Phonétisation automatique du dialecte tunisien

Abir Masmoudi[1,2], Mariem Ellouze Khemakhem[1], Yannick Estève[2] , Fethi Bougares[2], Sawssan Dabbar[1], Lamia Hadrich Belguith[1]

(1) ANLP group, MIRACL Lab., Université de Sfax, Tunisie
(2) LIUM, Université du Maine, France
masmoudiabir@gmail.com, Mariem.Ellouze@planet.tn, yannick.esteve@lium.univ-lemans.fr, fethi.bougares@lium.univ-lemans.fr, sawssendabbar@gmail.com, l.belguith@fsegs.rnu.tn

RÉSUMÉ

Un dictionnaire phonétique est un composant primordial d'un système de reconnaissance de la parole ou d'un système de synthèse de la parole. Notre travail cible la génération automatique d'un dictionnaire de prononciation du dialecte tunisien, en particulier dans le domaine du transport ferroviaire. Pour ce faire, nous avons créé deux outils de phonétisation de mots voyellés ou non voyellés en dialecte tunisien. La méthode proposée pour générer automatiquement des dictionnaires phonétiques est à base de règles et est présenté dans cet article. Cet article présente également les différentes étapes de création de notre propre corpus d'étude. Ensuite, il détaille les exceptions phonétiques et phonologiques du dialecte tunisien et donne des exemples de règles utilisées pour la construction des dictionnaires phonétiques.

ABSTRACT

A phonetic dictionary is an essential component of speech recognition system or speech synthesis system. Our work targets the generation of an automatic dictionary pronunciation of Tunisian Dialect, in particular in the field of railway transport. To do this, we created two tools of phonetic voweled and non voweled words in Tunisian Dialect. The proposed method to automatically generate phonetic dictionaries is based on rules and is presented in this article. This paper outlines the steps to create our own study corpus. Then it details the phonetic and phonological exceptions of Tunisian dialect and illustrates some rules used in the phonetic dictionary construction.

MOTS-CLÉS : Dialecte Tunisien, Phonétique, Phonologique, Dictionnaire phonétique, mot voyellé et non voyellé.
KEYWORDS: Tunisian Dialect, Phonetic, Phonologic, Phonetic Dictionary, voweled and non voweled words.

1 Introduction

Ce travail s'inscrit dans le cadre de la réalisation d'un système de reconnaissance automatique du dialecte tunisien utilisé dans le domaine du transport ferroviaire. Dans cet article, nous nous intéressons à l'un des composants clés des systèmes de reconnaissance de la parole, à savoir le dictionnaire phonétique. Le manque de ressources parlées et écrites est l'un des principaux problèmes rencontrés pour le traitement du dialecte tunisien. Ainsi, nous proposons dans cet article les efforts que nous avons déployés afin de créer nos propres ressources, ainsi que notre méthode de génération automatique d'un dictionnaire phonétique pour les mots voyellés et les mots non voyellés en dialecte tunisien.

Cet article comprend cinq sections. D'abord, nous présentons un aperçu de l'état de l'art. Puis, nous exposons quelques spécificités du dialecte tunisien. Ensuite, nous présentons les étapes de création de notre corpus d'étude. Nous détaillons aussi les exceptions phonologiques et phonétiques du dialecte tunisien. Dans la quatrième section de l'article, nous exposons le principe et les étapes de génération automatique des dictionnaires phonétiques. Enfin, nous donnons les résultats d'évaluation de nos outils de phonétisation.

2 Le Dialecte Tunisien

Le dialecte tunisien, généralement connu sous le nom de " Darija " ou " Tounsi ", est un sous-ensemble de dialectes arabes du groupe de l'Ouest. Il est utilisé dans la communication orale de la vie quotidienne des Tunisiens. On peut distinguer trois sous-dialectes suivant les appartenances sociales: les citadins, paysans/agriculteurs et bédouins. Ces sous-dialectes diffèrent considérablement les uns des autres. Ces différences affectent tous les niveaux de la langue, c'est-à-dire la prononciation, la phonologie, le vocabulaire, la morphologie et même la syntaxe. Malgré ces différences, le dialecte tunisien reste toujours compris par l'ensemble des Tunisiens. Parmi les différences phonologiques, nous pouvons noter que les Bédouins prononcent[1] la lettre " ق " /q/ comme " ڤ " /g/, mais la population urbaine prononce " ق " /q/. Au niveau du vocabulaire, les citadins utilisent le mot " يوقف " /i:u:qif/, mais les agriculteurs utilisent le mot "يحبس " /i:aħbis/. Ces deux mots ont le même sens : "arrêter".

