TArC : Incrementally and Semi-Automatically Collecting a Tunisian arabish Corpus

Elisa Gugliotta, Marco Dinarelli

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

TArC
Un corpus d'*arabish* tunisien

Elisa Gugliotta[1,2] Marco Dinarelli[1]

(1) LIG - Bâtiment IMAG, 700 Avenue Centrale 38401 Saint-Martin-d'Hères, Grenoble, France

(2) Université Sapienza di Roma, 82 Viale dello Scalo S. Lorenzo 00159, Roma, Italia

elisa.gugliotta@uniroma1.it, marco.dinarelli@univ-grenoble-alpes.fr

Résumé

Cet article décrit la procédure de constitution du premier corpus d'*arabish* tunisien (TArC) annoté avec des informations morpho-syntaxiques. L'*arabish* est la transcription spontanée des dialectes arabes en caractères latins et *arythmographies*, c'est à dire avec des chiffres utilisées comme lettres. Ce système d'encodage a été développé par les utilisateurs arabes des réseaux sociaux afin de faciliter l'écriture dans les communications informelles. L'*arabish* diffère pour chaque dialecte arabe et il est sous-doté en termes de ressources, de la même façon que la plupart des dialectes arabes. Dans les dernières années, l'attention des travaux de recherche en TAL sur les dialectes arabes est augmentée de façon remarquable. En prenant ceci en compte, *TArC* serait un support utile pour plusieurs types d'analyses, computationnelles ainsi que linguistiques, et pour l'apprentissage d'outils informatiques. Nous décrivons le travail fait pour mettre en place une procédure d'acquisition semi-automatique du corpus TArC, ainsi que certaines analyses faites sur les données collectées. Afin de montrer les difficultés rencontrées pendant la procédure de constitution du corpus, nous présentons également les caractéristiques principales du dialecte tunisien, ainsi que sa transcription en *arabish*.

Abstract

TArC : Incrementally and Semi-Automatically Collecting a Tunisian arabish Corpus

This article describes the collection process of the first morpho-syntactically annotated Tunisian arabish Corpus (TArC). Arabish is a spontaneous coding of Arabic Dialects (AD) in Latin characters and *arithmographs* (numbers used as letters). This code-system was developed by Arabic-speaking users of social media in order to facilitate the communication on digital devices. Arabish differs for each Arabic dialect and each arabish code-system is under-resourced. In the last few years, the attention of NLP on AD has considerably increased. TArC will be thus a useful support for different types of analyses, as well as for NLP tools training. In this article we will describe preliminary work on the TArC semi-automatic construction process and some of the first analyses on the corpus. In order to provide a complete overview of the challenges faced during the building process, we will present the main Tunisian dialect characteristics and its encoding in Tunisian arabish.

Mots-clés : Corpus d'arabish tunisien, Dialecte arabe, Arabizi.

Keywords : Tunisian arabish Corpus, Arabic Dialect, Arabizi.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 232–240.
HAL : hal-02784772.

1 Introduction

L'*arabish*[1] est la romanisation des dialectes arabes (DAs) utilisée pour des messages écrits informels, particulièrement dans les réseaux sociaux. Ce système d'écriture fournit un terrain très intéressant pour des recherches linguistiques, computationnelles, mais aussi socio-linguistiques, principalement grâce au fait qu'il s'agit d'une représentation écrite spontanée des DAs, et il est en constante expansion dans le web.[2] Malgré ce potentiel, peu de recherches de ce type on été dédiées à l'arabish tunisien (TA). Dans cet article nous décrivons notre travail pour le développement d'une ressource en TA riche, avec plusieurs niveaux d'annotation. Celle-ci inclut un corpus en TA annoté, mais aussi des outils informatiques pour à la fois analyser les données et étendre éventuellement le corpus avec des nouvelles données. En premier lieu, la ressource permet de donner une vue générale du TA. Au même temps elle constitue une représentation significative de l'évolution du dialecte tunisien (TUN) dans le cadre de la communication numérique (CMC, de l'anglais *Computer-Mediated Communication*) à travers les dix dernières années. En effet, les textes collectés datent de 2010 jusqu'à l'année courante.[3] Pour cette raison, notre corpus se prête à des études phonologiques, morphologiques, syntaxiques et aussi sémantiques, et à la fois dans un contexte de linguistique et de TAL. Nous avons donc décidé de construire un corpus qui peut mettre en évidence les caractéristiques structurelles du TA grâce à plusieurs niveaux d'annotation, comprenant parties du discours (POS) et lemmes. De plus, afin de faciliter la correspondance avec d'autres études, ainsi que des outils déjà existants pour le traitement du langage arabe, nous avons transcrit les tokens en arabish en caractères arabes du dialecte tunisien, suivant les lignes directrices du *Conventional Orthography for Dialectal Arabic* (CODA*) (Habash *et al.*, 2018).[4] Enfin, même si la traduction de notre corpus n'est pas l'objectif principal de nos recherches, nous avons décidé de traduire les textes de notre corpus en italien.[5]

