*Transcription des noms arabes en écriture latine**

Houda Saadane[1] – Nasredine Semmar[2]

[1]LIDILEM / Université Stendhal - Grenoble III
Domaine Universitaire - 1180, avenue centrale F-38400 Saint Martin d'Hères, France
[2]Institut CEA, LIST, DIASI / Laboratoire Vision et Ingénierie des Contenus
CEA Saclay – Nano-INNOV Bâtiment 861 – PC 173 F-91191 Gif-sur-Yvette Cedex, France

[1]houda.saadane@e.u-grenoble3.fr, [2]nasredine.semmar@cea.fr

Abstract: *In this paper, we focus our work on the transliteration of the Arabic names from their original written to the Latin one. This kind of technique is essential for retrieving inter-lingual information in order to obtain relevant results. Our approach is based on returning all spellings of a given word rather than computing the best solution. We have shown the interest of our technique by experiments using à web search engines and counting the number of results returned using our technique.*

Résumé : *Dans cet article, nous nous intéressons à la translittération des noms arabes de leur écriture originale vers l'écriture latine. Ce type de technique est nécessaire pour la recherche d'information inter-lingue afin de renvoyer des résultats pertinents. Le principe de notre approche est de proposer toutes les variantes orthographiques d'un nom arabe et non pas la meilleure solution. Nous avons montré l'intérêt de notre approche par une série d'expérimentations et en comptant à chaque fois le nombre de résultats obtenues par des moteurs de recherche après la transcription via notre technique.*

Keywords: *Transcription, transliteration, transcription standard, Arabic name, contextual rules.*
Mots clés *: Transcription, translittération, norme de transcription, nom arabe, règles contextuelles.*

**Latin transliteration of Arabic names*

© Revue RIST, vol. 20, n° 2, p.57-68

1 Introduction

d'écriture cible, est une tâche délicate qui nécessite un certain nombre d'opérations exigeant de prendre en considération un ensemble de propriétés – syntaxiques, phonétiques et sémantiques – caractérisant à la fois les systèmes d'écriture cible et source.

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, indépendamment de la prononciation.

Mais les problématiques soulevées par la transcription et la translittération ne relèvent pas tant de la nature de la convention adoptée que de la relation entre oralité et scripturalité lors du passage d'un système linguistique à un autre. En effet, l'oral et l'écrit obéissent à des règles différentes : l'un a un matériau sonore, l'autre a un matériau visuel, et chaque matériau possède une dynamique interne et des contraintes propres.

Parmi ces contraintes, citons les phénomènes de transformation morphologique qui affectent les mots en fonction de la nature de leur lettre initiale. Ainsi, si le mot contient l'article AL (ال), il faut faire la distinction entre les lettres «solaires» et les lettres «lunaires». Avec les premières (solaires), le «L» ne se prononce pas et la lettre qui le suit est dédoublée dans la prononciation et dans l'écriture. A l'inverse, avec les lettres lunaires, le «L» de l'article se prononce et la lettre qui le suit n'est pas dédoublée ni dans la prononciation ni dans l'écriture. Mais ces règles de la langue arabe ne sont pas toujours respectées dans la translittération usuelle comme en témoignent les exemples concernant la translittération des noms de journaux arabes en écriture latine : النهار (An-Nahar) au Liban, الزمان (AZZAMANE) à Londres, الصباح (AL-SABAH) en Palestine, اليوم (ELYAUM) en Algérie...

On constate un phénomène analogue d'écart par rapport à la norme phonologique du système d'origine dans la translittération de la lettre «T-attachée» (ة), dite en arabe « T-Marbouta ». Celle-ci se prononce (t) à l'état d'annexion exclusivement. Mais là encore, certaines translittérations de noms de journaux rendent compte de la graphie du mot arabe et non pas de sa prononciation effective : par exemple, الثورة (AL-TAWRA) au lieu de (ATH-THAWRAH).

Il est par conséquent important de mener un questionnement, préalable au traitement automatique, concernant des aspects qui peuvent paraître évidents au premier abord mais qui méritent une analyse approfondie. Citons les aspects suivants comme prioritaires pour le traitement automatique de la transcription et la translittération:

– Degré d'adéquation de la transcription à l'oral (phonème);

– Degré d'adéquation de la translittération à l'écrit (graphème);

– Degré d'adéquation de la notation symbolique à l'usage (social).