Parmi les spécificités remarquables du dialecte tunisien, on note la présence de mots empruntés au français, au berbère, à l'italien, au turc et à l'espagnol. Aussi, on note la présence de mots dont l'origine est l'arabe standard moderne(MSA). La présence de ces mots est le résultat de nombreux facteurs et événements historiques qui ont eu lieu tout au long des siècles, tels que : les invasions islamiques, la colonisation française, les migrations, les échanges commerciaux, etc.

Aujourd'hui, le dialecte tunisien est de plus en plus souvent utilisé lors d'interviews, de conversations téléphoniques, dans les services publics, etc. et il est très présent aujourd'hui dans les blogs, les forums et les commentaires d'utilisateur/lecteur sur internet.

Il est donc important de considérer le dialecte tunisien dans le contexte des nouvelles technologies telles que les systèmes de reconnaissance de la parole, les systèmes de dialogue Humain-Machine, etc. Cependant, le dialecte souffre de l'absence d'outils et de ressources

[1] Pour représenter les phonèmes, nous utilisons les symboles de l'alphabet phonétique international (http://fr.wiktionary.org/wiki/Annexe:Prononciation/arabe)

linguistiques. Les outils pour traiter le MSA sont très difficilement utilisables, compte tenu de la grande différence entre le MSA et le dialecte tunisien. Ainsi, il convient de noter que notre travail étudie le tunisien alors que nous manquons d'outils linguistiques adaptés à ce dialecte.

3 Etat de l'art

Dans un contexte de reconnaissance automatique de la parole, pour faire le lien entre le niveau lexical et le niveau acoustique, il est nécessaire d'associer à chaque mot du vocabulaire une ou plusieurs séquences d'unités acoustiques de base. Dans la littérature plusieurs approches sont utilisées pour obtenir ces séquences de phonèmes. Nous pouvons distinguer l'approche guidée par les données qui utilise généralement des données d'apprentissage et l'approche à base de règles qui nécessite une expertise linguistique (Béchet, 2001).Dans ce qui suit, nous allons expliquer davantage chacune de ces approches.

3.1 Approche guidée par les données

L'idée générale de cette approche est d'utiliser un dictionnaire de mots phonétisés manuellement. Des approches probabilistes basées sur des modèles joints (Bisani et Ney, 2008) ou encore sur des approches s'appuyant sur des outils de traduction automatique (Laurent et al. 2009), permettent de capturer par apprentissage automatique des liens entre les graphèmes utilisés pour écrire le mots et les phonèmes, unités de sons utilisées pour représenter la prononciation d'un mot.

Ces techniques statistiques nécessitent de posséder des données d'apprentissage. Lorsque ces données sont en nombre suffisant, elles permettent d'obtenir des résultats intéressants, tout en minimisant l'expertise humaine. En l'absence de données, ces approches ne sont pas applicables.

3.2 Approche à base de règles

La construction d'un système de phonétisation automatique à base de règles nécessitent une bonne connaissance de la langue et de ses règles de phonétisation qui, par ailleurs, ne doivent pas contenir trop d'exceptions : la phonétisation automatique utilisant des connaissances linguistiques ne nécessite pas de techniques innovantes, mais plutôt une expertise importante sur la langue.

L'avantage de cette approche est qu'elle permet de mieux contrôler la qualité de la construction des dictionnaires de prononciation : en cas d'erreur il est possible d'ajouter une nouvelle règle. De plus, le coût de développement peut être comparé au coût nécessaire à la construction manuelle du corpus d'apprentissage utilisé par l'approche statistique.

4 Corpus d'étude

Outre le manque d'outils pour traiter le dialecte tunisien, le manque de ressources linguistiques est également patent. Pour construire un système de reconnaissance automatique de la parole, nous avons besoin, a minima, d'un corpus constitué

d'enregistrements audio et de leur transcription. Nous avons été dans l'obligation de créer notre propre corpus d'étude.

La création de corpus repose sur trois phases : 1) la production d'enregistrements audio, 2) la transcription manuelle de ces enregistrements et 3) la normalisation de ces transcriptions (Masmoudi et al.,2014).