Le reste de l'article est organisé comme suit : dans la section 2 nous décrivons des études sur l'arabish tunisien (TA) et les corpus accessibles du dialecte tunisien (TUN) et TA ; la section 3 décrit le TA ; dans la section 4 nous présentons la procédure de construction du corpus *TArC* ; la section 5 montre des expériences préliminaires pour la transcription et l'annotation semi-automatiques des données, adoptés pour une construction plus facile et rapide du corpus ; nous tirons nos conclusions dans la section 6.

2 État de l'art

Le travail décrit dans cet article concerne surtout la transcription semi-automatique du corpus en caractères arabes. En revanche pour la réalisation d'un des niveaux d'annotation du corpus, nous décrivons ici quelques travaux représentatifs de l'état de l'art sur le codage de l'arabish tunisien. Nous décrivons également les corpus tunisiens accessibles librement. Pendant ces dernières années,

[1] Aussi appelé *Arabizi* (de la combinaison des mots arabes 'Arab', ['ʕarab] et 'English' [i ːn'ʒli ːzi ː], Franco-Arabe, *Arabic Chat Alphabet*, ACII-ized Arabic, entre autres. 'arabish' est probablement le résultat de l'union entre ['ʕarab] et 'English'. Ce mot a été choisi pour être lisible par tout le monde. Le terme 'arabizi' ne serait en fait compréhensible que par ceux qui connaissent la langue arabe.

[2] Pour des informations sur le pourcentage d'utilisation, veuillez vous référer à (Tobaili, 2016 ; Younes *et al.*, 2015).

[3] Cette sélection vise à observer le degré d'évolution diachronique d'une convention d'écriture arabish.

[4] Le CODA* est un ensemble unifié de lignes directrices pour la transcription de 28 dialectes arabes.

[5] Nous avons pris en considération une traduction à moindre coût, donc vers notre langue maternelle, laissant une traduction vers l'anglais comme éventuelle option future.

