Les nouveaux défis du TAL
Exploration des médias sociaux pour l'analyse des sentiments : Cas de l'Arabish

Par **Mohamed HASSOUN** /Université ENSSIB Lyon
Sinda BELHADJ / Université Lyon II

Résumé:
Les spécialistes du traitement automatique de l'arabe se sont toujours intéressé à l'arabe littéraire, le seul arabe écrit. La prolifération des médias sociaux dans le monde arabe a fait apparaître un nouveau type d'écriture appelé « Arabish » ou « arabizi ».Cet article présente une méthode pour l'analyse automatique de ce nouveau mode d'écriture en vue de la mise en place d'un système informatique pour l'analyse des sentiments en partant d'un recueil de corpus.
Mots clés: Arabish, réseaux sociaux, analyse de sentiments

Abstract:
Natural language processing has always been interested in Modern Standard Arabic, the only written Arabic. The proliferation of social media in the Arab world contributed in the emergence of a new writing style called «Arabish» or «arabizi». This article presents a method for the sentiment analysis of an Arabish corpus.

Key words: Arabish, social media, sentiment analysis

Introduction

Après l'émergence d'Internet et le développement de l'échange de données, la nouvelle ère des réseaux se tourne vers les réseaux sociaux comme Facebook et Twitterpour ne citer que les plus connus (en 2012, le nombre des utilisateurs a atteint 1.11 milliard pour Facebook[179]).Le monde arabe dans sa diversité prend une part importante dans l'utilisation des médias sociaux avec une affluence massive des diverses cultures langagières. La particularité de la communication des arabophones sur la toile c'est l'usage du dialecte transcrit en caractères latins (l'arabish). Plusieurs domaines et applications s'intéressent aux discussions et aux commentaires faits sur divers sujets. L'analyse des sentiments est une nouvelle application qui utilise les outils du TAL pour extraire des informations subjectives dans les sources de données textuelles.

Réseaux sociaux et apparition d'une nouvelle écriture

Les réseaux sociaux ont révolutionné les modes de communication et les modes d'écriture. Les plateformes, tels que Facebook, Twitter ou MySpace ont permis aux utilisateurs d'être producteurs de l'information. Les sites des réseaux sociaux sont devenus récemment la technologie la plus utilisée dans la communication médiée par ordinateur (Ku et al, 2013).C'est là où les communautés des utilisateurs en ligne développent des relations interpersonnelles et partagent un contenu généré par d'autres utilisateurs (V.R.JUNCO et al, 2011; R.JUNCO et A.W.CHICKERING, 2010). Les réseaux sociaux sont l'espace de prédilection des internautes pour l'expression des opinions. La prolifération de ces supports de communication numérique a contribué à l'essor d'un nouveau type d'écriture basé sur la phonétique, les raccourcis d'écriture, les abréviations et les émoticônes. En arabe, les claviers, qu'ils soient sur téléphone ou sur ordinateur, ne comportaient souvent pas de caractères arabes, et même s'ils existaient, il était fastidieux pour plusieurs communautés de faire la saisie en caractères arabes.

[179] http://www.prnewswire.com/news-releases/facebook-reports-first-quarter-2013-results-205652631.html

Le langage Arabish

Le flux d'échange de messages et d'écrits dans la téléphonie mobile et sur la toile a fait émerger une écriture arabe libre, nouvelle et unique dans sa forme, combinant à la fois des lettres et des chiffres. Le language Arabish, appelé aussi "Arabizi" (V.M.YAGHAN, 2008), «Franco-Arabe», «3ngleezy» (V.L.BAHRAINWALA, 2011), «Arabic Chat Alphabet» (V.M.ELMAHDY et al, 2012) ou encore «arabe latinisé» (V.M.ABOELEZZ, 2009) est un langage qui permet de compenser les lettres arabes qui n'ont pas d'équivalent dans l'alphabet latin. D'un point de vue terminologique, le mot arabish est une fusion hybride entre Arabic et English. Plusieurs définitions ont été données à ce terme. Nous retenons celle d'Elmahdy et al (2012) : « *L'Arabish est un système d'écriture de l'arabe dans lequel les lettres latines remplacent celles de l'arabe. Les phonèmes arabes qui n'ont pas d'équivalents en anglais sont remplacés par des chiffres proches en forme à la lettre arabe correspondante* ».Ce langage est une émanation de pratique d'écriture des arabophones et aussi une nécessité pour une communication urgente et rapide sur les réseaux de discussion.

