Une approche linguistique pour la détection des dialectes arabes

Houda Saadane, Damien Nouvel, Hosni Seffih, Christian Fluhr

Une approche linguistique pour la détection des dialectes arabes

Houda Saâdane[1] Damien Nouvel[2] Hosni Seffih[1, 3] Christian Fluhr[1]
(1) GEOLSemantics, 12 Avenue Raspail, 94250 Gentilly, France
(2) ERTIM, INALCO, 2 rue de Lille, 75007 Paris, France
(3) LIASD (EA4383), Université Paris8, 2 rue de la Liberté, 93526 Saint-Denis cedex, France
houda.saadane,hosni.seffih,christian.fluhr@geolsemantics.com,
damien.nouvel@inalco.fr

RÉSUMÉ
Dans cet article, nous présentons un processus d'identification automatique de l'origine dialectale pour la langue arabe de textes écrits en caractères arabes ou en écriture latine (arabizi). Nous décrivons le processus d'annotation des ressources construites et du système de translittération adopté. Deux approches d'identification de la langue sont comparées : la première est linguistique et exploite des dictionnaires, la seconde est statistique et repose sur des méthodes traditionnelles d'apprentissage automatique (n-grammes). L'évaluation de ces approches montre que la méthode linguistique donne des résultats satisfaisants, sans être dépendante des corpus d'apprentissage.

ABSTRACT
A linguistic approach for the detection of Arabic dialects.

In this work, we present the automatic identification process of the dialectal origin of the Arabic language in text written in Arabic characters and in Latin characters (Arabizi). We describe the annotation process of the constructed resources and the transliteration system. We compare two systems : the linguistic one is based on dictionary lookup, the statistical one is based on machine learning (ngrams). the evaluation of those systems shows that the linguistic approach obtains good results, without relying on a training corpus.

MOTS-CLÉS : dialectes arabes, arabizi, alternance codique, translittération, identification des dialectes, analyse morphologique.

KEYWORDS: Arabic Dialects, Arabizi, Code switching, transliteration, Identification of Dialects, Morphological analyzes.

1 Introduction

Le traitement du texte informel est devenu un domaine d'investigation extrêmement populaire (Yang & Eisenstein, 2013). Pour l'arabe, l'identification du dialecte est une tâche devenue centrale pour la plupart des applications de traitement de l'arabe, comme par exemple la traduction automatique ou l'analyse des médias sociaux. Selon (Zaidan & Callison-Burch, 2011), l'identification des dialectes peut être vue comme une identification des langues appliquée à un groupe de langues étroitement apparentées. Pour l'arabe, cette identification est rendue complexe par l'absence de conventions orthographiques et par la translittération en écriture latine «***arabizi***». Des travaux récents ont proposé des approches statistiques pour l'identification de langue. Cependant, les méthodes actuelles reposent sur l'hypothèse que des ressources existent (corpus et dictionnaires), ce qui n'est pas toujours le cas, notamment pour l'arabe maghrébin. L'objectif de notre travail est de construire des ressources adéquates pour l'algérien, le tunisien et le marocain, ainsi que pour le dialecte égyptien. Nous présentons une méthode linguistique pour l'identification de l'origine dialectale et discutons les résultats, en les comparant à ceux obtenus par une méthode statistique.

2 État de l'art

La constitution de ressources et de méthodes pour traiter les dialectes de l'arabe est devenue un enjeu de taille, et nécessite des ressources spécialisées. Le travail de (Zaidan & Callison-Burch, 2014) s'inscrit dans cette optique : les auteurs utilisent les sites de trois journaux arabes des dialectes levantin, golfe et égyptien. L'annotation, réalisée avec Amazon Mechanical Turk, permet une approche de modélisation linguistique et est exploitée afin de prédire le dialecte pour une phrase donnée. D'autres initiatives ont visé à créer des données dialectales afin de remédier au manque de ressources dédiées, comme par exemple l'extraction de commentaires, de journaux en ligne et de Twitter pour les dialectes égyptien, golfe, levantin, algérien et irakien (Cotterell & Callison-Burch, 2014).

Le travail de (Elfardy & Diab, 2012) propose des directives pour la formation de larges corpus de ressources arabes mixtes à code alterné pour le dialecte égyptien et a abouti au système *AIDA "Automatic Identification and glossing of Dialectal Arabic"* (Elfardy *et al.*, 2014), pour l'identification, la classification et l'interprétation des dialectes. Les auteurs de (Elfardy & Diab, 2013) ont ensuite présenté une approche supervisée pour l'identification des phrases en dialecte égyptien et ont étudié les effets des techniques de prétraitement sur la performance de classifieurs.

