La reconnaissance automatique des dialectes arabes à l'écrit

Dr. Saadane Houda
Université de Grenoble, (France)

Pr. Fluhr Christian
GeolSemantics, (France)

Pr. Guidère Mathieu
Université de Toulouse, (France)

Résumé

La langue arabe est caractérisée par une situation de diglossie‹ il existe une pluralité de l'arabe‹ d'une part‹ l'arabe classique et l'arabe moderne standard qui a été le sujet de nombreux travaux de recherche et d'autre part‹ l'arabe dialectal qui est une langue peu dotée et présente des variations d'un pays à un autre et voire d'une ville à une autre. Dans cet article‹ nous nous intéressons à la reconnaissance automatique des dialectes arabes à l'écrit. Dans un premiers temps‹ nous avons cherché à constituer un corpus de textes dialectaux rédigés en caractères latins. Ensuite‹ nous avons employé un outil de translittération pour les retranscrire en arabe. Enfin‹ nous avons extrait les traits morphosyntaxiques et sémantiques caractéristiques de deux principaux groupes de dialectes‹ à savoir le *Maghreb* et le *Machrek*.

Mots clés

Arabe standard‹ dialectes‹ langues peu dotées Maghreb‹ Machrek‹ transcription‹ corpus‹ translittération‹ traits linguistiques‹ TAL.

Introduction

La langue arabe est l'une des langues les plus parlées et utilisées dans le monde. Elle est la langue officielle de plus de 22 pays et parlée par plus de 320 millions de personnes. Elle est utilisée comme vecteur de transmission religieux pour tous les croyants musulmans au nombre de 1 milliard et demi à travers les cinq continents du globe. Elle constitue ainsi un élément principal dans la culture et la pensée d'une partie importante de l'humanité et du patrimoine mondial.

A l'origine‹ les peuples de la péninsule arabe tenait le monopole de cette langue qui est sémitique (comme l'hébreu ou l'araméen)‹ mais du fait qu'elle est la langue du coran elle s'est étendue au-delà du golf arabique‹ atteignant l'Afrique du nord et l'Asie mineur. De plus‹ l'expansion territoriale de l'empire musulman a fait de l'arabe une langue d'administration‹ de culture et de sciences à travers son utilisation dans la définition et la rédaction des contrats et des lois‹ la rédaction de manuscrits et de livres‹ la transmission et la formation‹ etc. Par ailleurs‹ la diversité des populations arabes et de leurs cultures ont fait émerger différentes variantes de l'arabe allant de l'arabe classique utilisé dans le coran‹ à

l'arabe standard moderne (ASM)، représentant l'arabe officiel employé actuellement dans la presse، les documents officiels، etc; en passant par l'arabe dialectal influencé par les spécificités historiques et culturelles locales des populations constituant le monde arabe. Historiquement، l'arabe tient ses origines au 2ème siècle et malgré son utilisation les premières traces écrites comme on la connait actuellement remontent au 6ème siècle. Ce fait peut être expliqué par l'analphabétisme des populations de l'époque qui communiquaient plus oralement que par écrits. L'apparition de l'islam a fait sortir l'arabe de son territoire d'origine et lui a donné une dimension internationale، en raison de son utilisation comme langue seule et unique pour tous les devoirs et rituels religieux، et du fait que le coran، comme texte sacré، ne peut être lu ou écrit qu'en arabe. Cette nouvelle dimension a multiplié considérablement l'utilisation de l'arabe dans les communications et échanges oraux et surtout écrits. Cette expansion à la fois géographique et fonctionnelle a rapidement généré des réflexions sur la structuration et l'organisation de cette langue، mais aussi des intégrations et des emprunts de mots depuis et vers d'autres langues comme le français، perse، turque، etc.