Le Tableau 1présente quelques statistiques de notre corpus.

Nombre d'heures	Nombre de dialogues	Nombre d'énoncés	Nombre de mots
20 h	4662	18657	71684

TABLE 1 - Statistiques de notre corpus d'étude

Lors de l'analyse de notre corpus, nous avons remarqué que les mots étrangers représentent 20% du vocabulaire. Voici quelques exemples d'emprunts étrangers en Dialecte Tunisien : le mot "تكاي " "Ticket " /tika:i:/ d'origine française, ou le mot "ترينو " "Trinou "/tri:nu:/ d'origine italienne. On peut noter aussi la présence des mots d'origine anglaise comme le mot "أوكاي " "Ok" /ʔʊ:kɑ:i:/.

Aussi, nous avons remarqué que certains mots étrangers s'utilisent avec des modifications par rapport à leurs origines. Ainsi, un mot étranger subit un ajout d'enclitiques ou de proclitiques de la langue arabe. Voici quelques exemples de mots :

- Le mot "سيبرميها " /sIprimIha:/ " la supprimer " est un mot emprunté du français, il subit un ajout de l'enclitique arabe "ها" /ha:/ "la" qui est attaché au mot.
- Le mot "ريزرڤيلي " /rIzirvIlI/ "réserver moi" est un mot emprunté du français, il subit un ajout de l'enclitique arabe "لي " /lI/ "moi" qui est attaché au mot.

5 Phonologie et phonétique du dialecte tunisien

Dans cette section, nous présentons d'abord le système d'écriture du dialecte tunisien et ensuite les exceptions phonétiques et phonologiques sa phonologie.

5.1 Le système d'écriture

Comme le MSA, le dialecte tunisien s'écrit de droite à gauche. L'alphabet se compose de trente et une lettres. Vingt-cinq d'entre elles sont des consonnes provenant de MSA, trois consonnes sont le résultat de la présence de mots étrangers : 'پ' "v", 'ڤ' "g" et 'ﭖ' "p" et trois lettres représentent les voyelles longues.

Chaque lettre possède au maximum quatre formes différentes, selon qu'elle apparaît au début, au milieu ou à la fin d'un mot, ou de manière isolée.

Les lettres sont la plupart du temps reliées les unes aux autres graphiquement et il n'y a pas de capitalisation. Une caractéristique distinctive du système d'écriture de l'arabe et ses dialectes est la présence de voyelles courtes appelées signes diacritiques. Ces signes diacritiques ne sont pas représentés par les lettres de l'alphabet, mais marqués par des traits courts placés au-dessus ou au-dessous de la consonne. On peut distinguer neuf signes

diacritiques: (a) trois signes diacritiques sont des voyelles courtes ; (b) trois signes diacritiques "de nunation" représentant une combinaison d'une voyelle courte et le marqueur / n / (ce type de signe diacritique est rarement utilisé dans le Tunisien), (c) un signe diacritique représente le doublement d'une consonne (appelé Shadda); (d) un signe diacritique pour marquer l'absence de voyelles courtes (appelé Sukun).

5.2 Quelques exceptions phonétiques et phonologiques

À partir de notre corpus d'étude, nous avons pu mettre en évidence qu'il existe plusieurs exceptions phonétiques et phonologiques dans le dialecte tunisien. Nous citons ci-dessous quelques-uns de ces traits phonétiques (Masmoudi et al.,2014) :

• Les voyelles courtes sont souvent omises. Par exemple, le verbe «écrire» est prononcé dans le MSA «كتب» /kataba/ mais en dialecte tunisien, nous disons /ktib/.

• L'intervention de la consonne «ش»/ʃ/ dans les régions rurales et dans les zones urbaines. Il semble en particulier dans la voix interrogative du dialecte.

• Il y a beaucoup de consonnes dans le dialecte tunisien qui peuvent être prononcées de différentes manières. Ci-dessous, nous présentons certaines d'entre elles :

- La consonne "ش" /ʃ/ peut être prononcée comme "ش" /ʃ/ ou "س" /s/.
- La consonne "ث" /θ/ peut être prononcée de deux manières : "ث" /θ/ ou "ف" /f/.
- La consonne "غ" /ɣ/ peut être prononcée de deux manières: "غ" /ɣ/ ou "خ" /x/.
- La consonne "ط" /tˤ/ peut être prononcée de deux manières : "ط" /tˤ/ ou "ت" /t/
- La consonne "أ" /ʔ/ au sein d'un mot peut être prononcé "أ" /ʔ/ ou "ه" /h/ .