L'absence de convention orthographique standardisée est une difficulté, objet de travaux proposant une convention de transcription nommée *CODA "Conventional Orthography for Dialectal Arabic"*. Elle a été proposée dans un premier temps pour le dialecte égyptien (Habash *et al.*, 2012), puis étendue à d'autres dialectes comme le tunisien (Zribi *et al.*, 2014), l'algérien (Saâdane & Habash, 2015) et le palestinien (Jarrar *et al.*, 2014).

Ces approches traitent des dialectes écrits en caractères arabes, or ces dernières années ont vu l'apparition de l'arabizi : l'écriture en langue arabe avec des caractères latins. L'arabizi est souvent utilisé dans des contextes informels comme les réseaux sociaux, et alterné avec d'autres langues étrangères, comme l'anglais ou le français. Les quelques outils TAL pour l'arabizi (Saâdane *et al.*, 2013; Darwish, 2013; Eskander *et al.*, 2014) visent au préalable à le détecter dans les textes, à translittérer les textes vers l'écriture arabe, ce qui permet de les traiter avec des systèmes TAL dédiés à l'ASM [1].

L'ensemble des travaux référencés ci-dessus sont focalisés sur l'égyptien et le levantin (jordanien, libanais, palestinien, syrien) et beaucoup moins de travaux ont été réalisés pour les langues maghrébines. Notons également que les ressources présentées ci-dessus ne sont généralement pas disponibles, ce qui rend difficile la reproduction des expériences et la comparaison des résultats. Dans notre travail, nous nous intéressons prioritairement à la constitution de ressources pour les dialectes maghrébins et plus spécifiquement à l'identification automatique du dialecte écrit en caractères arabes et latins : ce sont des problèmes bien réels qui appellent à des solutions concrètes.

1. ASM : Arabe standard moderne

3 Constitution des corpus

Nos ressources linguistiques ont été obtenues à partir d'une collecte de données faite sur les commentaires des lecteurs extraits des journaux arabes ainsi que ceux du réseau social Facebook. Nous avons choisi d'utiliser des sites et journaux spécifiques pour chaque dialecte étudié, comme exposé dans le tableau 2. Cette collecte nous a permis de constituer un corpus segmenté selon le dialecte et le mode d'écriture (arabe ou arabizi). Nous signalons que les commentaires rédigés en français ou en anglais ne contenant pas d'arabizi ont été exclus de nos corpus. Le processus suivi pour la construction de nos ressources est décrit dans (Saâdane *et al.*, 2013).

	ALGERIEN		**MAROCAIN**		**TUNISIEN**		**EGYPTIEN**	
Source	**Echorouk**	**Dzfoot**	**Yabiladi**	**Kifache**	**Aljarida**	**F.B**	**Misr5**	**F.B**
#messages	10.5K	67.5K	15.6K	103.1K	5.5K	13.6K	30.9K	4.3K
#phrases	20.3K	182.3K	100.7K	131.1K	13.7K	20.2K	36.2K	8.1K
#mots	400.9K	1.5M	2.4M	2.5M	70K	263K	31.8K	348K
#msg arabe	5.2K	60K	8.6K	3.3K	4K	5.1K	29.1K	2.1K
#msg latin	5.3K	7.4K	7.0K	99.8K	1.5K	8.4K	1.8K	2.2K

TABLE 1 – Volumes des corpus constitués

L'extraction de ces données, tirant parti de la localisation dialectale des sources, permet de constituer un corpus volumineux et relativement fiable de textes, pour lesquels le dialecte des messages est connu *a priori*. Ils seront ainsi exploités à la fois pour constituer un corpus de test pour les deux méthodes présentées ci-après, et un corpus d'apprentissage pour la méthode statistique.