Il est vrai que les messages en Arabish peuvent transcrire l'arabe littéraire mais l'influence du dialectal est beaucoup plus importante. Le dialecte arabe est une langue mélangée avec de nombreuses autres langues. La nature non-standard de l'alphabet tient compte de la variation régionale et personnelle de ses utilisateurs (Rodrigues, 2012).L'Arabish est, essentiellement, une forme écrite de l'arabe vernaculaire. Comme les dialectes arabes présentent de grandes divergences, il serait réducteur de considérer que l'Arabish est unique pour tous les pays arabes. Si dans sa définition le principe est le même –le fait d'utiliser l'alphabet latin et les chiffres pour écrire en arabe- il n'en demeure pas moins que la différence est grande quant à la manière dont ces outils sont mis en pratique. Un internaute tunisien écrirait «ch» pour désigner la lettre «ش » tandis qu'un égyptien ou un syrien la symbolisera par «sh». L'Arabish connaît plusieurs variantes dépendant du dialecte de son auteur et aussi de la seconde langue. A notre sens, le langage Arabish est spécifique à chaque pays arabe, il existe donc un Arabish tunisien, un Arabish marocain, un Arabish syrien ou encore un Arabish égyptien. Notre étude s'intéresse à l'Arabish tunisien.

Un corpus Arabish pour l'analyse de sentiments

La langue de prédilection des arabophones sur la toile c'est l'Arabish. C'est essentiellement dans cette langue qu'ils expriment leurs sentiments. La première étape de tout projet d'analyse de sentiments est le sourcing (V. D.BOULLIER et A. LOHARD, 2012). Etant donné que ce langage proche du dialectal ne peut être utilisé que dans les zones d'expression libre, nous avons choisi comme sources de données les réseaux sociaux généralistes de type Facebook où l'utilisation de l'Arabish est très fréquente. Ces réseaux sont en expansion continue dans tous les pays arabes (par exemple en Tunisie où 95.16% des internautes ont un compte Facebook[180]).

A l'issue de l'étape de sourcing, nous avons éliminé tous les bruits des pages et isolé les messages écrits par les internautes, par la suite nous avons identifié parmi les messages ceux en Arabish. Pour cela, nous avons effectué un balisage structurel des pages retenues qui va nous conduire à l'identification des messages. Dans notre étude, nous avons opté pour une démarche active de collecte du corpus en identifiant les sources et en téléchargeant l'information, et ce pour éviter d'orienter les discussions autour d'un sujet imposé. Nous avons donc constitué un corpus regroupant un ensemble de commentaires qui répondent à un post d'origine. Notons que dans d'autres projets sur les SMS en France « SMS 4 Science » (V.C.FAIRON et al, 2006b), la

[180] http://www.socialbakers.com/facebook-statistics/tunisia

collecte du corpus s'est faite par des campagnes de « dons » pour la recherche invitant les usagers à envoyer leurs SMS à un numéro particulier pour constituer le corpus de travail.

Structuration des données

Pour pouvoir isoler les messages des utilisateurs, il faut définir un format pouvant identifier toute la structure. Nous avons défini la structure XML des données que nous allons traiter (V.S.BELHADJ, 2013).

Pour extraire les messages, nous avons récupéré les balises du code source HTML de chaque page. Le langage PHP utilisé pour l'extraction des messages nous a permis d'identifier le début et la fin de balises HTML et de récupérer l'information existante entre les deux balises. Nous avons, par la suite, créé un fichier XML pour structurer les données. Pour améliorer l'affichage des informations, nous avons créé une feuille de style XSLT ainsi qu'une interface web qui permet de choisir les données à extraire.

