il est alors possible d'observer différentes formes de transcription du même terme anglais qui tiennent compte ou non de la présence de voyelles longues. Lors du passage des termes anglais à la langue arabe par transcription, les terminologues tentent de remplacer les voyelles latines par des voyelles arabes courtes, alors que d'autres mettent l'accent sur ces voyelles en les remplaçant par des voyelles arabes longues. Par exemple, le terme *oxygen* (*oxygène*) est transcrit trois fois sous la forme اوكسجين (*oksjin*) où la voyelle *o* est représentée par le caractère ا suivie par la voyelle longue و et six fois sous la forme اكسجين (*owksjin*) où la voyelle longue 'o' n'est pas représentée.

Même si les termes arabes translittérés proviennent d'autres langues étrangères, ceux-ci suivent les règles imposées par la langue arabe pour la production des formes du pluriel des noms et des adjectifs. Nous normalisons donc les termes anglais et arabes en les mettant au singulier. Par exemple, le passage du pluriel au singulier nous permet d'identifier le couple de termes *protein* - بروتين (*protéine*) dont le terme arabe apparaît dans le corpus sous la forme du pluriel بروتينات (*des protéines*). De même, pour le couple *hormone* - هورمون (*hormone*), chacun des termes apparaît dans le corpus sous la forme du pluriel ; *hormones* pour la partie anglaise et هرمونات pour celle en arabe (*des hormones*).

Pour calculer la correspondance entre un mot anglais et un mot arabe, nous commençons par translittérer le mot anglais à l'aide de notre table. Cette première étape tient compte des différentes correspondances pour chaque caractère. Plusieurs traitements sont également appliqués au mot arabe (désagglutination au niveau des proclitiques et des articles, passage de la forme du pluriel au singulier) ainsi qu'à notre translittération (comme l'ajout ou suppression des voyelles longues). Par exemple, le terme anglais *bacteria* (*bactérie*) est translittéré en arabe soit sous la forme بكتريا [*bktiria*], soit sous la forme بكتيريا [*bktiria*] en tenant compte de la voyelle longue. Dans la deuxième forme arabe, la voyelle 'e' de 'bacteria' est remplacée par la voyelle arabe longue 'ي'³

L'ensemble des étapes effectuées sert à tester si notre translittération peut correspondre, tout en appliquant certains traitements, au mot arabe en question. Si c'est le cas, nous considérons que le couple de termes anglais-arabe représente une translittération du terme anglais en caractères arabes.

²L'arabe standard moderne se caractérise par la présence de symboles optionnels non alphabétiques, appelés diacritiques. En leur absence, la connaissance de la langue permet de prononcer correctement le mot. Dans certains cas, l'absence de voyellation peut donc engendrer des ambiguïtés de compréhension. Ce constat est d'autant plus vrai lorsque les textes arabes non-voyellés sont analysés automatiquement.

³La différence graphique est liée au changement des allographes des caractères arabes selon leur position dans un mot ou leur position indépendante (lorsque le caractère est représenté seul).

5 Expériences et résultats

Nous avons évalué l'impact de l'analyse morphologique dans la qualité de l'alignement. Puis, nous évaluons la qualité de la translittération et l'utilisation des résultats obtenus pour détecter des couples de termes anglais-arabe. Puis nous présentons une analyse des erreurs.

Prise en compte de l'analyse morphologique lors de l'alignement L'alignement au niveau des mots est une étape préliminaire à notre méthode d'extraction des couples de termes anglais-arabe. Pour l'évaluer, nous avons considéré que le taux d'estimation de correspondance fourni par Giza++ reflète la qualité de l'alignement. Le taux de correspondance de l'alignement initial (pas d'analyse morphologique en préalable à l'alignement) est estimé à 75,98%. Nous observons une amélioration des taux de correspondance lorsque les enclitiques et les proclitiques sont désagglutinés (79,41%, +3,43) et encore après désagglutination des articles (82,45%, +3,07).