Par ailleurs، la langue arabe présente aujourd'hui une situation caractérisée par une polyglossie complexe (Dichy، 2009)، (Harry، 1996). Il existe، en effet، 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 trois groupes distincts de dialectes à l'intérieur du grand ensemble géographique que constitue le monde arabe. D'abord، les dialectes du Maghreb où l'on trouve : l'Algérie، la Mauritanie، le Maroc، la Tunisie et la Libye. Ensuite، les dialectes du Machrek où l'on trouve : l'Egypte، la Syrie، le Liban، la Jordanie et la Palestine. Enfin، les dialectes du Golfe où l'on trouve l'Arabie Saoudite، le Yémen، Oman، les Émirats arabes unis، le Qatar، le Bahreïn، le Koweït et l'Irak. Mais à l'intérieur de ces familles de géolectes، nous trouvons aussi bien des dialectes nationaux (natiolectes) que des dialectes régionaux (régiolectes) et même des dialectes locaux (topolectes)، parlés sur un espace limité (village، localité) (Saâdane، 2011).

Dans le cadre d'une recherche appliquée visant le développement d'un outil de reconnaissance automatique des dialectes arabes à l'écrit، nous avons d'abord constitué un corps de textes/discours dialectaux mais rédigés caractères latins. Nous avons ensuite employé un outil de translittération pour les retranscrire en arabe. Enfin، nous avons cherché à caractériser ces productions langagières écrites en analysant les traits morphosyntaxiques et sémantiques caractéristiques de chaque dialecte.

Notre traitement automatique a donc consisté en trois étapes distinctes et successives :

1) Constitution du corpus dialectal écrit en caractères latins: cette étape consiste à réunir، à partir des forums de l'Internet et des réseaux sociaux les productions langagières issus des dialectes arabes et rédigés en caractères latins.

2) Transcription du corpus dialectal: cette phase consiste à transcrire les messages de l'arabe dialectalisé en écriture latine vers l'écriture arabe.

3) **Transcodage du message**: cette étape consiste à présenter le message en arabe standard pour pouvoir mesurer l'écart existant entre l'arabe dialectal et l'arabe standard. A ce stade، nous proposons également، à titre indicatif، une traduction française du message.

Approche de reconnaissance des dialectes :

Dans cette section nous présentons notre approche de reconnaissance introduite dans la section 1. Cette reconnaissance est matérialisée par la détection des traits linguistiques caractérisant les dialectes arabes، et essentiellement ceux en relation avec les dialectes du moyen orient (Machrek) et du Maghreb. Notre approche se base sur les trois étapes suivantes :

Constitution du corpus dialectal écrit en caractères latins :

Le développement d'outils de traitement automatique des langues se heurte au manque des ressources de l'arabe dialectal، comme les corpus LDC et ELDA dédiés aux dialectes du Machrek qui sont écrits en arabe. En particulier il n'existe pas de corpus d'arabe dialectal tel que pratiqué en écriture latine dans les réseaux sociaux. C'est la raison pour laquelle nous avons défini comme première étape de notre approche la constitution d'un corpus de ressources dialectales éditées en écriture latine. Pour constituer ce corpus nous avons essentiellement considéré des ressources issues d'Internet، qui est en soit un gisement important d'information، en plus d'autres ressources générées par des applications comme les transcripteurs de paroles.

En effet، la ressource la plus viable de texte arabe dialectal est données en ligne، ce qui est plus individuel et moins institutionnalisée، et donc plus susceptibles d'inclure des contenus dialectaux. Les sources possibles de texte dialectal comprennent les blogs، les forums، les sites d'information (commentaires des articles et de l'actualité)، les réseaux sociaux (facebook، twitter، etc)، les diffuseurs de vidéos (youtube، dailymotion، etc) et les transcriptions de conversations.

Les ressources en ligne en général et les commentaires des utilisateurs en particulier présentent les avantages suivants pour la constitution d'un corpus (Zaidan et all،2011) :

- Une grande quantité de données، avec plus de données générées et disponibles quotidiennement.
- Les données sont publiquement accessibles، ayant un format cohérent et structuré، et qui sont faciles à extraire.
- Elles présentent des sujets ayant des niveaux élevés de pertinence.
- La dominance de l›arabe dialectal dans les échanges.