4 Système de détection des dialectes à base de dictionnaires

4.1 Format d'annotation dialectale des données

Une annotation automatique détermine, pour chaque mot du texte, les étiquettes suivantes :

— **lang1 :** mot en ASM dans un texte arabe, par ex. *AlĀn* [2] «maintenant». Elle annote également dans les textes arabizi les mots étrangers qui conservent leurs formes orthographique comme les mots «belle» du français ou «good» de l'anglais.
— **lang2 :** mot dialectal <AD> dans un texte arabe ou arabizi et information pour un texte arabe : <DZ> (*algérien*), <TN> (*tunisien*), <MA> (*marocain*) ou <EG> (*égyptien*). Cette information est rajoutée dans les textes arabizi après leur translittération.
— **entité Nommée :** une entité nommée, comme *AljazaAŷir* «l'Algérie».
— **autres :** ponctuations, chiffres, sons et émoticônes, URL, etc.
— **ambiguïté :** mot où la classe (lang1 ou lang2) ne peut être déterminée étant donné le contexte courant (par ex. *Tyb* peut être utilisé en ASM «bon» et en dialecte égyptien «ok»), observée aussi dans l'annotation de mots communs à plusieurs dialectes (comme la particule *wqtAš* «quand» qui est utilisée dans les dialectes maghrébins).

2. Translittération arabe présentée dans scheme Habash-Soudi-Buckwalter (HSB) (Habash *et al.*, 2007)

4.2 Translittération de l'arabizi

Après avoir annoté les textes dialectaux écrits en latin, les messages en arabizi sont convertis en écriture arabe, en suivant la convention CODA. Cette translittération arabizi-arabe se focalise d'abord sur les entités nommées en utilisant un système (Saâdane *et al.*, 2012) qui convertit l'arabizi en texte arabe, et inversement, par utilisation de transducteurs à états finis. Les translittérations possibles sont filtrées à l'aide d'un analyseur morphologique de l'arabe. Pour les mots qui ne sont pas des entités nommées, un translittérateur basé sur une approche à base de règles et de dictionnaires spécifiques pour chaque dialecte est utilisé. Les dictionnaires regroupent 24442 paires de correspondances (mots, mais aussi bigrames ou trigrammes) arabizi-arabe répartis entre les dialectes étudiés. Ces correspondances ont été établies par des natifs arabes et qui ont translittéré manuellement les motifs ayant le plus grand nombre d'occurrences dans des corpus. Les règles proposées permettent de définir les frontières des lettres dans un mot arabe et d'appliquer des règles d'exception pour chaque mot afin de supprimer les variantes orthographiques qui ne sont pas attestées dans l'usage. La liste générée est ensuite filtrée à l'aide de l'analyseur morphologique pour prédire l'appartenance d'un mot à un des dialectes étudiés.

Si la conversion de l'arabizi en écriture arabe est une étape importante du processus, cet article se focalise sur l'identification des dialectes et nous ne pouvons, faute de place, détailler plus avant cette partie du traitement. Notons cependant que ce processus conserve une information importante pour identifier le dialecte : la présence de voyelles en arabizi, permet d'indiquer les diacritiques en arabes (ils sont généralement omis), ce qui apportera un indice décisif pour déterminer le dialecte d'un texte.

4.3 Système ambigu d'annotation des mots à base de dictionnaire

L'originalité de notre approche d'annotation réside dans la production d'analyses morphologiques diacritiques. Les dialectes maghrébins produisent effectivement de nombreux morphèmes et éléments lexicaux qui sont tout à fait semblables, et ne diffèrent souvent que par les voyelles courtes. Si cette information n'est pas écrite en ASM, elle est en revanche présente par translittération de l'aribizi, ce qui va aider à la reconnaissance du dialecte.

Notre approche d'annotation est résumée comme suit :

— **Prétraitements :** séparation de la ponctuation et des nombres, normalisation des effets d'allongement de lettres, détection des URLs et des nombres, étiquetetage de la ponctuation, des émoticônes et des sons.

— **Analyseurs morphologiques :** des analyseurs morphologiques réalisent la segmentation, la lemmatisation et l'étiquetage pour déterminer des traits morphologiques et l'appartenance d'un mot à l'ASM, à un dialecte (DZ,TN, MA ou EG), au français ou à l'anglais. Nous utilisons par ailleurs le système présenté dans (Fluhr *et al.*, 2012) pour l'identification des mots étrangers, des entités nommées et des mots dialectaux (les mots hors vocabulaires) dans des textes arabizi. Après une phase de translitération et d'étiquetage des mots en arabizi et des noms propres, nous appliquons une analyse morphologique sur le texte arabe obtenu. Enfin, nous annotons les éléments obtenus et les textes en arabe via le système présenté dans (Saâdane, 2013; Saâdane, 2015).

— **Liste des entités nommées :** ressources proposées par ANERGazet (Benajiba & Rosso, 2007) et GeoNames [3] pour identifier les entités nommées arabes, français et anglais, dont :

3. http ://download.geonames.org/export/dump/

— **personnes :** 7387 entrées,
— **locations :** 73892 entrées (pays, villes, continents, etc.),
— **organisation :** 22772 entrées (entreprises, équipes de football, etc.).