Translittération et extraction des couples de termes anglais-arabe Nous avons utilisé les 133 documents à notre disposition afin d'avoir une bonne qualité d'alignement. Le système de translittération extrait 137 couples de termes anglais-arabe. Nous les avons évalué manuellement. Ainsi, 120 couples de mots sont jugés corrects (soit 87,59%). Parmi ceux-ci, 107 couples sont des couples de termes médicaux (soit 89,16% des translittérations correctes et 78,10% des résultats).

Nous avons observé que 7,29% des couples sont bien alignés mais ils ne représentent pas une translittération. Par exemple, le pronom relatif anglais *that* (*qui*) est associé au mot arabe *التي* (*alaty*, pronom relatif féminin singulier). Même si la translittération de certains d'entre eux donne un mot arabe qui en est la traduction, ceci n'est pas le résultat d'une écriture en caractères arabes des mots anglais sources. C'est le cas du mot *african* (*africain*) dont le mot arabe est *افريقي* (*afriqi*). 5,1% des couples de mots extraits ne sont ni alignés ni translittérés. Par exemple, le mot *grounds* (*marc*) est aligné avec un point ".", signe de ponctuation, ou aussi le mot *your* (*ton/ta/tes/votre/vos*) qui est aligné avec l'article arabe *ال* (*le/la/les*).

À partir des résultats obtenus, nous constatons que la terminologie arabe contient certains termes anglais translittérés même s'il existe déjà un terme arabe qui lui correspond. Autrement dit, un terme anglais peut à la fois avoir comme correspondant sa translittération en caractères arabes ainsi qu'un autre terme arabe ayant le même sens. Par exemple, le terme *ounces* (*onces*) peut correspondre au terme arabe *اوقيات* ou à sa translittération donnant le terme arabe *اونصة*.

Analyse des erreurs Comme il s'agit d'un alignement au niveau des mots, certains termes complexes sont alignés avec une partie du terme arabe qui serait le correspondant complet du terme anglais en question. Dans l'exemple suivant, nous constatons que le terme anglais *x-ray* (*rayon-X*) est aligné avec le terme simple arabe *اشعة* (*rayons*) qui représente une partie du terme complexe *اشعة اكس*. Dans le corpus, le tiret qui lie les deux parties du terme arabe n'existe pas. À cause de cela, l'alignement est fait avec une seule partie du terme. L'utilisation des acronymes est un phénomène fréquent lors de la construction de la terminologie médicale anglaise. Celle-ci influe sur les résultats de translittération obtenus. Par exemple, L'acronyme (*MRSA*) (*SARM : Staphylococcus aureus résistant à la méthicilline*) est préféré à la forme longue du terme *Methicillin-Resistant Staphylococcus Aureus*.

L'acronyme est alors aligné avec le mot arabe للمثيسيلين⁴ (à la *méticilline*). Le terme arabe extrait représente une partie du terme arabe complexe العنقوديات الذهبية المقاومة للمثيسيلين qui lui correspond. Une autre caractéristique de la terminologie anglaise est l'utilisation des abréviations. Ainsi, des couples de termes dont la partie arabe représente une translittération ne sont pas extraits par notre système. Par exemple, le couple des termes *flu* et انفلونزا (*grippe/influenza*) n'est pas retrouvé car le terme arabe correspond à la translittération du mot complet anglais *Influenza* et non à son abréviation *flu*.

Certains termes arabes translittérés ne figurent pas dans la liste produite par notre système suite aux erreurs produites lors de la phase d'alignement. Par exemple, le terme arabe زنك (*zinc*) est aligné avec le terme *oxide* alors que اكسيد (*oxide*) est aligné avec le terme *zinc*.

Comme nous l'avons déjà signalé, la construction de la terminologie médicale arabe repose sur l'assimilation du vocabulaire d'origine étrangère. L'origine des termes translittérés influe sur les résultats obtenus. Les résultats dépendent donc des langues utilisées dans le corpus. Si notre corpus était français-arabe, il aurait existé d'autres couples de translittérations. Par exemple, le terme arabe بوصة est une translittération du terme français *pouce*. Pour cette raison, le couple de termes anglais-arabe (*inch*, بوصة) ne figure pas dans la liste produite par notre système de translittération.