La construction de ce corpus nécessite une étape de collecte et d'extraction de ressources depuis les sites internet. Pour réaliser cette tâche‹ nous avons choisi l'outil HTTrack qui est un aspirateur de site Web open source permettant de copier tout le contenu d'un site sur un support local. Il permet de récupérer la structure originale du site ainsi que tous les fichiers (HTML‹ images‹ sons‹ etc) constituant le site analysé. Dans notre contexte‹ nous avons utilisé cet aspirateur afin de télécharger des pages Web des différents sites arabes. Ces contenus téléchargés nous permettront d'identifier la structure des mots et des phrases dans les différentes régions et les différents dialectes du monde arabe‹ ainsi caractériser les traits linguistiques propres à chaque groupe de dialecte.
Une fois téléchargés‹ les contenus des sites seront analysés par des algorithmes afin d'extraire les ressources du corpus tout en respectant une structure uniforme et cohérente pour tout le corpus. Le contenu du corpus est créé lors de cette étape en analysant le code HTML des contenus téléchargés afin d'extraire les informations suivantes si elles existent :

- L'URL et le nom du site téléchargé
- La date et l'heure du commentaire
- L'ID de l'auteur
- Le sous-titre
- L'adresse mail.
- La location géographique de l'auteur
- Le contenu du message.

Exemple : voici le contenu du corpus obtenu après l'analyse du code html suivant :

```
.... CURRENT_URL: http://www.elkhabar.com/ar/autres/athman_snadjki/240186.html ...
<div class=»comment_holder»>
  <a name=»comment_46854»> </a>
  <div class=»comment_header»>1 - RABIE</div>
  <div class=»comment_header_pays»>MARSEILLE</div>
  <div class=»comment_header_time»>2011-01-01 13:30 على م </div>
   <div class=»comment_body_holder»>
     <div class=»comment_body»>
       <div class=»comment_text»>
            ALLAH YARHMEK. INA LILLAH WA INA ILAYHI RAJI3OUN
        </div>
      </div>
   </div>
  </div>
```

Corpus obtenu :

<doc docid=»elkhabar_comment1»
articleURL=»http://www.elkhabar.com/ar/autres/athman_snadjki/240186.html»
author=»1-RABIE» pays=»MARSEILLE» date=»2011-01-01» time=»13:31»>
<comment> ALLAH YARHMEK. INA LILLAH WA INA ILAYHI RAJI3OUN</comment>
</doc>

Transcription du corpus dialectal

Cette phase consiste à transcrire les messages de l'arabe dialectalisé en écriture latine vers l'écriture arabe. Pour réaliser cette tâche nous mettons en œuvre les techniques de transcription et de translittération des messages. La transcription consiste à substituer à chaque son ou à chaque phonème d'un système phonologique‹ un graphème ou un groupe de graphèmes d'un système d'écriture‹ tandis que la translittération consiste à substituer à chaque graphème d'un système d'écriture‹ un autre graphème ou un groupe de graphèmes d'un autre système d'écriture‹ indépendamment de la prononciation. Dans le premier cas (transcription)‹ l'objectif est de reconstituer la prononciation originale à partir de l'écriture d'arrivée; dans le second (la translittération)‹ l'objectif est de retrouver le graphème d'origine à partir du système d'écriture d'arrivée.
Pour atteindre ces objectifs‹ il existe des systèmes de transcription phonologique et des normes de translittération graphématiques. Mais ces systèmes et ces normes conventionnels sont multiples et complexes‹ d'autant plus complexes que les langues mises en contact sont éloignées. Ainsi‹ pour l'arabe‹ il existe plusieurs normes de translittération‹ dont EI (1960)‹ ISO/R 233 (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 utilisées internationalement par la communauté scientifique : il s'agit de la norme DIN-31635 et de la norme adoptée par l'Encyclopédie de l'Islam (EI).
Dans le cadre de cette étape‹ notre objectif est de proposer un système automatique de translittération qui tient compte du lien entre phonologie‹ graphématique et dialectologie‹ dans la transcription des mots arabes écrits en latin vers l'écriture arabe. Pour ce faire‹ nous définissons un certain nombre de règles‹ issues d'une étude expérimentale‹ et qui rendent compte de la complexité du domaine.
Il existe‹ en effet‹ une multitude de cas de figure qu›il convient de traiter en fonction du niveau auquel l›on se situe. Nous récapitulons ces cas dans le tableau suivant‹ établi à partir des situations observées expérimentalement :