le nombre de travaux sur l'arabish a considérablement augmenté.[6] La plupart des recherches sur l'arabish se focalisent sur des variétés très différentes de l'arabe tunisien, par exemple l'égyptien (Darwish, 2014 ; Al-Badrashiny *et al.*, 2014). À cause d'un manque de place, nous nous focaliserons uniquement sur les travaux sur l'arabish tunisien. (Younes *et al.*, 2018b) est le seul travail qui propose une transcription automatique de l'arabish tunisien vers les caractères arabes et vice versa, basée sur des modèles d'apprentissage profond dits *sequence-to-sequence*. En transcrivant une collection de textes de Facebook, après l'avoir nettoyé des éléments non linguistiques (émojis, émoticônes, etc.) ou qui n'étaient pas dialectales, ils ont obtenu une précision de plus de 95.59%.[7] La même ressource a été utilisée dans (Younes *et al.*, 2020) pour une étude approfondie de la transcription automatique. En utilisant uniquement les mots tunisiens, une transcription en caractères arabes a été faites par des internautes tunisiens. La transcription d'un mot arabish donné est effectuée en utilisant une approche à base de modèles CRF, BLSTM et BLSTM-CRF, ce dernier donnant les meilleurs résultats quantitatifs. Le taux d'erreur global est de 2,7%, alors que le taux d'erreur de contexte est de 0,68%. Le seul travail qui prévoit une transcription de l'arabish tunisien dans la convention CODA est (Masmoudi *et al.*, 2015). Ce travail utilise un système à base de règles qui, pour chaque mot en arabish, génère toutes les transcriptions possibles en arabe, parmi lesquelles le meilleur candidat est ensuite sélectionné manuellement. Le système a un taux d'erreur de 10%. Cependant, à l'exception du recueil de textes de (Younes *et al.*, 2015), aucune de ces ressources n'est accessible au public. Les corpus en dialecte tunisien disponible au public sont cinq (Younes *et al.*, 2018a). Le *PADIC* (Meftouh *et al.*, 2015), composé de 6 400 phrases en six dialectes arabes, traduites en arabe moderne standard (MSA).[8] Deux autres corpus sont le *TuDiCoI* (*Tunisian Dialect Corpus Interlocutor*) (Graja *et al.*, 2010) et le *STAC* (*Spoken Tunisian Arabic Corpus*) (Zribi *et al.*, 2015), qui sont annotés avec des informations morpho-syntaxiques. Le premier est un corpus de dialogues entre les clients et le personnel collectés à la station des trains. Il est composé de 21 682 mots (Graja *et al.*, 2013).[9] Le corpus STAC est constitué de 42 388 mots transcrits de fichiers audio téléchargé du web (émissions télé et radio) (Zribi *et al.*, 2015). Le *TARIC* (Masmoudi *et al.*, 2014a) contient 20 heures de TUN oral, transcrites en caractères arabes et correspondant à 71 684 mots (Masmoudi *et al.*, 2014b). Un dernier corpus est le *TSAC*, composé de 17 000 commentaires de Facebook, annotés manuellement en polarité positive et négative pour la fouille d'opinion (Medhaffar *et al.*, 2017). Ce dernier corpus est le seul qui contient à la fois des textes transcrits en TA et en caractères arabes. À notre connaissance il n'y pas de corpus en TA transcrit aussi en caractères arabes et annoté avec des informations morpho-syntaxiques. Notre travail se propose de combler ce manque en offrant à la communauté scientifique une ressource d'arabish tunisien ouverte au public, c'est-à-dire le *TArC*, le premier corpus en TA annoté avec des informations morpho-syntaxiques (POS et lemmes) et transcrit aussi en caractères arabes en suivant la convention CODA*. Compte tenu également de l'indisponibilité des systèmes de transcription de l'arabish tunisien,[10] il est intéressant de construire un système *ad hoc* en fonction de nos besoins. Nous avons fait le choix, en effet, de ne pas exclure les termes étrangers ou les éléments para-linguistiques (émoticônes et emojis), et d'utiliser les conventions CODA.

[6] Pour une description détaillée veuillez consulter (Guellil *et al.*, 2019).

[7] Le corpus utilisé est disponible sous requête.

[8] Les dialectes arabes dans le corpus PADIC sont : le TUN (Sfax), deux dialectes de l'Algérie, le syrien, le palestinien, et le marocain (Meftouh *et al.*, 2018).

[9] L'annotation en revanche a été faite uniquement pour 7 814 mots.

[10] Pour la distinction entre transcription et translittération veuillez vous référer à (Coulmas, 2003).

3 Caractéristiques de l'*arabish* tunisien

Le dialecte tunisien (TUN) est la langue parlé dans la vie tunisienne de tous les jours, appelé généralement الدَّارِجَة, *ad-dārija*, العَامِّيَّة, *'āmmiyya*, or التُّونْسِي, *ət-tūnsī*. Conformément à la classification diatopique traditionnelle, le TUN appartient à la zone de l'arabe maghrébin, duquel il constitue une des variantes principales avec le libyen, l'algérien, le maroquin, et le hassanya de la Mauritanie (Durand, 2009). L'arabish est une transposition des DAs, qui sont des systèmes essentiellement oraux, dans une forme écrite qui n'est pas réalisé avec caractères arabes et par conséquence il n'est pas sujet aux règles orthographiques de l'arabe standard. Comme résultat, on peut considérer le TA comme un écrit spontané, fidèle à la réalisation orale de TUN, ou en d'autres termes un système quasi-oral.

La notion de *quasi-oralité* décrit des formes d'écriture typiques de la communication numérique (CMC), caractérisée par un ton informel, par la dépendance du contexte, par le manque d'attention sur la façon d'écrire, et ayant spécialement la capacité de créer un sens de collectivité (Hert, 1999). TA et TUN n'ont pas une orthographie standard, avec l'exception du CODA. Cependant, le TA est un système d'écriture utilisé depuis plus que dix ans, et il subit donc une conventionnalisation spontanée à travers son utilisation. Dans le tableau 1 nous montrons un schéma du système d'écriture TA. Il est possible d'observer qu'il n'y a pas une correspondance un à un entre les caractères en TA et ceux en arabes, et souvent le TA présente une ambiguïté dans la possibilité d'écriture.[11] Le problème principal est constitué par le manque d'une représentation propre en TA pour les phonèmes emphatiques : [ðˤ], [tˤ] et [sˤ]. D'un autre côté, puisque le TA n'est pas codifié à travers l'alphabet arabe, il peut