A cette dimension symbolique dans laquelle intervient la sémantique du nom, s'ajoute une dimension phonologique qu'il est nécessaire de prendre en considération lors du traitement automatique et qui tient à la situation linguistique du monde arabe. En effet, la langue arabe présente aujourd'hui une situation caractérisée par une polyglossie complexe (Dichy, 2009) Il existe ainsi une diversité de réalisations de l'arabe littéraire (classique, moderne, moyen, etc.) et une pluralité de dialectes (variétés d'arabe, régionales ou locales), dont les caractéristiques dominantes sont sensibles aux utilisateurs.

Dans cette optique, la dialectologie arabe distingue deux grandes familles de dialectes, celle du Maghreb (Maroc, Algérie, Tunisie et Libye) et celle du Machrek (Egypte-Syrie et

Moyen-Orient). Mais à l'intérieur de ces familles de géolectes, on trouve aussi bien des dialectes nationaux (natiolectes) que des dialectes régionaux (régiolectes) ou encore des dialectes locaux (topolectes), parlés sur un espace limité (village, localité).

Ainsi, lorsque l'on se propose de développer un translittérateur des noms arabes, l'on se trouve confronté aux spécificités phonologiques de ces variantes dialectales, car le locuteur-auditeur prononce différemment le même nom en fonction de son dialecte et reconnaît la variation dialectale en fonction de la prononciation qu'il entend. Cela est d'autant plus vrai que chaque dialecte arabe a subi l'influence, au cours de l'histoire moderne, d'autres langues comme le français, l'italien et l'espagnol (pour le Maghreb) ou encore de l'anglais (pour le Machrek). Cela fait qu'un même nom ou prénom en arabe peut avoir plusieurs prononciations différentes dans les dialectes et diverses translittérations en fonction des spécificités phonologiques et graphématiques des langues d'arrivée. Ainsi par exemple, le nom Kadhafi, qui possède une orthographe unique en arabe (معمر القذافي) mais plusieurs prononciations et accentuations en fonction des dialectes, est transcrit en écriture latine par plus de 60 formes différentes, parmi lesquelles : Muammar Qaddafi, Mo'ammar Gadhafi, Muammer Kaddafi, Moammar El Kadhafi, Muammar Gadafi, Moamer El Kazzafi, Mu'ammar al-Qadhdhafi, Mu'amar Qadafi, Muammar Gheddafi, Mu'ammar Al Qathafi, Muʿammar Al-Qaḏâfî…

Cette multiplicité des formes ne manque pas de poser problème tant au niveau de la recherche d'information pour une entité nommée (ici le nom d'un dirigeant politique) que pour l'enrichissement interlingue de données concernant un sujet particulier (par exemple, Tripoli en Libye). En effet, le fait de ne pas répertorier toutes les formes disponibles à un moment donné pour un même nom peut être préjudiciable à l'efficacité de la recherche.

Nous présentons dans la section 2 un état de l'art dans le domaine de la translittération suivi d'une description des 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 validons notre technique dans la section 3 en présentant des expérimentations utilisant des moteurs de recherche de référence. La section 4 conclut notre étude et présente nos travaux futurs.

2 Translittération

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 et plus particulièrement la recherche d'information interlingue. C'est le cas 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.

2.1 Etat de l'art

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 (Saadane *et al.*, 2012). Citons notamment (Yaser *et al.*, 2002) et (Sherif *et al.*, 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 al.*, 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 al.*, 1997).

Cette approche a été adaptée par (Stalls *et al.*, 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).

2.2 Translittération de noms arabes en écriture latine

Le système d'écriture de la langue arabe est constitué d'un alphabet de 28 lettres. Cet alphabet contient 25 consonnes et 3 voyelles longues. Il existe aussi des voyelles courtes

(Table 1) qui sont généralement présentes uniquement dans les textes religieux (Coran, Hadith, etc.) ou les manuels scolaires pour enfants. Cette particularité est l'une des principales sources d'ambiguïté pour les systèmes de translittération.