Type de traitement automatique	Unité de traitement source	Unité de traitement cible
Translittération	graphème latin : ḫ (FR); kh (EN)، j. (ES)...	graphème d'arabe standard : ex. خ
	graphème d'arabe standard : ex. خ	graphème latin : ḫ (FR); kh (EN)، j. (ES)...
	graphème arabe (d'un autre dialecte) : ex. Egypte g / ج	graphème arabe (d'un dialecte spécifique) : ex. Tunisie g / ڤ
	graphème latin (d'une autre langue) : ex. U (EN).	graphème latin (d'une langue) : ex. ou (FR)
Transcription	phonème latin : k (FR); q (EN)، k (ES)...	phonème arabe (en fonction des dialectes) : ex. ق
	phonème arabe (en fonction des dialectes) : ex. ر ; غ	phonème latin : ex. g (FR) / r (grasseyé)
	phonème arabe (d'un autre dialecte) : ex. ق	phonème arabe (d'un dialecte spécifique) : ex. ء
	phonème latin (d'une autre langue) : ex. z (FR)	phonème latin (d'une langue) : ex. th (EN)

Méthodologie de construction du translittérateur

Nous avons choisi une méthodologie «bottom-up» pour la construction de notre translittérateur. En d'autres termes، nous avons commencé par faire un état des lieux 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 une synthèse des équivalences graphématiques établies à partir de cette étude sur corpus (Saadane et all، 2012) :

Lettre arabe	Équivalent en écriture latine	Lettre arabe	Équivalent en écriture latine	Lettre arabe	Équivalent en écriture latine
ء	ʾ، a	ح	H، h، Ḥ، ḥ، ḥ، 7	س	S، s
ا	A، a، ä، â، á، ā، e، ê	خ	Kh، kh، ḫ، ẖ	ش	Ch، ch، Sh، sh، Š، š
ب	B، b	د	D، d	ص	S، s، Ş، ş، Ṣ، ṣ
ت	T، t	ذ	Dh، dh، D، d، Ḏ، ḏ، Ḍ، ḍ	ط	D، d، Ḍ، ḍ، Ḍ، ḍ
ث	Th، th، t، ṯ،	ر	R، r	ض	Z، z، Ẓ، ẓ، ٦'، Dh، dh، D، d
ج	J، j، Dj، dj، g، Ğ، ğ	ز	Z، z، Ẓ، ẓ	ظ	ʿ، ʻ، ʻ، 3، a، â

ع	ʾ، a	م	H، h، Ḥ، ḥ، ḥ، 7	ة	S، s
غ	A، a، ä، â، á، ā، e، ê	ن	Kh، kh، ẖ، ẖ	ى	Ch، ch، Sh، sh، Š، š
ف	B، b	ه	D، d	أ	S، s، Ṣ، ṣ، Ṣ، ṣ
ق	T، t	و	Dh، dh، D، d، Ḏ، ḏ، Ḍ، ḍ	ؤ	D، d، Ḍ، ḍ، Ḍ، ḍ
ك	Th، th، t، ṯ،	ي	R، r	إ	Z، z، Ẓ، ẓ، ٦'، Dh، dh، D، d
ل	J، j، Dj، dj، g، Ğ، ğ	آ	Z، z، Ẓ، ẓ	ئ	ʿ، ʿ، ʿ، 3، a، â
ك	G، g				

Tableau 1 : Équivalences graphématiques entre l'alphabet arabe et l'alphabet latin

L'étude sutr 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 constitue même la norme dans le langage SMS en usage en Europe et au Moyen-Orient. Le tableau suivant récapitule ces équivalences alphanumériques pour les lettres concernées de l'alphabet arabe :

Lettre de l'alphabet	Equivalence numérique	Lettre de l'alphabet	Equivalence numérique
ء	2	ط	6
ح	7	ظ	6'
خ	7'	ع	3
خ	5	غ	3'
ص	9	ق	8 ou 9
ض	9'		

Tableau 2 : Équivalences alphanumériques entre l'alphabet arabe et l'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...

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 (جـدة) 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 les plus répandus، celui du Prophète Muhammad (محمـد) – 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...}. Même lorsque ce prénom est voyellé (مُحَمَّـدْ)، 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 du corpus les noms suivants qui présentent des formes non conventionnelles en écriture latine : Mu`ammar، Mabrūk، aţ Ţulayḥah، Bū، Yaḥyá، Ḥammūdah، Musţafá،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 mots arabes [Sâadane et all].