1	[ðˤ]	[a ː]	[ʔ]	[b]	[θ]	[ʒ]	[ħ]	[x]	[d]	[ð]	[s]	[ʃ]	[sˤ]
2	ض	ة	ء	ب	ث	ج	ح	خ	د	ذ	س	ش	ص
3	dh th d	a e h	2	b p	th	j	7 h	5 kh	d	dh	s	ch, (sh)	s
1	[a][a ː]	[tˤ]	[ðˤ]	[ʕ]	[ɣ]	[q]	[k]	[l]	[m]	[n]	[h]	[w][u ː]	[j][i ː]
2	ا ى	ط	ظ	ع	غ	ق	ك	ل	م	ن	ه	و	ي
3	a e é è	6 t	th dh	3 a	4 gh	9 q	k	l	m	n	8, h	ou, w	i, y

Tab. 1: Exemples de correspondance entre TA et TUN. **1** indique la représentation phonétique des graphèmes. **2** représente les caractères arabe, **3** les caractères arabish correspondant.

bien représenter la réalisation phonétique du TUN, comme il est montré dans l'exemple suivant : **1.** L'alphabet arabe est généralement utilisé pour des conversations formelles en arabe moderne standard (MSA), l'arabe des contextes formels, ou pour l'arabe classique (CA), l'arabe du *Saint Qur'ān*. De la même façon que le MSA et le CA, les dialectes arabes également peuvent être écrits avec l'alphabet arabe, mais dans ce cas il est possible d'observer une auto-correction spontanée de l'orateur pour respecter les règles d'écriture du MSA. Par exemple, dans les textes en TUN écrits avec l'alphabet arabe, il est possible de trouver une voyelle muette (ʾalif <ا> épenthétique, ou additionnel) au début des mots qui commencent par la séquence '#CCv', ce qui n'est pas possible en MSA. **2.** En écrivant le TUN avec l'alphabet arabe, l'écriture de mots étrangers dans leur alphabet est très forcée, comme, par exemple, dans l'usage de mots empruntés d'autres langues. **3.** Comme nous le montrons dans le tableau 1, l'alphabet arabe fournit trois voyelles courtes, qui correspondent à leur version longue [a ː], [u ː], [i ː], mais le TUN présente un ensemble plus étendu de voyelles. En effet, l'ensemble des voyelles du TA offre une meilleure possibilité de représenter la phonétique du TUN.

[11] Pour ces raisons, la conversion de TA à TUN ne peut être traitée comme une simple translittération.

3.1 Collecte des données - trois étapes

1. Détection des catégories thématiques. Afin de construire un corpus qui soit représentatif du TUN, il nous semblait utile d'identifier des catégories thématiques larges, qui pourraient représenter les sujets de discussion plus courant dans les CMC. Dans cette perspective, nous avons employé deux instruments avec une organisation thématique similaire : Un dictionnaire arabe avec fréquence des mots (Buckwalter & Parkinson, 2014)[12] et la *Loanword Typology Meaning List* (LTML) (Haspelmath & Tadmor, 2009), une liste de 1 460 sens. 15 catégories ont été identifiées grâce à ces documents. Pour une description détaillée nous renvoyons à (Gugliotta & Dinarelli, 2020).

2. Construction des correspondances entre les catégories et les mots clefs du TA en relation sémantique. Nous avons associé à chaque catégorie un ensemble de mots-clefs en TA, appartenant au vocabulaire de base tunisien. Nous avons trouvé que trois sens pour chaque catégorie sémantique étaient suffisants pour obtenir un nombre significatif de mots-clefs pour chaque catégorie. Pour une analyse plus détaillée de cette procédure nous renvoyons à (Gugliotta & Dinarelli, 2020).

3. Extraction des textes et des méta-données. À travers ces mots-clés, nous avons effectué une recherche des textes sur les réseaux sociaux. Nous avons collecté des textes pour l'équivalent d'environ 40 000 mots, et leur méta-données associées, comme première partie de notre corpus.[13] Concernant les méta-données, nous avons extrait les informations publiées par les utilisateurs, en nous focalisant sur trois types d'information généralement utilisées dans les études ethnographiques : genre, tranche d'âge et ville d'origine.