• Le "Hamza" "أ " /ʔ/, accepte plusieurs variations de prononciations selon la position de cette lettre.

• "Ta-Marbouta" à la fin du mot est habituellement silencieux, mais il y a quelques exceptions.

• Il y a des lettres qui ne sont pas prises en compte. Le "alif" dans le mot "خرجوا" /xardʒʊ:/ "ils ont sorti" ne correspond pas à un son (silence).

• Les voyelles longues deviennent des voyelles courtes. Par exemple, "فِي التران"/fItran/ "dans le train" "فِ التران" / fi tran/.

• Nous avons remarqué l'élimination d'une consonne dans certains mots. Par exemple, " مانعرفش " /ma:naʕrafʃ/ "Je ne sais pas" peut être prononcé "مانعرش "/ma:naʕraʃ/. Dans cet exemple, la consonne "ف " /f/ est éliminée.

• Les nombres ont une caractéristique spécifique : les nombres entre "trois" et "neuf", acceptent une double prononciation dont l'une subit une élimination de quelques consonnes et une modification de la voyellation; les nombres à partir de "onze" acceptent aussi deux prononciations, dont l'une nécessite un ajout de la phonème "N" à la fin du nombre.

• Nous avons noté l'ajout de la phonème /E IH/ pour appuyer la prononciation de la première consonne muette d'un mot.

6 Construction automatique d'un dictionnaire phonétique

Nous avons créé deux outils de phonétisation : un pour les mots voyellés, et l'autre pour les mots non voyellés. Dans les sections suivantes, nous expliquons davantage le principe général de la phonétisation automatique puis nous détaillons les spécificités des deux outils de phonétisation.

6.1 Principe général de la phonétisation

Pour générer automatiquement un dictionnaire de prononciation, nous avons illustré, à partir de nos corpus d'étude, un ensemble de règles phonétiques et un lexique contenant des exceptions. Le dialecte tunisien se caractérise par la présence de nombreuses exceptions phonétiques. En effet, on peut trouver un mot qui peut être prononcé de deux ou plusieurs façons.

Le processus de phonétisation pour les mots voyellés ou non voyellés s'effectue en deux phases : la consultation de la base du lexique d'exceptions et l'application des règles de phonétisation.

6.1.1 Lexique d'exceptions

Il y a des mots qui ne peuvent pas suivre notre ensemble de règles phonétiques, il est donc nécessaire de définir un lexique des exceptions. Ce lexique est consulté avant que les règles soient appliquées. Si le mot est parmi les exceptions, il est encodé directement sous forme phonétique. Sinon, nous devons appliquer les règles pour générer sa forme phonétique. Dans notre lexique, nous avons plus de 30 exceptions. Notre base de lexique des exceptions a été validée par trois experts (de langue maternelle). Voici quelques exemples d'exceptions :

• Le mot "نصف" /nisˤf/ peut être prononcé de trois manières différentes: "نصف" /nisˤf/ , "نص" /nisˤ/ ou "نفص" /nifsˤ/.

• Le mot "هذا " /haða:/est prononcé comme ça "هاذا "/ha:ða:/.

6.1.2 Règles phonétiques

Nous avons développé un ensemble de règles phonétiques qui doivent être fournies pour chaque lettre. Chaque règle essaie de faire correspondre certaines conditions relatives au contexte de la lettre et de fournir un phonème ou séquences de phonèmes et parfois le silence . Nos règles ont également été validées par trois experts.

Chaque règle se lit de gauche à droite et suit ce format détaillé dans (Masmoudi et al.,2014):

{ Condition droite} + {Graphème} + { Condition gauche} ==> Phonétisation

Le nombre total de règles est d'environ quatre-vingts. Pour faciliter l'utilisation de ces règles et en se basant sur les exceptions phonétiques du dialecte tunisien, nous avons décidé de diviser les règles en cinq bases : une base pour les nombres, une base pour les mots qui commencent par la lettre "أ " /E/, une base pour les mots étrangers qui contiennent la lettre "R", une base pour les mots qui commencent par "Alif et Lam" et une base pour le reste des règles de phonétisation.