Pour plus de détails sur les techniques de la translittération، nous invitons le lecteur à consulter les travaux [Sâadane et all, 2012]

Exemple:

Message original	Transcription
aya chknou bech yehmik tawa minna nahna etwensa el kol enti ktalt bou3zizi wehed emma rahou tawa welw femma 10 999 998 bou3zizi e5er fi tounis !!!!:)	أيا شكون باش يحميك توى منا نحنا التوانسة الكل إنت قتلت بوعزيزي واحد إما راهو توى ولو فمة 10999998 بوعزيزي
ka3ed tkoul fi n'importe quoi et maniche hab noudekhoul fi la3ba dyalek،besah nekoulek hadja eneta kabyle wanna 3arbi، maniche khir menek et makech khir menni	قاعد تقول في نبورت كوا ومانيش حاب ندخل في لعبة ديالك،بصح نقولك حاجة إنت قبايلي وأنا عربي مانيش خير منك وما كش خير مني
iih el araf .da masri w da filastini w da algerien law ihtamitou bi elkathiya el filastinia law kan ahsan .bala kora bala araf	إيه الأرف دا مصري ودا فلسطيني ودا ألجريان لو اهتميتوا بالقضية الفلسطينة لوكان أحسن بلاكوورة بلا أرف

Transcodage du message

Cette étape consiste à trouver pour chaque mot issu de l'étape précédente le ou les mots qui lui correspondent en arabe MSA. L'établissement de cette correspondance permettra d'établir des classes de différences entre l'arabe standard et l'arabe dialectal afin d'extraire ensuite les trais linguistiques propres à chaque groupe de dialecte. Accessoirement à ce traitement، nous proposons à cette étape، à titre indicatif، une traduction française du message

Transcription	Arabe standard	Traduction
أيا شكون باش يحميك توى منا نحنا التوانسة الكل إنت قتلت بوعزيزي واحد إما راهو توى ولو فمة 10999998 بوعزيزي	من ذا الذي سيحميكي الآن منّا نحن التّونسيون إنت قتلت بوعزيزي واحد، لكن أصبح يوجد الآن 10999998 بوعزيزي.	Qui va vous protéger de nous tous les Tunisiens، vous avez tué un seul Bouazizi mais maintenant il y a 10 999 998 Bouazizi en Tunisie !!!! :)
قاعد تقول في نبورت كوا ومانيش حاب ندخل في لعبة ديالك، بصح نقولك حاجة إنت قبايلي وأنا عربي مانيش خير منك وما كش خير مني	إنك بصدد قول تفاهات، لا أريد أن أدخل في لعبتك،لكن أحب أن أقول لك أنت قبائلي وأنا عربي،لستُ بأحسن منك، ولستَ بأحسن مني	toi tu dis n'importe quoi je veux pas entrer dans votre jeu je te dis une seule chose toi le kabyle et moi l'arabe je suis pas mieux que toi et tu n'es pas mieux que moi
إيه الأرف دا مصري ودا فلسطيني ودا ألجريان لو اهتميتوا بالقضية الفلسطينة لوكان أحسن بلاكوورة بلا أرف	ماهذا القرف، هذا مصري ،هذا فلسطيني وهذا جزائري،لو إهتممتم بالقضية الفلسطينية لكان أحسن	C'est vraiment dégoutant، égyptien، palestinien et algérien، si vous vous étiez occupés de la cause palestinienne ça aurait été mieux

Traits de reconnaissance automatique des dialectes arabes

L'approche que nous avons développée ne vise pas une description exhaustive de chaque dialecte mais seulement la mise en évidence de traits linguistiques qui lui sont spécifiques et qui sont susceptibles d'être intégrés à un module de reconnaissance automatique de l'écrit dialectalisé.

L'objectif d'un tel traitement est le « criblage linguistique » qui consiste à passer un texte dialectalisé nouvellement recueilli au crible d'une base de données sémantique. Ce type d'opération vise à préciser les caractéristiques linguistiques et sociologiques de la production écrite par rapport aux données de la base.