4 Constitution du corpus *TArC*

Pour créer notre corpus, nous avons appliqué une annotation au niveau des mots. Cette phase a été précédée de quelques étapes de pré-traitement des données, en particulier la tokenisation. Chaque token a été associé à ses annotations et métadonnées (tableau 2). Afin d'obtenir la correspondance entre les transcriptions de morphèmes arabes et arabish, les tokens ont été segmentés en morphèmes. Cette segmentation a été effectuée manuellement pour un premier groupe de tokens.[14] Dans sa version finale, chaque token est associé à 11 annotations différentes, correspondant au nombre de niveaux d'annotation que nous avons choisi. Un extrait du corpus avec annotation est présenté dans le tableau 2.

Comme le TA est une écriture spontanée du TUN, nous avons jugé important d'adopter les directives CODA* comme modèle pour produire des lemmes et une transcription unifiées pour chaque token (colonnes *Lem* et *Tra* dans la tableau 2). Afin de garantir une transcription et une lemmatisation précises, nous avons annoté manuellement les 6 000 premiers tokens avec tous les niveaux d'annotation. Concernant les mots étrangers, nous avons transcrit les mots arabish en caractères arabes, à l'exception des termes étrangers. En fait, ces mots seront analysés dans un second temps, en faisant la distinction entre les mots étrangers intégrés et le mélange de langues. Les premiers seront transcrits en caractères arabes, les autres seront lemmatisés dans leur langue étrangère. Cette identification sera également utile pour la construction d'un analyseur morphologique qui, après une phase d'identifica-

[12] En particulier a été utilisé son vocabulaire thématique (TVL, de l'anglais *Thematic Vocabulary List*).

[13] Nous avons planifié d'augmenter la taille du corpus dans un second temps.

[14] L'arabe, en général, est une langue à haut niveau de synthèse, ce qui signifie qu'elle peut concentrer dans un token plusieurs informations grammaticales grâce à l'ajout de différents morphèmes.

A	B	C	D	E	F	G	H	I	J	K	L
Cor	Textco	Par	W	Arabiʃ	Tra	Ita	Lem	POS	Var	Age	Gen
3fE	150902	2	1	kifech	كيفاش	come	كيفاش	adv	Bnz	25-35	M
3fE	150902	2	2	tchou-fou	تشوفوا	vi pare	شاف	verb	Bnz	25-35	M
3fE	150902	2	3-4	l3icha	العيشة	la vita	عيشة	noun	Bnz	25-35	M
3fE	150902	2	3	l	ال	-	ال	det	Bnz	25-35	M
3fE	150902	2	4	3icha	عيشة	-	عيشة	noun	Bnz	25-35	M
3fE	150902	2	5-6	fil	فال	all'	في	prep	Bnz	25-35	M
3fE	150902	2	5	f	ف	-	في	prep	Bnz	25-35	M
3fE	150902	2	6	il	ال	-	ال	det	Bnz	25-35	M
3fE	150902	2	7	4orba	غربة	estero	غربة	noun	Bnz	25-35	M
3fE	150902	2	8	?	؟	?	؟	pct	Bnz	25-35	M

Tab. 2: Un extrait de la structure du corpus TArC. Parmi les colonnes plus significatives, la colonne E (*Arabiʃ*) correspond au token en arabish. La colonne F (*Tra*) est la transcription en caractères arabes. La colonne G (*Ita*) est la traduction en italien. La colonne H (*Lem*) est le lemme. La colonne I le *POS*, etc. Pour plus de détails voir (Gugliotta & Dinarelli, 2020).

tion des mots à transcrire et du *code-switching*, contribuera à la tâche de transcription en elle-même. Ce travail est toujours en cours.