Acquisition et traitement des données

L'acquisition s'est faite d'une manière automatique. Le crawling nous fournit un corpus brut affichant plusieurs données telles que les photos de profil, les pseudos, le nombre de « j'aime » ou la date et heure du commentaire. Nous avons balisé les données selon le format XML pour les commentaires.
Après la normalisation des données, nous disposons d'un corpus brut présentant des informations à anonymiser (noms, prénoms, surnoms) et du bruit (chiffres, symboles etc), des liens et des messages à titre commercial, des messages en plusieurs langues.

Via l'interface réalisée pour l'affichage structuré des données, nous avons procédé à l'anonymisation des commentaires en supprimant les noms (ou pseudos) des utilisateurs. Toutefois, dans les commentaires postés sur les réseaux sociaux, les noms et pseudos des utilisateurs existent même dans le contenu des messages. Ceci rend leur anonymisation plus critique, surtout automatiquement. Nous avons donc utilisé plusieurs expressions régulières de type (@ ? [A-Za-zأ-ي]* [A-Za-zأ-ي]* ?(:|@|;))

Certaines données telles que la date et l'heure du post des commentaires et le nombre de mentions de type «j'aime» par message sont intéressantes pour une étude des fils de discussion.

Répétition de lettres

Dans notre corpus, la répétition de lettres est un phénomène prédominant

(Exemple :raaabiiii m3aahaaaaaaaaaaaaa).

Nous trouvons qu'il serait judicieux d'éliminer cette répétition pour harmoniser le contenu lors de l'analyse. Nous gardons l'information sur l'intensité du commentaire pour l'analyse des sentiments.

Pour les mots contenant une répétition de la lettre «a», nous avons remplacé les répétitions qui sont égales ou supérieures à trois «a» par deux «a» :

(baaaashaaar = baashar).

Nous avons gardé la répétition de deux «a» car elle peut signifier la voyelle longue. De même pour les lettres «b», «c», «d», «e», «é», «è», «f», «g», «h», «i », «j», «k», «l», «m», «n», «o»,

«p», «q», «r», «s», «t», «w», «y», et «z», nous avons remplacé les répétitions à partir de 3 fois. Pour les lettres « u », «v» et «x», nous avons remplacé les répétitions au-delà de deux lettres par une seule lettre :

(bravvvo = bravo)

Onomatopées et interjections

Nous avons décidé de garder les onomatopées (hh, pff, hum, boff) ainsi que les interjections (aie, chut..) et supprimer les répétitions de bigrammes (exemple: on remplace «pffffff» par «pff»)

Identification de la langue

Les fils de discussion écrits sur les réseaux sociaux sont souvent un mélange de messages en arabe, en français, en anglais et en Arabish. Notre collecte de données a abouti à un corpus composé de messages hétérogènes écrits avec des scripts différents, utilisant une ou plusieurs langues à la fois.

L'identification des messages en caractères arabes est une tâche relativement simple contrairement à l'isolement des messages en Arabish. L'identification automatique des messages en Arabish s'avère très compliquée vue la ressemblance de cette dernière avec les deux autres langues à détecter, à savoir le français et l'anglais. Le point commun entre ces langues c'est qu'elles s'écrivent toutes en lettres latines. L'Arabish utilise largement les chiffres comme c'est le langage inventé par les internautes pour remplacer les lettres arabes qui n'ont pas d'équivalent phonétique dans les lettres latines. Le français et l'anglais ont rarement recours aux chiffres (sauf s'il s'agit d'une écriture en langage SMS comme dans 2m1= demain ou 4 u= for you). Le français utilise des caractères accentués, qui n'existent pas dans l'alphabet de l'anglais (é, è, à, ô…) mais ces derniers ne sont pas suffisamment discriminatoires.

Pour la distinction des trois langues (Arabish, français et anglais), nous avons utilisé la technique des n-grammes de caractère, largement utilisée dans l'identification des langues mais jusque-là non vérifiée sur l'Arabish.