Détection des dialectales par le biais des pronoms personnels isolés

Les pronoms personnels isolés se prononcent de façon différente selon les dialectales. Par exemple‹ la troisième personne du singulier en dialecte marocain est prononcée «hûwa» (masculin) et «hîya» (féminin)‹ alors qu'en dialecte libanais c'est «huwweh» (masculin) et «hiyyeh» (féminin). De même‹ la troisième personne du pluriel en marocain est prononcée «hûma»‹ alors qu'en libanais c'est «hinneh». Dans ce cas‹ les deux pronoms n'ont pratiquement plus de points communs‹ ce qui constitue un trait distinctif de ce dialecte dans le corpus.

Pour illustrer le caractère opérationnel de ce type d'indices‹ nous faisons figurer ci-après un tableau récapitulatif des usages du pronom personnel dans les dialectales du Yémen (Guidère‹ 2004). Celui-ci montre qu'on peut affiner la reconnaissance jusqu'au niveau «local» dans ce type d'étude dialectologique.

Régions du Yémen	Hadra Mawt	Shabwa	Mukeyras	Lahej	Dhâlef	Yâfif	ʾAden
Pro. 1er pers. Sg m.f	Ana	Ana	Ana	Anî	Ana	Ani	A n a / A n i

Détection des dialectes par les pronoms et les adverbes interrogatifs

La base contient également une série de pronoms interrogatifs qui permettent de distinguer les dialectes entre eux. Par exemple‹ le pronom «âsh» (que? Quoi?) peut former‹ en étant combiné à d'autres particules‹ des adverbes variés dans les dialectales du Maghreb : *lâsh* (à quoi)‹ *gaddâsh* (combien)‹ *ʿalâsh* (pourquoi)‹ etc.

Dans les dialectales du Machrek‹ ces adverbes sont beaucoup moins fréquents‹ sauf pour les interrogatifs «combien‹ pourquoi»‹ «édesh‹ lesh»‹ mais les pronoms se prononcent et se transcrivent différemment : par exemple‹ en dialectale libanais‹ le pronom se prononce et se transcrit comme une voyelle fermée «*é*»; c'est pourquoi on retiendra le suffixe «esh» au lieu de «ash» comme trait de distinction de ce dialecte.

Détection des dialectales par les pronoms personnels suffixes

Les locuteurs des dialectales du Maghreb (algérien، tunisien، marocain) prononcent la 2ème et la 3ème personne du singulier (masculin) différemment par rapport au dialecte du Machrek. Par exemple : le radical verbal « *na /si /ya* » (oublier)، on dira dans le dialecte maghrébin « nsit*ek* »، « nsit*eh* »، tandis qu'en dialecte du Machrek، on dira « nsit*ak* »، « nist*oh* ». Dans ce cas، pour la détection et la distinction entre les deux dialectales، l'accent sera mis sur *le suffixe* de la 2ème et la 3ème personne du singulier et non pas sur le radical du verbe.

Détection des dialectales par les particules

Les indices de la personne peuvent constituer un critère fiable pour faire la différence entre les dialectes. En effet، les dialectes du Maghreb sont caractérisés par l'utilisation de la particule « n » à la première personne du singulier، à l'inaccompli، alors que cette particule est presque absente dans les dialectes du Machrek، c'est le cas du dialecte libanais qui emploie à la place du « n » d'autres particules comme le « a »، « e »، « u ». Pour illustrer ces propos، prenons l'exemple du verbe « kataba » (écrire) qui est conjugué، à la première personne du singulier à l'inaccompli، comme suit dans les deux dialectes :

- ✓ « âna *n*ekteb » (j'écris) : dialectale du Maghreb
- ✓ « ana *e*kteb ou *u*kteb » (j'écris) : dialectale du Machrek

Détection des dialectales par le schème verbal et la forme passive

Le système phonétique des dialectes du Maghreb présentent la caractéristique de succession de deux consonnes au début du mot qui est rare voire inexistante dans son correspondant (système phonétique) dans les dialectes du Machrek. Cette caractéristique influence notablement sur le schème verbal « fa ʿala» en arabe standard qui se décline en « f ʿel » au Algérie et en « fa ʿal» en Egypte. Voici quelques exemples de l'effet de cette particularité :

Verbe	Arabe Standard	Dialecte Maghrébin	Dialecte du Machrek
frapper	*daraba*	*Dreb*	*Darab*
se taire	*sakata*	*Sket*	*Sakat*
Boire	*charaba*	*chreb*	*Charab*

- Notons que pour la reconnaissance، la conjugaison de ces verbes au passé، à la 3ème personne du singulier، est en soi un élément intéressant de classification.