Une étape de combinaison permet d'agréger plusieurs composants, (analyseurs morphologiques, dictionnaires des entités nommées) afin d'effectuer l'identification de la langue du texte en entrée. Chaque mot de la phrase peut obtenir différentes étiquettes selon chaque composant. L'étape de combinaison, en se basant sur les étiquettes générées, utilise de plus un ensemble de règles de décision qui affectent un tag final à chaque mot de la phrase d'entrée, en donnant la priorité aux règles les plus sûres (ponctuations, dictionnaires) tout en conservant de l'ambiguïté, en particulier si les mots sont reconnus comme pouvant appartenir à plusieurs dialectes. Les règles de décision utilisées sont présentées comme suit :

— Si le mot contient des numéros ou des signes de ponctuation, alors il est associé à la balise *Autre* (Ponct, NUM, etc) ;
— Sinon si le mot est présent dans l'un des dictionnaires ou si les analyseurs morphologiques assignent la balise entité nommée, alors le mot est étiqueté comme *Entité nommées <NE>* ;
— Sinon si le mot est identifié par les analyseurs que ce soit *Lang1* ou *Lang2*, le mot est alors associé à l'étiquette correspondante ;
— Sinon si le mot identifié est associé à la fois à *Lang1* et *Lang2*, alors nous attribuons au mot les balises *Lang1* et *Lang2*. Toutefois ce cas introduit une certaine ambiguïté.
— Sinon si les analyseurs n'étiquettent pas le mot, cela veut dire que le mot est considéré comme un mot hors vocabulaire, alors nous associons la balise *<UNK>* au mot analysé.

4.4 Désambiguïsation des messages

Après avoir annoté les mots des messages, le système de détection de dialecte a pour objectif de déterminer le dialecte du corpus analysé, en général ou texte par texte. Pour cela, les seuls indices disponibles sont les annotations dialectales présentées dans l'étape précédente. Deux méthodes ont été implémentée pour déterminer le dialecte.

La première est basée sur le nombre de mots discriminants, soit ceux qui n'appartiennent qu'à un seul des dialectes candidats. L'hypothèse est que la présence d'un mot discriminant suffit à donner une bonne idée du dialecte. Cette approche permet de renvoyer le nombre d'occurrences de chaque label ainsi que le dialecte détecté. L'inconvénient de cette méthode est que la détection est impossible si le message ne contient aucun mot propre à un dialecte donné, auquel cas le système fait appel à la deuxième méthode.

La seconde méthode est basée sur un système de notation pondéré. Pour un mot donné, le poids d'un dialecte est inversement proportionnel au nombre de dialectes détectés. Par sommation sur les mots d'un texte, cette méthode renvoie le dialecte qui a le poids le plus important.

5 Systèmes statistique de référence

Pour la méthode statistique, deux approches ont été mises en œuvre et évaluées. Elles diffèrent par le niveau d'analyse : la première s'appuie sur les mots (approche lexicale), la seconde sur les caractères (approche graphémique). Des modèles de langage N-grammes de mots ou de caractères sont construits pour chaque dialecte, et un modèle probabiliste détermine le langage le plus vraisemblable pour un texte donné. La qualité du modèle dépend donc entièrement du corpus d'apprentissage et de sa représentativité pour chaque dialecte, directement liée à la quantité de données disponibles. Lors de la phase de test, le système calcule le score de chaque message de test avec les modèles n-grammes de chaque dialecte, et renvoie celui qui obtient le meilleur score.

6 Résultats expérimentaux

6.1 Corpus de test

Nous avons sélectionné 820 messages issus des corpus dialectaux, écrits en caractères arabes et latins. Les deux systèmes linguistiques et statistiques sont comparés en calculant le taux d'erreur d'identification des dialectes (LER), c'est-à-dire la proportion de messages pour lesquels le dialecte est incorrectement détecté. Deux méthodes sont évaluées sur ces corpus : la première est purement statistique, la seconde exploite les dictionnaires comme décrits en section 4.1. Les résultats des deux méthodes sont présentés dans le tableau 2.