6.2 Phonétisation de mots voyellés

Lors de la transcription de notre corpus, nous avons choisi de voyeller les mots selon la prononciation du locuteur. La présence de ces voyelles permet une diminution substantielle du degré de l'ambiguïté phonétique.

Les mots voyellés en dialecte tunisien se terminent en général par un Sukun (consonne muette) ou bien par une Voyelle longue.

Pour générer automatiquement la forme phonétique d'un mot, l'application de règles phonétiques est faite dans le sens de lecture du mot, c'est à dire on commence par la premier lettre du mot et on respecte l'ordre des lettres.

Ci-dessous un exemple de phonétisation d'un mot voyellé :

Le mot «حْدَاشْ » /ħda: ʃ/ "Onze" appartient à la base des nombres, alors ce mot accepte une double prononciation dont l'une se termine par le phonème « N ».

Les règles utilisées sont les suivantes :

R1 : {C= ح /ħ/} + { Sukun } => {Phonème = HH }

Lorsque une consonne est suivie par "Sukun", alors on obtient toujours le phonème de la consonne.

R2 : {!= n'est pas pris en considération} + {C = د /d/ } + {VL = Fatha + alif} =>

{Phonème = D AE:}

Lorsque une consonne est suivie par une voyelle longue "Fatha et alif", alors on obtient les deux phonèmes : dont l'un correspondant du consonne "D" et l'autre du voyelle longue "AE:".

R3 : {!} + {ش/ʃ/ }+ { Sukun } => {Phonème = SH}

Lorsque une consonne est suivie par "Sukun", alors on obtient toujours le phonème du consonne.

La phonétique du mot est la suivante : «حْدَاشْ» /ħda: ʃ/ "Onze" 1. HH D AE: SH
2. HH D AE: SH N

6.3 Phonétisation de mots non voyellés

En partant de trois principes : (a) un mot en dialecte tunisien se termine soit par une consonne muette (avec Sukun) ou par une voyelle longue, (b) chaque voyelle longue est toujours suivie et précédée par une consonne muette et (c) on ne peut pas avoir deux consonnes successives qui portent une voyelle (longue ou courte) à l'exception des mots avec "Shadda" (dédoublement de consonnes), les règles de phonétisation pour les mots non voyellés peuvent être divisées en deux groupes : les règles d'appui et les règles secondaires.

Les règles d'appui sont appliquées dans un premier temps. La majorité de ces règles sont à l'origine de la production de phonèmes de voyelles longues dans le mot, ce qui facilite l'application des règles secondaires dans un deuxième temps. Les règles secondaires sont à

l'origine de la production de phonèmes de voyelles courtes ou "Sukun". Contrairement au mots voyellés, l'application de règles phonétiques ne respecte pas un ordre bien défini et donc ne suit pas le sens de lecture du mot.

En effet, la phonétisation commence par localiser dans le mot au moins l'une de ces quatre consonnes ("ا " /a:/ "Alif ", "ى " /a:/ "Alif Maksoura", "ي " /i:/ "Ya" et "وا " /u:/ "Waw et Alif") pour appliquer les règles d'appui. Ces règles permettent de donner la phonétisation de voyelles longues. Ensuite et en se basant sur le principe que "chaque voyelle longue est toujours suivie et précédée par une consonne muette", les règles secondaires sont appliquées afin d'ajouter les phonèmes des voyelles courtes ou "Sukun".

En absence de l'un de ces quatre consonnes, la phonétisation se repose sur deux principes : le mot se termine par une consonne muette (avec Sukun) et en deuxième lieu, on ne peut pas avoir deux consonnes successives qui portent une voyelle.

7 Evaluation

Le système de phonétisation a été testé sur un corpus de test constitué de 400 mots collectés à partir des blogs tunisiens de domaine variés (politique, sportifs, culture …). Le corpus a été dans un premier temps, normalisé selon la convention CODA (Zribi et al. 2014). Dans l'objectif d'évaluer les deux systèmes de phonétisation (avec et sans voyelles), nous avons créé une version voyellée du corpus de test (voyellation manuelle). Le tableau suivant montre les résultats obtenus pour la phonétisation du corpus de test avec et sans voyelles.