5 Procédure incrémentale et semi-automatique de constitution du corpus

Afin de rendre la collecte du corpus plus facile et plus rapide, nous avons adopté une procédure semi-automatique basée sur des modèles neuronaux séquentiels (Dinarelli & Grobol, 2019b,a). Puisqu'il est plus facile d'obtenir automatiquement certaines annotations une fois que les tokens arabish sont transcrits en caractères arabes, puisque la transcription de l'arabish en arabe est une information très importante pour étudier le système arabish, et puisque elle est aussi la plus coûteuse, la procédure semi-automatique ne concerne que la transcription de l'arabish en écriture arabe.[15] Pour cela, nous avons utilisé le premier groupe de 6 000 tokens transcrits manuellement comme ensemble de données d'entraînement et de test dans un cadre de validation croisée. Comme nous l'avons expliqué dans la section précédente, les tokens français ont été retirés des données puisque ils créent du bruit pour un modèle automatique et probabiliste basé sur l'orthographie arabish. Après l'élimination des tokens français, les données ont été réduites à environ 5 000 tokens. Nous constatons qu'en combinant l'index de la phrase, du paragraphe et l'index du token dans le corpus, des phrases entières, voire des paragraphes, peuvent être reconstruites. Cependant, à partir des 5 000 tokens seulement 300 séquences ont pu être reconstruites, ce qui n'est pas suffisant pour l'apprentissage d'un modèle neuronal.[16] Au lieu de cela, puisque les tokens sont transcrits au niveau des morphèmes, nous avons divisé les tokens arabish en caractères, et les tokens arabes en morphèmes, et nous avons traité chaque token comme

[15] En réalité nous sommes en train de développer des systèmes équivalant pour l'annotation en *POS* et pour la lemmatisation. Ce travail est en cours.

[16] Les expériences préliminaires ont donné des mauvais résultats : 50% de précision.

une séquence. Notre modèle apprend donc à transcrire les caractères arabish en morphèmes arabes. Dans ces conditions la validation croisée a donné une précision moyenne d'environ 65%. Ce résultat n'est pas satisfaisant dans l'absolu, mais il est plus qu'encourageant compte tenu de la petite taille de nos données. Ce résultat signifie que moins de 4 tokens, en moyenne, sur 10 doivent être corrigés manuellement. Avec ce modèle, nous avons automatiquement transcrit en morphèmes arabes environ 700 tokens supplémentaires. Ces tokens ont été corrigés manuellement et ont été ajoutés aux données d'entraînement de notre modèle neuronal, et une nouvelle validation croisée a été effectuée. Le résultat a été maintenant d'environ 70% en moyenne. Cette procédure a été ré-itéré 3 fois au total, pour transcrire 4 bloques de tokens sur 5. La précision moyennes sur le quatrième bloque a été d'environ 76%.[17]

6 Conclusions

Dans ce document, nous avons présenté TArC, le premier corpus d'arabish tunisien annoté avec des informations morpho-syntaxiques. Concernant la procédure de construction, nous avons décrit les étapes de constitution et notre effort visant à rendre le corpus le plus représentatif possible du TA et du TUN. Nous avons décrit l'étape de collecte des textes, ainsi que la construction du corpus et la procédure semi-automatique adoptée pour transcrire le TA en écriture arabe, en tenant compte des directives CODA*. Au stade actuel de la recherche, le TArC est constitué de 40 000 tokens, une partie desquels a été transcrite et annoté manuellement, le reste est en cours de transcription semi-automatique, qui a déjà montré des résultats encourageants avec une précision de transcription de 76% en moyenne.

Références

Al-Badrashiny M., Eskander R., Habash N. & Rambow O. (2014). Automatic transliteration of romanized dialectal arabic. In *Proceedings of the eighteenth conference on computational natural language learning*, p. 30–38.

Buckwalter T. & Parkinson D. (2014). *A frequency dictionary of Arabic : Core vocabulary for learners*. Routledge.

Coulmas F. (2003). *Writing systems : An introduction to their linguistic analysis*. Cambridge University Press.

Darwish K. (2014). Arabizi detection and conversion to Arabic. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, p. 217–224. DOI : 10.3115/v1/W14-3629.

Dinarelli M. & Grobol L. (2019a). Hybrid neural models for sequence modelling : The best of three worlds. *CoRR*. arXiv preprint 1909.07102.

Dinarelli M. & Grobol L. (2019b). Seq2biseq : Bidirectional output-wise recurrent neural networks for sequence modelling. *CoRR*. arXiv preprint 1904.04733.

Durand O. (2009). *Dialettologia araba*. 'Sapienza' University of Rome, 'Studi Orientali' Faculty.

[17] Au moment de la soumission finale, 2 bloques de tokens additionnels ont été téléchargés et ajoutés au corpus. Ces bloques sont donc à transcrire et à corriger avec le bloque 5. Nous sommes en train de développer un système d'apprentissage multi-tâche pour effectuer tous les niveaux d'annotations avec un seul modèle.