Dial. (#nb-ar/#nb-lat)	Approche lexicale	Approche graphémique	Approche dictionnaire
ALG (102/100)	38,6% (38/40)	39,1% (37/42)	20,3% (27/14)
EGY (105/103)	03,3% (0/7)	03,3% (0/7)	03,3% (2/5)
MAR (101/102)	15,3% (13/18)	11,3% (11/12)	14,8% (15/14)
TUN (100/107)	08,7% (10/8)	15,0% (22/9)	20,3% (26/16)
TOUS (408/412)	16,3% (61/73)	17,1% (70/70)	14,5% (70/49)

TABLE 2 – Taux d'erreurs de l'identification avec les méthodes linguistique et statistique

6.2 Résultats pour la méthode statistique

Pour la méthode statistique, le taux d'erreurs pour l'identification du dialecte (LER) obtenu sur les corpus de test est de 16.3% avec l'approche lexicale et de 17.1% avec l'approche graphémique. D'un point de vue global, les résultats selon que les algorithmes soient testés sur des textes écrits en arabe ou en latin sont assez similaire : 17,2% en arabe versus 17,0% pour le latin avec l'approche graphémique, et respectivement 15,0% et 17,7% pour l'approche lexicale. Pour ce qui concerne les dialectes, nous constatons d'emblée la difficulté de distinguer les textes algériens et marocains. L'égyptien est particulièrement bien détecté, ce qui s'explique de par la distance linguistique de l'arabe égyptien avec les autres dialectes. Les deux approches donnent des résultats assez proches, à l'exception du marocain (l'approche graphémique fonctionne mieux) et du tunisien (c'est l'approche lexicale qui l'emporte).

6.3 Résultats pour la méthode à base de dictionnaire

L'approche à base de dictionnaire donne un taux d'erreur de 14,5%. Sur le corpus en écriture latine, ce taux d'erreur descend même jusqu'à 11,8%, ce qui pourrait être lié à la présence de voyelles comme indice discriminant supplémentaire par rapport à l'écriture arabe. Il y a peu d'erreurs pour l'égyptien. L'algérien obtient un LER de 26,4% , important mais bien moins élevé que celui obtenu avec les approches statistiques (plus de 35%). Le marocain obtient des résultats proches de ceux des approches statistiques. Enfin, seul le tunisien présente de moins bons résultats, avec 20,3% (contre 11% en moyenne avec les approches statistiques).

En complément, nous avons étudié les résultats ambigus de la méthode à base de dictionnaire. Il est en effet possible que cette méthode ait un choix entre plusieurs dialectes avant de prendre une décision. Un oracle permettrait alors de choisir le bon dialecte parmi ceux envisagés. Tel est le cas pour 29/24 messages algériens, 5/4 égyptiens, 31/12 marocains et 30/21 tunisiens. Le même constat est dressé : la confusion porte sur l'algérien, le marocain et le tunisien, et une méthode plus fine serait bienvenue pour distinguer ces dialectes.

6.4 Discussion

Sans surprise, cette étude montre que la détection des dialectes est une tâche difficile lorsque l'on confronte des dialectes proches, comme cela a été constaté pour l'algérien, le marocain et le tunisien, et ceci quelle que soit la méthode choisie. Les résultats rapportés montrent que, dans les conditions d'expérimentation réalisées, la méthode statistique basée sur des n-grammes de mots ou de caractères est confrontée aux même difficultés qu'une approche basée sur des ressources lexicales (dictionnaires) et morphologiques (règles de flexion). Il semble pourtant que cette dernière donne de meilleurs résultats, ce qui pourrait être lié à l'insuffisance des corpus d'apprentissage par rapport aux lexiques constitués sur lesquels s'appuient l'approche à base de dictionnaire. Comme cela a déjà été constaté pour d'autres tâches, un meilleur contrôle de la qualité des ressources permet de garantir de meilleurs résultats en sortie. Une analyse plus détaillée des résultats nous a permis de constater quelques difficultés pour constituer un corpus de qualité : messages écrits dans un dialecte mais collectés dans un corpus d'un autre dialecte, présence de mots anglais ou français identiques à des translittération de mots dialectaux. Même si ces phénomènes restent marginaux, il reste impératif, pour évaluer correctement les résultats, de porter grande attention à la qualité des corpus.

Une avancée importante de notre travail porte sur le traitement de l'arabizi. Non seulement cette écriture peut être translittérée en arabe, mais nous constatons de plus que l'ajout des voyelles courtes par rapport à l'écriture arabe, aide à mieux distinguer des dialectes.