Dans le même cadre، la forme passive des verbes constitue aussi un élément de distinction des dialectes entre le Maghreb et le Machrek. En effet، en arabe standard la forme passive est dérivée par apophonie de la forme active avec un simple changement du timbre de la mélodie vocalique « a →u ». Cependant، dans les dialectes cette forme est obtenue en ajoutant au verbe à l'accompli le préfixe [t] dans le cas du dialecte du Maghreb et les

préfixes [it] ou [in] dans le cas du dialecte du Machrek. Par exemple la forme passive du verbe «kataba » (écrire) est « ***in***katab » ou « ***it***ktab » au Machrek، et « ***te***kteb » contre la forme« k***u***tiba » en arabe standard.

Le cas de la consonne « q » dans les dialectales arabes

Prononciation des consonnes: le meilleur exemple est l'utilisation de la consonne occlusive uvulaire sourde « ق » [q] dans certaines régions et l'occlusive palatale sonore « ڤ » [g] dans d'autres régions. La consonne « q » est l'un des sons qui méritent une attention particulière. En fait، selon les dialectales، les régions، les villes، et parfois les localités، ce son qui se prononce « q » en arabe littéral، peut être prononcé : [q،a،k، g ou kh]. Ce son est considéré comme une propriété qui traduit un clivage sociogéographique [D. Lajmi، 2009] entre parler citadin et parler rural et encore parler bédouin

Une distinction de base peut s'avérer utile pour un premier classement. En effet، dans la grande majorité des cas étudiés، on peut esquisser quelques tendances générales concernant la prononciation du son « q » :

q → ʾ**a** correspond en général aux parlers des citadins au Machrek;

qalb → ʾalb (coeur); qâla → ʾâl (il a dit)

q → **k** correspond en général aux parlers des ruraux :

qalb → kalb (coeur); qâla → kâl (il a dit)

q → **g** correspond en général aux parlers des bédouins :

qalb → galb; qâla → gâl

q → **q** correspond en général aux parlers de groupes qui sont restés plus ou moins fermés et qui continuent à prononcer le « q » à la manière classique.

qalb → qalb; qâla → qâl

Détection des dialectes par la transcription des lettres

Les systèmes d'écriture latine diffèrent d'une langue à une autre. Par exemple on ne trouve pas en anglais les lettres (é، è، ô، à، ù، â، ê، ç، î...) qui sont utilisées en français (Al-Balawi et all، 2009). Ce fait génère des différences lors de la transcription des mots arabes en écriture latine، car en général les gens du Maghreb sont influencés par la littérature française tandis que les gens du Machrek sont influencés par la littérature anglaise. Citons l'exemple de l'article (ال). Ses règles d'assimilation sont variables d'un dialecte à un autre. Les gens du Maghreb le transcrivent par (**El**) et les gens du Machrek le transcrivent par (**Al**). Le même problème se pose pour certaines lettres comme la lettre (ج) qui est transcrite en (Dj) en Algérie et (J) ou (G) dans une moitié du Maghreb et au Machrek. La lettre (ش)

est transcrite en (Ch) au Maghreb et en (Sh) au Machrek.
Tous ces phénomènes peuvent être observés à partir du corps dialectalisé que nous avons constitué tout au long de l'année 2012-2013. Il permet de reconnaître automatiquement les dialectes arabes écrits en caractères latins et de mieux cerner les spécificités morphosyntaxiques et sémantiques de chaque.

Conclusion

Dans cet article‹ nous avons décrit un outil de reconnaissance automatique des dialectaux arabes à l'écrit‹ dans un premier temps nous avons constitué un corpus de textes/discours dialectaux mais rédigés en caractères latins. Nous avons ensuite utilisé un outil de translittération que nous avons développé pour les retranscrire en écriture arabe. Enfin‹ nous avons cherché à caractériser ces productions langagières écrites en analysant les traits morphosyntaxiques et sémantiques caractéristiques de deux principaux groupes de dialectes (Maghreb et Machrek).