Pour l'utilisation des n-grammes, nous avons besoin de corpus d'apprentissage. Nous avons construit 3 corpus d'apprentissage pour chacune des langues (de taille 39 Ko chacun). Notons que ces corpus sont différents de notre corpus de validation des n-grammes. Nous avons calculé le nombre d'occurrence des bigrammes (n=2) et des trigrammes (n=3) de ces textes d'apprentissage ainsi que leur fréquence d'apparition. Ces tableaux serviront de listes de référence pour chaque langue. Par la suite, pour chaque message ou «texte de référence» nous avons construit son tableau de n-grammes que nous avons confronté avec le tableau de référence. Grâce à une mesure de similarité nous avons déterminé la langue du dit texte de référence.

Khi 2 pour la mesure de similarité

Pour identifier la langue sur un ensemble de textes, nous avons utilisé le test de similarité du $\chi 2$ qui s'avère le plus adapté pour mesurer la proximité entre des entités textuelles. Après l'extraction des bi-grammes, nous avons appliqué le test du Khi-deux ($\chi 2$) pour chaque texte de référence.

La classification des messages par les bigrammes donne de très bons résultats pour l'Arabish et le français (respectivement 94 et 93%), ce qui n'est pas le cas pour l'anglais avec 53%. Ceci n'empêche d'avoir un taux de classification total satisfaisant (94%).

Le taux de classification total des messages par les trigrammes est inférieur à celui fait par les bigrammes (on passe de 94% à 89%). Toutefois, on remarque que la classification par les trigrammes donne de meilleurs résultats pour le français (le taux passe de 93% à 95%), et pour l'anglais (de 53% à 70%) contrairement à l'Arabish où le taux diminue de 94% à 91%.

On note généralement que :

- Les trigrammes sont plus performants pour le français et l'anglais
- Les bigrammes sont plus performants pour l'Arabish

Particularité orthographique de l'Arabish

Nous avons étudié tous les cas de verbes et de noms qui figurent dans notre corpus en reprenant les transcriptions phonétiques selon les lettres arabes, consonnes et voyelles, et leurs prononciations en dialecte tunisien écrit en Arabish.

La caractéristique générale de l'orthographe de l'Arabish consiste à :

- Des consonnes qui ont un équivalent en langue latine ou anglo-saxonne. Celles-ci prennent la forme latine ou anglo-saxonne (d= د, m= م). Certaines consonnes nécessitent des fois plusieurs lettres.
- Des consonnes qui sont sans équivalents en langue latine/anglo-saxonne. Celles-ci sont transcrites sous forme de chiffres, choisis souvent pour leur ressemblance morphologique avec la lettre qu'ils représentent (3=ع, 7= ح).
- L'existence de voyelles: les voyelles qui sont absentes en arabe standard sont transcrites en Arabish.

L'analyse du langage Arabish tunisien nous a permis de présenter les caractéristiques suivantes:

- La lettre «ح» est transcrite par le chiffre «7» ou par «h».
- Al hamza « ء» est dans la plupart des cas transcrite par «a» ou par le chiffre «2» (a3tini).
- Alif al madd «ا» est généralement transcrite par une voyellation courte «a» ou bien par une voyellation longue «aa», et dans certains cas par «e».
- La lettre «ع» est transcrite dans la plupart des cas par le chiffre 3, elle est des fois non transcrite (عُولة=oula; شَعْب= chaab)
- Les lettres «ض » ou «ظ » s'expriment des fois par « th » et dans d'autres par «dh» ou bien par le chiffre « 8 ».
- La lettre « غ » est désignée par «gh» et dans certains cas par le chiffre «8».
- La lettre « ذ » est transcrite par «th», «dh» ou aussi par «d» (dawekni pour dire ذوّقني)
- La lettre «ق» est transcrite, dans la plupart des cas, par le chiffre « 9 » et des fois par «k».
- On ne fait pas la différence entre une «voyellation courte» et une «voyellation longue». La syllabe «ya» peut désigner la lettre « ي» à l'état accusatif, ce qui donne يَ et au même temps elle peut désigner « ي» avec alif al madd « ا» ce qui donne « يَا». Ceci est valable aussi pour « ها», « نا»... (ykallamha = يكلّمها). De même pour la voyelle longue (ي) comme dans tji= تجي
- Les mots en arabe finissant par تاء مربوطة ne la présentent pas à la fin quand il s'agit de l'Arabish (7orra= حُرّة, ta7founa= تحفونة).
- La « chadda » est des fois transcrite par la répétition des consonnes (ex: حُرّة =7orra; رَبِّي =rabbi) et des fois non représentée (خلّيها= 5aliha; مابنّك=mabanek). D'autre part, la répétition des consonnes en Arabish ne signifie pas toujours la présence d'un mot avec « chadda » (yasser=يَاسر)
- Les verbes conjugués commençant avec « ي» s'écrivent des fois avec «y» et par moments avec «i» (i3aychek/ y3aychek = يعيّشك)
- La conjonction de coordination «و » est parfois séparée par deux espaces «w» et des fois liée au mot qui la suit (wi3ayechhom)