Corpus de test	Taux d'erreur phonème
Voyellé	0 %
Non voyellé	12.9 %

TABLE 2 - Résultat de l'évaluation

Comme présenté dans la Table 2, le système de phonétisation de dialecte tunisien est 100% performant pour les textes voyellés. En ce qui concerne le texte non voyellé la phonétisation est une tâche plus complexe où notre système peut proposer plusieurs variantes de phonétisation. Le taux d'erreur phonème de 12,9% présenté dans la table 2 est celui de la première proposition du système.

8 Conclusion

Ce papier décrit nos efforts pour créer deux outils de phonétisation automatique des mots voyellés et non voyellés en dialecte tunisien dans le domaine de transport ferroviaire. Les outils génèrent un dictionnaire de prononciation en se basant sur des règles phonétiques du tunisien. Chaque règle essaie de faire correspondre certaines conditions relatives au contexte de la lettre et de fournir un remplaçant. Le nombre total de règles est environ 80.

Pour faire face à l'absence des ressources linguistiques en dialecte tunisien, nous étions amenés à créer notre propre corpus d'étude. La création de corpus repose sur trois phases à savoir, la production d'enregistrements audio, la transcription manuelle de ces enregistrements et la normalisation de ces transcriptions.

Références

Afify, M., Nguyen, L., Xiang, B., Abdou, S. et Makhoul.J., (2005), Arabic broadcast news transcription using a one million word vocalized vocabulary. *In Proceedings of Interspeech 2005*, Portugal.

Algamdi, M. (2000), Arabic Phonetics. Riyadh, Saudi Arabia.

Algamdi, M., (2003), KACST Arabic Phonetics Database. Fifteenth International Congress of Phonetics Science, (KACST'2003), Barcelona.

Algamdi, M., Elshafei, M., et Almuhtasib, H., (2002),Speech Units for Arabic Text-to-speech. *Fourth Workshop on Computer and Information Sciences*.

Antoine Laurent, Paul Deléglise, Sylvain Meignier: Grapheme to phoneme conversion using an SMT system. Interspeech 2009: 708-711

Baccouche , T., (1994), L'emprunt en arabe moderne, Beit Elhikmaet IBLV, Tunis.

Biadsy, F., Hirschberg, J. et Habash. N., (2009), Spoken Arabic Dialect Identification Using Phonotactic Modeling. *In Proceedings of EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece.

Bisani M. and Ney. H. (2008) "Joint-Sequence Models for Grapheme-to-Phoneme Conversion". Speech Communication, Volume 50, Issue 5, , Pages 434-451

Diehl, F., Gales, M. J. F., Tomalin, M., et Woodland, P. C., (2008), Phonetic pronunciations for Arabic speech-to-text systems. IEEE International Conference on Acoustics, Speech and Signal Processing.

El-Imam.Y., (2004).Phonetization of Arabic: rules and algorithms. In Computer Speech and Language.

Frédéric Béchet, LIA_PHON (2001) ,: un système complet de phonétisation de textes, Traitement Automatique des Langues - TAL - volume 42 numéro 1 - pp 47-67.

Gales, M. J. F., Diehl, F., Raut, C. K., Tomalin, M., Woodland, P. C., et Yu, K., (2007), Development of a phonetic system for large vocabulary Arabic speech recognition. IEEE Workshop on Automatic Speech Recognition & Understanding.

Habash, N., Diab, M., Rambow, O., (2012), Conventional Orthography for Dialectal Arabic. In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Masmoudi, A., Ellouze Khmekhem, M., Estève, Y., Hadrich Belguith, L., et Habash, N.,(2014), A corpus and phonetic dictionary for Tunisian Arabic speech recognition, In *19th edition of the Language Resources and Evaluation Conference (LREC'2014)*, Reykjavik, Iceland.

Masmoudi, A., Estève, Y., Ellouze Khmekhem, M, Estève, Y., Bougares, F et Hadrich Belguith, L., (2014), Phonetic tool for the Tunisian Arabic, *The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14),* Saint Petersburg, Russia.

Zribi, I., Boujelben, R., Masmoudi, A., Ellouze Khmekhem M., Hadrich Belguith, L., et Habash, N.,(2014), A Conventional Orthography for Tunisian Arabic, *In 19th edition of the Language Resources and Evaluation Conference (LREC'2014*), Reykjavik, Iceland.