Voyelle courte	Transcription	Nom
ـَ	a	fatha
ـُ	u	damma
ـِ	i	kasra
ـْ	e	sukun
ـّ	doublement	shadda
ـً	aa	fathatan
ـٌ	uu	dammatan
ـٍ	ii	kasratan

Table 1. *Les voyelles courtes en arabe*

2.2.1 Normes de translittération pour l'arabe

Il existe plusieurs normes de translittération, dont EI (1960), ISO/R 233 (International Organization for Standardization, 1961), UN (United Nations Group of Experts on Geographical names, 1972), DIN-31635 (Deutsches Institut für Normung, 1982), ISO 233 (International Organization for Standardization, 1984) ainsi que la norme ALA-LC (America Library Association, 1997). Parmi ces normes deux sont reconnues et utilisées internationalement par la communauté scientifique : DIN-31635 et la norme adoptée par l'Encyclopédie de l'Islam (EI).

2.2.2 Correspondances proposées pour la translittération des lettres arabes vers le latin

A partir des différentes normes de translittération, nous avons établi une nouvelle table qui réunit toutes les normes précédentes (Table 2).

Lettre	Équivalents en écriture latine	Lettre	Équivalents en écriture latine
ء	’, a	غ	Gh, gh, Ğ, ğ, 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, ṯ,	ل	L, l
ج	J, j, Dj, dj, g, Ğ, ğ	م	M, m
ح	H, h, Ḥ, ḥ, h, 7	ن	N, n
خ	Kh, kh, ḫ, h	ه	H, h
د	D, d	و	W, w, ou, o, u, ô, û, ū, ú, ü
ذ	Dh, dh, D, d, Ḏ, ḏ, Ḍ, ḍ	ي	I, i, y, ï, î, ī,
ر	R, r	أ	A, a, ā, ’ā, ’â
ز	Z, z, Ẓ, ẓ	ة	
س	S, s	ى	H, h, T, t, at, a, ẗ
ش	Ch, ch, Sh, sh, Š, š	أ	A, a, á, à, ā, ỳ
ص	S, s, Ṣ, ṣ, Ṣ, ṣ	ؤ	
ض	D, d, Ḍ, ḍ, Ḍ, ḍ	إ	I
ط	T, t, Ṭ, ṭ, Ṭ, ṭ	ئ	U, u, Ou, ou, Û, û
ظ	Z, z, Ẓ, ẓ, 6’, Dh, dh, D, d	گ	G, g
ع	‘, ', ‘, 3, a, â		’, (Blanc)

Tableau 2: Table de translittération des caractères arabes vers le latin

Certaines lettres arabes sont transcrites en chiffres. Cette translittération constitue la norme dans le langage SMS en Europe et au Moyen Orient. Cela est très utile pour supposer quelle est l'origine sociale de la personne qui écrit et quelle est l'origine

géographique des données extraites (géolocalisation). La table 3 suivante récapitule ces chiffres spéciaux :

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

Tableau 3: Equivalences 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 (Table 4) :

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...

Tableau 4: 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 (مُحَمَّدْ), 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.

2.2.3 Approche proposée pour la translittération de noms propres arabes en écriture latine

Le module de translittération de l'écriture arabe vers l'écriture latine est fondé sur les automates d'états finis constitués d'états et de transitions (Figure 1). 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 (Saadane *et al,* 2012).

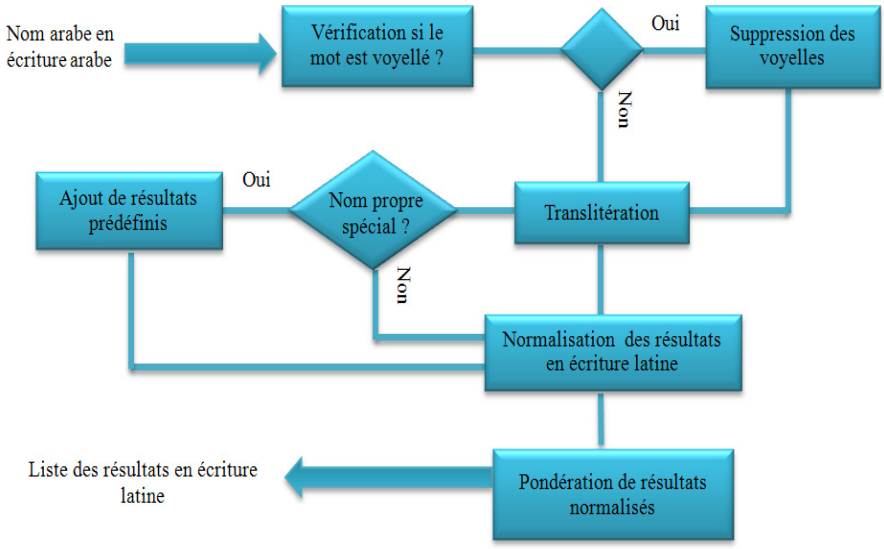

Figure 1: Organigramme de fonctionnement translittérateur de l'arabe vers le latin

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 :

- 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..
- 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..
- 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..
- 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

Tableau 5: 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.