- Les noms ou verbes finissant par «ء» représentent la fin par une voyellation qui diffère selon la prononciation de l'utilisateur. (جاء=ja/ jé/ jaa; ما شاءالله=ma chaallah/ ma cha2 allah).
- La voyellation par «kasra» est représentée par «i» ou «e» (مِل = mil/ mel; خاطِر = 5atir/ 5ater/ khater; يمْشِي = yemchi/ yemchy/ yimchi/ yimchy), c'est le cas aussi pour les voyellations longues «ا» et «ي». Généralement, le même mot peut être écrit différemment selon la prononciation ou le dialecte de l'utilisateur (يَاسِر = yesser/ yassir; نَاس = nes/ nas/ ness ; قالِك =9allek/ 9allik/ galik/ guellek)
- La lettre «ي » est parfois omise au début des mots (3aychek pour dire يعيّشك; fadala/ fadhalha pour dire يفضّلها).
- Le langage utilisé par les internautes Tunisiens dans les réseaux sociaux est généralement au présent, on remarque l'omniprésence de verbes commençant par «y» ou «n» (yestana, na3mlou)
- «th» peut signifier «ث », «ض», «ظ» comme dans thawra= ثورة, dho7k= ضحك, ou encore t+h pour transcrire les lettres «ت» et «ح» ou «ت» et «ه» comme dans theb= ثحِبّ ou thabbel= ثهَبِّل
- Le langage Arabish voyellise parfois les mots et parfois non ((korh/ koreh ;choghl/ choghel)
- L'utilisation du déterminant "ال" en langage Arabish:
    - Les lettres lunairesءبجحخعغفقكمهوي sont transcrites par: il, el, al, l, la (la3zé)
    - Les lettres solairesتثدذرزسشصضطظلن sont transcrites par: e (ettounsi), la (lasfer)
- Les déterminants «el», «al».etc. sont parfois liés aux noms et parfois non (el jahl, eljahl).

Au niveau des verbes, nous avons noté que**:**

- Certains verbes ne sont pas réflexifs (زلق)
- La forme réflexive devient imminente dans le cas du passif (ثضَرْبتْ)
- Règles d'accord de la négation selon le paterne syntaxique (شْ)
- La conjonction de subordination «wi» est compatible avec les proclitiques
- Redoublement de consonnes dans les verbes à racine augmentée

Étiquetage

Etiquetage lexical

Emoticônes et smileys: nous avons remplacé les émoticônes et smileys les plus fréquentes par leurs significations (exemple : ☺ par « Souriant »).

Mots ambigus : nous avons mis les mots ou les caractères ambigus entre crochets en les annotant.
Insultes : nous avons gardé les insultes et les gros mots en les étiquetant entre crochets [INSULTE], [GROS MOT]. Nous avons noté que les corpus Arabish contiennent énormément de mots de ce type.

Etiquetage morphosyntaxique

Nous avons attribué des étiquettes à chaque mot du corpus. Ces étiquettes concernent la morphosyntaxe et ont des libellés tels que : Nom, Verbe, Adverbe, Adjectif, Préposition, etc. Les aspects structurels ont été repris dans des formats de maquettage web de type XML pour structurer les messages. Dans les pages des réseaux sociaux, nous avions exploité des tags tels que : <lien>, <zone commentaire>, <message> afin de reproduire la structure de la page dans un format standardisé mettant en valeur le contenu des messages.