6 Conclusion et perspectives

Nous nous sommes intéressés à la construction d'un système de translittération afin d'extraire les couples de termes anglais-arabe translittérés de l'anglais en caractères arabes. Notre objectif est d'évaluer l'apport de la translittération dans la construction d'une terminologie bilingue à partir d'un corpus médical parallèle anglais-arabe. Nous avons proposé un processus de détection des translittérations de termes anglais en arabe. Pour cela, nous nous sommes appuyés sur une étude du corpus pour la création d'une table de correspondances des caractères anglais en arabe mais nous avons également pris en compte certaines particularités de la langue arabe comme les phénomènes d'agglutination et de non-voyellation. Nous avons ainsi défini des traitements supplémentaires qui tiennent compte de ces phénomènes. Les expériences réalisées sur un corpus anglais de 55 352 mots et un corpus arabe de 60 900 mots ont permis d'obtenir 87,59% de translittérations correctes. Parmi celles-ci, 78,10% représentent des couples de termes médicaux translittérés de l'anglais en caractères arabes.

Plusieurs perspectives de travail s'offrent à nous. D'une part, le processus d'alignement des textes au niveau des mots doit être amélioré. D'autre part, nous envisageons d'élargir notre corpus de test. Enfin, en fonction de la disponibilité de corpus parallèles en MSA, nous proposons d'évaluer notre travail dans d'autres domaines de spécialité.

Remerciements

Ce travail a bénéficié d'un soutien de l'action Horizon 2020 Marie Skłodowska-Curie Innovative Training Networks — European Joint doctorate (ITN-EJD) de l'Union européenne, projet N° 676207 (MiRoR) et de l'ANR dans le cadre du projet CLEAR (ANR-17-CE19-0016-01).

⁴Ce terme arabe n'est pas désagglutiné car il s'agit d'un mot étranger que MADAMIRA n'arrive pas à identifier.

Références

- AL-ONAIZAN Y. & KNIGHT K. (2002). Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 400–408.
- AL-SULAITI L. & ATWELL E. (2006). The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, **11**(1), 1–36.
- DARWISH K., MAGDY W. & MOURAD A. (2012). Language processing for arabic microblog retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, p. 2427–2430 : ACM.
- HABASH N. (2010). *Introduction to Arabic Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- LELUBRE X. (2008). La constitution de la terminologie arabe de la physique : aspects diachroniques. *Travaux du CRTT*. Article dans numéro spécial Travaux du CRTT.
- MASSOUD R. (2003). La terminologie au liban : réalités et défis. *Annales de l'Institut de langues et de traduction (ILT)*, **10**.
- MCDONALD R., PETROV S. & HALL K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 62–72.
- MUBARAK H. & ABDELALI A. (2016). Arabic to english person name transliteration using twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.
- OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–51.
- PASHA A., AL-BADRASHINY M., DIAB M., KHOLY A. E., ESKANDER R., HABASH N., POOLEERY M., RAMBOW O. & ROTH R. (2014). Madamira : A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).
- SAADANE H. & SEMMAR N. (2012). Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 127–140 : ATALA/AFCP.
- SAMY D., MORENO-SANDOVAL A., BUENO-DÍAZ C., GARROTE-SALAZAR M. & GUIRAO J. M. (2012). Medical term extraction in an arabic medical corpus. In *Proceedings of LREC'12*, p. 640–645.
- SEMMAR N. & SAADANE H. (2014). 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). In *Traitement Automatique des Langues Naturelles (TALN 2014)*, p. 268–279, Marseille, France.

SHERIF T. & 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*, p. 864–871, Prague, Czech Republic : Association for Computational Linguistics.

WULFF H. (2004). The language of medicine. *Journal of the Royal Society of Medicine*, **97**, 187–188.