Graja M., Jaoua M. & Hadrich-Belguith L. (2010). Tunisian dialect corpus interlocutor (tudicoi). In *Arabic Natural Language Processing Research Group (ANLP)* : MIRACL Laboratory, Sfax (Tunis).

Graja M., Jaoua M. & Hadrich-Belguith L. (2013). Discriminative framework for spoken tunisian dialect understanding. In *International Conference on Statistical Language and Speech Processing*, p. 102–110: Springer.

Guellil I., Saâdane H., Azouaou F., Gueni B. & Nouvel D. (2019). Arabic natural language processing : an overview. *Journal of King Saud University-Computer and Information Sciences*. doi : 10.1016/j.jksuci.2019.02.006.

Gugliotta E. & Dinarelli M. (2020). Tarc : Incrementally and semi-automatically collecting a tunisian arabish corpus. *cs.CL*. arXiv preprint 2003.09520.

Habash N., Eryani F., Khalifa S., Rambow O., Abdulrahim D., Erdmann A., Faraj R., Zaghouani W., Bouamor H., Zalmout N., Hassan S., Al-Shargi F., Alkhereyf S., Abdulkareem B., Eskander R., Salameh M. & Saddiki H. (2018). Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

Haspelmath M. & Tadmor U. (2009). *Loanwords in the world's languages : a comparative handbook*. Walter de Gruyter.

Hert P. (1999). Quasi-oralité de l'écriture électronique et sentiment de communauté dans les débats scientifiques en ligne. *Réseaux*, **17**(97).

Masmoudi A., Ellouze Khemakhem M., Estève Y. & Hadrich-Belguith L. (2014a). Tunisian arabic railway interaction corpus. In *Arabic Natural Language Processing Research Group (ANLP)* : MIRACL Laboratory, Sfax (Tunis).

Masmoudi A., Habash N., Ellouze M., Estève Y. & Hadrich-Belguith L. (2015). Arabic transliteration of romanized tunisian dialect text : A preliminary investigation. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), p. 608–619: Springer-Verlag. doi : 10.1007/978-3-319-18111-0_46.

Masmoudi A., Khmekhem M. E., Estève Y., Belguith L. H. & Habash N. (2014b). A corpus and phonetic dictionary for Tunisian Arabic speech recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 306–310, Reykjavik, Iceland : European Language Resources Association (ELRA).

Medhaffar S., Bougares F., Estève Y. & Hadrich-Belguith L. (2017). Sentiment analysis of Tunisian Dialects : Linguistic ressources and experiments. In *Proceedings of the third Arabic Natural Language Processing Workshop*, p. 55–61: Association for Computational Linguistics. doi : 10.18653/v1/W17-1307.

Meftouh K., Harrat S., Jamoussi S., Abbas M. & Smaili K. (2015). Machine Translation Experiments on PADIC : A Parallel Arabic DIalect Corpus. In *The 29th Pacific Asia Conference on Language, Information and Computation*, shanghai, China. hal : hal-01261587.

Meftouh K., Harrat S. & Smaïli K. (2018). PADIC : extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, Antalya, Turkey. hal : hal-01718858.

Tobaili T. (2016). Arabizi identification in twitter data. In *Proceedings of the ACL 2016 Student Research Workshop*, p. 51–57.

Younes J., Achour H. & Souissi E. (2015). Constructing linguistic resources for the tunisian dia-

lect using textual user-generated contents on the social web. In *International Conference on Web Engineering*, p. 3–14: Springer.

Younes J., Achour H., Souissi E. & Ferchichi A. (2018a). Survey on corpora availability for the tunisian dialect automatic processing. In *2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing (JCCO : TICET-ICCA-GECO)*, p. 1–7: IEEE.

Younes J., Achour H., Souissi E. & Ferchichi A. (2020). Romanized tunisian dialect transliteration using sequence labelling techniques. *Journal of King Saud University-Computer and Information Sciences*.

Younes J., Souissi E., Achour H. & Ferchichi A. (2018b). A sequence-to-sequence based approach for the double transliteration of tunisian dialect. *Procedia computer science*, **142**, 238–245.

Zribi I., Ellouze M., Hadrich-Belguith L. & Blache P. (2015). Spoken tunisian arabic corpus "stac" : Transcription and annotation. *Research in Computing Science*, **90**, 123–135.