D'autre part, nous avons attribué des catégories pour mieux désambiguïser ce langage:

> Noun (noms), Verb (verbes), VNoun (nom verbal), PCL (proclitique), Part (particule), Adv (adverbe), Adj (adjectif), Prn (pronom), PersPrn (pronom personnel), NomGrp (groupe nominal), Interj (interjection), entN (entité nommée), nbr (nombre), smiley

Une observation profonde de notre corpus nous a permis d'identifier les informations importantes à garder pour l'analyse morphologique. Ces champs qui décrivent les mots sont:

- Le mot ou la chaîne de caractère tel que ça existe dans le corpus
- Le lemme du mot dans sa forme translittérée.

Un des problèmes auxquels nous avons fait face c'est que l'orthographe de notre corpus est irrégulière. Nous avons donc mené les différentes formes d'écriture d'un mot vers une seule.

- La forme harmonisée du mot en Arabish, pour formaliser conventionnellement notre corpus (ex: 9olli, golli→ 9olli.. khoukha, 5ou5a, 5o5a →5ou5a)

La forme harmonisée que nous avons choisie revient aux formes les plus utilisées. Toutes les manipulations ont pris en compte ce format. Ainsi, la correction orthographique des mots s'est faite en amont.

- La traduction française du mot
- Le mot écrit en alphabet arabe

Étiquetage des parties du discours

L'étiquetage de parties de discours est souvent utilisé en arabe pour désambiguïser les mots. Dans le cas de l'Arabish, l'usage est souvent de séparer le mot de ses clitiques. On trouve souvent la particule «wa» isolée. Nous avons donc adapté l'outil de Part Of Speech (POS) pour le regroupement du mot avec ses clitiques et rajouté les informations de détermination, de liaison, etc.

1ère analyse :

La première couche de traitement du POS serait le regroupement des constituants immédiats du mot c'est à dire les proclitiques (w, el ..) et les enclitiques (V. HASSOUN.M et al, 2008). Ces derniers sont généralement liés au mot mais peuvent être dans certains cas séparés, nous les avons donc regroupé (noun + hom).

Pour certains enclitiques, comme le « ha », le traitement est légèrement complexe car même si dans la plupart des cas il joue le rôle d'enclitique, il prend la fonction de pronom démonstratif dans le dialecte tunisien (ha denia).

Dans le cas des verbes exprimant le futur, nous avons regroupé les particules «bech» et «ataw» avec le verbe qui les suit: bech+ verbe = futur; ataw+ verbe = futur
La négation des verbes en Arabish tunisien est rajoutée à la fin des mots grâce à l'enclitique «ch» comme le cas de «ma3maltech»

2ème Analyse:

Dans la deuxième couche de traitement POS nous avons regroupé des mots avec des mots outils auxquels ils sont fortement liés comme le «ya» et le nom pour les interjections.
Nous avons construit des groupes prépositionnels on reliant les prépositions de l'Arabish avec les noms comme par exemple lier «3la»,«mel», «fi», «men», «bel» avec les noms qui les suivent pour former des groupes prépositionnels.

- L'article de détermination «al», noté «el» en Arabish tunisien, est toujours collé à la préposition «mel», «bel», «fel»
- Il existe plusieurs phénomènes de fusion spontanée, comme la suppression de « l », du «el» quand le mot commence par «e», comme par exemple «fi eddar» au lieu de «fel

eddar» cela s'apparente à la fusion des «al» dans le mot arabe الليل. Ce qui n'est pas le cas des mots tels que «makteb» qui respecte la notation «préposition + el + espace + nom» «felmakteb»

La phase d'harmonisation qui a consisté à éliminer les répétitions des lettres nous a permis de réduire le nombre de règles de POS.