- 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

Tableau 6 : 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.

3 Résultats expérimentaux

Au stade actuel du développement du système, le processus de validation des hypothèses et des résultats a été réalisé essentiellement avec des moteurs de recherche sur Internet.

Grâce aux stratégies mises en place, le moteur de recherche permet de récupérer tous les documents pertinents, et cela quelle que soit la langue du document d'origine et quelle que soit la forme du nom propre employée. Grâce aux procédures de désambiguïsation

mises en œuvre, le système permet également de valider s'il s'agit ou non de la même personne que celle recherchée.

Le processus de recherche de l'existence de la translittération par un moteur de recherche (Google par exemple) peut être décrit par l'algorithme suivant :

```
Fonction Rech-Google( NOMS : Mots) : Entier
 Résultat : Entier
 Ouverture de la connexion vers la machine distante
 www.google.fr, dans la socket S.
 Envoie de la requête de recherche (suite de mots) à
  travers la socket S.
 Lecture de la page Web résultat depuis la socket S
 dans un buffer B.
 Si (la chaîne de caractères "Aucun résultat trouvé"
 existe dans B) Alors
   Résultat ← 0;
 Sinon
   Résultat ← N : le nombre de pages Web
    renvoyées  par Google dans le buffer B
 Fin Si
 Retourner Résultat (les pages web)
   Fin
```

Cette vérification se fait concernant la corrélation entre le mot translittéré (requête) et les documents récupérés pour ce nom. Une translittération est considérée comme pertinente si le résultat des requêtes pour une forme translittérée n'est pas nul, et si, pour chaque forme translittérée, le moteur de recherche récupère au moins une réponse à chaque fois pour une même personne.

Considérons l'exemple suivant : requête sur le nom du président algérien.

Nom en entrée : بوتفليقة

Sortie : Formes translittérées indiquées en gras dans les documents récupérés à partir d'une dizaine de langues différentes :

1. KING_ SADDRESS AT ARAB SUMMIT IN ALGIERS-By:IMRA algiers, march 23 (petra-jordan news agency)--his majesty king abdullah ii said that the roadmap peace plan is the only available means to settle the palestinian ... thanks and appreciation for his excellency president abdul aziz **botafliqah** and to the algerian people for their kind hospitality ... israpost.com/Community/articles/show.php?articleID=5361>Cached
2. The Angry Arab News Service " As'ad the angry arab news service. a source on politics, war, the middle east, arabic poetry, and art by as'ad abukhalil ... posted by as'ad at 6:52 am 04/10/09. **butufliqa** wins. the algerian president wins re-election with 99.99% of the ... angryarab.net/author/falastin>Cached
3. كونا: Arab League congratulates Algeria"s **Boutfalika**... cairo, april 11 (kuna) -- the arab league congratulated saturday president abdelaziz **boutaflika** for his win in the algerian presidential elections, hoping that he would continue the development process in the north african nation. arab league kuna.net.kw/NewsAgenciesPublicSite/...&Language=en>Cached
4. Times of Oman

it comes in implementation of the directives of his majesty sultan qaboos bin said and president abdulaziz **boutafliqah** aimed at cementing bilateral relations. later at a press conference, alawi said that the two sides signed a number of agreements and mous. ... timesofoman.com/innercat.asp?detail=33983>Cached

5. Abdelaziz **Bouteflika** - Wikipedia, the free encyclopedia bouteflika lived and studied in algeria until he joined the front de libération nationale ... on boumédienne's unexpected death in 1978, bouteflika was seen as one of the two main ... en.wikipedia.org/wiki/Abdelaziz_Bouteflika>Cached

6. Maliki : If Sultan Hsahim is not executed I will resign kurdish aspect covers issues related to kurds and kurdistan within the larger context of middle eastern concerns. the website offers readers a ... he revealed to them that in the opec meeting the algerian prime minister abd-al-aziz **botafliqa** had asked him: are you from an iranian origin? ... kurdishaspect.com/doc020208AWENE.html>Cached