7 Conclusion

Dans cet article, nous avons décrit un système de détection de l'origine dialectale de textes écrits en caractères arabes ou latins (arabizi). Nous avons constitué des corpus pour quatre dialectes auxquels nous nous sommes intéressés : l'algérien, le tunisien, le marocain et l'égyptien. Un translittérateur de l'arabizi vers l'arabe et des analyseurs base de dictionnaires ont été développés. Les expériences montrent qu'une approche contrôlée à base de dictionnaires obtient de meilleurs résultats qu'une approche statistique, même si le système reste à améliorer : l'algérien, le tunisien et le marocain restent difficiles à distinguer. Nos futurs travaux s'orientent, d'une part, vers une évaluation à large échelle de notre outil d'identification des dialectes en vue de consolider les résultats déjà obtenus, et d'autre part, vers la constitution d'outils et de ressources pour d'autres dialectes arabes.

Remerciements

Le projet a été soutenu par la DGE (Ministère de l'Industrie) et par la DGA (Ministère de la Défense) : projet RAPID "ORELO", référencé par le N :142906001.

Références

BENAJIBA Y. & ROSSO P. (2007). Conquering the ner task for the arabic language by combining the maximum entropy with pos-tag information. In *Proceedings of Workshop on Natural Language-Independant Engineering*.

COTTERELL R. & CALLISON-BURCH C. (2014). A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*, p. 241–245.

DARWISH K. (2013). Arabizi detection and conversion to arabic. p. 217—-224, Doha, Qatar. Association for Computational Linguistics.

ELFARDY H., AL-BADRASHINY M. & DIAB M. (2014). Aida : Identifying code switching in informal arabic text. p.9̃4 : Citeseer.

ELFARDY H. & DIAB M. T. (2012). Simplified guidelines for the creation of large scale dialectal arabic annotations. In *LREC*, p. 371–378 : Citeseer.

ELFARDY H. & DIAB M. T. (2013). Sentence level dialect identification in arabic. In *ACL (2)*, p. 456–461.

ESKANDER R., AL-BADRASHINY M., HABASH N. & RAMBOW O. (2014). Foreign words and the automatic processing of arabic social media text written in roman script. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, p. 1—12, Doha, Qatar. Association for Computational Linguistics.

FLUHR C., ROSSI A., BOUCHESECHE L. & KERDJOUDJ F. (2012). Extraction of information on activities of persons suspected of illegal activities from web open sources. p.1̃9.

HABASH N., DIAB M. T. & RAMBOW O. (2012). Conventional orthography for dialectal arabic. In *LREC*, p. 711—-718, Istanbul, Turkey.

HABASH N., SOUDI A. & BUCKWALTER T. (2007). On arabic transliteration. In *Arabic computational morphology*, p. 15–22 : Springer.

JARRAR M., HABASH N., AKRA D. & ZALMOUT N. (2014). Building a corpus for palestinian arabic : a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Ar-abic Natural Langauge Processing (ANLP)*, p. 18—-27, Doha, Qatar. Association for Computational Linguistics.

SAÂDANE H. (2015). Traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmiques. In *PhD thesis, Université Grenoble Alpes*.

SAÂDANE H. (2013). Une approche linguistique pour l'extraction des connaissances dans un texte arabe. Les Sables d'Olonne, France : TALN-Récital.

SAÂDANE H., GUIDERE M. & FLUHR C. (2013). La reconnaissance automatique des dialectes arabes à l'écrit. In *colloque international «Quelle place pour la langue arabe aujourd'hui*, p. 18–20.

SAÂDANE H. & HABASH N. (2015). A conventional orthography for algerian arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 69–79, Beijing, China : Association for Computational Linguistics.

SAÂDANE H., ROSSI A., FLUHR C. & GUIDERE M. (2012). Transcription of arabic names into latin. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, p. 857–866.

YANG Y. & EISENSTEIN J. (2013). A log-linear model for unsupervised text normalization. In *EMNLP*, p. 61–72.

ZAIDAN O. & CALLISON-BURCH C. (2014). Arabic dialect identification. p. 171–202.

ZAIDAN O. F. & CALLISON-BURCH C. (2011). The arabic online commentary dataset : An annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Short Papers - Volume 2*, HLT '11, p. 37–41, Stroudsburg, PA, USA : Association for Computational Linguistics.

ZRIBI I., BOUJELBANE R., MASMOUDI A., ELLOUZE M., BELGUITH L. H. & HABASH N. (2014). A conventional orthography for tunisian arabic. In *LREC*, p. 2355—-2361, Reykjavik, Iceland.