7. Conclusion

Dans cet article, nous avons exploré, à travers une étude des nouvelles formes de communication écrite, les réseaux sociaux généralistes pour l'analyse des sentiments. Les applications modernes du TAL ne peuvent ignorer l'importance de l'utilisation des dialectes sur internet et l'apparition de l'Arabish comme nouvelle langue d'expression de sentiments sur le web.Pour mener cette première exploration nous avons mis les bases de la constitution d'un corpus pour l'analyse des sentiments. Après l'identification de la source d'information, nous avons rapatrié les données publiques libres de toute propriété intellectuelle ou personnelle. Nous avons procédé par la suite à un dépouillement fin pour identifier la structure des pages web.

L'identification de l'Arabish a été faite grâce à la méthode des n-grammes pour séparer les messages en quatre lots: messages en français, messages en anglais, messages en arabe et messages en Arabish. Les mutations du marketing et de la communication ainsi que les nouvelles approches du consommateur propulsent l'analyse de sentiments dans les applications leaders du TAL de nos jours. A l'issue de notre étiquetage, nos données sont prêtes à une étude d'opinions. Nous préconisons pour des futures voies de recherche d'enrichir notre démarche et de la prolonger par une étude des groupes d'influence et des leaders d'opinion.

Référence Bibliographique:

1) ABOELEZZ, M. 2009. « Latinised Arabic and Connections to Bilingual Ability ». *In* S. Disney, B. Forchtner, W. Ibrahim, N. Millar (eds.) The Lancaster University *Postgraduate Conference in Linguistics & Language Teaching* (LAEL PG 2008*), volume 3*. Department of Linguistics and English Language, Lancaster University.
2) BAHRAINWALA, *L 2011. You say Hello*, I *say Mar7aba*: *Exploring* the *digi-speak* that *powered* the *Arab revolution.*,thèse de doctorat, Michigan State University
3) BELHADJ. S., 2013. « Analyse des sentiments et nouvelles formes de communication écrite : cas de l'Arabish », Mémoire de Master en Systèmes d'informations multilingues, *Ingénierie linguistique et traduction*, sous la direction de M. Mohamed Hassoun, Université Lumière Lyon 2, 67 pages.
4) BOULLIER. D et LOHARD.A., 2012. *Opinion Mining et Sentiment Analysis, méthode et outils.* Éditeur: OpenEdition Press. Collection: Sciences Po │ médialab. ISBN: 9782821812260
5) ELMAHDY. M, Gruhn R, and Minker W, 2012. « Novel Techniques for Dialectal Arabic Speech Recognition », Springer (Boston) pp 71-80.
6) FAIRON, C. 2010. « *Constitution et étude de corpus spécialisés sur le Web ». In: Le discours et la langue*, Vol. 2, no.1, p. nchttp://hdl.handle.net/2078.1/120781
7) FAIRON, C, Klein J-R et Paumier S. 2006b. *Le langage SMS, révélateur d'1 compétence ?*, Cahiers du CENTAL, vol. 1.
8) HASSOUN .M, Dichy, J et Abbès, R, 2008 :« Traitement de l'arabe écrit et web arabe, l'apport de l'équipe lyonnaise SILAT *», contribution à l'Atelier sur les contenus arabes sur la Toile (Arabic Content on the Internet)* organisé par la Société syrienne d'informatique (Syrian Computer Society), Damas, les 13-14Avril 2008.
9) JUNCO.R. et CHICKERING, A. W. 2010: « Civil discourse in the age of social media ». *About Campus*, 15(4), 12-18

10) JUNCO. R., Heiberger, G., etLoken, E. 2011. « The effect of Twitter on college student engagement and grades ».*Journal of Computer Assisted Learning*, 27(2), 119-132.
11) KU Y., Chu T., Tseng C., 2013.«Gratifications for using CMC technologies: A comparison among SNS, IM, and e-mail ». *Computers in Human Behavior*, Volume 29, Issue 1, Pages 226–234
12) RODRIGUES. P, 2012. *Processing highly variant language using incremental model selection.*Thèse de doctorat, Indiana University
13) YAGHAN. M. 2008. « Arabizi: A Contemporary Style of Arabic Slang ». *Design Issues*, 24 (2): 39-52. USA: MIT Press