Nos travaux futurs s'orientent‹ d'une part‹ vers une évaluation à une large échelle de notre outil de reconnaissance automatique des dialectales arabes à l'écrit en vue de consolider les résultats déjà obtenus‹ et d'autre part‹ vers une géolocalisation pour identifier comment les différents lexiques et traits linguistiques peuvent fournir des indications sur l'origine et/ou sur le profil de celui qui les rédige (francophone ou anglophone‹ du Maghreb ou du Macherek‹ du nord ou du sud...).

Références

Al-Balawi I.، Al-Baya A. (2009) «Différentes façons d'écrire les noms arabes en latin : formes et raisons». Research Report of King Saud (In arabic)، 2009.

Alsalman Abdulmalik، Mansour Alghamdi، ، Khalid Alhuqayl and Salih Alsubay (2007). «A Computerized System to Romanize Arabic Names». In *Proceedings of The First International Symposium on Computer and Arabic Language* (ISCAL – 07)، 25-28/3/2007، Riyadh، pages 214–227. Bahloul Noureddine (2009)، « L'arabe dialectal، un outil pour une intercompréhension en classe de langue »، in *Synergies*، n° 4، pp. 255-263.

Dichy Joseph (2009) « 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.

Dichy Joseph (2003) « La variation linguistique comme fait culturel : l'exemple de l'arabe et de son enseignement en France. In *Les contenus culturels dans l'enseignement des langues vivante*، Ministère de l'éducation nationale، Académie de Versailles : CRDP، 79–101.

Guidère Mathieu (2004) « Le Traitement de la parole et la détection des dialectes arabes ». In *Langues stratégiques et Défense nationale*، Publications du CREC Saint-Cyr، 53–75.

Harry Benjamin (1996) « The importance of the language continuum in Arabic multiglossia». In Alaa Elgibali، (ed.)، *Understanding Arabic. Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi,* Cairo: the American University of Cairo Press، 69–90.

Lajmi Dhouha (2009) « Spécificité du dialectes Sfaxiens »، Synergies Tunisie n1، pp. 135-142، Tunisie، 2009.

Lentin Jérôme (2008) « Middle Arabic»، In Kees Versteegh et al.، (eds.)، *Encyclopaedia of Arabic Language and Linguistics* , Leiden : Brill، vol. III، 215–224.

Medfaï Ammar (1998) « Réalisations tunisiennes de l'arabe moyen، à partir d'un corpus télévisé»، *thèse de Doctorat en Sciences du Langage*، Université Lumière-Lyon 2.

Saadane Houda (2011)، « Dialectologie arabe et transcription automatique des noms»، Actes des *IX e Rencontres des Jeunes Chercheurs en Parole*، Grenoble 25-27 Mai 2011، page 91-93.

Saadane Houda، Nasredine Semmar، Ouafa Benterki and Christian Fluhr (2012) « Using Arabic Transliteration to Improve Word Alignment from French-Arabic Parallel Corpora». *The fourth Workshop on computational Approaches to Arabic Scrip-based languages*، AMTA 2012، San Diego، CA، USA.

Saadane Houda and Nasredine Semmar (2012) «Utilisation de la translittération arabe pour l'amélioration de l'alignement de mots à partir de corpus parrallèles français - arabe». (2012). In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*، volume 2: TALN، publisher ATALA، AFCP، pages 127–140، Grenoble، 4 au 8 juin 2012.

Saadane Houda، Aurélie Rossi، Christian Fluhr et Mathieu Guidère (2012)، «Transcription of Arabic names into Latin ». The 6th international conference (SETIT) : Sciences of Electronic، Technologies of Information and Telecommunications، du 21 au 24 Mars 2012 à Sousse en Tunisie. Publisher IEEE.

Zaidan Omar F. et Chris Callison-Burch (2011)، «The Arabic Online Commentary Dataset: An Annotated dataset of informal Arabic with high dialectal content. In *Proceedings of ACL*، pages 37 – 41، 2011a.