7. ...النهار الجديد- ثورة في عالم الإعلام - لأول مرة ..أسرار عن - 1 | عرض: 66المجموع: 0. i love you **boutafli8a** ... .لأول مرة ..أسرار عن بوتفليقة الرجل . أضف تعليقك. اسمك: أضف تعليقاتك: اقرأ أيضا فى: الوطني. وزارة التربية الوطنية تلغي التسجيل في 66 ...بكالوريا النظام القديم. المتهم الرئيسي في مقتل سارة يواجه تهم الخلوة ennaharonline.com/ar/national/29309.html>Cached

8. **Butaflika** fires Benflis, brings back Oyahya, due to Algerian ... butaflika fires benflis, brings back oyahya, due to algerian presidential elections, algeria, politics. arabicnews.com - your source for daily news about the arabic world. ... algerian news agency quoted benflis as saying after a meeting with butfalika that he did not take part in taking the decision of ...

9. YouTube - viva l'algerie , algeria is back HMD algeria mon amour algeria is back no matter what algeria is still standing ya rab hamdolah , maghrab united ya **botaflika** , les marocain khawatna toujour m3a jazair youtube.com/watch?v=IAF1AOmnQlM>Cached

10. The Angry Arab News Service " As'ad the angry arab news service. a source on politics, war, the middle east, arabic poetry, and art by as'ad abukhalil ... posted by as'ad at 6:52 am 04/10/09. **butufliqa** wins. the algerian president wins re-election with 99.99% of the ... angryarab.net/author/falastin>Cached

4 Conclusion

Dans cette étude, nous avons expliqué le cadre et la méthodologie qui ont permis de construire un système de translittération des noms arabes de l'écriture latine vers l'écriture arabe et inversement. Ce système génère toutes les formes orthographiques possibles pour un nom en s'appuyant sur des expressions régulières et sur des règles contextuelles.

Il est possible d'améliorer le système de translittération du latin vers l'arabe en affinant les résultats obtenus grâce à un calcul de fréquence des formes constatées dans chaque langue considérée. Mais cela suppose d'étendre la couverture linguistique du système, et notamment l'analyse des noms d'origine non arabe.

Dans l'immédiat, nous envisageons d'orienter nos recherches vers la translittération géolocalisée pour répondre à la question de savoir 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...). Cette orientation répond à la fois à une demande industrielle urgente et à une problématique de recherche intéressante.

La prochaine étape consiste à étendre le système de translittération aux noms d'origine non arabe.

5 Réferences

Abdulhay A., « Constitution d'une ressource sémantique arabe à partir d'un corpus multilingue aligné », Thèse de Doctorat de l'Université Stendhal – Grenoble III, 2012.

Abduljaleel N., Larkey L., "Statistical transliteration for English-Arabic Cross Language Information Retrieval", *Proceedings of the Twelfth ACM International Conference on Information and Knowledge Management*, New Orleans, Louisiana, 2003.

Alghamdi M., "Alghorithms for Romanizing Arabic names", Computer Sciences and Information, n° 17, 2005, Riyadh, 2005.

Alsalman A., Alghamdi M., Alhuqayl K., Alsubai S., « Romanization System for Arabic Names », *Proceedings of The First International Symposium on Computer and Arabic Language (ISCAL – 07)*, Riyadh, Novembre 2007, p. 214-227.

Barkat-Defradas M., Hamdi R., Pellegrino F., « De la caractérisation linguistique à l'identification automatique des dialectes arabes », *Proceedings of MIDL* 2004.

Dichy J., «La polyglossie de l'arabe illustrée par deux corpus», In M. Bozdemir et L.-J. Calvet (eds), Politiques linguistiques en Méditerranée, Paris: Honoré Champion, 85–102.

Guidère M., « Le traitement de la parole et la détection des dialectes arabes », Langues stratégiques et défense nationale, Publications du CREC, Saint-Cyr, pages 53–75, 2004.

Knight K., Graehl J., « Machine transliteration », Journal version Computational linguistics, 24(4), 1997, p. 599-612.

Saâdane H., Semmar N., « Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parallèles français-arabe», *Actes TALN 2012*, Grenoble, France.

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

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

Stalls B., Knight K., « Translating names and technical terms in Arabic text », *Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages, 1998*, Montreal, Québec.

Tao T., Yoon S. Y., Fister A., Sproat R., Zhai C., « Unsupervised named entity transliteration using temporal and phonetic correlation », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'06), Sydney, pages 250–257,* 2006.

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